UNIVERSITÀ DEGLI STUDI DI NAPOLI
"L' ORIENTALE"

DIPARTIMENTO DI STUDI LETTERARI, LINGUISTICI E COMPARATI

DOTTORATO DI RICERCA IN STUDI LETTERARI, LINGUISTICI E COMPARATI

XXXVII CICLO

SAFEGUARDING MINORS: A STUDY ON AUTOMATIC AGE-BASED CLASSIFICATION OF ONLINE MULTIMEDIA CONTENT USING TEXT TRANSCRIPTS

Tutor:
Ch.ma Prof.ssa
Johanna Monti

Candidata:
Paone Antonietta
DLLC/00135

Coordinatore:
Ch.mo Prof.
Alberto Manco

ANNO ACCADEMICO 2023/2024

UNIONE EUROPEA
Fondo Sociale Europeo
Ministero dell'Università e della Ricerca
PON RICERCA E INNOVAZIONE
REACT EU

Abstract

English. When we talk about online multimedia content, we refer to those elements that can enrich the user experience through the use of audio and visual material, such as videos, podcasts, live, and webinars. It often happens that videos, with content or language not suitable for children, are seen by minors who can be attracted and influenced by what they see.

Automatic content classification would make the identification of content dangers quick and applicable to a large amount of data. Furthermore, considering that all online multimedia content contains textual components, a linguistic approach is particularly useful.

The aim of this thesis is to explore the topic of automatic classification of texts related to multimedia content (such as subtitles and transcriptions) to assess the appropriateness of such content for different age groups, with a focus on protecting minors from exposure to unsuitable material. This study investigates machine learning-based approaches, comparing various classifiers, as well as semantic and lexical methods. Furthermore, resources have been developed for both Italian and English languages to support this classification process.

Chapter 1 outlines the background and current state of multimedia content classification research, providing an overview of what has already been done in the field of computational linguistics on this problem and identifying the elements that still need to be developed.

In Chapter 2, the methodology for carrying out the work is illustrated. Two corpora have been created: one for Italian and one for English. The initial idea was to focus exclusively on TED Talks, a particular type of content that covers a wide range of topics presented by experts in various fields. These talks are generally of high quality and well-curated, making them ideal for linguistic and content analysis. In addition, TED Talks are available in many languages, facilitating the creation of multilingual corpora and comparative analysis in different languages. However, after some experiments, it was decided to add other texts to the corpora, specifically for children, extracted from YouTube Kids and other children's channels, and specifically for adults, as they deal with themes related to sex, drugs, violence, and fear, or are characterized by vulgar language.

In order to conduct supervised classification, the corpus has been annotated following the AGCOM (Authority for Communications Guarantees) *Guidelines on the Classification*

of Audiovisual Works Intended for the Web and Video Games (AGCOM 2017).

High-specialized lexicons have been used to extract the linguistic indicators of the main features used to rate multimedia online content. The dictionaries created are the dictionary of *Violence*, *Drugs*, and *Bad language*.

In addition, a study of emotions has been conducted, focusing particularly on emotions that indicate negative sentiment such as *Anger*, *Disgust*, and *Fear*. The emotions have been analyzed thanks to the *NRC Emotion Intensity Lexicon* (Saif M Mohammad 2017): a list of words that have been assigned a value ranging from 0 to 1 in relation to the eight main emotions theorized by Plutchik (1984).

The semantic information contained in the texts played a fundamental role in the classification. All texts considered for the analysis were represented in a network based on their semantic proximity. Specifically, a distributional semantic matrix was created to extract similarity values between words and to perform a semantic expansion of the more significant words in the transcripts. Once the texts were vectorized, the similarity values between text vectors were calculated using the Cosine Similarity algorithm. After comparing all texts and generating a large network of text-per-text edges, this network was graphically represented with Gephi (Bastian et al. 2009), assigning the similarity value as weight. Modularity Class (Blondel et al. 2008) was applied to the network to generate a classification.

The experimental procedure is detailed in Chapter 3. An evaluation of machine learning algorithms was conducted to determine the best algorithm for text categorization. Specifically, the algorithms tested include the *Support Vector Machine*, *Naive Bayes classifier*, *Decision Tree*, *Random Forest*, *Convolutional Neural Network*, *Recurrent Neural Networks*, *BERT*, *DistilBERT*, *mBERT*, *RoBERTa*, and, specifically for Italian, *ELECTRA*, *AlBERTo*, *GilBERTo*, and *UmBERTo*. Additionally, several large language models such as *Flan T5*, *GPT-2*, and *Minerva* were also tested.

To further improve the results, a filtering rule based on a badwords dictionary was also added. This rule allows for the filtering of content with language that is not suitable for minors, ensuring that if the system misclassifies certain texts, those containing inappropriate language are still categorized as adult content.

As discussed in Chapter 4, the results are interesting and encouraging; in a future prospective, the system could be applied to other types of online multimedia content, and an integrated analysis with other User Generated Content, such as comment and descriptions, could be developed. Other elements, such as the language complexity and the structure of the phrase, could be integrated in order to enrich the analysis, and other languages could be taken in account. Furthermore, the studies developed in this thesis regarding the textual component of multimedia content can be integrated with approaches that consider additional elements, such as metadata, images, and frames, to develop comprehensive content

classification and monitoring systems.

Italian. Quando si parla di contenuti multimediali online, ci si riferisce a quegli elementi che arricchiscono l'esperienza dell'utente utilizzando materiale audio e visual, come video, podcast, dirette e webinar. Tuttavia, accade frequentemente, che contenuti multimediali destinati a un pubblico adulto, o con un linguaggio non adatto ai bambini, vengano fruiti da minori che possono esserne attratti e influenzati. La classificazione automatica dei contenuti multimediali diffusi online, renderebbe l'identificazione di probabili pericoli rapida e applicabile a una grande quantità di dati. Inoltre, considerando che tutti i contenuti multimediali online contengono componenti testuali, un approccio linguistico risulta particolarmente utile.

L'obiettivo della tesi è quello di esplorare il tema della classificazione automatica dei testi relativi ai contenuti multimediali (come sottotitoli e trascrizioni) per valutarne l'adeguatezza per le diverse fasce d'età, con un focus sulla protezione dei minori dall'esposizione a materiali non adatti a loro. Questo studio esamina approcci basati sul machine learning, confrontando vari classificatori, oltre a metodi semantici e lessicali. Inoltre, ha previsto lo sviluppo di risorse per le lingue italiana e inglese a supporto di questo processo di classificazione.

Il Capitolo 1 delinea il background teorico e lo stato dell'arte in riferimento alla classificazione dei contenuti multimediali, fornendo una panoramica di quanto è stato già realizzato nel campo della linguistica computazionale e identificando gli elementi che necessitano di ulteriore sviluppo.

Nel Capitolo 2 viene illustrata la metodologia. Sono stati creati due corpora: uno per l'italiano e uno per l'inglese. L'idea iniziale era di concentrarsi esclusivamente sui TED Talks, un particolare tipo di contenuto che copre una vasta gamma di argomenti presentati da esperti in vari campi. Questi discorsi sono generalmente di alta qualità e ben curati, rendendoli ideali per l'analisi linguistica e di contenuto. Inoltre, i TED Talks sono disponibili in molte lingue, facilitando la creazione di corpora multilingue e un'analisi comparativa. Tuttavia, dopo alcuni esperimenti, è stato deciso di aggiungere altri testi ai corpora: testi specificamente per bambini, estratti da YouTube Kids e altri canali per bambini, e testi specificamente per adulti, contenenti tematiche quali sesso, droga, violenza e paura, o caratterizzati da un linguaggio volgare.

Per condurre una classificazione supervisionata, il corpus è stato annotato seguendo le *Linee guida sulla classificazione delle opere audiovisive destinate al web e dei videogiochi* dell'AGCOM (Autorità per le Garanzie nelle Comunicazioni) (AGCOM 2017).

Per estrarre gli indicatori linguistici utilizzati nelle linee guida, sono stati realizzati dei lessici altamente specializzati, al fine di valutare i contenuti multimediali. I dizionari creati sono il dizionario della *Violenza*, *Droga* e *Linguaggio scurrile*.

Inoltre, è stato condotto uno studio sulle emozioni, concentrandosi in particolare sulle emozioni che indicano sentimenti negativi come *Rabbia*, *Disgusto* e *Paura*. Le emozioni sono state analizzate grazie al *NRC Emotion Intensity Lexicon* (Saif M Mohammad 2017): una lista di parole a cui è stato assegnato un valore che va da 0 a 1 in relazione alle otto emozioni principali teorizzate da Plutchik (1984).

Le informazioni semantiche contenute nei testi hanno svolto un ruolo fondamentale nella classificazione. Tutti i testi considerati per l'analisi sono stati rappresentati in una rete basata sulla loro prossimità semantica. Nello specifico, è stata creata una matrice semantica per estrarre i valori di similarità tra le parole e per eseguire un'espansione semantica delle parole più significative nei testi. Una volta vettorizzati i testi, i valori di similarità tra i vettori di testo sono stati calcolati utilizzando l'algoritmo della Similarità del Coseno. Dopo aver confrontato tutti i testi e generato una vasta rete di connessioni testo per testo, questa rete è stata rappresentata graficamente con Gephi (Bastian et al. 2009), assegnando il valore di similarità come peso. Al fine di generare una classificazione è stato utilizzato l'algoritmo di Modularity Class (Blondel et al. 2008).

Nel capitolo 3 viene illustrata la parte sperimentale del lavoro. Per determinare il miglior algoritmo per la categorizzazione dei testi su questo specifico task, è stata eseguita una valutazione degli algoritmi di machine learning comunemente usati per la classificazione. Nello specifico, gli algoritmi testati sono stati il *Support Vector Machine*, il *Naive Bayes classifier*, l'*Albero Decisionale*, la *Random Forest*, le *Reti neurali convoluzionali* e le *Reti neurali ricorrenti*, alcuni trasnformers, tra cui: *BERT*, *DistilBERT*, *mBERT*, *RoBERTa* e, alcuni specifici per l'italiano tra cui: *ELECTRA*, *AlBERTo*, *GilBERTo* e *UmBERTo*. Infine alcuni Large Language Model come *Flan T-5, GPT-2* e *Minerva*.

Al fine di migliorare i risultati, è stata aggiunta una regola che funge da filtro, basata su un dizionario di dominio contenente le bad words (parole offensive). Questa regola ha consentito di filtrare i contenuti con un linguaggio non adatto ai bambini classificati erroneamente dal sistema, categorizzandoli come adatti agli adulti.

Dal capitolo 4 emerge che i risultati sono interessanti e incoraggianti; in una prospettiva futura, il sistema potrebbe essere applicato ad altri tipi di contenuti multimediali online. Potrebbe, inoltre, essere sviluppata un'analisi integrata con altri contenuti generati dagli utenti, come commenti e descrizioni. Per rendere completa l'analisi, potrebbero essere integrati altri elementi come la complessità linguistica e la struttura della frase, e potrebbero essere prese in considerazione altre lingue. Inoltre, gli studi sviluppati in questa tesi riguardanti la componente testuale dei contenuti multimediali potrebbero essere integrati con approcci che considerano ulteriori elementi, come metadati, immagini e frame, per sviluppare sistemi completi di classificazione e monitoraggio dei contenuti.

Declaration

I hereby certify that this material which is subject to submission for assessment leading to the award of PhD in Literary, Linguistics and Comparative Studies, where I am affiliated with the UNIOR NLP Research Group, from the University of Naples 'L'Orientale' is entirely my own work. I was careful to ensure that this work is original, and does not, to the best of my knowledge, breach any law of copyright. I also declare that I have not used any type of generative Artificial Intelligence tools for the writing of this manuscript nor for the creation of images, graphics, tables, or their corresponding captions.

Naples, November 28, 2024 Antonietta Paone

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Prof. Johanna Monti, for her invaluable guidance, understanding, and humanity throughout these years. Her exceptional expertise has been a source of inspiration for my growth, and I am deeply grateful for the numerous opportunities she has offered me during my doctoral journey.

I am also profoundly grateful to Prof. Maria Pia di Buono for her insightful advice and inspiring suggestions, which have been invaluable to my research.

I would also like to extend my sincere thanks to Theuth Linguistic Engines, particularly to Prof. Emeritus Annibale Elia and Dr. Alessandro Maisto, who encouraged and believed in me from the very beginning. Working under their guidance during my six months at the company has been an invaluable experience, one from which I have learned immensely. I hold deep admiration for their dedication and expertise.

I am also deeply thankful to all the colleagues of the UNIOR NLP Research Group, always ready to offer advice and support. Special thanks to Khadija, whose solidarity has been a constant source of reassurance, and to Argentina and Carmen, for the shared adventures during conferences, which have made this journey even more memorable.

A special thank you goes to all my friends, especially Sara and Enza Maria for their support, generosity, and dedication to this project. Their willingness to share their time and expertise for the annotation work was invaluable, and I am sincerely grateful for their contributions.

To my family, whose unwavering support has been the foundation of my journey, I am profoundly thankful. To my parents, Roberto and Patrizia, and my brother Rinaldo, thank you for being my pillars of strength. Thank you, for being my crutch when thoughts invaded, preventing me from carrying out even the simplest daily tasks, thank you for every sacrifice, for accompanying me, and standing by me, never leaving me alone.

I would like to express my deep gratitude to my partner, Marco, for standing by me throughout the thesis process. Thank you for being my number one supporter, for believing in my worth even when I couldn't see it myself, and for patiently listening to countless presentations as my dedicated test audience.

Finally, I gratefully acknowledge *PON Ricerca e Innovazione 2014-2020*, "Dottorati Innovativi", for co-funding this research and supporting my doctoral studies.

Contents

List of Figures

List of Tables

List of Abbreviations

AGCOM Authority for Communications Guarantees

BiLSTM Bidirectional Long Short-Term Memory

BoW Bag of Words

CCFC Campaign for a Commercial-Free Childhood

CDD Center for Digital Democracy

CNN Convolutional Neural Network

DL Deep Learning

DT Decision Tree

FTC Federal Trade Commission

IAA Inter-Annotator Agreement

k-NN k-Nearest Neighbors

LDA Latent Dirichlet Allocation

LDA Latent Dirichlet Allocation

LLM Large Language Model

LR Logistic Regression

LSTM Long Short-Term Memory

ML Machine Learning

MPAA Motion Picture Association of America

NB Naive Bayes

NLP Natural Language Processing

NN Neural Network

PEGI Pan European Game Information

RF Random Forest

RI Rule Induction

RNN Recurrent Neural Network

SPID Sistema Pubblico di Identità Digitale

SVM Support Vector Machine

TF-IDF Term Frequency-Inverse Document Frequency

UGC User Generated Content

YTP YouTube Poop

Introduction

0.1 Purpose and Research Questions

As YouTube's Vice President of Engineering, Cristos Goodrow, wrote on the video streaming service's own blog: *there is an audience for almost every video, and the job of our recommendation system is to find that audience. Think about how hard it would be to navigate all of the books in a massive library without the help of librarians* (Goodrow 2021).

This statement encapsulates the essence of my research. Going beyond a single platform, my goal is to lay the groundwork for a system that leverages the potential of language to automatically classify online multimedia content. The purpose of such a system could be to assist those who guide younger users, much like the library envisioned by Goodrow, in supporting individuals who may feel overwhelmed by the vast array of content presented daily in the online realm. Specifically, my work aims to support those guiding children who navigate swiftly through this vast sea of content but may lack the critical discernment needed to protect themselves from potential online risks.

Consider an amusement park where a certain number of attractions are only meant for specific individuals: those who are at least one meter tall, those who are older than 10 years, those who weigh less than 100 kg, and so on. Imagine that in this park, there's a roller coaster train with seat belts designed to secure those who are at least one meter tall to their seats. If a child measuring 80 centimeters gets on that train, there is nothing anyone can do to ensure they come down safe and sound; it might go well, but there is also a chance they could fall and suffer severe injuries. The Internet is like that amusement park, but much, much larger. This metaphor of Guido Scorza, a member of the College of the Personal Data Protection Authority, as emphasized in *Tempi Digitali, Atlante dell'Infanzia (a rischio) in Italia*, explains why the problem of verifying age has become central for those who deal with online activities.

In the same volume published by Save The Children, edited by Vichi de Marchi 2023, various data are highlighted to understand the importance of research headed in this direction, with the aim of protecting the youngest ones. In Italy, 78.3% of children aged 11 to 13 years use the Internet every day, mainly through smartphones. The age at which children

own or use a smartphone is decreasing, with a significant increase in the number of children aged 6 to 10 using a mobile phone every day after the pandemic: from 18.4% to 30.2% between the two-year periods 2018-19 and 2021-22. Although the law stipulates that a user can have access to social networks only after turning 13, reality shows a significant presence of preadolescents who have opened a profile indicating a higher age or have used that of an adult, often a more or less aware parent: 40.7% of 11-13 year olds in Italy use social networks, with a higher prevalence among females (47.1%) compared to males (34.5%) In the same report, discussing the realm of video games, it is highlighted how, especially for the youngest individuals, those who do not yet have strong emotional and cognitive structures, the role of guidance and participation of the adult, usually the parents, is crucial. The adult is capable of assessing the type of content, and imposing rules on its use, such as exposure time, etc. However, this alone is not sufficient. The participation of other stakeholders, such as the platforms themselves, is necessary to reinforce control systems that already exist in part. Even more so, impartial external verification processes are needed to assess the level of content safety for minors and their protection, according to Marianna Ganapini, a professor at Union College in New York (Vichi de Marchi 2023).

For video games, there is a pan-European classification called Pan European Game Information (PEGI), which categorizes the product by age groups and content to guide consumers and prevent potentially inappropriate video games from reaching younger audiences. According to this classification, in 2022, 64.4% of the video games sold in Italy were labeled as suitable for audiences between 3 and 12 years old. However, a more rigorous evaluation by a third-party independent entity would be important. Guidelines verification and standardization, that are crucial for implementation, face numerous obstacles, including a global video game market, where the norms vary depending on the geographic area (Vichi de Marchi 2023).

My work aims to help fill this gap by exploring, through experiments and resource building, the feasibility of implementing a classification system for online multimedia content to determine the appropriate age for consuming specific content, focusing on its textual aspects.

The decision to focus primarily on the linguistic aspect, despite the multimedia nature of the content, stems from the unique advantages that textual analysis offers in content classification. Texts, such as transcriptions and subtitles associated with online multimedia content, allow for a comprehensive and objective analysis of the entire material, independent of its length. Unlike images, which capture attention immediately and can lead to quick judgments, text-based analysis provides a deeper understanding of the content's language and tone. For instance, a parent may easily recognize potentially violent, sexual, or frightening content through visual previews and intervene accordingly. However, they could be

deceived by videos with seemingly child-friendly visuals but inappropriate language, as seen in amateur-dubbed cartoons or AI-generated videos that use language unsuited to young audiences. This research focuses on testing and evaluating existing text classification systems to determine the most effective approach for age-based content assessment. Although a combined analysis of text and images could yield even more accurate results, this initial focus on linguistic aspects offers a solid foundation for potential future developments that may incorporate visual elements.

I argue that particular importance should be given to text analysis initially to avoid falling into certain traps. The main one among these is the trap of YouTube Poop (YTP), where online-distributed videos take frames from children's videos, mix them, dub them, and give them an entirely different meaning. A famous case is that of the poet Michael Rosen, who had to post a warning on his website in 2012, stating *quite a few people have fun taking my videos and making new versions of them, known as 'YouTube poops.' Many of these are not suitable for young children. I am not responsible for either the words or pictures of these*[1]. In 2021, a British teacher accidentally sent her students an extremely vulgar YouTube poop of Rosen's poem "The Car Trip" instead of the original poem, mistaking it for the original[2]. In 2019, Rosen stated on his Twitter channel there were "about 4,000 YTPs" of Rosen performing his poems and stories.

A significant need has emerged, in recent years, for a system that safeguards children; although parents constantly monitor their children's online activity, it is often challenging for them to determine whether the content is suitable for their child's age. This need became evident from a survey I conducted and distributed among parents of minors, in the context of the project I presented during the *European Researchers' Night* event in Italy (2022), entitled *Horror or funny?*[3], and later circulated through word of mouth in targeted WhatsApp groups (mothers' class groups). 86.4% of the respondents, while monitoring the online activity of a minor, noticed that autoplay displayed videos unsuitable for the viewer's age. The importance of language also emerged from the survey, as a majority of respondents believed that content, to be classified as suitable for minors, should have appropriate language. 72.7% argued that it should not display strong images, and 59.1% believed it should not show scary images. Similary, 72.7% stated that a video should be educational to be considered suitable. The survey results highlight the real problem related to evaluating a video based only on preview images. Specifically, when shown a video featuring the famous character from the horror video game Poppy Playtime, Huggy Wuggy, 66.7% of respondents would never show the video based on the preview. However, the remaining 33,3% find the

[1]Rosen, Michael (May 29, 2012). "News - For Adults". michaelrosen.co.uk. Archived from the original on March 11, 2015.

[2]Lovell, Joanna (January 14, 2021). "Hull teacher shares vile parody video to class of nine-year-olds." Hull Daily Mail.

[3]*Horror or Funny?* survey: https://forms.gle/VYs53fKLQb7SbW8g6

character to be cute and sweet, so they would show it. After reading the transcription of a part of the video's song ("*My name is Huggy Wuggy, sharp teeth leave you bleeding, never call me ugly, hug me until you die.*"), 100% of the respondents declared that they would no longer show the video to a minor. Reasons include inappropriate language, excessive violence and horror, creepy and threatening song, fear-inducing content, violence, lack of reassurance, unsuitability for minors, and overall unpleasantness. Another phenomenon on which I asked for parents' opinions is that of challenges, of which the phenomenon is known. The challenge phenomenon is extremely concerning for 50% of the respondents and very concerning for 31.8%. Save The Children's Atlas (Vichi de Marchi 2023) also addressed this phenomenon, listing online challenges among the risks of the internet (alongside cyberbullying, sexting, morphing, doxing, cognitive overload, and compulsive shopping). Challenges are shared social media trends that can be fun and harmless, but in some cases, they can be harmful and extremely dangerous, often involving one's own body or body parts being subjected to injury to complete the challenge. Participating in online challenges can be perceived as being part of a large community. In the survey, I showed a video challenge. In this specific case, the challenge involved opening the car trunk automatically and positioning oneself at the right distance without moving: if the trunk hit the person while opening, the challenge was lost. 69.6% believed that a video of this kind would not be suitable for a minor, while the other 30.4% believed that it could be seen by a minor without consequences. Those who would not show it find it silly, useless, and they thought that a minor could imitate the actions and get hurt. Those who would show it believed that it would not be dangerous if well-executed, does not show strong or scary scenes. 8.7% thought it was suitable for an audience aged 6-10 years; 21.7% for an audience aged 11-14 years, 30.4% for an audience aged 15-18 years, and 39.1% for an adult audience.

Over the years, especially after 2017, following the Elsagate scandal, which will be discussed in the first chapter of this work, various attempts have been made to address the problem. Platforms primarily relied on users, asking them to declare whether the videos they uploaded online were suitable for minors or not. The Atlas (Vichi de Marchi 2023) includes a review of various strategies evaluated to enhance the online experience for minors. Among the various measures considered, one proposal involved restricting access to social networks for users under the age of 13, but this limit is easily circumvented by declaring a different age or using a parent's mobile phone. There was also contemplation of using a system like the Sistema Pubblico di Identità Digitale (SPID), also known as the Public Digital Identity System in English, but this would pose a problem related to privacy and the accumulation of personal data (Vichi de Marchi 2023). It was proposed to have the videos manually reviewed by dedicated teams, but this would expose individuals to hours of potentially unpleasant content and therefore be a source of stress (Wilson 2020). Therefore, the current preferred

solution is to utilize the classification capabilities of automatic systems, like those that will be explored in this thesis.

The definition of the problem and the setting of the research objectives give rise to questions regarding the feasibility and potential approaches for addressing the issue. It becomes essential to understand:

1. What has already been done in the field of computational linguistics on this problem and what elements need to be developed for the realization of such a system?
2. What resources can be exploited and how existing frameworks can be improved?
3. Which of the currently existing algorithms commonly used for classification tasks works best for this specific task?
4. How could a classification approach based on semantics work instead?
5. What are the results, and what could be the future developments of the project?

0.2 Thesis Structure

This thesis is structured as follows:

1. **Chapter 1.** In the first chapter, the theoretical background in which the work will be situated is described. It begins with a brief overview of the issue of the sharing of inappropriate videos for minors online and the danger posed by their viewing by an unsuitable audience. One of the possible solutions to mitigate the problem is to rely on automatic systems; therefore, systems utilizing Distributional Semantics will be analyzed, as well as systems using Machine Learning (ML) or a combination of both. Considering the contemporaneity and success of OpenAI's ChatGPT, the latest research on automatic text classification through Large Language Model (LLM) will be explored. At the end of the chapter, the state of the art concerning work on emotions and text complexity will be analyzed.
2. **Chapter 2.** In the second chapter, the methodology for carrying out the work will be illustrated, including the construction of corpora and electronic dictionaries, which will be described in detail.
3. **Chapter 3.** In the third chapter, the experiment conducted to evaluate the ML algorithms on this specific classification task will be described, along with the experiments related to the classification performed using semantic analysis.
4. **Chapter (4)** The fourth chapter of the thesis will be dedicated to the analysis of results and possible future developments of the project.

0.3 Derived Outcomes

The work discussed in this thesis is based on research activities and investigations published in national and international peer-reviewed conferences, both during the three years of the PhD and previously.

The publications related to the thesis are:

- Cirillo N., **Paone A.**. The Role of Semi-Productivity in Multiword Expression Identification: Why can BERT Capture Novel MWEs? In Proceedings of the International Conference EUROPHRAS 2022.
- Maisto A., Martorelli, G., **Paone A.**, Pelosi, S.. *Automatic Classification and Rating of Videogames Based on Dialogues Transcript Files.* In Advances in Internet, Data and Web Technologies: The 9th International Conference on Emerging Internet, Data & Web Technologies (EIDWT-2021) (Vol. 65, p. 301). Springer Nature.
- Maisto A., Martorelli G., **Paone A.**, Pelosi S.. *Extracting video games rating labels from transcript.* In Internet of Things; Engineering Cyber Physical Human Systems, volume 15. Elsevier, 2021.

Presentations at conferences and at other events are:

- **Paone A.**, *Horror or Funny?.* Project presented during the 2022 European Researchers' Night in Italy.
- **Paone A.**, Monti J., *Safeguarding Minors: Evaluating Machine Learning and Transformer-Based Approaches to Age-Appropriate Content Classification.* Poster presented during the AILC Lectures on Computational Linguistics, in June 2024.

0.4 Partnership and Collaborations

This thesis is the result of a three years Industial Innovative PhD Project developed at the University of Naples "L'Orientale", Department of Literary, Linguistic and Comparative Studies.

Since this is an Industrial PhD, the project included six months of work in a company, in my case, an Italian one. I collaborated with Theuth Linguistic Engines[4], an academic spin-off composed of professors and researchers from the Department of Political and Communication Sciences at the University of Salerno. Its goal is to transition from basic research in computational linguistics and data visualization to applied research aimed at the market.

During these months, I have learned a lot about the topic of classification, and the work carried out in the company has been a great inspiration for my thesis work.

[4] https://web.unisa.it/terza-missione/trasferimento-tecnologico/spin-off/unisa/elenco?id=67

During my time at the company, I was involved in the development of hybrid classification systems: automatic, supervised, and unsupervised classifications based on automatic algorithms, complemented by algorithms that leverage the synergy between syntactic rules and electronic dictionaries. The method developed in the company involved the use of POS Taggers and electronic dictionaries during the pre-processing phase, also utilizing domain-specific semantic information and rule-based algorithms to identify meaningful structures, then employs existing automatic classification algorithms to categorize texts into classes or identify a network of similarities between texts and cluster them.

I was involved in the development of a Multidomain Sentiment Analysis and Emotion Detection system, specifically designing an approach to identify dyads within user-generated texts. This included the conception and implementation of a method for calculating the values of complex emotions on the Plutchik (1984) wheel. Additionally, I worked on cleaning and updating the dictionaries and participated in writing reports.

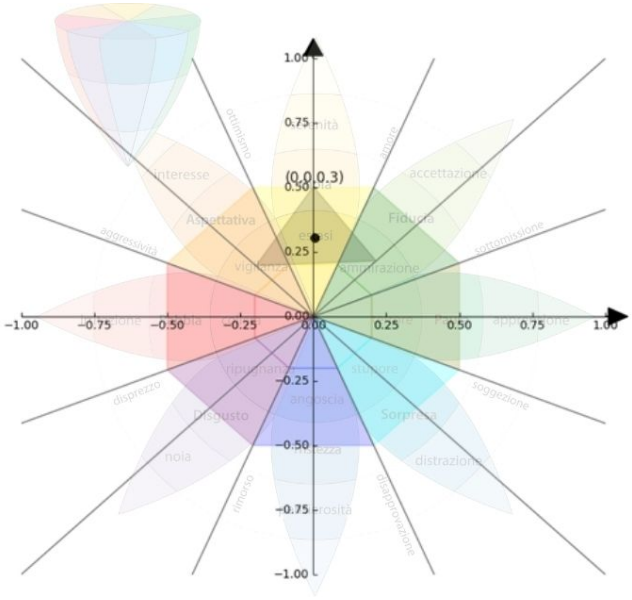


Figure 1: Scheme for identifying secondary emotions

The created system, designed to assess the varying intensities of the 8 main emotions, is based on a geometric principle (see Figure 1) whereby:

- If there is only one emotion present, its position is evaluated on the axis on which it appears.
- If there are two emotions, a line is drawn, and the midpoint is calculated.
- If there are three emotions, the predominant emotion is given by the area where the centroid of the generated triangle falls (is the case represented in Figure 1).
- If there are 4 emotions, the centroid of the 4 points is calculated.

The system considers the identification of the immediate context of the word (3 preceding and 3 following words) to search for elements of intensification or negation, and the relationship between this value and the number of emotional words found in the sentence. This way, the score is normalized and is more easily comparable in cases of sentences with extremely different lengths.

The system has been integrated with a syntactic module that modifies the emotional values. Moreover, each emotion is calculated in relation to its opposite emotion, ensuring that the lesser value is nullified and always resulting in a maximum of 4 values out of the 8 on the wheel.

The PhD project *Sviluppo di un sistema per la classificazione automatica dei testi relativi ai contenuti multimediali online* has been co-funded by PON Ricerca e Innovazione 2014-2020, "Dottorati Innovativi".

Chapter 1

Theoretical background and State of The Art

1.1 The Problem of Harmful Online Videos

In modern times, children are heavily involved with technology, dedicating a significant portion of their time to interacting with devices such as tablets and mobile phones (Ishikawa et al. 2019). Screen media consumption on mobile devices accounted for only 4% of a child's screen time initially, but by 2017, this figure had increased to 35%. This rise is attributed to the growing prevalence of these devices in households (Wilson 2020). The primary activity children engage in with these devices is watching animated shows and children's programs. Typically, they access these through platforms that host age-appropriate video content. However, there are instances where system vulnerabilities expose them to unsuitable content for their age group (Ishikawa et al. 2019). As Wilson (2020) explains, YouTube has both revolutionized and dominated the online video streaming market over the past decade. The company's rapid growth has made it challenging to keep inappropriate content away from children. It has been observed that around 80% of children between the ages of 6 and 12 watch multimedia videos through the YouTube platform. While the platform hosts content from reputable television networks (such as Cartoon Network), it also features content from less reliable sources.

Numerous efforts have been made to address this issue from a legislative standpoint. As Alghowinem (2019) underlined, regulatory bodies across countries like the United States, Australia, and the United Kingdom have established guidelines for children's television programs and advertising. Common regulatory measures shared among these nations include defining specific time slots for family-friendly programming, during which content must be suitable for children, and the imposition of restrictions on both the quantity and content

of advertisements within these time slots. Children's programs are evaluated against several criteria: they must be tailored exclusively for a youthful audience, well-produced with adequate resources, employ language that is comprehensible to children, and possess educational and entertaining objectives. Material strictly prohibited within these programs encompasses any form of demeaning content directed at individuals or groups, scenes featuring fear or violence, content of a sexual nature, unsafe depictions, as well as any references to drugs or alcohol. Nonconforming to these regulations could result in the channel losing its broadcasting license, although this does not apply to online content.

In 2019, the Authority for Communications Guarantees (AGCOM), an independent administrative authority, developed guidelines in Italy. These guidelines were in compliance with resolution no. 74/19/CONS and covered the classification of audiovisual works intended for the web, as well as video games, focusing on their conscious consumption. The classification of works intended for the web is based on thematic descriptors that identify the content areas that may influence the sensitivity of minors and, therefore, require caution in audiovisual representation. These guidelines take into consideration the specificities of the various services available to end-users related to the dissemination of web content and the respective consumption methods in the market at the time of their adoption. To enhance the user experience, technical measures can be applied to either individual content or the entire service, website, mobile, or app, based on the content classification category (AGCOM 2017). We will discuss these guidelines more thoroughly in the methodology chapter as they will be very useful in defining the annotation guidelines for this project.

Increasing awareness of the need to protect minors online has captured researchers' attention. Examining the relevant scientific literature reveals a body of work addressing the challenge of inappropriate video viewing by minors, proposing innovative approaches and targeted strategies. One notable phenomenon is *Elsagate*, where familiar characters like Disney figures or superheroes are depicted in disturbing and inappropriate scenarios. These scenarios involve actions like engaging in alcohol theft, inflicting harm upon one another, doing disgusting things, as well as depicting instances of sexual and violent nature. The term *Elsagate* is composed of *Elsa*, a character from the 2013 Disney animated film Frozen, and *-gate*, a suffix for scandals. Such videos were widespread on YouTube and YouTube Kids and they were at the center of numerous discussions on Twitter and Reddit in 2017 (Ishikawa et al. 2019). YouTube uses a system called "search and discovery system" to help users search for the most suitable video for them. The company has released some information about what data is fed into the algorithm. Users are suggested videos based on their watch history and the topic or channel of the current video, but the recommendations are also influenced by other unspecified 'signals'[1] (Wilson 2020). Sometimes this system gen-

[1] *L'algoritmo, Come funziona il sistema di ricerca e scoperta di YouTube*, YouTube Creators, in www.youtube.com/watch?v=hPxnIix5ExI#strategies-zippy-link-3, last accessed: 2023/09/15.

erates problems, especially for preschool-aged children who can not yet read. They rely on autoplay or the recommendation panel to select their next video (Wilson 2020). As Bridle (2017) pointed out, "For example, a child who begins a viewing session with a Peppa Pig video on the official Peppa Pig channel can eventually be recommended a video that shows Peppa Pig eating her dad or drinking bleach".

One of the challenges of this thesis is precisely to study these particular aspects of videos to understand how automatic classifiers perform in their presence and to theorize and test systems to mitigate them.

Another issue that emerged from the study of these phenomena, related to text, concerns comments. Many predators leave sexualized comments under otherwise harmless children's videos (Schwartz 2019).

However, comments were not specifically addressed in this thesis project but were considered for potential future developments. One reason for this is that the texts used in this thesis were extracted from YouTube Kids, where comments have been disabled as a preventive measure (Wilson 2020).

YouTube Kids is a child-oriented video app developed by YouTube. The app offers curated content, parental controls, and filtering to ensure suitability for children under twelve years old[2]. YouTube launched its YouTube Kids app in 2015 (Alba 2015). Unfortunately, YouTube Kids has not succeeded in preventing children from seeing disturbing content online. In May 2015, the Campaign for a Commercial-Free Childhood (CCFC) and Center for Digital Democracy (CDD) submitted a letter to the Federal Trade Commission (FTC), expressing concern that YouTube Kids was featuring videos that were both inappropriate and potentially harmful to children (Wilson 2020). To solve the problem, YouTube deploys three tactics: ML, user flagging, and human content moderation. In this regard, developing ML systems to address these issues is particularly important, and this is where my thesis contributes.

Google also announced it would hire 10,000 workers to "address violative content" on YouTube[3]. As Wilson (2020) underlines, the cost of repeatedly subjecting human beings to violent and conspiratorial content is high. Moderators can experience symptoms of secondary traumatic stress, resulting from the observation of others' firsthand trauma (Lekach 2018). This could be another good reason to make machines more efficient and to prevent such stressful work from being done by humans.

The *Elsagate* phenomenon has opened a new line of research on this topic (Brandom

[2]Youtube Kids on Wikipedia in https://it.wikipedia.org/wiki/YouTube_Kids, last accessed: 2023/09/15.

[3]*YOUTUBE, More Information, Faster Removals, More People. An Update on What We're Doing to Enforce YouTube's Community Guidelines*, Official YouTube Blog (Apr. 23, 2018), in https://youtube.googleblog.com/2018/04/more-information-faster-removals-more.html, last accessed: 2023/09/15.

2017). This line of inquiry includes studies on Natural Language Processing (NLP), as thoroughly examined by Alqahtani et al. (2023). The authors conducted a systematic review of ML, Deep Learning (DL), and NLP techniques to provide an overview of the existing methods, tools, and approaches developed by scholars. These methods aim to protect children from inappropriate or illegal video content. The authors identified several risks associated with exposing minors to online multimedia content: (1) exposure to inappropriate content, (2) consumption of misleading information, and (3) the creation of false perceptions. To protect minors, the authors list several methods used by scholars, including textual data analysis, metadata analysis, video frame analysis, and audio-based analysis. These methods are implemented through two main approaches: AI-applied techniques and content analysis using statistics derived from YouTube video metadata.

It is precisely within this line of research, and with the aim of addressing this significant and timely issue, that this thesis positions itself through the study of text-based automatic classification systems.

1.2 Text classification with Distributional Semantics

This thesis focuses on specific text types, particularly those related to multimedia content such as subtitles and transcriptions. One of the approaches explored in this thesis involves distributional semantics for text classification tasks. Exploring the literature on text classification with distributional semantics, an understanding of the concept of distributional semantics becomes essential before delving into practical applications. Distributional semantics comprises a set of theories and computational linguistic methods for studying the semantic distribution of words in natural language (Lenci 2008). As highlighted by Lenci and Sahlgren (2023), this approach focuses on quantifying and categorizing the semantic properties of linguistic items based on their distributional properties in text corpora.

Each language can be described in terms of a distributional structure that does not take into account historical characteristics or the meaning of the words (Harris 1954). According to Harris, the distribution of an element can be understood as the sum of all its contexts, where the context of an element A is the actual arrangement of its co-occurrences. In other words, the verb *to eat* (the element A of which Harris speaks) selects to its right, that is in direct object position, a nominal element that belongs to the class of everything that is «edible, solid» (Vietri 2004). If two elements appear in the same context, they are equivalent; from this, it can be inferred that their meanings are also equivalent. Conversely, if they do not select the same class of elements, there is no distributional equivalence between them.

Over time this type of methodology presented by Harris was gradually abandoned in favor of Chomsky's Generative Grammar and then in favor of the Lexicon-Grammar model theorized by Gross, which, following some descriptive analysis, begins to detach itself from

the Chomskian theory and to criticize the elements to the base of such theory, like the concept of grammaticality, the adoption of the subject-predicate schema and the notion of verbal syntagma. While Chomsky regarded the lexicon and syntax as two unrelated elements, Gross argues that these two elements are to be considered as related.
Parallel to the reflection on generative grammar, cognitive linguistics has approached semantics from a conceptual point of view that considers the meaning of a word as a conceptualization of an entity. Conceptualization includes abstract and sensory concepts and physical, social, cultural, and linguistic contexts. From this particular vision comes the fact that even distributional constraints are destined to receive a functional explanation in terms of the principles that govern our processes of conceptualizing the world. This has led to a rejection of distributive semantics by cognitive approaches as meaning needs to be anchored to extra-linguistic entities and cannot be explained in terms of language word distribution (Lenci 2008).

In the field of psychology, Miller and Charles (1991)'s studies assumed great importance, arguing that speakers learn to use words by observing how others use them, which involves attention to the context since words are used together in sentences. The fact that a word is found in various linguistic contexts, determines the formation of a contextual representation that includes syntactic, semantic, paradigmatic, and stylistic conditions. Miller and Charles (1991) used contextual representations to provide an empirical description of semantic similarity in psycholinguistic research. From their point of view, the semantic similitude must be considered as a variable dependent on the context in which the words are used.

This is coherent with the importance assigned to the association of words (for example, the first word that a subject produces in response to a word that acts as a stimulus), considered evidence for the analysis of the organization of language. In turn, word associations are interpreted concerning the distribution between stimuli and responses in linguistic contexts, while the force of the stimulus-response association is related to the distributional similarity between the two words. From their point of view, the association between words turns out to be a purely lexical phenomenon in which the associative mechanism alone is not enough to explain the cognitive dynamics: this led the psycho-linguists to argue that true semantics reside in conceptual systems and not in the associations between words (Lenci 2008).

The discussed theoretical principles translate into practical application in computational linguistics through two distinct approaches: Topic Modeling and Word Embedding.

Topic Modeling, is a type of statistical model for discovering abstract "topics" that occur in a collection of documents. These are frequently used for the uncover latent thematic structures in a text or in a text collection. Intuitively, since a document is about a particular topic, one would expect that particular words more or less frequently appear in the document: *dog* and *bone* will appear more often in documents about dogs, *cat* and *meow*

will appear in documents about cats and *the* and *is* will appear approximately the same way in both. A document typically covers more topics in different proportions; therefore, in a document that talks about 10% of cats and 90% of dogs, there would probably be about 9 times more dog words than cat words. These algorithms look for groups of similar words. A topic model captures this insight in a mathematical framework, allowing you to examine several documents and discover, based on the statistics of the words in each, what the topics could be and what the balance of the arguments of each document is (D. M. Blei 2012).

Word Embedding, instead, also known as a distributed representation of words, allows storing both semantic and syntactic information of words from an annotated corpus and building a vector space in which word vectors are closest if words occur in the same linguistic contexts, that is, if they are recognized as semantically more similar (Turian et al. 2010).

Despite being two different approaches, Topic Modeling and Word Embeddings have gained popularity for the semantic representation of text (H. Zhu and Lei 2022).

When applying ML and DL techniques, input features must be represented as fixed-length vectors. Methods to extract features from text can be broadly categorized into Topic Modeling and Word Embeddings, each serving different purposes.

Topic Modeling methods aim to extract latent thematic structures from a corpus of text. These techniques analyze word co-occurrence patterns to identify groups of related words that define specific topics. A well-known approach is Latent Dirichlet Allocation (LDA) (Blei et al. 2002), which assigns probability distributions over topics for each document, enabling dimensionality reduction and text categorization. However, these methods do not explicitly encode the semantic meaning of words.

Word Embeddings, on the other hand, generate dense vector representations that capture both semantic and syntactic relationships between words. Unlike topic models, word embeddings position words in a high-dimensional space where similar words are closer based on their contextual usage. This approach gained popularity with Word2Vec (Mikolov et al. 2013), and today, semantic representations of text have become essential in computational linguistics, with advanced techniques such as GloVe (Pennington et al. 2014), FastText (Bojanowski et al. 2017), and BERT (Devlin et al. 2018) providing state-of-the-art solutions for DL applications. A clear example is provided by da Costa et al. (2023), where the authors summarized the literature on this topic. Figure 1.1 illustrates the most effective combinations of classification techniques (white rectangles) and embeddings (rounded rectangles).

The methods that use ML and DL will be discussed in the next section; for now, we will review the works concerning the use of distributional semantics applied to multimedia content classification. Several works have been reviewed in order to understand where the work carried out in this thesis can be positioned.

Some researchers used topic modeling approaches to identify inappropriate content in

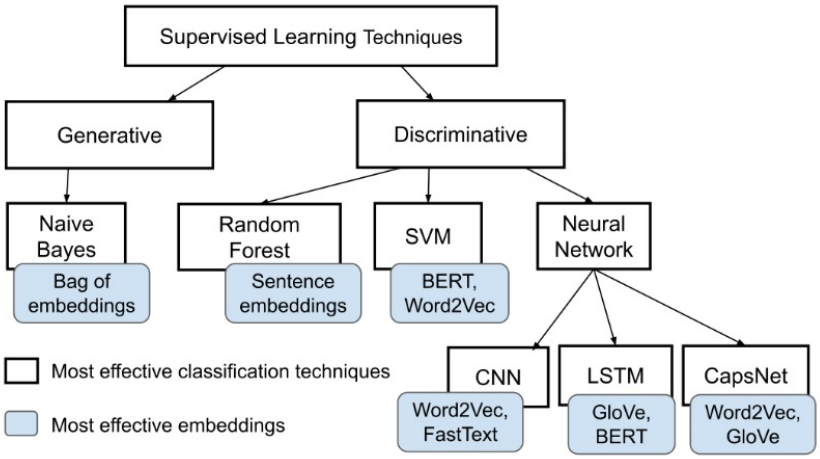


Figure 1.1: Most effective combinations of classification techniques and embeddings identified in the text classification literature, in da Costa et al. (2023)

YouTube videos. Obadimu et al. (2019), for example, applied LDA to detect inappropriate content in YouTube user comments in English; The user comments were given a score between 0 and 1 under five proposed classes: threats, hate, sexually explicit, insults, and identity-based attack. They performed an analysis and visualization for the frequency of the words. However, this only focused on the existing issues and showed the data analysis. Alshamrani, Abuhamad, et al. (2020) studied the detection of toxic behaviors presented in YouTube video comments and new video captions. Their objective was to detect toxic behaviors in comments. Their datasets, consisting of English-language comments and captions, were manually annotated as toxic, safe, hate, and obscene. The pre-processing of comments included transforming words to numerical vectors using the pre-trained Word2Vec, which was also trained on English corpora. The authors used a deep neural network ensemble model to classify comments into three categories: obscene, toxic, and identity hate. The association between toxic behavior and extracted topics from their captions was identified with the topic modeling using LDA. The LDA model achieved a coherence score of 0.55.

Elhoseiny et al. (2016) used distributional semantics to detect zero-shot events within multimedia content. Although this approach leverages distributional semantics for a different purpose than this work, which is focused on text classification, it is based on the same principle of semantic proximity between vectors, in this case, the vectors of the query and the vectors of the video. The authors demonstrated the effectiveness of their work within the TRECVID MED (Multimedia Event Detection) challenge. Using only the event title as a query, their method outperformed the state-of-the-art, which uses long descriptions, showing an improvement from 12.6% to 13.5% in the MAP metric and from 0.73 to 0.83 in the ROC-AUC metric. Additionally, it is an order of magnitude faster. This experiment

was conducted using English-language corpora, specifically the Google News and Wikipedia corpora, to train the distributional semantic models.

In Maisto et al. (2021b) and Maisto et al. (2021a) we focused on the automatic classification of video game transcripts to propose a system of video game rating based on automatic classification of the products performed through the use of the full transcription of dialogues in a video game. The automatic classification, in this case, relied on specialized dictionaries enriched with a vector semantics algorithm. The system was capable of identifying an age rating and a genre classification of video games. In particular, the methodology focused on two tasks: the automatic identification of linguistic indices and the semantic classification of video game transcripts. In this work, as in my thesis work, we decided to focus on the linguistic sphere, abandoning the claim to grasp elements that belong to other spheres. The input data were the video game transcripts in English, which contained the dialogues and description or menu texts. The three indicators we have focused on were: the presence of Bad Language, Violence, and Drugs. Once we extracted the three indices, we performed a semantic expansion of each transcript's frequent words to classify the products automatically. The work included the construction of dictionaries (specifically the Dictionary of Slang, the Dictionary of Violence, the Dictionary of Drugs, and the Dictionary of Discriminatory Terms) and the extraction phase, that started with the matching of dictionary entries in the game transcripts. The values have been normalized, by dividing the number of matches of the respective dictionary entries by the dimension of the text and combined, to calculate Bad Language, Violence, and Drugs indexes. For the semantic classification, we used a distributional semantic matrix to extract similarity values between words and search for the nearest neighbors of a single transcript's most frequent terms. To build the matrix we used COALS (Rohde et al. 2006), a Word-based distributional semantics algorithm, that uses Pearson Correlation over word co-occurrence vectors to calculate the conditional rate of word pairs. The phases of the work include Text pre-processing, Term Frequency-Inverse Document Frequency (TF-IDF), the addition of Slang Terms to the Text-Vectors, and Semantic Expansion of Text-Vectors. After obtaining the vectorized text, we performed a classification task by generating values of similarity between text vectors with the Cosine Similarity algorithm. Our idea was to build a network of documents using similarity values as weighted edges and represent the network with Gephi (Bastian et al. 2009). All the methodology has been tested on a corpus of 50 video game user-generated transcriptions, that we built for that purpose. To calculate the precision of our methodology we compared the indices we extracted from the corpus with indicators of Bad Language, Violence, and Drugs given to a video game from the Rating System. In the conclusion of our works, we underlined that the results we achieved could be considered a good starting point for future analysis and this could be considered the starting point for my research.

A relevant approach in this area is presented by Binh et al. (2022) who developed a classifier called *Samba* to detect inappropriate videos for young children on YouTube. This model integrates subtitles in English and metadata (such as video titles, tags, and comments) for improved classification accuracy. The use of BERT-based word embeddings, allows for more nuanced semantic understanding of the video's content. Their results show that using a combination of these features significantly improves classification performance, achieving 95% accuracy compared to 88% for metadata-only approaches. This highlights the potential for incorporating semantic representations alongside traditional metadata in content classification tasks.

1.3 Traditional and Deep Learning Algorithms for Text Classification

In addition to the unsupervised approach based on distributional semantics, it is also important to understand how ML models perform on this specific classification task. Therefore, understanding the supervised approach becomes essential for the purposes of this work. ML techniques involve a general inductive process that automatically constructs a classifier by learning from a collection of pre-classified documents, discerning the defining characteristics of each category. In this approach, the inductive process builds a classifier for a specific category by analyzing the characteristics of a set of documents manually classified by domain experts. From these discerned characteristics, the inductive process establishes the criteria that a new, unseen document should possess to be classified within that category. The effectiveness of ML relies on an initial corpus of documents, and its accuracy, comparable to that achieved by human experts, is evaluated after implementation. This approach proves valuable in saving time and reducing labor requirements, as it eliminates the need for interventions from knowledge engineers or domain experts (Sebastiani 2002).

As Thangaraj and Sivakami (2018) well explained in their review of the literature on text classification, supervised classification algorithms can be parametric and non-parametric, based on the supremacy of parameters in the data. Among the parametric algorithms, Logistic Regression (LR) and Naive Bayes (NB) are commonly employed:

- LR is useful for estimating the probability of an event and works well with high-dimensional data, especially when the number of features exceeds the number of observations.
- NB is a probabilistic classifier based on Bayes' theorem, often used in text classification due to its ability to handle large, sparse feature spaces efficiently.

Among non-parametric algorithms we can find Support Vector Machine (SVM), Decision

Tree (DT), Rule Induction (RI), k-Nearest Neighbors (k-NN) and Neural Network (NN):

- SVMs are highly effective for high-dimensional text data and are commonly applied in text categorization tasks.
- DTs classify data by making decisions at each node based on feature weights. They are simple to understand but can overfit without proper pruning.
- RI uses a set of rules to classify documents, often leveraging domain-specific lexical patterns to optimize performance.
- k-NN classifies documents based on the majority class of the k closest training examples in the feature space. It relies on similarity measures like cosine similarity or Euclidean distance to compare text representations.
- NN, including various DL models such as Recurrent Neural Network (RNN)s, Long Short-Term Memory (LSTM), and Transformers, are particularly effective for modeling complex, non-linear relationships in text. While RNNs and LSTMs are traditionally used to model sequential dependencies, Transformers have revolutionized text classification by enabling a better understanding of context through self-attention mechanisms, making them ideal for analyzing nuanced content such as subtitles and transcriptions.

Within this thesis, all the described algorithms will be tested on a specifically created corpus of subtitles and transcriptions from online multimedia content. A comparative study will be conducted to determine which approach yields the most effective results for classifying such content.

1.4 Text classification of Multimedia Online Content with Machine Learning

The models described above have been widely used for multimedia content classification tasks, taking text into account, as intended in this thesis. The first surveys on the topic of automatic classification of texts related to online multimedia content were conducted in the early 2000s, and even in these, the possibilities that working on text offers were highlighted. An example is the work of Brezeale and Cook (2008), who in their survey on the automatic classification of videos, underlined the advantages and disadvantages of an analysis connected to each element (audio, video, text). Concerning the text, in the form of closed captions or subtitles, it is contended that this kind of text displays information about other types of sounds such as sound effects (e.g. *[BEAR GROWLS]*), onomatopoeias (e.g *grrr*) and music lyrics. Among the advantages of using text, the authors highlighted the

possibility of accessing the large body of research conducted on document text classification. Another advantage is that the relationship between the features (i.e. words) and specific genre is easy for humans to understand. For example, few people would be surprised to find the words *stadium*, *umpire*, and *shortstop* in a transcript from a baseball game. Among the disadvantages, the authors highlighted the prevalence of dialogues over descriptions and the fact that generating the feature vectors of terms and learning from them can be computationally expensive since the feature vectors can have tens of thousands of terms. Another difficulty in using text-based features is that text has fairly high error rates, especially in the case of real-time generated closed captions for news broadcasts, which may suffer from misspellings and omissions. Although transcription technologies have progressed since 2008 and computers now possess enhanced computational capabilities, transcription-related issues indeed persist. However, these problems can be readily identified and rectified through a brief manual review, or they occur so rarely that they do not undermine the classification process.

Among the very early works, we also find one of the first studies that focused on video-sharing websites, instead of blogs and forums, which is the study by Huang et al. (2010). The authors utilized diverse user-generated data, including titles, descriptions, and comments, aiming to gain insights into cyber communities associated with online videos. The research employed a multi-faceted approach, incorporating three text features: lexical, syntactic, and content-specific. While lexical features alone proved inefficient for video classification, the inclusion of content-specific features significantly enhanced performance across all classification techniques. SVM, with selected features from all types, achieves the best performance. According to the authors, an accurate video classification could have been very useful for identifying implicit cyber communities on video-sharing websites: the interaction between users generates additional text information that could be analyzed. These data often contain explicit information about the video content and can be utilized to classify videos.

Eickhoff and de Vries (2010) presented an automatic method for determining a shared video's suitability for children based on non-audio-visual data. They evaluated its performance on a corpus of web videos that was annotated by domain experts, in particular, they worked on users' comments. Their sentiment analysis was based on the sentiment corpus by B. Pang and L. Lee (2004). In their study, the probability of a specific comment being intended positively was quantified as the mean value of the probabilities associated with its individual constituent terms. The probability of a single term, denoted as t, being positively inclined was subsequently determined as the ratio of occurrences where 't' was observed in a positive comment to the overall instances of t. The analogous score for the negative case is computed as well and both are reported as features. The authors trained a range of ML classifiers (SVM, Random Forest (RF), Decision Table, LR, DT, Ada Boost, MLP, NB) and

evaluated the performance in terms of precision, recall, and the F0.5-measure. Authors assumed that showing as few as possible unsuitable videos to children (high precision) is more important than retrieving all suitable videos (high recall). The work showed that Naïve Bayesian approaches yielded convincing performance but the best-performing model was an SVM classifier. The authors observed that among video information, author information, meta-information, and community-generated information, the latter consistently yielded the most satisfactory outcomes in discerning videos deemed inappropriate for minors. This assumption is also very important in the context of this thesis work.

To look into the specifics of this study and understand how the present thesis aligns with recent developments, it is essential to examine recent works related to text classification in the realm of multimedia content. Specifically, my intention is to focus on studies that aim to determine whether the content is suitable for minors or not. A noteworthy starting point is the work of Alqahtani et al. (2023), which provides a concise overview of the text classification process before delving into the literature.

Alqahtani et al. (2023) conducted a systematic review of the existing methods, techniques, tools, and approaches used to protect kids and prevent them from accessing inappropriate content on YouTube videos. They have taken into consideration research papers that were published between January 2016 and December 2022 in international journals and international conferences. Authors started from the assumption that traditionally, kids around the world like to spend hours in front of the screens, but if traditional ways (such as TV) are easy to control by the government or parents based on kids-friendly program regulations, videos uploaded on modern platforms are more difficult to control. In the last few years, ML, DL, and NLP researchers have contributed to proposing methods, approaches, techniques, and tools to ensure the safety of children on online videos. In their research summarized all the work on the argument, based on the paper's main objective; the focus given to the approaches used (ML or DL or content analysis focusing on statistics); the database construction method (using the YouTube API or other used publicly available datasets); the language and the type of methods or models used in the representation of textual data, the analysis of the meta-data collection / extraction frame or audio conversion and the accuracy of the performance of the model. The authors also conducted a critical analysis of each article. They were able to identify patterns and methodologies common across various works. They concluded that the typical phases for AI-applied techniques, such as ML and DL, include: (1) data preparation, (2) data cleaning, (3) annotation, (4) data pre-processing, (5) feature extraction and representation, (6) model building, (7) model evaluation, and (8) performance analysis. In the data preparation phase, as emerges from their analysis, the data can be collected from targeted YouTube channels using the YouTube API or datasets that are available to the public. The form of the data can be textual, meta-

data of videos, audio, and image frames from YouTube videos. After a cleaning phase, the annotation process is required to label the data into binary classes or multi-classes.

NLP methods and pre-trained models that use word embedding techniques such as Word2Vec and GloVe were utilized in some of the reviewed articles, while others used traditional word representations like Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), which outperformed them in certain cases.

In a ML model, the dataset can be divided into two parts: one for data training to train the model and one for data testing. For the evaluation phase, several metrics can be used to evaluate the model, such as accuracy, area under the curve (AUC), and F1-score. The final phase is the performance analysis, which is used to compare the results and form the discussion part.

From their analysis of the reviewed articles, it was inferred that the classification of multimedia online content, based on text analysis, can be performed by collecting the user comments or converting video manuscripts to textual data through a transcription method. Then, a dataset is built (or a publicly available dataset is used) and annotated. Several NLP techniques can be utilized to analyze the contents, such as NB, LR, DT, RF, k-NN, and SVM.

Many works that used ML for the classification of multimedia content have focused on the analysis of comments: Mouheb et al. (2019) worked on user comments in Arabic from YouTube and Twitter and applied an NB classifier achieving 95% of accuracy in detecting inappropriate information; M. T. Ahmed et al. (2021) and Awal et al. (2018), worked on comments for the Bengali language. M. T. Ahmed et al. (2021) classified the texts of the three datasets into "bullying" and "non-bullying". Their model achieved an accuracy of 76%, 84%, and 80% using the first dataset with SVM, the second and the third dataset with NB. Awal et al. (2018) used a corpus of comments collected from Youtube to classify them into two classes: "abusive" and "non-abusive" and achieved an accuracy of 95%.
Alsubait and Alfageh (2021) used three classifiers: multinomial NB, complement NB, and LR classifiers, to compare the automatic detection of cyberbullying in two different feature representations using TF-IDF and count vectorizer. The authors compared different ML algorithms for their performance in cyberbullying detection based on a labeled dataset of Arabic YouTube comments. The results showed that LR outperforms the other two classifiers in vectorizer, which achieved 78.6% in the F1-score, while complement NB recorded 78.5%.

The works analyzed above, detected inappropriate content based on binary classes. Other authors have worked on identifying ranges. Alshamrani (2020), for example, also worked on comments, in this case in English, but the users' comments were divided into five groups based on age. They found that the group aged 13–17 was the most exposed to inappropriate comments, followed by the 6–8 age group.

Alshamrani, Abusnaina, and Mohaisen (2020) studied the presence of inappropriate content in comments of YouTube videos for kids and teenagers by investigating the existence of malicious and inappropriate URLs inside the comments posted. The authors pointed out the great likelihood that people blindly click on the URL.

Regarding the works on the classification of multimedia content using transcriptions, we find the studies of Alghowinem (2019) and Jagtap et al. (2021). Alghowinem (2019) proposed an advanced automated method to filter YouTube videos for young viewers. The author opted for video content analysis, extracting images, audio, and transcripts from random one-second slices for analysis of content appropriateness. The investigation began by assessing media appropriateness for children, combining multimodal aspects of media. To reduce computational time, the author applied the thin-slicing theory, analyzing several random one-second slices of a video. Real online videos from YouTube Kids, passing the filtering mechanism, were collected to provide an additional layer of safety for kids. The goal was to explore the fusion of three modalities (audio, video, and transcript) for analyzing video content appropriateness for children. In Transcribed Content Analysis, the author referenced various articles (Eickhoff, Serdyukov, and De Vries (2011), Gossen and Nürnberger (2013), Eickhoff, Serdyukov, and de Vries (2010)) that focused on filtering content in web pages. These works highlighted features to consider, including text complexity measures such as readability, language models designed for children, and analysis of children's references (e.g., if the child is directly mentioned). According to these methods, videos passing the YouTube Kids filtering mechanism were collected and annotated into two classes (appropriate and not appropriate for kids). From each extracted slice, image frames, audio signals, and transcribed text were analyzed. The transcribed text from audio slices proved useful for detecting violent, sexual, drug, and alcohol language, as well as hate speech, and for measuring language complexity. These factors serve as indicators of a video's unsuitability for children. Texts underwent pre-processing before extracting features for analysis. To detect violations and extract complex features, the author used large datasets (including annotated hate speech and offensive language, a dataset for adult content filtering, and another for detecting child grooming). The complexity and readability formula were employed to calculate the complexity level of the transcribed language, considering syntactic and lexical measures, such as characters per word, syllables per word, words per sentence, etc. Higher values in these measures indicate greater text complexity for children.

Jagtap et al. (2021) used NLP techniques to extract features from video captions (subtitles). To evaluate their approach they utilized a publicly and labeled dataset for classifying videos as misinformation or not, analyzing captions in English. In their work, they used the Caption Scraper script to download subtitles of the videos present in the dataset. The authors proposed a model for classifying videos into three classes: Misinformation, Debunking

Misinformation, and Neutral. During the experiment phase, they used a set of classifiers from the Python library *LazyPredict* that contains 28 classifiers such as Support Vector Classifier (SVC), k-NN, RF, and XGBoost. To evaluate the performance of these classifiers, they used standard metrics such as F1-score (weighted averaging), AUC-ROC, Precision (weighted averaging), Recall (weighted averaging), and Accuracy. Their analysis showed that they could classify videos among three classes with 0.85 to 0.90 F1-score. To emphasize the relevance of the misinformation class, they re-formulate the classification problem as a two-class classification - Misinformation vs. others (Debunking Misinformation and Neutral). In this case, the model could classify videos with 0.92 to 0.95 F1-score and 0.78 to 0.90 AUC ROC. Authors underlined the limitations of their work: they used embeddings trained with Wikipedia articles, Google News, and Twitter for the word-to-vector representations but they claimed that, in the future, it would be better to a new embedding based on YouTube captions that can be trained using the GloVe algorithm, which might work best for YouTube video classification compared to pre-trained embeddings.

1.4.1 Text classification of Multimedia Online Content with Deep Learning

Most recent studies on text classification of online multimedia content have employed DL techniques such as Convolutional Neural Network (CNN), RNN, LSTM, Bidirectional Long Short-Term Memory (BiLSTM), and hybrid models, as well as Transformer-based architectures like BERT. These models, used individually or in combination with other approaches, have been widely applied to classify textual content associated with multimedia shared online, particularly to protect minors, demonstrating excellent results. For this reason, in this thesis, these models were tested on the specific classification task of determining content age appropriateness.

Among the studies that have addressed the topic of multimedia content classification by testing neural networks, is the work of Martinez et al. (2019), which focused on identifying violence in movies. The authors collected 945 Hollywood movie scripts in English, categorizing them into three groups based on the level of violence: low, medium, and high. They employed an SVM classifier and RNN, achieving a model performance of 60% accuracy. Shafaei et al. (2019) used the Motion Picture Association of America (MPAA)[4] as a reference to classify the content. They worked on the classification of texts related to online multimedia content. In particular, they worked on the prediction of the suitability of the movie content for children and young adults based on scripts. To measure suitability they used the MPAA rating and proposed an RNN-based architecture that jointly models the

[4]The Motion Picture Association of America (MPAA) is an American trade association representing the five major film studios of the United States.

genre and the emotions in the script to predict the rating. Their goal was to classify movies into one of the MPAA ratings based on the content of the movie. To achieve this goal, they used three types of resources: raw data (conversations between characters), emotion vectors of these conversations, and genre vectors of the movies. For the classification model, the authors achieved a weighted F1 score 78%. These studies demonstrate the effectiveness of NNs in transcription classification tasks. The first study addresses a key topic explored in this thesis, namely the detection of violence; however, in my case, I employ a lexicon-based approach for this purpose rather than a ML model. In the case of the second study, both the methodology, using the MPAA as a reference, similar to the AGCOM guidelines used in this thesis, and the consideration of emotional factors are closely aligned with the work conducted in this thesis.

Similar to approaches based on classical classifiers, studies on NNs and DL often prioritize text analysis in multimedia content by focusing primarily on comments rather than video transcripts. When transcriptions are used, they are often limited to binary classifications (appropriate/inappropriate) without considering a detailed age-based categorization, as proposed in this thesis. This underscores the need for research that is focused directly on the content by examining its textual component. However, integrating User Generated Content (UGC) remains an asset, especially given the nature of online shared media.

Among the most relevant studies that have taken comments into consideration, together with other elements, is Anand et al. (2019) that suggested a technique to find inappropriate YouTube videos using DL architectures and ML classifiers. To collect YouTube videos using the YouTube API, they first built an offensive dictionary. Based on the sentiment analysis approach, the system was used to find inappropriate YouTube comments in English and rate them on a scale of 0-100% for how inappropriate they were. Based on keyword searches, they compiled a dataset of more than 20.000 videos, collected all pertinent data, and divided the videos into offensive and non-offensive categories. They conducted several trials utilizing different combinations of SVM, CNN, LR, and LSTM features, including sentiments from comments, thumbnail images, text-based features, and video and frame-level features. The model used GRU and Hierarchical Attention Network with GloVec to get the best accuracy of 86.4%.

Within the body of studies focusing on comments, the work of Mohaouchane et al. (2019) stands out for its research on identifying inappropriate content for Arabic-speaking social media users. The dataset they employed was built by Alakrot et al. (2018), and contains 1050 comments extracted from YouTube videos about Arab celebrities. They were labeled as offensive or not by 3 annotators. They evaluated the performance of four different NN architectures: CNN, BiLSTM, Bi-LSTM with attention mechanism, and a combined CNN-LSTM architecture. The authors trained and tested the system on a labeled dataset of

Arabic YouTube comments. The CNN-LSTM achieves the highest recall (83.46%), the CNN (82.24%), the Bi-LSTM with attention (81.51%) and the Bi-LSTM (80.97%).

These studies highlight that, beyond testing simple NNs, combined approaches can sometimes achieve better results. In this thesis, NNs were not tested in combination with one another; however, the addition of a rule-based approach using a specialized lexicon, similar to Anand et al. (2019)'s offensive dictionary for bad words, brought significant improvements. Furthermore, the study on Arabic raises a fundamental issue addressed in this thesis: the creation of resources for Italian, which remains under-resourced compared to English.

Also leveraging comments Mollas et al. (2022) proposed two dataset of comments in English, to detect inappropriate content: one for binary classification and the other for multi-classes. To assess the proposed datasets they used several ML classifiers such as: LR, SVM, RF, and NB. Additionally, several DL architectures such as CNN and BiLSTM, and transfer learning concepts such as BERT and DistilBERT were also used. DistilBERT achieved the best accuracy (80%). While, Vasantharajan and Thayasivam (2022) examined the detection of offensive content/speech on YouTube based on textual data from multiple languages, specifically Tamil and English. They used a public dataset of user comments on YouTube and categorized it in six classes: non-offensive, offensive-targeted-insult–individual, offensive—targeted–insult–group, offensive—targeted–insult–other, offensive—untargeted, and not-Tamil. They used multiple DL models such as CNN-BiLSTM, mBERT-BiLSTM, DistilBERT, XLM-RoBERTa, and ULMFiT. The best-performing model was ULMFiT, which had a recorded accuracy of 74%. This work is particularly relevant to the present thesis as it also involves analyzing offensive content across multiple languages, highlighting the complexities of multilingual text classification. The use of various advanced DL models, including transformer-based architectures like mBERT and XLM-RoBERTa, shows how multilingual approaches can achieve significant results in content classification.

These studies demonstrate the efficiency of transformers in these tasks, which motivated me to test them in this thesis work. Additionally the work of Vasantharajan and Thayasivam (2022) underscores the importance of a multilingual approach, such as the one presented in this thesis, to gain a broader perspective on the models tested.

Papadamou et al. (2019) developed a classifier able to detect toddler-oriented inappropriate content on YouTube with 82.8% accuracy. They categorized these videos into four groups: suitable, disturbing, restricted (similar to the NC-17 and R categories used by the MPAA), and irrelevant videos. To gather the necessary data, they utilized YouTube APIs and Reddit, resulting in a dataset of 12,097 videos and their corresponding recommended videos (approximately 800). The objective was to train a classifier capable of identifying disturbing content online. A subset of 5,000 videos was manually reviewed and annotated by three annotators. For each video, their model processed: title, tags, thumbnail, statistics

(number of views, likes, dislikes, and comments), and style features (e.g., duration, number of bad words, description length, number of tags). The initial classifier had four branches, each processing a distinct feature type. The outputs of these branches were combined into a two-layer NN for final classification. They focused on determining which feature types contributed most to identifying disturbing videos. The dataset labels were collapsed into two categories: "appropriate" (combining suitable and irrelevant) and "inappropriate" (disturbing and restricted). The study revealed that their DL model outperformed baseline models across all performance metrics. The authors examined the prevalence of inappropriate videos within different subsets of their dataset and analyzed YouTube's likelihood of recommending inappropriate content. Their findings demonstrated that even toddler-oriented search terms could lead to recommendations of inappropriate videos. The study also discussed limitations, such as the small training dataset and the inability to collect YouTube Kids data due to API restrictions. This study is interesting because, although it takes into account various elements, not only textual but also statistical, it still demonstrates the effectiveness of neural network-based approaches. More importantly, it highlights an issue also encountered in this thesis, namely the unavailability of YouTube Kids' APIs. In this work, an effort was made to mitigate the issue: the transcriptions were collected in a semi-automated manner, videos were manually located on YouTube after being identified on YouTube Kids, and the transcriptions were then automatically extracted using a Python scraping script and manually corrected for accuracy.

Many studies on this topic are multimodal, incorporating a limited textual component, typically derived from comments, while NNs are primarily applied to the visual elements. This is the case with the work by Kaushal et al. (2016), which, although it considered the textual component from comments, used CNNs for visual analysis. This work, while not using NNs for text classification, is still relevant as it highlights the importance of analyzing cartoons uploaded to YouTube, a focus shared with this thesis. Furthermore, as in this thesis, the textual component was manually annotated for the presence or absence of unsafe content, such as nudity or abusive language. The authors also emphasize a key point discussed in Chapter 1.1, namely, the limitations of current platform efforts, which still rely on user feedback. These observations underscore the pressing need for automated classification approaches to address this significant issue effectively.

The same issue has been highlighted by Tahir et al. (2019). They observed that unsuitable content for children often escapes control and ends up on children's platforms. Disturbing videos are being uploaded to these platforms (YouTube Kids and others), exploiting the popularity of well-known cartoon characters like Elsa from Frozen, Spiderman, Snow White, etc., without being flagged. The authors developed a DL architecture to flag such videos and report them. Authors manually watched around 5k videos to curate the dataset. They

divided their target content into three categories of fake videos: Benign Fake Videos (BFV), Explicit Fake Videos (EFV), and Violent Fake Videos (VFV) to compare these videos with corresponding real videos. The work of Tahir et al. (2019) focuses on visual content rather than directly on text, but it is useful to consider it within the context of this thesis as it highlights the very issue discussed in Section 1.1, concerning manipulated videos like YouTube Poop. In these cases, integrating a text-based analysis could be particularly valuable for identifying inappropriate content.

Among other studies that also incorporate a visual component, we find the work of Yu and E. Park (2023) and the work of N. Pang et al. (2023). In order to avoid the potential risks of a traditional age-restriction procedure, that involves the biased subjective judgments of creators of multimedia content shared online, Yu and E. Park (2023) proposed automatic and objective multimedia (in particular, they worked on webtoon, digital comics) age-restriction approach based on an automatic prediction model by employing SVM and CNN in investigating whether a specific webtoon content has an age restriction. They performed a multimodal analysis. As regards the text, they extracted primary words (nouns, verbs, and adjectives, which include substantive text meaning) from the descriptions in Korean, by using open-source text segmentation library, 'Okt' of Konlpy (E. L. Park and Cho 2014). They generated a vector of each extracted word with Word2vec technique by using a pre-trained Word2Vec model for Korean, Kor2Vec5, and they observed a similarity between the words. For the multimodal analysis, they utilized the concatenated dense layer of text/image classification. Their CNN classifier achieved an F1 score of 81.88% and an accuracy of 79.09%.

N. Pang et al. (2023), in their study, underscored the significance of text analysis, alongside other elements like video frames, to attain more satisfactory outcomes. The authors argued that processing video frame by frame fell short of meeting the accuracy requirements in specific scenarios. In response to these challenges, they introduced a short video categorization architecture centered around cross-modal fusion in visual sensor systems. This approach jointly leverages video features and text features for short video classification. Firstly, the image space was extended to three-dimensional space–time using a self-attention mechanism, and a series of patches were extracted from a single image frame. Each patch underwent linear mapping into the embedding layer of the Timesformer network and was augmented with positional information to extract video features. Second, the text features of subtitles were extracted through bidirectional encoder representation from the Transformers (BERT) pre-training model. Finally, cross-modal fusion was executed based on the extracted video and text features, resulting in an enhanced accuracy for short video classification tasks (achieving an F1 score of 91.7%).

These studies demonstrate the effectiveness of deep leraning approaches even in mul-

timodal analyses, where the textual component is essential. Although this work does not consider images and instead aims to study the textual part in depth as a starting point, it is clear from the examined literature that integrating both approaches could yield excellent results, paving the way for future developments.

In the recent work by Balat et al. (2024), the issue of the rise of short-form videos on platforms like TikTok and the associated challenges in protecting young viewers from inappropriate content is addressed. The authors introduce TikGuard, a transformer-based DL approach designed to detect and flag content unsuitable for children on TikTok. By utilizing the specially curated TikHarm dataset and employing advanced video classification techniques, TikGuard achieves an accuracy of 86.7%. While their work primarily focuses on frame analysis, the authors conclude by emphasizing the potential of a multimodal approach that incorporates text analysis, offering a more comprehensive understanding of video content. This highlights the continued relevance of text-based approaches, such as the one presented in this thesis, which could be further explored for integration into multimodal analysis.

1.4.2 Text classification of Multimedia Online Content with Large Language Models

In recent years, LLM, such as GPT-3 and its predecessors, have revolutionized NLP, initially designed for text generation tasks. However, their versatility has led to widespread adaptation for a range of NLP tasks, including text classification. LLMs acquire language comprehension and generation capabilities by utilizing extensive volumes of data to train on billions of parameters, requiring substantial computational resources during both their training and operation phases. LLMs are artificial NNs, predominantly based on Transformers, and they undergo (pre-)training through self-supervised and semi-supervised learning methods. These methods operate by taking an input text and iteratively predicting the subsequent token or word (Bowman 2023). Luccioni and Rogers (2023) proposed three criteria that a model must meet to be considered LLM:

- (1) LLMs model text and can be used to generate it based on input context provided as input.
- (2) LLMs receive large-scale pretraining, where 'large-scale' refers to the pre-training data rather than the number of parameters (Luccioni and Rogers (2023) proposed setting it to 1B tokens).
- (3) LLMs make inferences based on transfer learning: LLMs are meant to be adaptable to many tasks, given that their language modeling objective encodes information that can then be leveraged in other tasks.

Before 2020, fine-tuning was the sole method through which a model could be customized to perform specific tasks. However, larger models, such as GPT-3, can now be engineered with prompts to achieve similar outcomes (Brown et al. 2020). It is believed that they acquire embedded knowledge regarding syntax, semantics, and the "ontology" inherent in human language corpora, while also inheriting the inaccuracies and biases present in those corpora (Manning 2022).

LLMs can be broadly divided into three categories based on the model architecture (Sun et al. 2023). The first category is the encoder-only model like BERT and its variants (Devlin et al. 2018) based on a two-stage approach: first, training a language model on masked texts, and then customizing it to perform specific tasks using annotated data. The second category is the decoder-only models like GPT (Radford et al. 2019). GPT uses the decoder of an autoregressive transformer (Vaswani et al. 2017) to predict the next token in a sequence, and follow the pre-training then fine-tuning paradigm. GPT-3 (175B) Brown et al. (2020) proposes to formalize all NLP tasks as generating textual responses condition on the given prompt. The third category is the encoder-decoder models like T5 (Raffel et al. 2020) that are encoder-decoder transformer models, which generate new sentences depending on a given input, following the pre-training then fine-tuning paradigm.

In the context of this thesis, for consistency with the rest of the work, the models were fine-tuned on the specially created corpus, after which metrics were calculated to evaluate their performance. However, in the literature, we find many studies that utilize prompting to test these models even on classification tasks.

As Sun et al. (2023) summarize, the prompt consists of the following three components: (1) Task description that generally describes the task (for example: *Classify the overall sentiment of the input as positive or negative*); (2) Demonstration that consists of a sequence of annotated examples; and (3) Input that is the test text sequence to classify. The models are based on e a progressive reasoning strategy that involves clue collection, reasoning, and decision-making. Clues are local fact evidence such as keywords. The example reported by Sun et al. (2023) is:

> **Input**: *Steers turns in a snappy screenplay that curls at the edges; it's so clever you want to hate it.*
> **Clues**: "*snappy*", "*clever*", "*want to hate it*"

Sun et al. (2023), in their paper, introduced Clue And Reasoning Prompting (CARP) for text classification tasks. CARP adopts a progressive reasoning strategy to address linguistic features such as keywords, tones, semantic relations, references, etc. The model employs a fine-tuned model using supervised data for kNN-based query search within the context of in-context learning. This approach enables the model to leverage the generalization capabilities of LLM as well as the task-specific evidence derived from the fully labeled dataset. Authors

found that CARP delivers impressive abilities on low-resource and domain-adaptation setups and using 16 examples per class, achieves comparable performances to supervised models with 1,024 examples per class.

In the context of content moderation, especially concerning minors, the precision of classification models is paramount to prevent exposure to inappropriate material. Fine-tuning LLMs on domain-specific data enhances their ability to accurately identify and filter such content. Ma et al. (2023) demonstrated that even models based on prompting for text classification can be improved through a fine-tuning process, as applied in this thesis work. Specifically, they explored a combined approach to content moderation for web content using LLMs, which involves an initial light fine-tuning followed by advanced Chain of Thought (CoT) prompting, a method where the model is guided through a sequence of structured reasoning steps to make decisions in a more logical and interpretable manner. Their method starts with a preliminary fine-tuning phase to help the model become familiar with the content domain, followed by CoT prompting, which encourages thoughtful decision-making and minimizes the use of oversimplified reasoning. The use of CoT prompting enhances both accuracy and interpretability while reducing the risks of overfitting, without requiring extensive data annotation.

Additionally, Z. Li et al. (2023) explored the potential and limitations of LLMs specifically in text classification, and they also noted that fine-tuning strategies increase the model's ability to adapt to task-specific classification data. Furthermore, they observed that data augmentation (i.e., increasing the dataset synthetically) can improve the performance of these models, which tend to work better with very large datasets. However, they caution that attention must be given to the potential amplification of any biases present in the training data.

These findings underscore the importance of fine-tuning LLMs with domain-specific data to enhance their precision in sensitive applications like content moderation for minors.

Given their potential in text understanding and categorization, LLMs are increasingly used for automatic classification of multimedia content. It is useful to examine how they have already been applied to the task of multimedia content classification for child protection, to understand where this thesis work fits within the existing research.

S. H. Ahmed et al. (2023) in their paper evaluate the performance of CLIP (Contrastive Language-Image Pre training), on both supervised and zero-shot settings for video content moderation, examining how context-specific language prompts affect performance, especially in cases where harmful content may mimic popular cartoon styles. This method emphasizes the use of foundation models in a vision-language framework, where prompt engineering plays a key role in refining the model's ability to identify inappropriate material. The study concludes by highlighting the limitations of prompt-based moderation for nuanced content

like children's cartoons, noting the difficulty in verbally defining inappropriate content and suggesting the potential of learnable prompt parameters for future improvements. However, the approach adopted in this thesis addresses this challenge from a different angle by leveraging fine-tuning for a deep linguistic understanding of text content, which may be less constrained by the challenges of visual prompting and more adaptable for nuanced, age-specific content classification.

1.5 Analysis of emotions

In the domain of textual analysis, the inherent capacity of words to express specific emotions can be utilized for the purpose of classification. As explained by Saif M Mohammad and Turney (2013b), words are associated with emotions. For example, the words *delightful* and *yummy* indicate the emotion of joy. Thanks to taxonomies theorized by psychologists, we can distinguish emotions in simple and complex ones. Ekman (1992) and Plutchik (1984) have theorized that there are some basic emotions. Many of the basic emotions are also instinctual (Saif M Mohammad and Turney 2013a). Ekman (1992) argues that there are six basic emotions: *joy*, *sadness*, *anger*, *fear*, *disgust*, and *surprise*. For Plutchik (1984) the basic emotions are eight, to those theorized by Ekman, he adds *trust* and *anticipation* and organizes the emotions in a wheel where the radius indicates intensity—the closer to the center, the higher the intensity (Saif M Mohammad and Turney 2013a). The work of Saif M Mohammad and Turney (2013a) builds on the theory of Plutchik (1984). According to this theory, the eight primary emotions can be categorized as positive (four) and negative (four). Each positive emotion has a corresponding negative counterpart: joy is opposed to sadness, surprise is opposed to anticipation, trust is opposed to disgust, and anger is opposed to fear. In categorizing emotions, Plutchik analyzed each emotion in detail and divided them into subgroups that contain secondary and tertiary emotions, giving rise to a wheel-like mechanism to describe the relationship between emotions, their intensity, and polarity. Specifically, the intensity of an emotion is high when it is located at the center of the wheel and decreases as the distance from the center increases (Abbasi and Beltiukov 2019).

In recent years, computational linguistics has seen a rising interest in feelings and emotions. Many works have been made on the development of novel methods to automatically classify the emotions expressed in a text and many lexical resources have been built (Passaro et al. 2015).

Strapparava, Valitutti, et al. (2004) developed the WordNet Affect Lexicon (WAL). The lexicon contained a few hundred words manually annotated with the emotions they evokes. Their system was based on the idea that each WordNet synonym had the same emotion. In this case, the theoretical basis from which the author started is that of the six emotions of

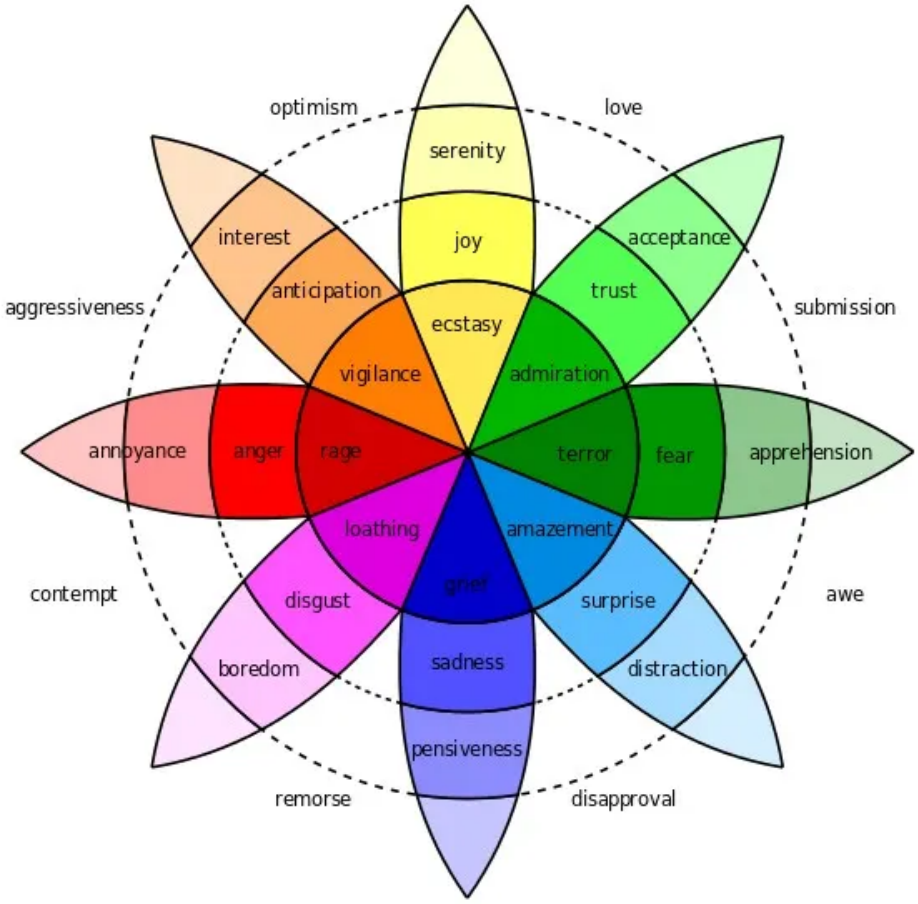


Figure 1.2: Plutchik's Wheel of Emotions

Ekman (1992).

Saif M Mohammad and Turney (2013b) presented a large word–emotion association lexicon, initially called the NRC Affect Intensity Lexicon (NRC-AIL). This earlier version (v0.5) included intensity scores for only four primary emotions: *anger*, *fear*, *sadness*, and *joy*. Later, in S. Mohammad et al. (2018) the lexicon was expanded to include intensity scores for eight primary emotions: *anger*, *fear*, *anticipation*, *trust*, *surprise*, *sadness*, *joy*, and *disgust*, and was renamed the NRC Emotion Intensity Lexicon (version 1) to better reflect its alignment with the NRC Emotion Lexicon. This lexicon provides real-valued scores between 0 and 1 for words associated with these emotions.

To build the corpus, Saif M Mohammad and Turney (2013b) used Roget's Thesaurus[5] as the source for terms. The thesaurus is in the public domain. They annotated only those words that occurred more than 120,000 times in the Google n-gram corpus[6]. Each

[5]Roget's Thesaurus: http://www.gutenberg.org/ebooks/10681

[6]The Google n-gram corpus is available through the Linguistic Data Consortium: https://www.ldc.

term was annotated by five different people; for 74.4% of the instances, all five annotators agreed on whether a term was associated with a particular emotion or not. For 16.9% of the instances, four out of five people agreed with each other. The data from multiple annotators regarding a given term was consolidated through a majority vote mechanism. The lexicon encompasses approximately 24,200 word-sense pairs. Information from various senses of a word was consolidated by merging all emotions associated with the diverse senses of the words.

The corpus has had several applications, including: tracking sentiment towards politicians, movies, products, countries, and other target entities (Saif M Mohammad et al. 2013; B. Pang and L. Lee 2004) and depicting the flow of emotions in novels and other books (Boucouvalas 2002; Saif M Mohammad 2012).

In the experimental part of my thesis, I will work on the expansion of dictionaries for Italian: starting from the NRC's dictionaries, considering that the authors themselves have highlighted the limitations of dictionaries in languages other than English (Saif M Mohammad 2020), I decided to expand them by extracting synonyms from Wordnet and assigning them, for each emotion, the same values as the starting word.

In Maisto et al. (2021a) and Maisto et al. (2021b), we worked on the automatic classification of video game scripts, with an automatic classification algorithm able to provide an age rating and a genre classification of video games, based on large, specialized dictionaries, semantic vector spaces, and sentiment analysis. Emotions have been processed through the exploitation of the NRC Affect Intensity Lexicon. In that case, the original list of words, available for both the English and the Italian language, has been lemmatized and inflected. The enriched lexical database contained more than 20,000 entries (9,921 English and 10,334 Italian lemmas). Specifically have been taken into account only the negative emotions (fear, anger, and disgust), since these indicators, combined with the other dictionaries, can produce a value that can be used as an index of the presence/absence of specific features inside the texts.

Another relevant approach in the field of emotion classification is DEGARI (Dynamic Emotion Generator And Reclassifier), introduced by Lieto et al. (2021). This system integrates commonsense reasoning and conceptual blending with Plutchik's emotion model and the NRC lexicon for emotion detection and classification. DEGARI distinguishes itself from lexicon-based approaches by leveraging structured knowledge-based reasoning to infer complex emotional states rather than relying solely on word frequency and predefined lexical entries. It dynamically generates new commonsense representations of compound emotions (e.g., Love as a combination of Joy and Trust, as defined by Plutchik) and applies them to content classification tasks. DEGARI has been successfully used to reclassify emotion-

upenn.edu

related contents in artistic domains, including datasets from art collections and multimedia platforms such as RaiPlay, the Italian national broadcaster's digital content service. The system's explainability component allows for a transparent assessment of its classifications, making it a valuable tool for contexts where human interpretability is crucial.

While DEGARI represents a state-of-the-art system in emotion analysis, its methodology differs from more traditional lexicon-based approaches, which rely on pre-annotated word lists and frequency-based scoring. In contrast, DEGARI enriches sentiment analysis through ontology-driven reasoning and conceptual combination techniques.

Mohammadi et al. (2023) used deep neural network models to accurately predict the range of human emotions experienced while watching movies. They focused on visual, and auditory components, speech, and music, and also linguistic elements encompassing actors'dialogues. To extract linguistic features, authors used BERT, finding that language significantly influences experienced arousal (the intensity or excitement level of an emotion), while sound emerges as the primary determinant for predicting valence (whether an emotion is positive or negative). In contrast, the visual modality exhibits the least impact among all modalities in predicting emotions. These findings support the central idea of my thesis: that text is a fundamental aspect to be leveraged in an automatic classification system for online multimedia content. In their case, bidirectional linguistic semantic data was extracted from the subtitles of each movie. They applied a pre-processing step to remove stop words and employed BERT and Word2Vec to embed each word from the subtitles and understand the relationships between words, the meaning of sentences, and the structure of the subtitle text. The authors inspected diverse combinations of modalities (video, sound, and text), and each modality was treated as a separate input to the architecture in this way the authors have obtained information on the effectiveness of each modality in capturing and predicting emotions. As said, text features provide a classification accuracy for arousal emotion labels higher than the other modalities (image and motion) confirming that text features, such as the semantic content derived from subtitles, play a more important role in conveying information related to arousal emotions.

A methodology for creating Italian-specific resources that is scalable and therefore applicable to new languages and expandable was presented in Passaro et al. (2015) where the authors proposed an emotion lexicon for Italian called ItEM. In their lexicon, each target term was provided with an association score with eight basic emotions. ItEM was not a static lexicon but provides also a dynamic method to continuously update the emotional value of words, as well as to increment its coverage. The lexicon required a minimal use of external resources to collect the seed terms, a little annotation work to build the lexicon; and its update is mostly automatized.

Another interesting work worth mentioning is the work of Kant et al. (2018), who have

demonstrated that large-scale unsupervised language modeling combined with fine-tuning can offer a solution to multi-emotion sentiment classification on difficult datasets. Their model achieved a 0.69 F1 score on the SemEval Task 1:E-c multi-dimensional emotion classification problem (S. Mohammad et al. 2018)[7], based on the Plutchik wheel of emotions (Plutchik 1984).

A focused study was conducted on works related to the understanding of emotion dyads that is combinations of emotions (Rachele Sprugnoli et al. 2020). Although emotion dyads were not included in this specific thesis work, they could be considered for future developments, based on the research described in Section 0.4. Given the relative lack of research on emotion dyads, as highlighted by Saif M Mohammad and Turney (2013a), exploring this area could help address an existing gap in the literature and provide valuable insights for more nuanced emotion classification models.

Pearl and Steyvers (2010) focused on politeness, rudeness, embarrassment, formality, persuasion, deception, confidence, and disbelief by developing a game-based annotation project; Belainine et al. (2020) worked on complex emotions, in their paper proposed a new dataset for complex emotion detection using a combination of several existing corpora to represent and interpret complex emotions based on Plutchik's theory. They proved that a rich emotional corpus facilitates the detection of complex emotions. Their goal was to create an emotion classifier capable of detecting complex emotions based on implicit rules to detect either an emotion like joy or sadness but not both at the same time. The process consisted of three phases: (1) collection of annotated data; (2) detection of basic emotions and representation with a four-dimensional emotion vector and (3) learning and interpretation of complex emotions using multi-label classification. All eight basic emotions according to Plutchik's theory are considered, plus three complex emotions: *love* (Joy + Trust), *guilt* (Fear + Joy), and *shame* (Fear + Disgust).

Even if there is no direct correlation between the emotions and the age to which multimedia content refers, in this thesis the analysis of emotions is useful because the identification of the emotions present in the text, especially negative emotions, could be additional information that allows you to identify the presence or not of certain elements. For example: splatter (disgust), frightening (fear), or elements of violence (anger) are also reported in the AGCOM resolution (2017) as determining elements of the target audience of a given content. In the annex to the resolution, which explains the meaning of *threats*, is reported as follows: *per "minacce" deve intendersi la presenza di contenuti suscettibili di includere nello spettatore stati di tensione emotiva, paura, angoscia derivate dalle dinamiche strutturali del racconto* ("threats" must mean the presence of content likely to include in the viewer states of emotional tension, fear, anguish derived from the structural dynamics of the story). These

[7]The SemEval Task 1:E-c problem (S. Mohammad et al. 2018) offers a training set of 6,857 tweets, with binary labels for the eight Plutchik categories, plus Optimism, Pessimism, and Love.

indicators, combined with other dictionaries, can produce a value that can be used as an index of the presence/absence of specific characteristics within the texts.

1.6 Previous Research on Text Complexity

For assessing age-appropriate content, it is valuable to review studies examining the complexity of text associated with multimedia content shared online. Text complexity serves as an additional factor enhancing automated classification systems aimed at determining the suitable age group for specific content. Importantly, even when texts related to videos lack explicit inappropriate language, they may be unsuitable for young audiences due to inherent complexity.

In relation to assessing the suitability of multimedia content for children Eickhoff and de Vries (2010) conducted a study on the suitability of websites for children. Their automatic classifier reached a performance close to that of human agreement. The dataset for the work was acquired using the Open Directory Project[8], which is an Internet directory created especially for children and teenagers. It includes both sites designed specifically for children and/or teens as well as sites designed for general audiences. In this Internet directory, the Kids & Teens section contains web pages annotated by human annotators according to the following criteria: kid (12), teen (13-15), and mature teen (16-18). The authors extracted a training set of 20,778 web pages (6,225 for children and 14,553 for adults) from a range of 1,350 distinct topics. Authors expected child-friendly web pages to rely on language use and general design of textual resources that respect these specific aspects: shallow features as a measure of syntactic text complexity, examples of features in this category are the average number of words per sentence, the number of complexes (3+ syllables) words, or the average word length; readability scores; part-of-speech features since there are syntactic differences between adult and children's texts on a higher linguistic level than could be captured by mere shallow features; entity occurrences because children's text also semantically simpler than general text; OOV rates since child-friendly websites have simpler text structure in terms of shorter sentences containing less named entities, but also the use of more basic language. To reflect this difference, the authors constructed 7 distinct vocabularies of the most frequent/most basic English words since they expected that the texts for children commonly use a smaller and simpler range of words, they also used an additional vocabulary of academic terms for which the opposite tendency is expected; Wiktionary features, according to the authors, also are a possible way of capturing textual complexity such as the ambiguity of the words; Presentation since child-friendly websites should be presented in a way that is appealing to children for example with the use of videos, images and colours; HTML features since scripts were expected to be an indicator of child-friendliness; visual features

[8]Directory of Kids and Teens Resources in http://odp.org/Kids_and_Teens/

since child-friendly web pages often rely on great numbers of images to convey their message; navigational given that research in the topical classification of web pages has found that a page's neighbors are strong indicators of its topic, moreover children prefer to explore and browsing over searching; link neighborhood features in order to analyze a web page's outgoing links. Authors also do ethical considerations, in particular as regards advertisement because, as stated in the ODP's content guidelines, child pages should not be commercial.

Gossen and Nürnberger (2013) in a work about the importance of understanding how to design information retrieval systems for children, described the basic theories of human development to explain the difference between young and adult users from a point of view cognitive, emotional, and knowledge. When it comes to children it is important to consider that there is a need to target very narrow age groups, children, are human beings whose cognitive abilities are not fully formed. Piaget and Inhelder (1972) and after Ormord and Davis (1999) have reflected on the differences in human abilities at different age stages, claiming that human development occurs in sequential order and later abilities and knowledge are built upon the previous ones. Piaget has theorized four developmental stages: sensorimotor (0-2 years) during which the child begins to recognize cause-and-effect relationship but is not yet able to think about objects other than those directly in front of it (Ormord and Davis 1999), pre-operational (2-7 years) during which child learn to use language and have illogical thinking and difficulties in classification, concrete operational (7-11 years) during which children use trial and error approach and begin to think logically, their understanding is eliminated but can classify objects (Ormord and Davis 1999) and formal operational (from 11 years on) during which learn to think logically about abstract concepts and learn about hobbies. Gossen and Nürnberger (2013) specifies that the neo-Piagetian explained cognitive growth from an information processing perspective. They underlined that children's information processing requires much more memory work. As the child grows up, he acquires new information by completing various tasks and therefore an older child needs less time to process the information.

In this regard, as Gossen and Nürnberger (2013) also pointed out, it was useful to create child-safe content, it can be done automatically with ML techniques such as classification to identify inappropriate content. It must be considered that text or web documents can be written using language complexity, so providing safe content is not enough: children should also be able to understand the meaning (Gossen and Nürnberger 2013).

In the context of this thesis, text complexity was addressed in a preliminary experiment, applying an initial metric to evaluate vocabulary frequency through resources like the *Nuovo vocabolario di base* (NVdB) (New Basic Vocabulary) of De Mauro for Italian and the Oxford 3000 for English. While this approach offers valuable insights, it has limitations, as it considers vocabulary alone, without accounting for syntactic factors. Future work could en-

hance this metric by developing a value-age association scale to better align text complexity with age-specific needs. This thesis builds on established research, laying the groundwork for refined complexity assessment methods tailored to age-appropriate multimedia content.

Chapter 2

Methodology and resources

2.1 Rating of multimedia content

The methodology of this research is centered on a comprehensive evaluation of the performance of various advanced classification methods, including traditional classification algorithms, NNs, Transformers, and LLMs. The research also focuses on the automatic identification of linguistic indices and a semantic classification of texts related to multimedia content shared online. The research was conducted in English and Italian.

The goal of evaluating the classifiers on this specific task, along with the lexical and semantic analysis, is to determine the best method, or combination of methods, to classify multimedia content based on the age of the intended audience. The results of this research could potentially be applied in fields such as content moderation. Furthermore, the study conducted on the textual component of the content could be integrated with other elements, such as images or frames, to develop multimodal classification systems.

The primary focus is always to protect minors and prevent them from being exposed to inappropriate content. As we saw in Chapter 1, content moderation is a field where there is still little legislation and a lack of clarity from platforms. In fact, these platforms often leave it up to the users themselves to classify their content, leading to classification errors, such as the case mentioned in the previous chapter regarding the "Elsagate" incident.

The first step of the methodology was to build dictionaries for the extraction of indices. The second step involved creating a corpus of texts for training the models to be tested. The third step focused on the annotation of the corpora, followed by the fourth step, which concerned the evaluation of the classifiers. The fifth step was related to semantic analysis, and the final step concerned lexical analysis.

In this methodology, the input data are multimedia content transcripts, which mainly contains the dialogues and, in some cases, description of what is happening.

2.2 Classification algorithms evaluation

The methodology adopted for the evaluation of the classifiers was based on a comparative approach aimed at identifying the best method, or combination of methods, for classifying texts related to multimedia content. The ultimate goal of the classification was to determine the appropriate age group for the intended audience to prevent minors from being exposed to content that was not suitable for their age.

Metrics such as accuracy, recall, and F1-score were selected to evaluate the performance of the models, taking into account the specific challenges posed by the linguistic data used.

Given the focus of this research on linguistic data, the selection of classifiers was informed by a comprehensive review of the state of the art, focusing on approaches that had been successfully applied in similar tasks.

For traditional ML models (e.g., SVM, NB, DT), pre-processing steps included stop-word removal using appropriate modules from the `NLTK` Python library, tokenization, and text normalization (conversion to lowercase and removal of punctuation). Texts were then represented numerically using the TF-IDF (Term Frequency-Inverse Document Frequency) method. The `TfidfVectorizer` class from the `scikit-learn` library was used to compute term frequencies (TF) and apply an Inverse Document Frequency (IDF) transformation to reduce the weight of terms common across multiple documents. The resulting vectors were normalized before being used as input for the models.

For NNs, the pre-processed texts were tokenized into sequences of integers using the *Keras* Tokenizer, which assigned unique indices to words based on their frequency. These sequences were standardized to a fixed length and passed through an embedding layer that converted them into dense vector representations. These embeddings were subsequently processed by convolutional layers in CNNs or recurrent layers in RNNs, such as LSTMs, to capture patterns and relationships in the text.

For Transformers and LLMs, tokenization and text representation were managed using pre-trained tokenizers specific to each architecture, leveraging the models' inherent ability to process linguistic data effectively.

To evaluate the models, k-fold cross-validation was employed with k set to 5. The dataset was divided into five equal parts, with the model trained on k-1 parts and tested on the remaining fold for each iteration. This process was repeated five times, and the results were averaged to provide a reliable estimate of model performance.

Bad Words Rule

To improve classification accuracy and adhere to AGCOM guidelines, a lexicon-based rule was implemented.

Specifically, if a Bad Word was detected in a text labeled as "for kids", the label was

revised to "for adults". This rule was consistently applied across all tested models to ensure accurate categorization. As shown in the experimental results (Chapter 3, Table 3.1), this approach yielded improved performance across all classifiers.

The Bad Words lists for both English and Italian were derived from Discriminative-terms dictionaries, described in Section 2.4, applying a selection methodology that excluded low-impact expressions as well as more generic or regionally-used insults. Through this process, words with a distinctly offensive or discriminatory character were identified, while terms with colloquial usage or lower levels of intensity were excluded.

This filtering process confirmed that the Bad Words list is indeed a refined and concentrated version of the broader insult dictionary, meeting the criteria for focusing on highly offensive content. The selection was carefully conducted to avoid excessive exclusion, ensuring that only terms genuinely regarded as Bad Words were retained after thorough manual review. This approach prioritized terms with a clear offensive or discriminatory nature, creating a list centered on highly impactful language without losing essential entries due to overly restrictive filtering criteria.

The final Bad Words list, therefore, focuses on direct, high-intensity insults, which are most relevant to this research.

The list includes a total of 305 entries for English and 459 for Italian. This difference reflects the distinct lexical and cultural nuances of each language.

2.2.1 Support Vector Machine

The Support Vector Machine[1] (SVM) was chosen as one of the classifiers to be tested for this research. SVM is a supervised ML model that finds the optimal hyperplane that maximizes the margin between different classes. The margin represents the distance between the separating hyperplane (decision boundary) and the closest training examples, known as support vectors (Raschka and Mirjalili 2020).

For this study, SVM was extended to handle multi-class classification problems using the One-vs-One (OvO) approach, which involves training multiple binary classifiers for each pair of classes. This method allows the SVM to handle classification tasks with more than two classes. A total of three binary classifiers were trained: (1) distinguishing between kids and teenagers, (2) distinguishing between kids and adults, and (3) distinguishing between teenagers and adults. When classifying a new data point, each of the three binary classifiers provides a vote for one of the two classes it distinguishes. The final classification of a new data point was based on the majority vote of these classifiers. This voting mechanism ensures that the classifier can make a robust decision by leveraging the combined strengths

[1]Support Vector Machine. In https://scikit-learn.org/stable/modules/svm.html#classification, last accessed: 2024/06/26.

of multiple binary classifiers.

To visualize and explore the distribution of the different categories in the dataset, Principal Component Analysis[2] (PCA) was used as a dimensionality reduction technique. PCA helps in emphasizing variation in the data and highlights the patterns by projecting the dataset into a lower-dimensional space.

Support Vector Machines (SVM) are also popular because they can be kernelized. A kernel is a function that transforms the data into a higher-dimensional feature space to make the classes linearly separable. These methods are used when a linear classification is not possible. When this is the case, the training data need to be transformed into a higher-dimensional feature space through a mapping function. To save computational costs, the kernel trick is used. Specifically, a kernel function is defined, with the most commonly used being the Radial Basis Function (RBF) or Gaussian kernel. The parameter gamma determines the cut-off of the Gaussian sphere (Raschka and Mirjalili 2020).

To ensure optimal performance, hyperparameter tuning was performed using `GridSearchCV` with a 5-fold cross-validation, exploring different values for key parameters such as C (regularization strength) and kernel type. The process involved systematically searching for the best combination of hyperparameters based on validation accuracy.

The chosen hyperparameters for the SVM model were:

- $C = 1$ (regularization strength)
- Kernel type = linear

These parameters were selected because they offered the best trade-off between model complexity (bias) and generalization (variance), ensuring balanced performance across the dataset.

2.2.2 Naive Bayes Classifier

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable (H. Zhang 2004).
Bayes' Theorem can be written as:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

where:

- $P(C|X)$ is the posterior probability of class C given the feature vector X.

[2]Principal Component Analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It's often used to make data easy to explore and visualize. In: https://setosa.io/ev/principal-component-analysis/, last accessed: 2024/06/27.

- $P(X|C)$ is the probability of the feature vector X given it belongs to class C.
- $P(C)$ is the prior probability of class C.
- $P(X)$ is the probability of the feature vector X.

For this study, the Multinomial Naive Bayes algorithm was used, which is suitable for discrete data, making it appropriate for text classification where the features represent word occurrences. This algorithm calculates the probability of each word appearing in a given category (e.g., kids, teenagers, adults) and uses this information to classify new texts. In this case, the texts were divided into three categories, and the model computed the probabilities of the words for each class to make predictions.

Hyperparameters were tuned using `GridSearchCV` with a 5-fold cross-validation, ensuring that the best configuration was selected based on validation accuracy.

2.2.3 Decision Tree

Another algorithm that was chosen to be tested as part of this thesis is the Decision Tree (DT). DTs are a non-parametric supervised learning method used for classification and regression. The main objective is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The model works by recursively splitting the dataset into subsets based on the feature that provides the best split according to a metric like Gini impurity or entropy. The splitting process continues until all the leaves of the tree represent a pure class, meaning that all training examples at a leaf node belong to the same class (Breiman 2017; Raschka and Mirjalili 2020).

For this study, the DT algorithm was used to classify texts into the three categories: Kids, Teenagers, and Adults. The tree iteratively selected the features (words) that most effectively separated the classes in the training dataset. The model evaluated each feature at every node based on metrics like Gini impurity or entropy, and chose the one that resulted in the greatest reduction of impurity at each node. This recursive process continued until the entire decision tree was constructed.

Additionally, hyperparameter tuning was performed using GridSearchCV with a 5-fold cross-validation. The key parameter tuned in this study was the maximum depth of the tree. This tuning process ensured that the model did not overfit or underfit the data by controlling the depth of the tree. GridSearchCV systematically searched for the best configuration of hyperparameters to maximize validation accuracy across the dataset.

2.2.4 Random Forest

Among the various *ensemble* models, the Random Forest algorithm was selected for this task due to its ability to reduce overfitting, which is a common issue with individual decision

trees. Random Forest can be considered an "ensemble" of Decision Trees, where the idea is to average the results of multiple trees to create a more robust model that generalizes better to new data. This ensemble method is less sensitive to overfitting, as it reduces the variance that individual trees might suffer from, thus improving generalization performance overall (Raschka and Mirjalili 2020).

During the training of a RF, a random sample with replacement (bootstrap sample) is drawn from the original training dataset to construct each decision tree. This method, known as bootstrap aggregation or bagging, ensures that each tree is trained on a slightly different dataset, which helps to decorrelate the trees. The second step involves growing the tree from the bootstrap sample. At each node, a subset of features is randomly selected, and the feature that provides the best split according to an objective function, such as maximizing information gain, is chosen to split the node. This process is repeated across multiple decision trees, with the final prediction being determined by a majority vote across all trees (Raschka and Mirjalili 2020).

In this study, hyperparameter tuning was performed to optimize the performance of the RF model. The primary hyperparameter of interest was the number of trees, k, which was set to 300 for English and 200 for Italian, based on `GridSearchCV` results. Other hyperparameters, such as the number of features randomly selected for each split and the size of the bootstrap sample, were set to default values in `scikit-learn`. Specifically, the size of the bootstrap sample was equal to the number of training examples in the dataset, and the number of features selected at each node was calculated as is $d = \sqrt{m}$, where m, where m is the total number of features in the dataset (Raschka and Mirjalili 2020).

2.2.5 K-Nearest Neighbor

The k-NN algorithm is distinct from the other classifiers tested in this research because it is a "lazy learner." Unlike models that learn a discriminative function from the training data, k-NN simply memorizes the dataset and classifies new instances based on similarity to the memorized data. It is categorized as a non-parametric model, meaning it is not characterized by a fixed set of parameters, as discussed in 1.3. k-NN is also known as an instance-based learner, where the training process incurs zero cost, as no explicit model is built. Instead, the model defers the classification decision until new data is introduced (Raschka and Mirjalili 2020).

The functioning of k-NN can be summarized into three key steps:

1. Select the number of k and a distance metric.
2. Identify the k nearest neighbors of the point to be classified within the training data.
3. Assign the class label to the point based on a majority vote.

The model classifies new points by finding the k most similar examples in the training set and assigning the class label that occurs most frequently among these neighbors (Raschka and Mirjalili 2020).

For this study, the optimal hyperparameters were selected using `GridSearchCV`. Specifically, the number of neighbors (k) chosen was 7 for English and 9 for Italian. The model used a `'uniform'` weighting scheme, meaning that all neighbors contribute equally to the prediction, without giving more weight to those closer to the query point.

2.2.6 Neural Networks

In addition to the previously mentioned models, NNs were also selected for testing, specifically CNNs and LSTM networks.

A NN is a ML model partially and functionally inspired by the human brain (on the differences between biological and artificial NN models and mechanisms see Lieto (2021)), using interconnected layers of artificial neurons to process data and make decisions. Each NN consists of layers of nodes, or artificial neurons: an input layer, one or more hidden layers, and an output layer. Each node connects to others and has an associated weight and threshold. When the output of a node passes through an activation function and exceeds the threshold, the node is activated and data is passed to the next layer[3].

Although CNNs are commonly used for image analysis, they can also be applied to text classification, as demonstrated in this thesis.

The CNN model was implemented using the `TensorFlow` and `Keras` libraries. `TensorFlow` is a scalable, multi-platform programming interface for implementing ML algorithms, while `Keras` is a high-level NN API that offers a user-friendly and modular interface for rapid prototyping of complex models (Raschka and Mirjalili 2020).

Unlike simpler ML models, word representation in this case was performed using an embedding layer. Specifically, each word in the text was converted into a numerical vector of 64 dimensions, as defined in the embedding layer. This enables the NN to capture more nuanced relationships between words, going beyond simple frequency-based representations.

For this thesis experiment, the maximum vocabulary size was set to 10,000 words, meaning that only the 10,000 most frequent words in the dataset were considered, while all others were ignored. Additionally, each text sequence was either padded or truncated to a length of 100 tokens to ensure uniform input size across all texts. Padding, in this case, involves adding zeros at the beginning of sequences, ensuring that even the shortest sequences are 100 tokens long, preventing the loss of important information from the edges of the sequence (Raschka and Mirjalili 2020). This setup strikes a balance between computational efficiency

[3] *Cos'è una rete neurale?* in https://www.ibm.com/it-it/topics/neural-networks, last accessed: 2024/09/23.

and the ability to capture meaningful patterns in the text data.

After the embedding layer, the text data is passed through a convolutional layer, where filters are applied to the numerical vectors to detect significant patterns within sequences of words.

Given that the data is sequential in nature, specifically textual data, a `Conv1D` layer was chosen for the convolutional layer. `Conv1D` is particularly well-suited for one-dimensional sequences, such as sequences of words or characters in text. This layer effectively captures local patterns, such as common phrases or word combinations, by applying convolutional filters along the sequence.

A moderate number of filters (64) were used to balance capturing sufficient features while maintaining computational efficiency. The `kernel_size` was set to 3, allowing the model to capture patterns across small groups of words, a common approach in NLP tasks. The `ReLU` (Rectified Linear Unit) activation function was used to introduce non-linearity, enabling the model to learn more complex patterns from the data.

To reduce computational complexity, a `GlobalMaxPooling1D` layer was employed. Pooling layers, or subsampling, reduce the dimensionality of the input by selecting the most important features[4]. Specifically, `GlobalMaxPooling1D` selects the most important features from the entire text sequence by taking the maximum value, unlike `GlobalAveragePooling`, which averages values. Furthermore, unlike regular MaxPooling, which operates on small regions of text, `GlobalMaxPooling` applies to the entire sequence (Raschka and Mirjalili 2020).

The final layer of the model is a Dense layer with 3 neurons, each representing one of the target classes. A softmax activation function was applied to transform the raw output into a probability distribution. The softmax ensures that the sum of the probabilities across all three classes equals 1, with the highest probability indicating the predicted class.

In addition to CNNs, Recurrent Neural Networks (RNNs) were also tested, as mentioned earlier. RNNs work with sequential data and are typically used in NLP. Unlike other NNs, the output of RNNs depends on the previous elements in the sequence. Another key feature of RNNs is that they share parameters across every layer[5].

Specifically, the LSTM network, introduced by Hochreiter (1997), was tested. LSTMs account for long-term dependencies by using cells in the hidden layers equipped with three gates—an input gate, an output gate, and a forget gate—that control the flow of information necessary to predict the output (Raschka and Mirjalili 2020).

Like the CNN model, the LSTM was implemented using TensorFlow and Keras. The

[4] *Cosa sono le reti neurali convoluzionali?* in: https://www.ibm.com/it-it/topics/convolutional-neural-networks, last accessed: 2024/24/09.

[5] *Cosa sono le reti neurali ricorrenti?* in: https://www.ibm.com/it-it/topics/recurrent-neural-networks, last accessed: 2024/09/30.

embedding layer plays a key role in representing words as numerical vectors, mapping each word to a 64-dimensional vector to capture semantic relationships within the text.

The pre-processing steps for the LSTM mirrored those for the CNN model: a vocabulary limit of 10,000 words and sequences padded or truncated to 100 tokens. This ensured consistency across all input sequences and allowed both models to learn from the same structured data.

The main difference between the models lies in how they handle text sequences. CNNs excel at capturing local patterns, while LSTMs are designed to capture long-term dependencies. The LSTM layer consists of 64 units, each equipped with memory cells and gates that regulate the flow of information, allowing the network to selectively retain or discard information over longer sequences (Raschka and Mirjalili 2020).

After the LSTM layer, a fully connected Dense layer with 64 units and `ReLU` activation function is applied, similar to the CNN structure. The final layer, as with CNN, is a Dense layer with 3 neurons, each representing a target class, using `softmax` activation to output a probability distribution across these categories.

In addition to the LSTM, a BiLSTM model was also tested. The BiLSTM processes text sequences in both directions—forward and backward—by combining two LSTM layers running in opposite directions. This allows the model to capture context not only from past words but also from future words in the sequence, making it particularly useful in tasks where the meaning of a word is influenced by both its preceding and succeeding context[6].

2.2.7 Transformers

Following the evaluation of NN models, the analysis was extended to include key Transformer models to gain further insights into their performance on this classification task.

The transformer architecture, introduced by Google researchers in 2017, is based on the concept of multi-head attention. It processes text by converting it into numerical representations, called tokens, which are then transformed into vectors via a word embedding table. The transformer uses a multi-head attention mechanism to contextualize each token in relation to others within the sequence, emphasizing key tokens while de-emphasizing less relevant ones (Vaswani et al. 2017).

The models tested in this study were imported from the Hugging Face[7] Transformers library, which offers a wide range of pre-trained transformer models. To handle the tensor operations and data calculations, `PyTorch` was used, while performance metrics were computed using `Scikit-learn`.

To ensure consistency across all models, key parameters were fixed: the number of epochs was set to 3 (since further increases showed no performance improvement), the batch size

[6]Bidirectional LSTM, in: https://paperswithcode.com/method/bilstm, last accessed: 10/03/2024
[7]Hugging Face: https://huggingface.co/huggingface

was 8 for training and 16 for evaluation to optimize GPU usage (T4 on Google Colab), and the number of warm-up steps was set at 500. A weight decay of 0.01 was applied to prevent overfitting by penalizing large weights.

To adapt these pre-trained models for the specific task, fine-tuning was performed using the corpora described in Section 2.5. The K-fold cross-validation method was used, with the dataset divided into five folds. For each iteration, four folds were used for training and fine-tuning, while the remaining fold was reserved for testing. This process was repeated for all folds, ensuring each part of the dataset was used for both training and testing across different iterations.

The first model tested was BERT, which stands for Bidirectional Encoder Representations of Transformers, introduced in 2018 by researchers at Google. BERT represents text as a sequence of vectors through self-supervised learning, utilizing the encoder-only architecture of the transformer model (Devlin et al. 2018).

For this classification task, the pre-trained `BertForSequenceClassification` model was imported from the Hugging Face Transformers library and fine-tuned on the specific corpus. Text was converted into vectors using BERT's tokenization mechanism, which is based on WordPiece tokenization[8]. This approach splits out-of-vocabulary words into smaller, more common sub-units that are often reusable across words. This technique allows the model to handle new or rare words effectively, preserving context and ensuring an accurate representation without requiring an extensive vocabulary.

To ensure optimal model performance, preliminary tests were conducted to determine whether the cased or uncased versions of BERT and related models would yield better results. As will be discussed in section 3.3, these tests revealed minimal differences, informing the choice of version for the final analysis.

The `bert-base-cased model` was selected, and tokenization was performed using `BertTokenizer`, also from the Transformers library. The model was fine-tuned for both English and Italian text.

Preliminary tests were also conducted to assess the impact of applying the "bad words" filter, which had been used in prior models. The results showed that for English, performance remained consistent with or without the filter, whereas for Italian, the application of the filter led to a consistent improvement in accuracy. Based on these findings, the bad-words filter was applied to all reported results.

To address the language issue, the multilingual version of BERT (mBERT[9]), specifically `bert-base-multilingual-cased`, was included in the analysis. This model, trained on 104 languages, including Italian, was chosen to potentially improve performance on Italian texts

[8]WordPiece tokenization in https://huggingface.co/learn/nlp-course/en/chapter6/6, last accessed: 2024/10/07.

[9]BERT Multilingual: https://huggingface.co/google-bert/bert-base-multilingual-cased

by leveraging its multilingual training.

The selection of additional models to test was guided by prior studies on children's safety on YouTube, as referenced in Chapter 1. Many studies reported satisfactory results using DistilBERT[10], a smaller, faster, and more efficient variant of BERT developed through distillation of the base BERT model (Sanh 2019). DistilBERT has 40% fewer parameters than bert-base-uncased, operates 60% faster, and retains over 95% of BERT's original performance, making it well-suited for this task.

For this study, the pre-trained `distilbert-base-cased` model from the Hugging Face Transformers library was used, and fine-tuned on both English and Italian texts in the corpus. The DistilBERT tokenizer, based on WordPiece tokenization, was used to convert the text into vectors, and as with other models, the bad-words filter was applied.

Another model tested in this study was RoBERTa[11] (Y. Liu 2019), a Transformer-based architecture developed as a variant of BERT with modified hyperparameters. RoBERTa removes the next-sentence prediction objective and uses larger mini-batches and higher learning rates during training, which enhances its performance in certain tasks. For this study, the pre-trained `RoBERTaForSequenceClassification` model from the Hugging Face Transformers library was used. This model is specifically designed for sequence classification tasks.

The text was tokenized using the AutoTokenizer, which applies byte-level Byte-Pair Encoding[12] (BPE) to tokenize words into subword units. This encoding method efficiently handles rare and unknown words by breaking them down into smaller, more common units. Initially, the tokenizer starts with a vocabulary of individual characters, merging the most frequent pairs of characters or subwords to form new tokens until reaching the predefined vocabulary size.

For model evaluation, a 5-fold cross-validation strategy was employed, shuffling the data with a consistent random seed to ensure reproducibility. The cross-validation provided averaged precision, recall, and F1 scores across folds, giving robust metrics for each class. An accuracy score was calculated as the mean across folds. Inappropriate language detection was also applied to this model to adjust classification when necessary.

To enhance classification performance specifically for Italian, several Transformer-based models fine-tuned for Italian language tasks were selected. These included `dbmdz/bert-base-italian-cased`[13] and `dbmdz/electra-base-italian-xxl-cased`[14] discriminator, both developed by the MDZ Digital Library team at the Bavarian State Library. These models draw

[10] DistilBERT: https://huggingface.co/docs/transformers/model_doc/distilbert
[11] RoBERTa: https://huggingface.co/docs/transformers/model_doc/roberta
[12] BPE: https://huggingface.co/learn/nlp-course/chapter6/5
[13] dbmdz Italian Bert: https://huggingface.co/dbmdz/bert-base-italian-cased
[14] Electra: https://huggingface.co/dbmdz/bert-base-italian-cased

from substantial Italian corpora, including the OPUS[15] collection and the OSCAR[16] corpus, making them well-suited for tasks requiring nuanced understanding of Italian language contexts. The Italian ELECTRA model was trained on the XXL corpus for 1 million steps, using a batch size of 128.

For the classification task, the BERT-based model (`dbmdz/bert-base-italian-cased`) and the ELECTRA-based model (`dbmdz/electra-base-italian-xxl-cased discriminator`) were fine-tuned using the Transformers library. The `BertTokenizer` and `ElectraTokenizer` were applied to tokenize text into subword units, enabling efficient handling of rare or compound words by breaking them into smaller components.

Additionally, other Italian-specialized Transformer models were tested: AlBERTo[17] (Polignano et al. 2019), designed for Italian social media text, specifically on Twitter, and two RoBERTa-based models, GilBERTo[18] and UmBERTo[19] (Parisi et al. 2020). GilBERTo is an Italian pre-trained language model based on Facebook RoBERTa architecture and CamemBERT text tokenization approach. UmBERTo is a Roberta-based Language Model trained on large Italian corpora and uses two innovative approaches: SentencePiece and Whole Word Masking. Each of these models was fine-tuned on the corpus using a 5-fold cross-validation strategy for robust evaluation, with the "bad words" detection mechanism integrated to improve classification reliability. Results for each model were evaluated across several metrics, including precision, recall, and F1-score, for all three target categories.

2.2.8 Large Language Models

In the scope of this work, an analysis of LLMs was considered appropriate to gain insights into their performance on this specific classification task. Three models—Flan-T5, GPT-2, and Minerva—were selected to assess their effectiveness. Given their recent advancements, LLMs leverage a transformer architecture (Vaswani et al. 2017) that processes vast amounts of text data, capturing complex contextual relationships within language. This architecture, combined with billions of parameters, enables LLMs to perform a wide range of tasks, from text comprehension to generation. Typically, these models are pre-trained through self-supervised or semi-supervised learning, allowing them to predict missing parts of sentences or generate text from prompts, making them highly adaptable for general-purpose language applications[20].

[15] OPUS corpora: https://opus.nlpl.eu/

[16] OSCAR corpus: https://oscar-project.org/

[17] AlBERTo, https://github.com/marcopoli/AlBERTo-it?tab=readme-ov-file

[18] GilBERTo, https://huggingface.co/idb-ita/gilberto-uncased-from-camembert/blob/main/pytorch_model.bin

[19] UmBERTo https://github.com/musixmatchresearch/umberto

[20] *Better language models and their implications*, in https://openai.com/index/better-language-models/, last accessed: 2024/10/16.

For each model, fine-tuning was preferred over prompting techniques to ensure consistent comparisons across the dataset. These models were fine-tuned on the custom corpus created for this study, ensuring optimal alignment with the classification task. The selection of these models was also influenced by API availability and hardware capabilities, specifically the Tesla T4 GPU with 16 GB RAM.

To ensure uniformity across models, the same fine-tuning parameters were applied for each LLM tested. The training process for each model involved 8 epochs and batch sizes of 8 for training and 16 for evaluation, which allowed for a fair comparison by standardizing the training conditions. Additional parameters, including 500 warmup steps, a weight decay of 0.01, and logging every 10 steps, were consistently applied. Evaluation was conducted after each epoch, and 5-fold cross-validation was used to assess performance robustness.

Moreover, the bad words detection filter was included in each model, ensuring that inappropriate language was flagged and reclassified as necessary, a decision made following a preliminary test, which will be discussed in section 3.4. This standardized setup across all tested LLMs helps in understanding each model's relative effectiveness in handling this classification task.

Flan-T5[21], an enhanced version of T5 (Raffel et al. 2020), was fine-tuned for a mixture of tasks (Chung et al. 2024). For the classification task, the Flan-T5 small model was configured using Hugging Face's `AutoModelForSequenceClassification` and AutoTokenizer for proper text tokenization.

Another model tested was the OpenAI GPT-2[22] model, introduced by Radford et al. (2019). GPT-2 is a large, unidirectional transformer with 1.5 billion parameters, pre-trained on a language modeling task using approximately 40 GB of text data, drawn from a corpus of 8 million web pages. Although GPT-2 was primarily trained to predict the next word in a sequence, Its potential for sequence classification was explored by applying fine-tuning to the target task. Tokenization was conducted using the GPT-2 tokenizer via the AutoTokenizer class from the Hugging Face Transformers library.

The final model, Minerva, was pre-trained on Italian by Sapienza NLP[23] in collaboration with FAIR[24] and CINECA[25]. Each Minerva model was trained on an extensive dataset of online and documented Italian and English sources, totaling over 500 billion words, the equivalent of more than 5 million novels.

For this experiment, the `Minerva-350M-base-v1.0` model was selected. Tokenization was performed using the AutoTokenizer from Hugging Face's Transformers library, specifically configured for the Minerva model.

[21] Flan T5: https://huggingface.co/docs/transformers/model_doc/flan-t5
[22] OpenAI GPT-2: https://huggingface.co/docs/transformers/model_doc/gpt2
[23] Sapienza NLP: https://nlp.uniroma1.it/
[24] FAIR: https://fondazione-fair.it/
[25] CINECA: https://www.cineca.it/it

2.3 Semantic Analysis

The second experiment carried out in this work focuses on the classification of online multimedia content using a semantic approach. This type of analysis allows for result comparison, while also exploring an unsupervised classification approach closely related to the semantics of the texts. This experiment is valuable for thoroughly exploring various possibilities in classifying online multimedia content. The analysis was carried out on English and Italian content, using distinct resources customized for each language.

Specifically, the Word2Vec word embedding model was used to identify the words most semantically similar. For Italian, the model was trained on the PAISÀ corpus (Lyding et al. 2014), a comprehensive collection of approximately 380,000 Italian-language documents sourced from websites protected by Creative Commons licenses, totaling around 250 million words. Of these, about 260,000 documents come from Wikipedia, approximately 9,300 from Indymedia, and around 65,000 from blogs. For English, the British National Corpus (BNC) (Leech 1992) was used, containing 100 million words spanning a wide range of genres, including spoken, fiction, magazines, newspapers, and academic texts. Before training the models, tests were conducted to determine the appropriate level of pre-processing to apply to the texts. After conducting various tests with more intensive pre-processing methods, it was ultimately decided to apply only tokenization to the corpus in order to preserve the essential information within the texts.

A distributional semantic matrix was used to extract similarity values between words and identify the nearest neighbors of the most frequent terms in each transcript. This choice allowed for an analysis independent of predefined context, which would otherwise be required when testing a pre-trained model. However, this context-independent approach could pose a challenge for future project developments.

The `Gensim`[26] library was chosen to implement the Word2Vec model. For this analysis, a fixed set of parameters was used, which included a window size of 5 (indicating the maximum distance between the current and predicted words within a sentence), a minimum word frequency count of 5, four workers for parallel processing, and five training epochs. This configuration ensured consistency across both English and Italian corpora, with each word represented by 100 vectors, preserving the desired semantic depth for analysis.

The window size, set to 5, defined the maximum distance between the current and predicted word within a sentence, while the `min_count` parameter set a minimum frequency threshold for words to be included in the model[27].

Additionally, a comprehensive pre-processing pipeline was applied to the corpus: texts

[26] Gensim: https://radimrehurek.com/gensim/models/word2vec.html

[27] *A Dummy's Guide to Word2Vec*, in https://medium.com/@manansuri/a-dummys-guide-to-word2vec-456444f3c673, last accessed: 2024/07/18.

were converted to lowercase, punctuation was removed, and English and Italian stopwords were eliminated using the `NLTK`[28] library. Tokenization and lemmatization were then performed with the `SpaCy` library.

The process of classifying multimedia content involved various features related to different aspects of the material, such as the language used in dialogues and the presence of themes like violence, drugs, or gambling. While some of these features are linked to the visual or non-linguistic dimension—which cannot be assessed through linguistic analysis alone—this study focuses exclusively on the linguistic attributes of each text.

Ultimately, a lexical analysis was conducted to automatically identify the presence of certain relevant themes in the texts, aiming to evaluate the content's suitability for a younger audience. For this purpose, three specific indicators were selected: Bad Language, Violence, and Drugs, which are critical for determining the themes addressed within the texts.

The lexical analysis was performed using specialized dictionaries for each indicator. These dictionaries, developed specifically for this research, allowed for the identification of the frequency and distribution of key terms associated with each selected theme within the corpus. This approach involved measuring both the absolute and relative frequencies of each indicator term, enabling a score to be assigned to each text based on the presence of the selected indicators.

The scores obtained were then used to classify the content automatically and were subsequently compared with manual annotations made during the corpus preparation phase. This comparison helped evaluate the accuracy of the lexical analysis and identify potential issues, such as ambiguous terms.

2.4 Construction of electronic dictionaries

The extraction of indices for *Violence*, *Bad Language*, and the presence of *Drugs* in texts is performed by searching for specific lexical indicators within the texts. These indicators are stored in electronic dictionaries.

The dictionaries for English, described in Maisto et al. (2021a,b), were developed during a previous project that I conducted. These dictionaries were expanded using the COALS algorithm and leveraging the BNC. This expansion increased each dictionary's word count by adding the ten most semantically similar words for each entry. The expansion was achieved using a Distributional Semantic Matrix produced by COALS (Rohde et al. 2006), applied to the BNC (Leech 1992). COALS is a word-based distributional semantics approach that leverages Pearson Correlation on word co-occurrence vectors to compute the conditional association between word pairs, and it has demonstrated excellent results across a range of semantic tasks (Jurgens and Stevens 2010).

[28] NLTK: `www.nltk.org/`

We employed COALS to extract semantically similar terms for each basic entry in the dictionaries. However, as many of the resulting terms were not relevant to the dictionary's focus, a manual verification process was necessary to eliminate common or ambiguous words that could lead to inaccurate outcomes.

All dictionaries were subsequently updated and manually reviewed for the specific purpose of this thesis. For Italian, however, some dictionaries, such as those for slang, discriminatory terms, and drugs, were built from scratch by exploring various resources. Others, such as the violence dictionary, were derived by automatically translating the English dictionaries using Google Translate's API.

The dictionaries were created by examining and extracting terms from resources published on various web pages and online thesauri, including `urbandictionary.com`, `wikipedia.com`, `urbanthesaurus.org`, and `slengo.it`.

After preliminary tests, it was decided not to expand the Italian dictionaries, opting to keep them more contained and selective to avoid ambiguity and prioritize higher precision. Although this choice limited lexical coverage, it ensured that the Italian dictionaries remained more focused.

In the literature, various lexicons have been developed for detecting offensive and discriminatory language, such as HurtLex (Bassignana et al. 2018), a multilingual lexicon containing offensive, aggressive, and hateful words in over 50 languages. This lexicon categorizes words into 17 classes, along with a macro-category indicating whether a stereotype is involved. However, in this study, custom dictionaries were created from scratch instead of relying on general-purpose resources like HurtLex. This decision was made to ensure that the dictionaries were specifically tailored to the scope of this research and to minimize lexical ambiguity during classification. One of the challenges of using broad lexicons is that they often include terms that can appear in neutral or non-discriminatory contexts, potentially increasing false positives. For instance, in HurtLex's OM category (words related to homosexuality), terms such as *finocchio* (fennel), *ragazzo* (boy), *diverso* (different), and *strano* (strange) are included, despite having multiple meanings unrelated to discrimination in Italian. Given the focus of this study, creating dedicated dictionaries allowed for a more controlled selection of terms, prioritizing precision while acknowledging the inherent challenges of lexical ambiguity.

As will be discussed in Section 3.5.3, this approach confirms the initial hypothesis: While the expanded English dictionaries show broader term coverage, they suffer from lower precision due to false positives. In contrast, the Italian dictionaries, being more streamlined, demonstrate higher precision but somewhat lower recall. These experimental results underscore the importance of balancing coverage and accuracy when creating specialized lexicons, suggesting that future work could focus on refining both sets of resources to better manage

ambiguity and improve overall reliability.

The dictionaries include both single words and Multi-Word Expressions (MWEs), which enhance the performance of semantic tagging, as MWEs are generally less ambiguous.

The characteristics of the four dictionaries are as follows:

- **Slang Dictionary**: This dictionary includes vulgar slang terms generally associated with offensive language, featuring unambiguous insults and derogatory terms, particularly those related to discrimination or topics such as sex and sexism.

  A brief excerpt from the English insult dictionary:

  > `'o' face, anal, ass, bufu, blowjob, bondage, fetish, fingering, fisting, french kiss`.

  A brief excerpt from the Italian insult dictionary:

  > `affanculo, allocco, anale, analmente, arteriosclerotico, babbeo, bagascia, bagascione, baggiano, baldracca, bamboccio`.

- **Violence Dictionary**: This encompasses terms connected to violence, including references to weapons, war-related terminology, and similar themes.

  A brief excerpt from the English violence dictionary:

  > `attack, conflict, detest, fight, h8, hate guts, hate on hate with a passion, injure, injury, kill`.

  A brief excerpt from the Italian violence dictionary:

  > `abuso, addestramento, aereo da guerra, aggirarsi furtivamente, aggressione, aggressività, aggressore, agitatore, allerta, anarchia`.

- **Drug Dictionary**: This dictionary contains general terms, commercial and scientific names of drugs, as well as slang terms commonly used to refer to them.

  A brief excerpt from the English drug dictionary:

  > `amyl nitrite, amphetamine, joint, marijuana, methylphenidate, pookie, pentobarbital, r-ball, roofies, weed`.

  A brief excerpt from the Italian drug dictionary:

  > `allucinogeni, anfetamine, barbiturici, canna, cannabis, chetamina, cocaina, codeina, crack, ecstasy`.

- **Discriminatory Terms Dictionary**: This focuses on words related to discrimination based on sexuality, ethnicity, regional identity, or religion.

  A brief excerpt from the English discrimination dictionary:

`boogie, burrhead, charlie, chukhna, greaser, haji, macaca, nipper, ogay, okie`.

A brief excerpt from the Italian discrimination dictionary:

`barbaro, beduino, beota, bifolco, checca, deficiente, frocio, guercio, mona, mongoloide`.

For this analysis, two sets of four dictionaries were used for each language, resulting in a total of eight dictionaries. In Table 2.1 and Table 2.2 are showed the dimension of each dictionary, for English ones and Italian ones.

Name	**Number of entries**
Dictionary of Slang and Insults	6612
Dictionary of Violence-related terms	507
Dictionary of Drugs	737
Dictionary of Discriminative-terms	728

Table 2.1: List of English Electronic Dictionaries.

Name	**Number of entries**
Dictionary of Slang and Insults	636
Dictionary of Violence-related terms	490
Dictionary of Drugs	137
Dictionary of Discriminative-terms	93

Table 2.2: List of Italian Electronic Dictionaries.

2.4.1 Expansion and Translation of the NRC Emotion Intensity Lexicon

Another aspect of this research focused on adapting emotion lexicons to identify whether specific negative emotions, such as anger, sadness, disgust, and fear, are expressed within the language of the analyzed content. These emotional indicators are essential for refining classification criteria. According to AGCOM guidelines, content that is shocking (disgust) or horror-based (fear) must be classified as suitable only for audiences over eighteen years of age.

Emotions have been analyzed using the *NRC Emotion Intensity Lexicon* (NRC-EIL) (Saif M. Mohammad 2018). The NRC-EIL[29], formerly called *NRC Affect Intensity Lexicon*

[29] https://saifmohammad.com/WebPages/AffectIntensity.htm

(NRC-AIL), is a lexicon of English words with real-valued intensity scores for eight basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust).

In Maisto et al. (2021a,b), the original word list, available in both English and Italian, underwent lemmatization and inflection, resulting in an expanded lexical database with over 20,000 entries (9,921 in English and 10,334 in Italian). Each entry is associated with one of Plutchik's eight primary emotions (Plutchik 1984) and annotated with an intensity score, ranging from 0 (lowest emotional expression) to 1 (highest emotional expression).

The NRC EIL Lexicon provides affect annotations for English words. While some cultural differences exist, previous studies suggest that many affective norms show a degree of stability across languages. According to Saif M Mohammad (2021), this claim is supported by linguistic and cognitive science research (Bender and Koller 2020; Bisk et al. 2020; Chomsky 2014; Harris 1954; Hovy and D. Yang 2021; Moore 2014). Based on this premise, versions of the lexicon in over 100 languages were created by translating the English terms via Google Translate, assuming a level of cross-linguistic affective stability.

After analyzing the dictionary's performance on an Italian language processing task conducted at Theuth Linguistic Engines, an expansion of the dictionary was undertaken.

Each entry went through lemmatization and inflection via two key steps:

- Synonym Expansion Using MultiWordNet (Pianta et al. 2002): Synonyms were retrieved through WordNet for each word in the dictionary. If a synonym was not already included, it was added to broaden lexical coverage.

- Inflection of Each Lemma: To ensure accuracy, each lemma's inflected forms were manually verified and added to the dictionary. The process followed these steps:

    1. Confirmation of Inflected Forms: The primary inflected forms (e.g., verb tenses, moods, noun declensions) were checked against authoritative linguistic resources, such as uch as the Reverso Conjugation function[30]. Special attention was given to irregular verb forms to ensure completeness.

    2. Manual Addition to the Dictionary: Once confirmed, each inflected form was manually inserted into the dictionary, ensuring that all variations (e.g., present, past, future, subjunctive, conditional) were included. This process was repeated for synonyms to maintain consistency across semantically related words.

    For example: Starting with the lemma *cominciare* (to start), various inflected forms were added, such as *comincio* (I start), *cominci* (you start), *comincia* (he/she/it starts), *cominciamo* (we start), and so on, including conjugations like *comincerei* (I would start) and *cominciasse* (he/she/it started, subjunctive mood). The same ex-

[30] https://coniugazione.reverso.net/coniugazione-italiano.html

pansion was applied to synonyms, such as *iniziare* (to begin), yielding forms like *inizio* (I begin), *inizi* (you begin), *iniziamo* (we begin), etc.

The original emotion scores were retained for each synonym and its inflected forms. Following this process, the resource now contains 43,606 entries.

For this research, I used a limited set of indicators focusing on Fear, Anger, and Disgust. These indicators, combined with other dictionaries, can generate values that index the presence or absence of specific features within texts.

For this experiment, after careful consideration, an essential analysis was chosen, omitting morphological details and utilizing the same dictionary for both English and Italian. The aim was not to identify which emotion a text might represent but rather to confirm the presence or absence of specific negative emotions: anger, fear, and disgust. Moreover, I was working on lemmatized texts.

For English, I used available resources by downloading separate files for each emotion, each containing a list of emotionally connoted words with scores ranging from 0 to 1. For Italian, the file was translated using DeepL's API[31].

For future developments, a resource tailored from scratch for each specific emotion in Italian could be created, building upon the expansion work conducted.

2.5 Corpora-building

To train and evaluate ML models, it was necessary to create and annotate a comprehensive corpus. Since this work addresses both Italian and English, two separate corpora were required. Initially, the objective was to develop parallel corpora to enable a precise comparison of system performance across the two languages. However, after careful consideration, this parallel approach was ultimately set aside to better serve the project's primary aim: establishing a foundation for a classification system that operates effectively for both English and Italian content. This decision is grounded in the recognition that digital platforms, including social networks, commonly display content in multiple languages, exposing children to Italian and English material. To capture this multilingual reality, a balanced and representative corpus was assembled, drawing on texts from diverse resources in their original language rather than relying on translations.

For the initial corpus construction, I began with the TED2020 corpus[32] as described in Reimers and Gurevych (2020). This resource includes nearly 4000 TED and TED-X transcripts from July 2020, translated into over 100 languages by volunteers. From this collection, approximately 200 texts were selected for each language—196 texts in English and 202 in Italian—to provide a baseline of varied content suitable for language analysis.

[31] https://www.deepl.com/

[32] https://www.ted.com/participate/translate

The TED2020 corpus offered significant advantages: it supported a parallel analysis of both languages and represented a multimedia dataset available online. Using TED Talks as a corpus brings notable benefits. Firstly, the talks cover a wide array of topics, reflecting a broad spectrum of vocabulary and discourse styles, which is beneficial for training models intended to classify diverse content. Additionally, TED Talks are professionally transcribed and translated, ensuring high consistency and quality in both English and Italian. Finally, as a widely accessible and recognizable source of content, this corpus aligns well with real-world applications, where users encounter a variety of subjects and perspectives.

During the initial phase of annotation, it was observed that many texts were primarily suited for teenage or adult audiences, consistently maintaining a polite and formal tone. Only a few were explicitly intended for children, often presented by child speakers. This posed challenges in creating a balanced corpus with content specifically targeted toward children or adults, both essential for effectively training the classification system. Additionally, annotators sometimes found it challenging to confidently determine whether a TED Talk was more appropriate for an adult or adolescent audience.

To address this, was decided to supplement the corpora with texts explicitly aimed at children and adults.

For Italian, children's texts were extracted from *YouTube Kids*[33] as subtitles, with an equal amount of material collected across each age group designated by the platform (Preschool: up to 4 years; Younger: 5-8 years; Older: 9-12 years). The subtitles were manually reviewed and adjusted to correct the frequent errors found in automatically generated captions.

For adult content, some texts were downloaded from `OpenSubtitles.org`[34] with filters set to language: `Italian`, genre: `Horror`, and `Adult`. Additional material was gathered from YouTube, searching for "YTP in italiano" (YTP in Italian). These subtitles, like those for children's content, were manually reviewed and adjusted; censored words were restored to accurately reflect the original content.

For English, children's texts were collected from *Tvo Kids*[35], a site hosting videos for children categorized by age (Preschool and School-age). Transcripts were manually collected for each video, ensuring accuracy in content for this age group.

Adult English content was similarly obtained from `OpenSubtitles.org` using filters set to language: `English`, genre: `Horror`, and `Adult`. Additionally, subtitles were collected from YouTube videos, particularly English YTPs, by searching for "English YTP." These subtitles were then manually reviewed and adjusted to restore any censored language, ensuring fidelity to the original content.

[33] https://www.youtubekids.com
[34] OpenSubtitles.org
[35] https://www.tvokids.com

The final corpora consist of 446 texts in English and 514 texts in Italian, as shown in Tables 2.3 and 2.4. This structure is designed to capture representative linguistic variation, supporting the training of a classification system that functions effectively across both languages.

Source	Number of Texts	Number of Tokens	Number of Types
TED2020	196	536,895	22,717
Tvo Kids	200	270,329	9,660
Opensubtitles.org	30	60,242	6,085
English YTP	20	15,438	2,776
Total	**446**	882,904	28,768

Table 2.3: Corpus composition for English content.

Source	Number of Texts	Number of Tokens	Number of Types
TED2020	202	504,817	32,312
YouTubeKids	236	59,069	8,229
Opensubtitles.org	54	56,363	8,510
Italian YTP	22	8,062	2,501
Total	**514**	628,311	38,495

Table 2.4: Corpus composition for Italian content.

2.6 Annotation Procedure

In order to employ the described corpora for the training and testing of the models within a supervised analysis framework, it was essential to perform a systematic annotation process, adding information to the text through specific labels of a tagset. In this case, the information added consists of labels assigned to the entire text, indicating the appropriate age for the multimedia content's transcript (e.g., 3, 6, 12, 15, or 18 years) and the presence or absence of specific indicators (such as violence, drugs, and discrimination) within the text.

To establish objective criteria for annotating the corpus, annotation guidelines were created, i.e., a document that explains to annotators how to apply the annotation schema in practice, providing examples (Ježek and R. Sprugnoli 2023). These guidelines were tested in collaboration with two female annotators, who were selected to enhance their robustness. Both annotators were native Italian speakers and were either graduates or current students in the humanities.

After testing the guidelines, identifying their limitations, and making improvements based on the test results, I proceeded to annotation of the corpora described above.

To facilitate the annotation procedure, an annotation platform was created in Google Sheets, featuring drop-down menus for efficient data entry (Figure 2.1). The annotation schema consists of several key columns, each designed to capture specific information related to the text and its classification criteria:

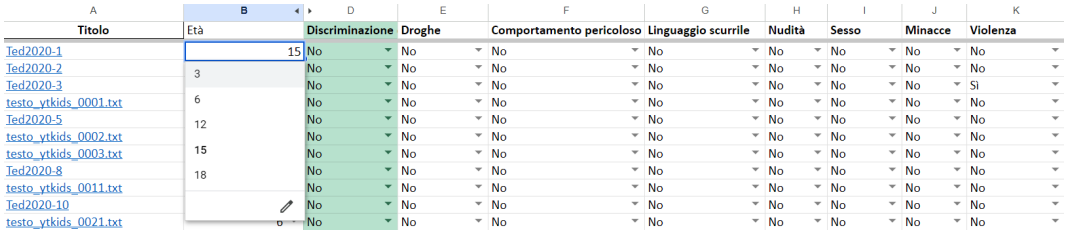
- **Title**: This column lists the title of each text, ensuring that the annotators correctly match each text file with its corresponding row in the spreadsheet.
- **Audience Age**: In this column, annotators selects and records the recommended minimum age for suitable audiences from a drop-down menu for each text
- **Descriptors**: Separate columns are provided for each thematic descriptor, including "Discrimination," "Drugs," "Dangerous Behavior," "Offensive Language," "Nudity," "Sex," "Threats," and "Violence." Each descriptor column included a drop-down menu with "Yes/No" options, allowing annotators to indicate the presence or absence of words or phrases that reference these specific themes within the text.

In addition, the annotation schema includes several pre-filled or automatically generated columns:

- **Language Complexity**: Reserved for automatic annotation of the complexity of the text, obtained from the calculation of word frequency in the NVDB (for Italian) and the Oxford 3000 (for English), normalized by the total number of words in the text; annotators were instructed to leave this column empty.
- **Language**: This pre-filled column specifies the language of the text, either "English" or "Italian."
- **Annotator ID**: A pre-filled column included in each annotator's Excel sheet to identify the annotator responsible for each text. While all annotators worked on the same set of texts independently in separate sheets, this identifier facilitated the subsequent analysis of the dataset.
- **Notes**: A dedicated column for annotators to add comments or observations as needed, providing space for any relevant additional insights.

For the annotation guideline testing process, the annotators were provided with oral instruction, a demonstration of the annotation platform and a detailed instructions in a two-page guideline.

The guidelines were inspired by the *AGCOM guidelines for the classification of audiovisual works for the web*, but were revised and adapted to focus solely on the textual aspect



A	B	D	E	F	G	H	I	J	K
Titolo	Età	Discriminazione	Droghe	Comportamento pericoloso	Linguaggio scurrile	Nudità	Sesso	Minacce	Violenza
Ted2020-1	15	No	No	No	No	No	No	No	No
Ted2020-2		No	No	No	No	No	No	No	No
Ted2020-3		No	No	No	No	No	No	No	Sì
testo_ytkids_0001.txt		No	No	No	No	No	No	No	No
Ted2020-5		No	No	No	No	No	No	No	No
testo_ytkids_0002.txt		No	No	No	No	No	No	No	No
testo_ytkids_0003.txt		No	No	No	No	No	No	No	No
Ted2020-8		No	No	No	No	No	No	No	No
testo_ytkids_0011.txt		No	No	No	No	No	No	No	No
Ted2020-10		No	No	No	No	No	No	No	No
testo_ytkids_0021.txt		No	No	No	No	No	No	No	No

3
6
12
15
18

Figure 2.1: Partial excerpt of the annotation platform, showing age categories and thematic descriptors annotated for each text.

of multimedia content. In adapting the AGCOM guidelines for text-based content, certain descriptors were adjusted to suit the nature of written language. For example, the original guideline states, "*Absence of representation of sexual activities or violence.*" In the context of text, this was modified to "*Absence of descriptions of sexual activities or violence.*" This adjustment reflects the shift from visual or multimedia content to purely textual descriptions, ensuring the guidelines remain relevant and applicable to text-based annotations.

This guide included the following elements:

- Introduction: A brief overview explaining the objectives of the project and providing the context for the annotation process.
- Annotation Interface: The annotation platform (created in Google Sheets) was explained in detail, with instructions on how to use the interface properly. Drop-down menus were made available to facilitate data entry.
- Annotation Categories: All age categories were outlined, along with specific characteristics to consider when assigning each text to a particular category. Annotators could refer to AGCOM indicators for classifying content according to age suitability.
- Thematic Indicators: Thematic indicators, such as "Drugs", "Violence", "Discrimination" and others, were listed and described, specifying how each should be annotated. The “Yes / No” options were used to indicate the presence or absence of these themes within the texts.
- Helpful Suggestions: The guide concluded with practical tips for addressing potential ambiguities and ensuring consistency across annotations.

References for categorization were the following:

- *Opere per tutti* (Works for all): the content of the text is considered suitable for all ages. The text must not contain any language or descriptions of discriminatory behavior, incitement to hatred. There must be no reference use of tobacco, gambling, drugs or drug addiction (drugs, alcohol, etc.). There can be no hint of dangerous or easily

mimicked behavior. Language cannot be obnoxious. Descriptions of complete nudity are not allowed unless it is attributable to natural, scientific, and/or everyday contexts. The text must not contain any explicit sexual descriptions or references, emotional representations must be limited and without erotic references. Violent behaviors must be absent but are allowed if attributable to unrealistic contexts (e.g. comic gags).

- *Opere non adatte ai minori di anni 6* (Works not suitable for children under 6 years): the text shall not contain any discriminatory language or behavior unless (a) is clearly disapproved of or inserted in historical contexts, (b) has educational purposes. The representation of drugs and drug addiction (drugs, alcohol, etc.) is only allowed with obvious educational purposes, or if linked to social initiatives (e.g. prevention, awareness and information campaigns). There may be an occasional, hinted, or unrealistic description of dangerous behaviors, with evident disapproval. The presence of vulgar terms commonly used in everyday life is allowed only if sporadic. Descriptions of partial nudity are allowed if they are not attributable to a sexual context. Displays of affection are allowed as long as they are not excessive. There can be the description of scenes that cause agitation and emotional tension. There may be the description of scenes of mild violence related to fiction in which the consequences (e.g., blood and injuries) are not mentioned.

- *Opere non adatte ai minori di anni 12* (Works not suitable for children under 12 years): the text may contain discriminatory language and/or behaviors justified by the narrative context, provided that: they are explicitly condemned and included in contexts for educational purposes, and they are attributed to negative characters with whom minors do not identify. The text must not contain incitement to hatred. The representation of drugs and substance addiction is allowed provided that: (a) it is connected to educational contexts, (b) it is condemned. The description of easily imitable dangerous behaviors may be present, provided that they are not presented as glamorous, not encouraged, and there is a positive final message. The use of bad language is allowed only if it is included in comedic contexts. Description of full nudity is permitted if not in sexual contexts, while description of partial nudity is allowed even in sexual contexts. Descriptions of sexual activities should be rare, and expressions of affection should remain limited and free from explicit or provocatively suggestive language. Descriptions of scary and/or dangerous events, capable of inducing agitation or tension, may be included only if: (a) they are brief, and (b) they are appropriate to the comprehension abilities of the corresponding age group. References to mild violence are allowed only if: (a) they are contextualized, (b) they are morally sanctioned, and (c) they avoid morbid focus on graphic or brutal details, which may be included only if justified by the context.

- *Opere non adatte ai minori di anni 15* (Works not suitable for children under 15 years): the text can contain discriminatory language and behaviors, as long as they are not encouraged. Incitement to hatred is allowed only with clear condemnation. The representation of drugs and substance addiction is permitted as long as it is not portrayed positively and lacks legitimization or glorification. The presence of easily imitable dangerous behaviors can occur frequently but must be discouraged or depicted as difficult to imitate. The frequent use of vulgar or offensive terms and profanity is allowed only if they are not portrayed as positive standards and they are morally sanctioned. Description of full nudity is allowed if not in sexual contexts, even accompanied by explicit language, while description of partial nudity is permitted even in erotic contexts. Descriptions capable of inducing agitation and high emotional tension are allowed, provided they are not excessively prolonged and avoid morbid focus on brutal details. The presence of violent scenes is allowed only if is not portrayed in a positive light and is devoid of morbid insistence on brutal details, which may be present only if justified by the context. Horror sequences may be present. Elements referring to the use of chance games depicted as typical gambling games may be included, as long as they are of secondary importance compared to the overall gaming experience.

- *Opere non adatte ai minori di anni 18* (Works not suitable for children under 18 years): the text may contain unrestricted depictions of intolerant, racist, or sexist behaviors. Incitement to hatred, by negative characters, must be reconstructed and balanced within the narrative. Unrestricted description of the use, acquisition, and manufacturing of drugs is allowed. The description of dangerous behaviors is unrestricted. The frequent use of vulgar and/or offensive terms and bad language is permitted. Non-occasional blasphemy is excluded. Description of full nudity is allowed, including in sexual contexts, as long as it is not part of pornographic scenes. Description of sexual and erotic activities may be present, accompanied by explicit language, justified by the plot and context. The presence of distressing or frightening sequences with high emotional impact is unrestricted. The explicit description of violence, even gratuitous, excessive, and brutal, is allowed only if it is not legitimized or glorified. The description of sadistic violence is excluded. Horror sequences may be present. Elements referring to the use of chance games, portrayed as typical gambling games, may be included, but they should be of secondary importance compared to the overall narrative.

The specific characteristics that the annotators had to consider for each specific category have been summarized in the tables in the appendix (see Table 1, and Table 2). Specifically, the tables summarize the extent to which thematic indicators can be present in works intended for the web.

Annotators have also annotated the presence or absence of words or phrases within the

text that refer to the following thematic descriptors: *Discriminazione* (Discrimination), *Comportamento pericoloso* (Dangerous behavior) *Droghe* (Drugs), *Linguaggio scurrile* (Bad language), *Nudità* (Nudity), *Sesso* (Sex), *Minacce* (Threats), and *Violenza* (Violence). To annotate this information, the annotators have selected "Yes" or "No" from a dropdown menu. For example, the text titled *Ted2020-171* included phrases such as: *And what the hell do you know about the National Guard?*; *Sons of bitches!*; *If one of those incompetent doctors had told me to stop the treatment, I would have slit his throat on the spot.* In this case, the annotator would select "Yes" in the columns for "Offensive Language," "Threats," and "Violence." If no other indicators were present in the text, "No" would be selected in the remaining columns.

Annotators were encouraged to maintain a report during annotation to note any doubts, concerns, or errors encountered. This report was to be periodically emailed to me; to maintain objectivity and validity, annotators were instructed not to discuss their annotations of specific texts with other annotators; annotators were advised not to open any links to external sites that might be included in the text; annotators were instructed not to alter fields in the Google Sheets file without permission, even if errors were noticed. Any such issues should be reported; if a text was written in a language other than Italian or English, annotators were asked to flag it.

For the testing phase, the two annotators and I annotated the same sample of 30 texts in Italian to assess the guidelines. This annotation task involved determining the appropriate age suitability for each content piece and indicating the presence or absence of terms related to AGCOM thematic descriptors.

Table 2.5 summarizes the distribution of annotations by label (age suitability) and annotator during the testing phase, providing an overview of the inter-annotator contributions and the total annotations for each label.

Label (Years)	**Annotator 1**	**Annotator 2**	**Annotator 3**	**Total Annotations**
3	4	7	3	14
6	7	4	8	19
12	7	17	10	34
15	12	2	8	22
18	0	0	1	1

Table 2.5: Annotation distribution by label and annotator.

Once the guidelines were evaluated and refined, I proceeded to annotate the entire corpus on my own.

Annotation evaluation

As Ježek and R. Sprugnoli (2023) explained, with the right availability of funds, personnel, and time, it would be optimal to annotate all data with multiple people; unfortunately, this procedure is not easily feasible. This is the case in this thesis work, where the decision was due to a lack of personnel who, without compensation, would have had to annotate a moderate amount of texts, some of which were very lengthy. Consequently, I was forced to adhere to the minimum requirement of having a portion of the data annotated by at least two annotators.

However, I decided to perform a quantitative analysis on these test data by calculating Inter-Annotator Agreement (IAA), in order to identify any flaws in the design and to guide any revisions before proceeding with the complete annotation of the data. Given the presence of more than three annotators, Fleiss' kappa (Fleiss 1971), has been used to calculate IAA:

$$Fleiss'kappa(\kappa) = \frac{P_{observed} - P_{random}}{1 - P_{random}}$$

In this equation, $P\ observed$ is the observed agreement of the raters and $P\ random$ is the expected agreement of the raters. The expected agreement is given if the raters judge completely randomly.

Kappa Metric	**Value**	**Interpretation**
Fleiss' Kappa (k)	0.25	Fair agreement

Table 2.6: Inter-annotator agreement calculated using Fleiss' Kappa.

In this way, we obtained a k equal to 0.25, which based on the grid for the interpretation of Landis and Koch (1977) (where < 0.0 is poor and 0.81 - 1.00 almost perfect), corresponds to a fair score. The result is not unexpected considering that each annotator could rely on one's own experience as a native Italian speaker (which is a subjective element) to determine the suitable age to be exposed to that specific content, in addition to considering the objective parameters of the AGCOM guidelines.

Low IAA values are not uncommon in tasks where subjectivity plays a significant role, particularly when dealing with complex and interpretative annotations. Similar results have been observed in studies where annotators had to categorize content based on implicit criteria or contextual understanding. As shown in Leonardelli et al. (2021), subjective judgments often lead to lower agreement scores, even among trained annotators, due to the inherent variability in human interpretation. This aspect is especially relevant in tasks involving age classification, where differences in cultural background, personal experience, and individual

perception of appropriateness can influence the assigned labels. Therefore, while the IAA score in this study may appear low, it aligns with findings from related works dealing with similarly subjective annotation tasks.

Thematic descriptors	**Match rate**
Discrimination	100%
Drug	96%
Dangerous behaviour	100%
Bad language	96%
Nudity	100%
Sex	96%
Threats	100%
Violence	96%

Table 2.7: Raw agreement percentage for each descriptor.

If, instead, additional indicators are considered, where annotators may choose between "Yes" or "No", based on whether certain elements in the text fall within the categories specified by the AGCOM resolution, the raw agreement in the annotations is nearly perfect (Table 2.7): a match rate between annotators of 100% for Discrimination, Dangerous behavior, Nudity and Threats and a match rate of 96% for Drugs, Bad language, Sex and Violence.

Following the evaluation of the guidelines, several challenges emerged during the annotation process, particularly regarding the interpretation and application of age classification criteria. Annotators reported difficulties when assigning age categories to texts that did not explicitly contain the thematic elements outlined in the directives (such as terms related to drugs, violence, or threats). One annotator, for example, questioned how to handle cases where, even in the absence of these indicators, she felt that the content was still unsuitable for minors. The annotator also explained that she was annotating based on the principle of "Would I show this to a child of a certain age?" whereas I considered the age that, in my view, the viewer should have to access the content. This resulted in significant differences in annotations, especially between neighboring classes.

The observations raised a significant issue: initially, it was thought that annotating based on themes, such as marking topics like economics or politics as unsuitable for children, would provide a straightforward guideline. The assumption was that, even in the absence of explicit indicators, these themes would inherently not interest children and thus should not be classified as suitable. However, this approach still involves a degree of subjectivity. Firstly, no topic can be deemed inherently inappropriate for children, as suitability depends heavily on how the content is presented. Secondly, even if a theme is not typically child-

friendly, determining the appropriate age range (e.g., 12 years, 15 years, or 18 years) remains subjective.

Title	**001A**	**002A**	**003A**	**Title**	**001A**	**002A**	**003A**
Ted2020-1	15	12	15	Ted2020-20	12	12	12
Ted2020-2	15	12	15	testo_ytkids_0047	12	12	12
Ted2020-3	15	15	18	Ted2020-22	15	12	15
testo_ytkids_0001	12	6	6	testo_ytkids_0057	6	3	6
Ted2020-5	15	12	15	testo_ytkids_0063	6	3	6
testo_ytkids_0002	3	3	3	Ted2020-26	15	12	15
testo_ytkids_0003	3	3	3	canzone_per_bambini_03	3	3	6
Ted2020-8	15	12	15	Ted2020-28	12	12	12
testo_ytkids_0011	3	6	3	canzone_per_bambini_09	6	3	6
Ted2020-10	15	12	15	Ted2020-30	15	12	12
testo_ytkids_0021	6	12	6	Ted2020-31	15	12	12
Ted2020-12	15	15	15	Ted2020-32	15	12	12
testo_ytkids_0024	6	12	12	canzone_per_bambini_10	6	3	6
testo_ytkids_0026	12	6	6	testo_ytkids_0035	6	6	12
testo_ytkids_0030	12	12	12	testo_ytkids_0046	12	12	12

Table 2.8: Sample annotation ratings for different texts by three annotators (001A, 002A, 003A).

When excluding the theme as a basis for age classification, we considered assessing the language complexity. For instance, if a text contains technical terms, it might be marked as unsuitable for younger audiences. Yet, this also introduces subjectivity, as a 15-year-old with a keen interest in a certain topic may find the content accessible despite its technical nature. Thus, judging based on technical language remains subjective, too.

Consequently, I chose to rely on the annotators' judgment, allowing them to decide based on their experience, while setting aside a more detailed study of theme and language complexity for future development. As noted earlier, I incorporated an automatic text complexity calculation column in the annotation schema to add another layer of evaluation. An initial experiment on a small sample of texts calculated complexity by assessing the frequency of words found in the NVdB (for Italian) and the Oxford 3000 (for English), normalized by the total number of words in the text. While this approach provides some insight, it has limitations; for instance, it considers only vocabulary, not syntax. A possible future improvement could be developing a value-age association scale to better link text complexity with age brackets.

Alternatively, in the future, a perspectivist approach (V. Basile et al. 2021) could be considered, one that captures the different viewpoints of annotators and treats disagreements as informative data.

In general, I suggested to the annotators relying on their own judgment by asking themselves, "Even if these elements are absent, do you think it's appropriate? Would you show it to a child? ', since the primary goal is to have a classifier that indicates the most suitable age to view the content, preventing minors from being exposed to inappropriate material, and offering an indicative age for which the content may be appropriate. This speaker experience allows human annotators to judge both the tone and intended audience of the content in a way that goes beyond literal words and phrases. By relying on this judgment, a human annotator can ensure that content suitability reflects not only the presence or absence of sensitive terms but also the broader appropriateness of themes and linguistic complexity for the intended age group.

However, as we can see from the Table 2.8, this subjectivity has led to a low IAA. The table shows the annotation results, highlighting a comparison of the classifications made by the three annotators. The reasons for the low agreement among annotators become clear. We see that disagreement often arises when one annotator has classified the content as suitable for viewers aged 12, while the other two have classified it as suitable for those aged 15 (as in the cases of *Ted2020-1*, *Ted2020-2*, *Ted2020-5*, *Ted2020-8*, and *Ted2020-10*). Conversely, in other cases, two annotators have marked the content as suitable for viewers aged 15, while one has marked it as suitable for viewers aged 12, as in cases *Ted2020-30*, *Ted2020-31*, and *Ted2020-32*. This pattern also occurs with other neighboring age classes; for example, for some YouTube Kids content, two annotators classified it as appropriate for age 6, while one classified it as appropriate for age 3.

In conclusion, based on the annotations and discussions regarding the 30 texts reviewed with the annotators, I adopted the majority vote approach. I annotated the remainder of the corpus independently, thereby turning the initial limitation (lack of personnel) into an advantage for the project by ensuring a consistent human judgment across the entire corpus. Although, as noted, this process could be further refined in the future with appropriate considerations and resources.

2.7 Final Corpus Composition

After the preliminary tests with the five age categories defined in the annotation guidelines, it became apparent that both the annotators and the models experienced similar challenges in distinguishing between adjacent age labels. This difficulty highlighted a consistent pattern of confusion when classifying texts in categories with close age ranges, such as those under 6, 12, and 15 years.

Given that the primary objective is to prevent minors from being exposed to unsuitable content, I decided to merge neighboring age groups to improve the accuracy and robustness of the classification model. The final categories are as follows:

- 0_kids (0-11): Content deemed suitable for children under 12.
- 1_teenagers (12-17): Content with moderate themes suitable for teenagers.
- 2_adults (18+): Content restricted to adults due to explicit themes.

Labels	**Age Range**	**Number of Texts**	**Number of Tokens**
0_kids	0-11	200	336,805
1_teenagers	12-17	146	449,439
2_adults	18+	100	257,016
Total		**446**	**1,043,260**

Table 2.9: English Corpus Composition.

Tables 2.9 and 2.10 summarize the number of texts and total tokens for each consolidated label, with 446 texts in the English corpus and 514 in the Italian corpus. These differences in token counts reflect inherent linguistic and structural variations between the two languages, which may influence certain classification aspects.

Labels	**Age Range**	**Number of Texts**	**Number of Tokens**
0_kids	0-11	209	62,409
1_teenagers	12-17	179	440,661
2_adults	18+	126	234,349
Total		**514**	**737,419**

Table 2.10: Italian Corpus Composition.

By consolidating the labels, the corpus is now organized into three broader categories that align more effectively with the classification objective. This reorganization minimizes misclassifications and emphasizes the critical task of age-appropriate content filtering.

The annotated corpus described in this study is not currently available for public access. However, there are plans to make it accessible through a GitHub repository in the near future.

Chapter 3

Experiments

3.1 Experimental Evaluation of Traditional Machine Learning Models

Following the methodology outlined in Section 2.1, the annotated corpora described in Section 2.5 were used to evaluate various ML models. The datasets were divided into training and testing sets using a k-fold cross-validation strategy to ensure robust evaluation. This section focuses on the evaluation of traditional ML classifiers, including SVM, NB, DT, RF, and k-NN.

The annotated texts were preprocessed as described in Section 2.1. Specifically, stopwords in both English and Italian were removed using the `NLTK` library, texts were normalized by converting them to lowercase and removing punctuation, and tokenization was applied to split the text into individual words. The texts were then numerically represented using the TF-IDF method, implemented through the `TfidfVectorizer` class from the `scikit-learn` library. This process calculated term frequencies (TF) and applied an Inverse Document Frequency (IDF) transformation to reduce the weight of common terms. Finally, the resulting vectors were normalized to serve as input for the models.

The models were evaluated using k-fold cross-validation, with $k = 5$. The dataset was divided into five equally sized folds, with the model trained on four folds and tested on the remaining fold during each iteration. This process ensured that every data point was used for both training and testing. The results were averaged over the five iterations to provide reliable estimates of the model performance. Metrics such as accuracy, precision, recall, and F1-score were calculated to assess the classifiers.

To improve classification and adhere to AGCOM guidelines, the lexicon-based rule described in Section 2.2 was applied. Specifically, if a bad word (contained in the dictionary) was found in the text and the model labeled it as "for kids", the label was changed to "for

adults". Before implementing this rule, an evaluation was conducted to determine whether it significantly affected the model's accuracy. For English, the accuracy remained unchanged except for the NB model. For Italian, the initial results showed lower accuracy compared to English but improved with the application of this rule across all classifiers.

The results are shown in the Table 3.1

Classifier	**ITA**	**BW_ITA**	**ENG**	**BW_ENG**
SVM	0.73	**0.78**	0.86	0.86
Naive Bayes	0.73	**0.78**	0.87	**0.88**
Decision Tree	0.67	**0.72**	0.85	0.85
Random Forest	0.74	**0.79**	0.87	0.87
K-Nearest Neighbors	0.68	**0.70**	0.87	0.87

Table 3.1: Accuracies of various classifiers before and after applying the bad words rule in both Italian and English.

This rule proved particularly useful in cases where the model failed to correctly identify borderline instances, such as YouTube Poop (YTP) content. Error analysis revealed that these cases were among those where the label was adjusted during the prediction phase. YTP content often mimics children's videos in structure (e.g., featuring the same characters) but includes inappropriate language, making accurate detection and classification challenging. Indeed, the majority of errors were associated with this type of content.

For example, in the case of English, the text *YTP_EN_16*, labeled as "for adults" in the corpus, was initially classified as "for kids" by the NB classifier. However, the rule applied during the prediction phase successfully corrected the label.

Text: [...] SpongeBob, I got diarrhea, wait spiderbob this a serious, Patrick you're right, we got to convince Squidward to come back home I have an idea, let's leave! [Music] good [Music] night split word speaking I hope you've been showing your **pussy** tonight baby. [...]
Title: YTP_EN_16.txt
Detected Bad Words: 'pussy', 'ass', 'cock', 'penis'
Original label: 2_adults
Predicted label: 0_kids
Changed to: 2_adults

3.1.1 Support Vector Machine

The SVM model was implemented using the One-vs-One (OvO) approach, which is particularly useful for multi-class classification tasks like this one, as it allows the model to

effectively distinguish between each pair of categories (Kids, Teenagers, Adults) by training separate classifiers for each pair. Specifically, for this task, three binary classifiers were trained, each focused on distinguishing between two of the three categories:

- Classifier 1: Distinguishes between class 0 (kids) and class 1 (teenagers).

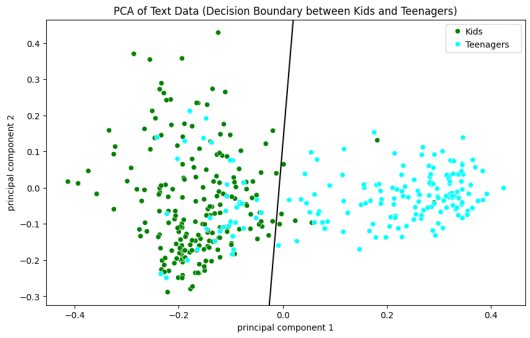


Figure 3.1: PCA of Italian Texts for the labels "Kids" and "Teenagers".

- Classifier 2: Distinguishes between class 0 (kids) and class 2 (adults).

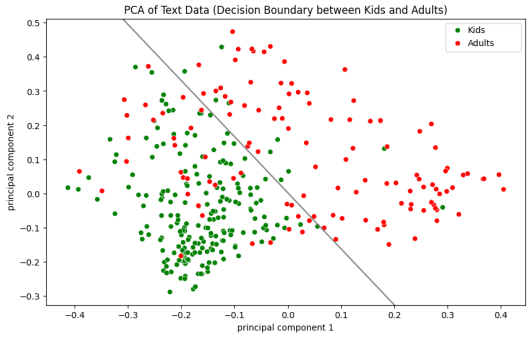


Figure 3.2: PCA of Italian Texts for the labels "Kids" and "Adults".

- Classifier 3: Distinguishes between class 1 (Teenagers) and class 2 (adults).

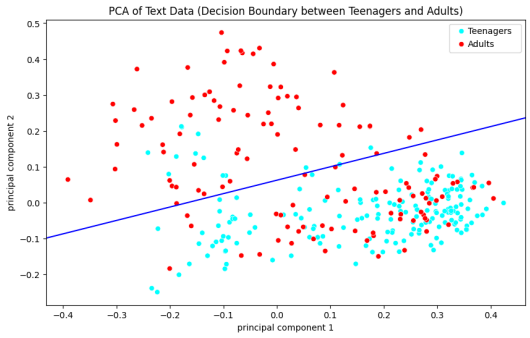


Figure 3.3: PCA of Italian Texts for the labels "Teenagers" and "Adults".

Each binary classifier provided a vote for one of the two classes it distinguished, and the final class label was assigned based on the majority vote.

To account for non-linearly separable data, kernel methods were employed. The Radial Basis Function (RBF) and linear kernels were tested. `GridSearchCV` was used for hyperparameter tuning, where the following parameters were tested:

- kernel: the parameter that specifies the kernel type to be used in the algorithm. The options considered were `linear` and `rbf`.
- C: the parameter that represents the regularization strength. A higher value of C implies less regularization. The explored values were: 1, 10, 100.

The grid search was also set with a 5-fold cross-validation (cv=5). The dataset was divided into 5 parts, and the model was trained on four parts and tested on the remaining part, cycling through all parts.

For each combination, the model was trained and its performance was evaluated on the validation set. The process involved fitting the model 30 times in total (6 combinations $\times$ 5 folds).

After completing the grid search, the combination of hyperparameters that yielded the highest average validation accuracy was selected as the best parameters. The results indicated that the optimal hyperparameters for the SVM model, both for English and Italian, were:

- $C = 1$
- Kernel type = 'linear'

These parameters were chosen because they provided the best balance between bias and variance, minimizing the overall error in the validation sets[1]. The regularization parameter $C = 1$ ensures that the model is neither too rigid (which would result in underfitting) nor too flexible (which would result in overfitting), and the linear kernel was found to be the most suitable for the given data set, probably due to the nature of the data distribution and the problem at hand.

The results obtained from applying the SVM model are summarized in Tables 3.2 for English and 3.3 for Italian.

Observing the results, we notice that in the "Kids" category, for the English dataset, the model achieved perfect results, with a precision, recall, and F1-score of 1.00. However, in the Italian dataset, the model performed slightly lower in this category, achieving a precision of 0.82, recall of 0.97, and an F1-score of 0.89.

[1]Tuning the hyper-parameters of an estimator in https://scikit-learn.org/stable/modules/grid_search.html#grid-search, last accessed: 2024/07/14.

	Precision	Recall	F1-Score	Support
0_kids	1.00	1.00	1.00	200
1_teenagers	0.71	0.98	0.82	146
2_adults	0.93	0.41	0.57	100
Accuracy	0.86			
Macro avg	0.88	0.80	0.80	446
Weighted avg	0.89	0.86	0.85	446

Table 3.2: SVM classification report for the English language.

	Precision	Recall	F1-Score	Support
0_kids	0.82	0.97	0.89	209
1_teenagers	0.70	0.74	0.72	179
2_adults	0.89	0.54	0.67	126
Accuracy	0.78			
Macro avg	0.80	0.75	0.76	514
Weighted avg	0.79	0.78	0.78	514

Table 3.3: SVM classification report for the Italian langauge.

For the "Teenagers" category, the English model achieved a precision of 0.71, recall of 0.98, and an F1-score of 0.82. In contrast, the Italian model had more balanced results, with a precision of 0.70, recall of 0.74, and an F1-score of 0.72.

In the "Adults" category, the English model showed a precision of 0.93 but a low recall of 0.41, resulting in an F1-score of 0.57, indicating that the model missed many correct predictions. Similarly, the Italian model had comparable difficulties, with a precision of 0.89, recall of 0.54, and an F1-score of 0.67.

The overall accuracy of the SVM model was 0.86 for the English dataset and 0.78 for the Italian dataset. The macro average precision, recall, and F1-Score for the English dataset were 0.88, 0.80, and 0.80, respectively, while for the Italian dataset, the macro averages were 0.80, 0.75, and 0.76.

As shown in Table 3.4, where some of the misclassification errors for the English dataset are detailed with index, filename, true label, and predicted label for instances incorrectly predicted, the model frequently misclassified instances of class 2 (adults) as class 1 (teenagers). This suggests that there may be similarities between the features of these classes that caused confusion for the model. One potential reason for the suboptimal results in the "Adults" class could be the presence of borderline cases, such as YouTube Poop videos, which often contain content that blurs the lines between typical adult and teenager categories. This

phenomenon became evident upon examining the errors in detail.

Index	Filename	True Label	Predicted Label
57	YTP_EN_06.txt	2	1
58	YTP_EN_12.txt	2	1
59	YTP_EN_14.txt	2	1
60	YTP_EN_16.txt	2	1
61	YTP_EN_18.txt	2	1

Table 3.4: Misclassification errors made by the SVM model for the English language.

This is a particular case because, for Italian, the same type of video was consistently misclassified, but not as "for teenagers"; rather, it was labeled as "for kids". Thus, the confusion, as shown in Table 3, occurred between label 0 and label 2. This issue was effectively resolved by the bad words rule, a resolution that did not occur with the "teenagers" label.

Filename	True Label	Predicted Label	Identified Bad Words
YTP_001.txt	2	0	'cazzo', 'checca', 'cazzi', 'merda', 'stronzetti', 'puttana'
YTP_002.txt	2	0	'coglione', 'cazzo', 'merda', 'culo'
YTP_003.txt	2	0	'cazzetto', 'piscio', 'troia', 'cazzata'
YTP_004.txt	2	0	'cazzone', 'scopare', 'fighetta', 'cazzo', 'fica', 'merda', 'culo'
YTP_006.txt	2	0	'cesso', 'cazzo', 'merda'

Table 3.5: Misclassification errors made by the SVM model for the Italian language.

The error made by the model tested on English is less severe, as the primary goal is to prevent children from being exposed to such content. Moreover, this type of video is typically created by teenagers for a teenage audience. While this kind of language cannot be considered appropriate for minors, their exposure to it is less critical. Considering the guidelines used for annotation (see Section 2.6), a small amount of profanity is permissible in videos aimed at teenagers.

A similar issue could be addressed in the future by establishing additional rules to distinguish between content unsuitable for teenagers and adults. For instance, this could include criteria such as the percentage of profanity present or other comparable measures.

The overall accuracy of the SVM model was 0.86 for the English dataset and 0.78 for the Italian dataset. The macro average precision, recall, and F1-Score for the English dataset were 0.88, 0.80, and 0.80, respectively, while for the Italian dataset, the macro averages were 0.80, 0.75, and 0.76.

In conclusion, the SVM model performed exceptionally well in the Kids category, achieving perfect results for the English dataset and high results for the Italian dataset. However, both datasets showed challenges in the Adults category, where the recall was particularly low, leading to a lower F1-score. These issues, especially the misclassification of adult content as teenager content, could be further addressed with additional rules or features, such as handling borderline cases like YouTube Poop videos.

3.1.2 Naive Bayes classifier

The second model, applied to the task of classifying text related to multimedia content into three categories, was the NB model. Specifically, the Multinomial Naive Bayes[2] classifier (`Multinomial NB`) was implemented. This model uses Bayes' Theorem to calculate the posterior probability for each class based on the features (words) present in the text. The category with the highest posterior probability is selected as the predicted label for the text.

For this experiment, the Multinomial Naive Bayes algorithm was chosen due to its effectiveness in handling text data and word frequencies, which are crucial for distinguishing between the three target categories.

The classifier's performance was evaluated on both the English and Italian datasets. The results for the English dataset are presented in Table 3.6, while the results for the Italian dataset are summarized in Table 3.7.

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.99	1.00	1.00	200
1_teenagers	0.74	0.98	0.84	146
2_adults	0.96	0.48	0.64	100
Accuracy	0.88			
Macro avg	0.90	0.82	0.83	446
Weighted avg	0.90	0.88	0.87	446

Table 3.6: Naive Bayes classification report for the English language.

For the Kids category, the NB classifier performed exceptionally well in both languages, although the English dataset yielded slightly higher results. For the English dataset, the model achieved a precision of 0.99, a recall of 1.00, and a F1-score of 1.00. This indicates near-perfect classification of texts targeted at children. In comparison, for the Italian dataset, the model reached a precision of 0.85, a recall of 0.89, and an F1-score of 0.87, which, while still high, reflects a slight drop in accuracy compared to the English results.

[2] 1.9.2. Multinomial Naive Bayes in https://scikit-learn.org/stable/modules/naive_bayes.html, last accessed: 2024/07/14.

	Precision	Recall	F1-Score	Support
0_kids	0.85	0.89	0.87	209
1_teenagers	0.67	0.82	0.74	179
2_adults	0.91	0.54	0.68	126
Accuracy	0.78			
Macro avg	0.81	0.75	0.76	514
Weighted avg	0.80	0.78	0.78	514

Table 3.7: Naive Bayes classification report for the Italian language.

These results suggest that the model can effectively identify children's content in both languages, but the higher accuracy in English may be attributed to the nature of the dataset or linguistic differences between the two languages.

In the Teenagers category, the NB classifier showed a more noticeable difference between the two languages. For the English dataset, the model achieved a precision of 0.74, a recall of 0.98, and an F1-score of 0.84. The high recall indicates that the model correctly identified the majority of texts meant for teenagers, though the lower precision suggests that some texts not intended for teenagers were misclassified as such.

For the Italian dataset, the results were lower across all metrics, with a precision of 0.67, a recall of 0.82, and an F1-score of 0.74. This indicates that, also in this case, the model struggles more to distinguish teenage content in Italian than in English.

The Adults category presented the most significant challenge for the NB classifier, with both languages showing similar difficulties. For the English dataset, the model achieved a precision of 0.96, which is quite high, but the recall was only 0.48, resulting in an F1-score of 0.64. This means that while the model is good at identifying adult content when it does so, it often fails to recognize a large portion of such texts, leading to a lower recall.

Similarly, for the Italian dataset, the model achieved a precision of 0.91 and a recall of 0.54, leading to an F1-score of 0.68. Despite the slight improvement in recall for the Italian dataset, both languages exhibit the same pattern: high precision but low recall, indicating that the classifier misses a considerable number of texts intended for adults.

When comparing the overall performance between the two languages, the English dataset showed better results in both the Kids and Teenagers categories, with higher precision, recall, and F1-scores. However, for the Adults category, the Italian dataset achieved a slightly higher recall (0.54 compared to 0.48), suggesting that the model performed marginally better in identifying adult content in Italian.

The overall accuracy of the NB model for the English dataset was 0.88, while for the Italian dataset, it was 0.78. This disparity in accuracy is reflected in the macro averages: for English, the model achieved a precision of 0.90, a recall of 0.82, and an F1-score of 0.83; for

Italian, these metrics were lower, with a precision of 0.81, a recall of 0.75, and an F1-score of 0.76.

In conclusion, while the NB classifier performed well overall, especially in the Kids category, it faced challenges in the Adults category in both languages, with slightly better recall in Italian. These results highlight the difficulties of classifying content aimed at adults, likely due to the more nuanced and variable nature of adult-oriented texts.

In comparison with the SVM model previously tested, the NB classifier performed better in the Kids category, particularly in terms of recall and F1-score for both languages. However, in the Adults category, both models faced similar difficulties, with NB achieving slightly higher recall than SVM in Italian (0.54 versus 0.41). For English, NB also showed a better overall performance, with an accuracy of 0.88 compared to SVM's 0.86.

In this case, analyzing the results, we observe that the English model misclassified many of the TED Talk texts, labeling them as intended for children instead of adults. On the other hand, the Italian model made errors in classifying YTP texts, categorizing them as suitable for children rather than adults, similar to what occurred with the SVM.

Overall, while NB outperformed SVM in certain categories, particularly for Kids, both models demonstrated challenges in correctly identifying content for Adults, with NB showing a slight advantage in recall for the Italian dataset.

3.1.3 Decision Tree

As discussed in the methodology section (see 2.2.3), the DT model works by selecting the most important features from the training dataset—words that best separate the three target classes: Kids, Teenagers, and Adults. The model uses metrics such as Gini impurity or entropy to evaluate each feature and chooses the one that provides the greatest reduction in impurity at each node. This recursive process continues until the entire DT is constructed, allowing the model to effectively classify texts based on these decision rules. Figure 3.4 shows the DT obtained for the Italian dataset, highlighting the features selected by the model to make its classifications.

With the model fully trained, we proceeded to evaluate its performance on both the English and Italian datasets. The results are summarized in Table 3.8 for English and Table 3.9 for Italian. Similar to the other models, the DT performed better on the English dataset, achieving an accuracy of 0.85, compared to an accuracy of 0.72 for the Italian model. For the English model, `GridSearchCV` selected a maximum depth of 10 nodes, while for the Italian model, it selected a maximum depth of 4 nodes.

As with the other models, the DT achieved its best performance in the Kids category. In English, the model reached a precision of 0.98, a recall of 0.96, and an F1-score of 0.97. For Italian, precision was 0.84, with a recall of 0.88, and an F1-score of 0.86. These results

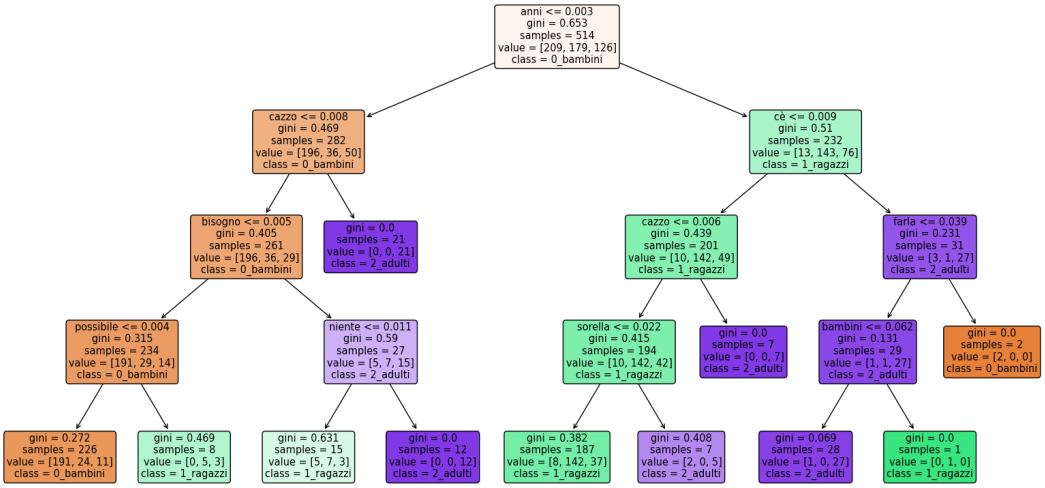


Figure 3.4: Decision Tree graphical representation for the Italian language.

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.98	0.96	0.97	200
1_teenagers	0.74	0.82	0.78	146
2_adults	0.74	0.64	0.68	100
Accuracy	0.85			
Macro avg	0.82	0.81	0.81	446
Weighted avg	0.85	0.85	0.84	446

Table 3.8: Decision Tree classification report for the English language.

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.84	0.88	0.86	209
1_teenagers	0.61	0.72	0.66	179
2_adults	0.71	0.48	0.57	126
Accuracy	0.72			
Macro avg	0.72	0.69	0.70	514
Weighted avg	0.73	0.72	0.72	514

Table 3.9: Decision Tree classification report for the Italian language.

indicate that the DT was highly effective in identifying content aimed at children, although the performance was slightly better in English.

In the Teenagers category, the English model achieved a precision of 0.74, a recall of 0.82, and an F1-score of 0.78. In comparison, the Italian model had lower metrics, with a precision of 0.61, recall of 0.72, and an F1-score of 0.66. These results suggest that the model struggled more to classify teenage content in Italian than in English.

For the Adults category, both models encountered the most difficulties. The English model obtained a precision of 0.74, a recall of 0.64, and an F1-score of 0.68, while the Italian model reached a precision of 0.71, but had a much lower recall of 0.48, leading to an F1-score of 0.57. This indicates that the DT model, similar to other classifiers, struggled to correctly identify all adult content, particularly in the Italian dataset.

Despite these challenges, the DT model produced relatively balanced results across the categories. In comparison with the NB and SVM models, the DT achieved similar performance for the English dataset, with a macro average precision of 0.82, recall of 0.81, and F1-score of 0.81, aligning closely with the SVM's results. However, in Italian, the DT showed lower macro averages compared to both NB and SVM, with a precision of 0.72, recall of 0.69, and F1-score of 0.70, highlighting some challenges in distinguishing between categories, especially in the Adults class.

Further analysis of the DT for the Italian dataset revealed some interesting insights. By analyzing the graphical representation of the DT for the Italian dataset (Figure 3.4), we can observe the specific words that it selected as key features. In particular, among the chosen features is the word *cazzo* (translated as *fuck* in English). The model correctly classified texts that contained this word frequently (more than 0.008 occurrences) into the "Adults" category. However, it also misclassified some texts containing this word at a lower frequency into the "Kids" category.

```
Title: YTP_001.txt
Detected Bad Words: {'cazzi', 'stronzetti', 'merda', 'cazzo', 'checca', 'puttana'}
Original Prediction: 0
Changed to '2_adulti'

Title: YTP_004.txt
Detected Bad Words: {'cazzone', 'cazzo', 'merda', 'culo', 'scopare', 'fica', 'fighetta'}
Original Prediction: 0
Changed to '2_adulti'

Title: YTP_007.txt
Detected Bad Words: {'cazzi', 'fanculo'}
Original Prediction: 0
Changed to '2_adulti'
```

Figure 3.5: Screenshot illustrating classification errors made by the DT model on Italian texts.

This pattern is also evident when the model errors are examined and the presence of inappropriate language is identified (as illustrated in Figure 3.5). The rule implemented to

explicitly check for such bad words, previously discussed in the methodology, proved to be effective, as it not only increased the accuracy from 0.67 to 0.72, but also ensured more accurate categorization of content.

By applying this rule, the model was able to reduce the risk of minors being exposed to inappropriate material, thereby enhancing the overall reliability of the classification system.

3.1.4 Random Forest

As discussed in the methodology, the RF model was selected due to its ability to reduce overfitting and generalize better compared to a single DT by averaging the predictions of multiple trees. For this experiment, the main hyperparameter that required tuning was the number of trees, k. Using `GridSearchCV`, the optimal configuration for the RF model was determined to be 300 trees for the English dataset and 200 trees for the Italian dataset. Additionally, the size of the bootstrap sample and the number of features randomly selected at each split were automatically set by `scikit-learn`. For both datasets, the `max_features` parameter was set to "auto", meaning that the model selected is $d = \sqrt{m}$, where m, is the number of features in the training set.

With these hyperparameters selected, the RF model was evaluated on both the English and Italian datasets. The results, shown in Table 3.10 for English and Table 3.11 for Italian, indicate that the RF model outperformed the DT model previously tested, achieving an accuracy of 0.87 for English and 0.79 for Italian. Notably, the Italian model performed better than all other classifiers tested thus far, demonstrating significant improvement across various categories.

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.99	1.00	0.99	200
1_teenagers	0.73	0.99	0.84	146
2_adults	1.00	0.45	0.62	100
Accuracy	0.87			
Macro avg	0.90	0.81	0.82	446
Weighted avg	0.90	0.87	0.86	446

Table 3.10: Random Forest classification report for English language.

Analyzing the performance across categories, we can see that the model achieved its best results in the Kids category. In English, the RF model obtained near-perfect results with a precision of 0.99, a recall of 1.00, and an F1-score of 0.99. For Italian, the model also performed strongly, with a precision of 0.82, recall of 0.98, and an F1-score of 0.89. This demonstrates the model's effectiveness in identifying children's content across both

	Precision	Recall	F1-Score	Support
0_kids	0.82	0.98	0.89	209
1_teenagers	0.70	0.75	0.72	179
2_adults	0.92	0.52	0.67	126
Accuracy	0.79			
Macro avg	0.81	0.75	0.76	514
Weighted avg	0.80	0.79	0.78	514

Table 3.11: Random Forest classification report for the Italian language.

languages.

For the Teenagers category, the English model achieved a precision of 0.73, recall of 0.99, and an F1-score of 0.84. The Italian model's performance was slightly lower, with a precision of 0.70, recall of 0.75, and an F1-score of 0.72. The high recall in both cases shows the model's capability to correctly identify most teenage content, though the lower precision suggests that some non-teenage texts were misclassified as teenage content. A detailed analysis of the errors made by the model for the English dataset shows that it predominantly misclassified numerous TED Talks, and some YTP videos, intended for adults as content for teenagers.

The Adults category presented the greatest challenge for the RF model. In English, while it achieved a perfect precision of 1.00, the recall was much lower at 0.45, leading to an F1-score of 0.62. This suggests that although the model rarely misclassified texts as adult content, it struggled to identify all adult-related texts. For Italian, the model achieved a precision of 0.92 but a recall of 0.52, resulting in an F1-score of 0.67. This pattern indicates the difficulty in accurately classifying adult content, a trend consistent across both languages.

Overall, in the English case, the RF model demonstrated superior performance compared to the DT, particularly in its ability to generalize across multiple categories, though its performance was slightly lower than that of the NB and SVM models. For Italian, as mentioned earlier, it emerged as the best-performing model. However, like other models, it struggled with the Adults category, suggesting that further refinement is needed to enhance classification accuracy for adult content.

3.1.5 K-Nearest Neighbor

The K-Nearest Neighbor (k-NN) model was evaluated on both the English and Italian corpora, with the results summarized in Tables 3.12 and 3.13. Figures 3.6 and 3.7 visually represent the decision regions for English and Italian, using Principal Component Analysis (PCA). These figures show how the model classified the test data (represented by triangles) based on proximity to the training data points (circles).

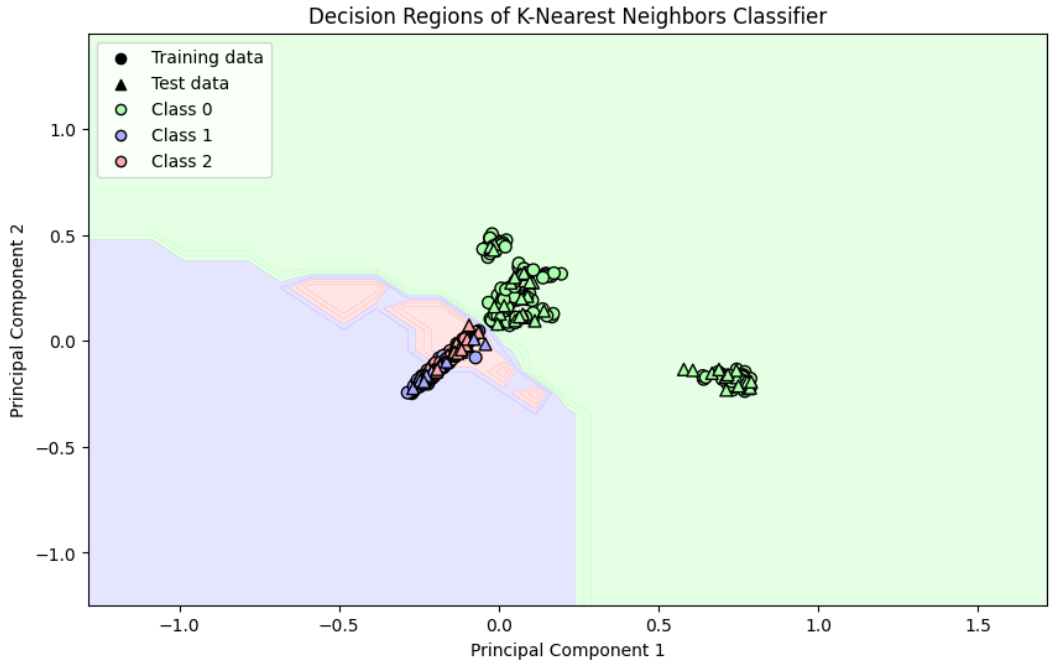


Figure 3.6: Graphical representation of the classification with KNN for English, obtained through PCA.

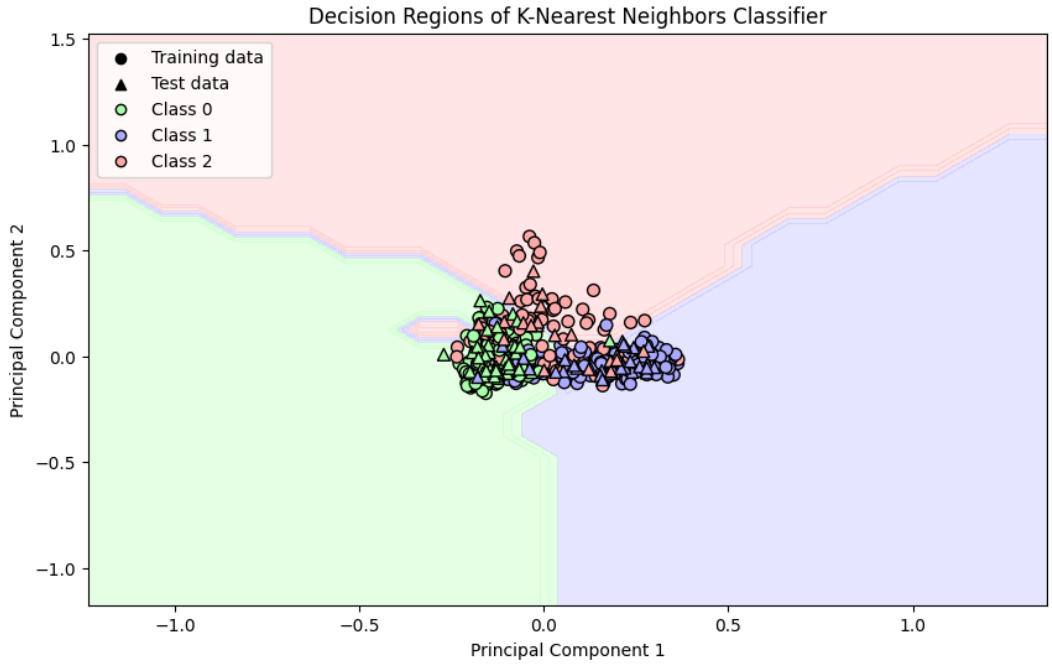


Figure 3.7: Graphical representation of the classification with KNN for Italian, obtained through PCA.

For this experiment, `GridSearchCV` selected the optimal number of neighbors as 7 for English and 9 for Italian, with the "uniform" weighting scheme. This means that the model gave equal weight to all k nearest neighbors when determining the class label.

The results show that the k-NN model performed best in the Kids category. For English, the model achieved perfect results, with a precision, recall, and F1-score of 1.00. However, for the Italian dataset, k-NN faced more difficulty, achieving a precision of 0.91, a recall of 0.69, and an F1-score of 0.78. These differences reflect the model's varying ability to classify children's content accurately across languages.

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	1.00	1.00	1.00	200
1_teenagers	0.75	0.90	0.82	146
2_adults	0.80	0.55	0.65	100
Accuracy	0.87			
Macro avg	0.85	0.82	0.82	446
Weighted avg	0.87	0.87	0.86	446

Table 3.12: k-NN classification report for the English language.

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.91	0.69	0.78	209
1_teenagers	0.62	0.83	0.71	179
2_adults	0.59	0.56	0.57	126
Accuracy	0.70			
Macro avg	0.71	0.69	0.69	514
Weighted avg	0.73	0.70	0.71	514

Table 3.13: k-NN classification report for the Italian language.

In the Teenagers category, k-NN showed higher recall than precision for both languages. In English, the model achieved a precision of 0.75, recall of 0.90, and an F1-score of 0.82. For Italian, it obtained a precision of 0.62, a recall of 0.83, and an F1-score of 0.71. The higher recall indicates that the model correctly identified most teenage texts, but the lower precision suggests that it misclassified some non-teenage texts as teenager. Indeed, a closer analysis of the results reveals that, in the case of Italian, the model erroneously classified both texts intended for children, primarily TED Talks, and texts intended for adults as content for teenagers.

The Adults category presented the greatest challenge for k-NN. For English, the model achieved a precision of 0.80 and a recall of 0.55, leading to an F1-score of 0.65. In Italian,

the model performed worse, with a precision of 0.59, recall of 0.56, and an F1-score of 0.57. This pattern shows that, like other models tested, k-NN struggled to correctly identify adult content, particularly in Italian.

Overall, the k-NN model produced strong results for English, with an accuracy of 0.87. For Italian, the model performed less effectively, achieving an accuracy of 0.70, the lowest result among all the models tested. These results indicate that, while k-NN works well for certain categories like Kids and Teenagers, it encounters difficulties in the Adults category, particularly in the Italian dataset.

3.1.6 Comprehensive Evaluation of Traditional Machine Learning Algorithms

After evaluating all classifiers (SVM, NB, DT, RF, and k-NN) on both the English and Italian datasets, we can draw several key insights into their performance and suitability for this text classification task across different languages.

For the English dataset, NB emerged as the top performer with an accuracy of 0.88, closely followed by RF and k-NN, both achieving an accuracy of 0.87. While SVM performed slightly lower with an accuracy of 0.86, it was still highly competitive. DT exhibited the lowest accuracy for English at 0.85, although its performance was still balanced across the categories, particularly in the Kids category.

In terms of precision, recall, and F1-score, NB and RF outperformed the other models, with NB having the edge in overall precision and F1-score, while RF demonstrated superior generalization abilities across categories. k-NN displayed a strong performance in the Kids and Teenager categories, but struggled slightly in the Adults category, where recall was particularly low, reflecting the model's difficulty in identifying adult content.

For the Italian dataset, the results were notably different. RF achieved the highest accuracy at 0.79, followed by NB and SVM, both at 0.78. k-NN and DT struggled more, with k-NN achieving the lowest accuracy at 0.70 and DT scoring 0.72. These differences suggest that some models, such as RF, were more robust to the complexities of Italian text, while others, like k-NN, encountered difficulties, particularly in the Kids and Adults categories.

Across both datasets, the Adults category consistently proved to be the most challenging for all models, particularly in terms of recall. Models often failed to capture all relevant instances of adult content, with k-NN and DT showing the greatest difficulties in this regard. In contrast, the Kids category yielded the best results across the board, with most models achieving high precision, recall, and F1-scores, indicating a strong ability to classify children's content accurately.

In conclusion, NB and RF were the most reliable models overall, with NB excelling in

precision and RF showing balanced performance across categories and languages. SVM also performed well, especially for English texts, while k-NN and DT exhibited weaker results, particularly for the Italian dataset.

These findings underscore the importance of selecting the most suitable classifier based on the language and unique challenges of classifying online video scripts, particularly in determining their suitability for minors.

3.2 Neural Networks

Following the methodology outlined in Chapter 2, specifically in Section 2.2.6, the experiments were conducted using the NN models described. CNN, LSTM networks, and BiLSTM networks were tested.

Also in this case, the impact of the badwords rule on the results was preliminarily evaluated. Table 3.14 shows that, once again, the rule improves the results obtained for Italian, while the results for English remain unchanged.

Classifier	**ITA**	**BW_ITA**	**ENG**	**BW_ENG**
CNN	0.75	**0.78**	0.86	0.86
LSTM	0.75	**0.77**	0.86	0.86
BiLSTM	0.72	**0.74**	0.85	0.85

Table 3.14: Accuracies of neural networks before and after applying the bad words rule in both Italian and English.

Each model was implemented using the `TensorFlow` and `Keras` libraries, with the parameters and settings specified in Section 2.2.6.

Specifically, the vocabulary size was limited to 10,000 words, and each text sequence was padded or truncated to a length of 100 tokens to ensure a uniform input size. The same pre-processing pipeline was applied to both English and Italian datasets to maintain consistency across the experiments. The models were then trained and evaluated using the created corpora, and key performance metrics, such as accuracy, precision, recall, and F1-score, were recorded for each category: kids, teenagers, and adults.

These settings were maintained for both English and Italian. The results are reported in Tables 3.15 and Table 3.16, respectively.

In the case of CNN, the accuracy score for English (0.86) was higher than for Italian (0.78). For the "Kids" category in both languages, precision and recall achieved very high scores, indicating that the model is highly capable of identifying texts aimed at children.

For the "Teenagers" category, recall remained high (0.92 for English and 0.83 for Italian), but precision decreased (0.74 for English and 0.68 for Italian). This meant that, while the

	Precision	Recall	F1-Score	Support
0_kids	1.00	0.98	0.99	200
1_teenagers	0.74	0.92	0.82	146
2_adults	0.78	0.52	0.62	100
Accuracy	0.86			
Macro avg	0.84	0.81	0.81	446
Weighted avg	0.86	0.86	0.85	446

Table 3.15: CNN classification report for the English language.

	Precision	Recall	F1-Score	Support
0_kids	0.86	0.90	0.88	209
1_teenagers	0.68	0.83	0.75	179
2_adults	0.83	0.50	0.62	126
Accuracy	0.78			
Macro avg	0.79	0.74	0.75	514
Weighted avg	0.79	0.78	0.77	514

Table 3.16: CNN classification report for the Italian language.

model correctly identified most texts belonging to the "Teenagers" class, some texts were incorrectly classified as such. This trend was consistent across both Italian and English datasets. Specifically, a significant number of texts intended for adults were misclassified as content for teenagers.

In contrast, the "Adults" category showed higher precision (0.78 for English and 0.83 for Italian) but lower recall (0.52 for English and 0.50 for Italian), meaning that while the model was more conservative in classifying texts as "adult" content, it was often correct when it did so. However, it tended to miss some texts that should have been classified as "adult".

Overall, the macro-average results for both English and Italian can be considered satisfactory.

In addition to CNNs, as mentioned at the beginning of the paragraph, recurrent neural networks (RNNs) were also tested, specifically the LSTM and BiLSTM architectures.

The results for the LSTM model are shown in Table 3.17 for English and in Table 3.18 for Italian.

When analyzing the English results, the model showed excellent performance in the children category, achieving very high precision and recall, resulting in an F1-score of 0.99. In the teenagers category, the model achieved a precision of 0.75 and a recall of 0.84, indicating

	Precision	Recall	F1-Score	Support
0_kids	1.00	0.97	0.99	200
1_teenagers	0.75	0.84	0.79	146
2_adults	0.67	0.58	0.62	100
Accuracy	0.84			
Macro avg	0.81	0.80	0.80	446
Weighted avg	0.84	0.84	0.84	446

Table 3.17: LSTM classification report for the English language.

	Precision	Recall	F1-Score	Support
0_kids	0.91	0.83	0.86	209
1_teenagers	0.71	0.79	0.75	179
2_adults	0.64	0.63	0.63	126
Accuracy	0.77			
Macro avg	0.75	0.75	0.75	514
Weighted avg	0.77	0.77	0.77	514

Table 3.18: LSTM classification report for the Italian language.

that while most texts aimed at teenagers were correctly identified, a small portion was misclassified. For the Adults category, the model exhibited weaker performance, with a precision of 0.67 and a recall of 0.58, showing that the model tended to confuse adult texts with other categories. Overall, the LSTM model achieved an accuracy of 84%.

For the Italian dataset, the model followed a similar pattern. The children category yielded the best results, with an F1-score of 0.86, although slightly lower than the English results. The Teenagers category performed similarly to the English case, with a precision of 0.71 and a recall of 0.79. The adults category remained the most challenging, with a slightly better F1-score of 0.63 compared to English, but it still showed that the model struggled with this class. Overall, the model achieved an accuracy of 77% for the Italian dataset.

In addition to the LSTM, the BiLSTM model was tested.

Looking at the results for Italian, we can observe a similar pattern, where the children category was again the one where the model performed best, achieving an F1-score of 0.86. Although this was slightly lower than the score for English, the result was still satisfactory. The teenagers category yielded similar results to English, with a precision of 0.71 and a recall of 0.79, leading to an F1-score of 0.75. Once again, the adults category presented the most challenges, although the performance was slightly better than in English, with an F1-score of 0.63, underscoring the difficulties the model faced with this category. Overall,

the model achieved an accuracy of 77% for Italian.

The results are shown in Table 3.19 for English and Table 3.20 for Italian.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0_kids	1.00	0.99	1.00	200
1_teenagers	0.74	0.86	0.80	146
2_adults	0.72	0.57	0.64	100
Accuracy	0.85			
Macro avg	0.82	0.81	0.81	446
Weighted avg	0.85	0.85	0.85	446

Table 3.19: BiLSTM classification report for the English language.

Class	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.90	0.80	0.85	209
1_teenagers	0.72	0.70	0.71	179
2_adults	0.55	0.68	0.61	126
Accuracy	0.74			
Macro avg	0.73	0.73	0.72	514
Weighted avg	0.75	0.74	0.74	514

Table 3.20: BiLSTM classification report for the Italian language.

For the English dataset, the BiLSTM model achieved excellent performance in the Kids category, with an F1-score of 1.0. The Teenagers category achieved a precision of 0.74 and a recall of 0.86, indicating that the model correctly identified most texts in this class, though some were misclassified. As with the LSTM model, the adults category was the most problematic, with a precision of 0.72 and a recall of 0.57. Overall, the model achieved an accuracy of 0.85.

For the Italian dataset, the results were slightly lower, with an accuracy of 0.74. The Kids category still showed strong performance, with a precision of 0.90 and a recall of 0.80. This results in an F1-score of 0.85, reflecting a strong balance between precision and recall. The Teenagers category achieved a precision of 0.72 and a recall of 0.70, which indicates that 28% of the texts classified as "Teenagers" were actually from another category, while the adults category remained the most challenging, with a precision of 0.55 and a recall of 0.68. The F1-score of 0.61 suggests that the model's performance in this category is the weakest, primarily due to the low precision. Despite these challenges, the model achieved an overall accuracy of 0.74, which is lower than in English, but the performance remains consistent with the results obtained from other tests and is still satisfactory.

3.2.1 Comprehensive Evaluation of Neural Networks

Looking at the overall performance of the NNs tested, we can draw the following conclusions: For both English and Italian, the convolutional model (CNN) performed the best, although the difference from the recurrent model (LSTM) was minimal. The most significant difference is observed with the BiLSTM model, which achieved lower accuracy than the other two models for Italian. In contrast, for English, the LSTM model recorded the lowest accuracy, though the differences between the models remain minor.

One plausible explanation for why CNNs outperform RNNs, despite this being a text classification task, could be related to the size and nature of the dataset. CNNs are highly efficient at capturing local patterns, such as short sequences or phrases, which might be sufficient for classifying the content in this case. Additionally, the relatively small dataset may reduce the need to capture long-term dependencies, which LSTMs and BiLSTMs typically excel at. As a result, CNNs might benefit from their ability to quickly detect key features, making them more effective for this specific task.

3.3 Transformers

Based on the methodology detailed in section 2.2.7, the Transformers models were tested to evaluate their performance on the classification task.

The first model tested was BERT. Before selecting which version of BERT to use, several preliminary tests were performed on some models to determine whether the cased or uncased version would yield better results. These tests indicated minimal differences in accuracy for Italian (0.71 for the uncased version and 0.70 for the cased version) and no difference for English (0.87 for both versions). This demonstrated that models of this complexity could capture word meanings from context, regardless of case. To ensure consistency with models that only offer a cased version, results were reported using the cased version across all classifiers.

The preliminary results are reported in Table 3.21.

Further preliminary tests assessed the impact of adding a "bad-words" filter, previously used in other models, to see if it would improve performance. These tests showed that for English, results remained stable with or without the filter, while for Italian, accuracy consistently improved with its application, as we can observe in Table 3.22. This suggests that English models are capable of accurate classification by relying on broader contextual features, while in Italian, models continue to face challenges and exhibit confusion without this filter. As a result, the bad-words filter was applied to all reported results.

Accordingly, all reported results include the application of this filter. Examining the data also reveals that most label changes from 0 (Kids) to 2 (Adults) occur in the so-called

Model	Language	Cased	Uncased
BERT	English	0.87	0.87
BERT	Italian	0.70	0.71
DistilBERT	English	0.86	0.87
DistilBERT	Italian	0.71	0.70
mBERT	English	0.87	0.86
mBERT	Italian	0.70	0.72
dbmdz italian BERT	Italian	0.72	0.68

Table 3.21: Preliminary test results for the version of the transformer models.

Model	ITA	BW_ITA	ENG	BW_ENG
BERT	0.70	**0.72**	0.87	0.87
DistilBERT	0.71	**0.75**	0.86	0.86
mBERT	0.70	**0.74**	0.87	0.87
RoBERTa	0.77	**0.80**	0.89	0.89
dbmdz italian BERT	0.72	**0.74**	-	-
ELECTRA	0.73	**0.76**	-	-

Table 3.22: Accuracies of various models before and after applying the rule, for both Italian and English.

YouTube Poops (YTP) cases. As previously mentioned, these are mashups of pre-existing material, often involving cartoons re-dubbed with inappropriate language. This type of content poses a challenge even for advanced models like transformers, particularly in Italian, just as it did for other models, further underscoring the importance of integrating filters to prevent minors from accessing such material.

The results obtained by testing BERT for English are presented in Table 3.23, while the results for Italian are shown in Table 3.24.

Examining the results obtained from BERT, we find that the English model significantly outperforms the Italian model, likely due to BERT's initial training on English text; thus, its performance in Italian depends entirely on the fine-tuning applied. The overall accuracy for English is relatively high at 0.87, with the model performing best in the children category (F1-score of 1.0). However, for teenagers, while recall is high at 0.92, precision drops to 0.74, indicating that although the model identifies most texts suitable for teenagers, it also misclassifies some content as suitable for this age group.

To address the language issue, mBERT, was tested. The results of the experiment are

	Precision	Recall	F1-Score	Support
0_kids	1.00	1.00	1.00	200.0
1_teenagers	0.74	0.92	0.81	146.0
2_adults	0.88	0.49	0.59	100.0
Accuracy	0.87			
Macro avg	0.87	0.80	0.80	446
Weighted avg	0.89	0.87	0.85	446

Table 3.23: BERT classification report for the English language.

	Precision	Recall	F1-Score	Support
0_kids	0.84	0.87	0.85	209.0
1_teenagers	0.66	0.69	0.63	179.0
2_adults	0.71	0.55	0.57	126.0
Accuracy	0.72			
Macro avg	0.73	0.70	0.68	514
Weighted avg	0.75	0.72	0.70	514

Table 3.24: BERT classification report for the Italian language.

reported in Table 3.25 for English and Table 3.26 for Italian.

	Precision	Recall	F1-Score	Support
0_kids	1.00	1.00	1.00	200.0
1_teenagers	0.72	0.93	0.81	146.0
2_adults	0.93	0.46	0.58	100.0
Accuracy	0.87			
Macro avg	0.88	0.80	0.80	446
Weighted avg	0.89	0.87	0.85	446

Table 3.25: mBERT classification report for the English language.

mBERT's performance in English was consistent with BERT, achieving an accuracy of 0.87. For Italian, however, mBERT showed a slight improvement over BERT, with an accuracy increase from 0.72 to 0.74. Examining the English results, the F1-scores for the Kids and Teenagers categories closely mirrored those achieved by BERT. In the adults category, mBERT displayed a slight improvement in precision, but a slight decrease in recall (0.46 with mBERT compared to 0.49 with BERT), indicating continued difficulty in accurately classifying adult-targeted texts. For Italian, mBERT also showed some improvement in

	Precision	Recall	F1-Score	Support
0_kids	0.85	0.91	0.88	209.0
1_teenagers	0.64	0.80	0.71	179.0
2_adults	0.86	0.40	0.51	126.0
Accuracy	0.74			
Macro avg	0.78	0.70	0.70	514
Weighted avg	0.78	0.74	0.72	514

Table 3.26: mBERT classification report for the Italian language.

the children category, with the F1-score increasing from 0.85 to 0.88, and in the Teenagers category, where the F1-score rose from 0.63 to 0.71. However, the model struggled more in the adults category, with a notable drop in recall from 0.55 with BERT to 0.40 with mBERT, indicating difficulty in recognizing a significant portion of texts intended for adults.

Following the testing of mBERT, the results for DistilBERT were also examined. The DistilBERT results are shown in Table 3.27 for English and Table 3.28 for Italian.

	Precision	Recall	F1-Score	Support
0_kids	1.00	1.00	1.00	200.0
1_teenagers	0.75	0.88	0.80	146.0
2_adults	0.79	0.56	0.63	100.0
Accuracy	0.86			
Macro avg	0.85	0.81	0.81	446
Weighted avg	0.87	0.86	0.86	446

Table 3.27: DistilBERT classification report for the English language.

	Precision	Recall	F1-Score	Support
0_kids	0.89	0.88	0.88	209.0
1_teenagers	0.62	0.87	0.72	179.0
2_adults	0.86	0.36	0.50	126.0
Accuracy	0.75			
Macro avg	0.79	0.71	0.70	514
Weighted avg	0.79	0.75	0.73	514

Table 3.28: DistilBERT classification report for the Italian language.

For English, DistilBERT's accuracy was slightly lower than BERT's (0.86 vs. 0.87). Similar to BERT, DistilBERT achieved an F1-score of 1 in the children category, though it

showed challenges in the teenagers and adults categories. However, DistilBERT did perform better in the adults category, with an F1-score improvement from 0.59 with BERT to 0.63.

For Italian, DistilBERT achieved an accuracy of 0.75, slightly higher than BERT's 0.72, showing fewer classification challenges in the children category. In the teenagers category, DistilBERT performed significantly better than BERT, with an F1-score increase from 0.63 to 0.72. However, for the adults category, DistilBERT's performance slightly declined, with an F1-score of 0.50 compared to BERT's 0.57. In this category, DistilBERT's precision was 0.86, but the recall dropped to 0.36, indicating a limited ability to recognize texts meant for adults.

Also the Transformer-based model RoBERTa was tested. As outlined in the methodology (see Section 2.2.7), RoBERTa, was selected for its enhanced capabilities in capturing nuanced language patterns essential for text classification tasks. The results are reported in Table 3.29 for English and Table 3.30 for Italian.

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	1.00	1.00	1.00	200.0
1_teenagers	0.74	1.00	0.85	146.0
2_adults	1.00	0.49	0.65	100.0
Accuracy	0.89			
Macro avg	0.91	0.83	0.83	446
Weighted avg	0.92	0.89	0.88	446

Table 3.29: RoBERTa classification report for the English language.

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.87	0.90	0.88	209.0
1_teenagers	0.69	0.86	0.76	179.0
2_adults	0.94	0.54	0.67	126.0
Accuracy	0.80			
Macro avg	0.84	0.76	0.77	514
Weighted avg	0.83	0.80	0.79	514

Table 3.30: RoBERTa classification report for the Italian language.

Overall, RoBERTa achieved the highest accuracy among the models tested, with an accuracy of 0.89 for English and 0.80 for Italian. The model excelled in the classification of children's content, achieving an F1-score of 1.00 in English and 0.88 in Italian, indicating near-perfect precision and recall for texts in this category.

In the teenagers category, the English model showed high recall (1.00), indicating that

it correctly identified most texts intended for teenagers. However, the precision for this category was slightly lower (0.74), suggesting that some non-teenager texts were misclassified as teenagers. For Italian, the model achieved a precision of 0.69 and recall of 0.86, indicating similar behavior in this category across both languages.

For the adults category, the model achieved a high precision of 1.00 for English, meaning that when the model classified a text as adult content, it was usually correct. However, the recall was lower (0.49), revealing that the model missed many adult-targeted texts. This pattern was mirrored in the Italian model, where the precision for adults was high (0.94), but the recall was lower at 0.54. This suggests a consistent challenge in identifying all relevant adult texts, possibly due to overlapping language features between categories.

In summary, the RoBERTa model demonstrated the highest performance across all models tested, achieving top accuracy scores in both languages. However, as seen in previous models, it encountered challenges with the teenagers and adults categories, particularly in accurately identifying all adult-oriented texts. The lower recall for adults and the lower precision for teenagers indicate that the model occasionally misclassifies adult texts as teenager content and fails to recognize some teenager-targeted texts, possibly due to overlapping language patterns.

To improve performance in Italian, additional models directly specialized in Italian were tested.

The first model tested was the `dbmdz/bert-base-italian-cased`, used within the `BertForSequenceClassification` architecture for fine-tuning on the specific task. The final results are presented in Table 3.31.

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.85	0.89	0.87	209.0
1_teenagers	0.64	0.81	0.70	179.0
2_adults	0.80	0.39	0.49	126.0
Accuracy	0.74			
Macro avg	0.76	0.70	0.69	514
Weighted avg	0.77	0.74	0.72	514

Table 3.31: dbmdz classification report.

The results are comparable to those achieved with other BERT-based models fine-tuned on this corpus. This model also reached a satisfactory accuracy of 0.74, aligning with the other results but still lower than the performance of models tested on English. Consistent with the findings for all other models, this classifier performs well in the children's category but encounters challenges in the teenagers and adults categories. In the teenagers category, the model has low precision (0.64) but high recall (0.81), while in the adults category, it

shows high precision (0.80) but low recall (0.39).

The second model tested was `ELECTRA-based model`, specifically the `ElectraForSequenceClassification` pre-trained on Italian (`dbmdz/electra-base-italian-xxl-cased-discriminator`. The results are shown in Table 3.32

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.87	0.91	0.88	209.0
1_teenagers	0.65	0.89	0.74	179.0
2_adults	0.96	0.36	0.52	126.0
Accuracy	0.76			
Macro avg	0.82	0.72	0.71	514
Weighted avg	0.82	0.76	0.74	514

Table 3.32: Electra classification report.

The results show that this model achieves a slightly higher accuracy (0.76) than the previous one, though the RoBERTa model remains the best performer overall. Similar to previous cases, this model performs very well in the children's category but struggles with the teenagers and adults categories, showing low precision in the teenagers category and particularly low recall in the adults category.

As part of the experiment, specific Italian transformers were also explored, namely AlBERTo (based on BERT), and GilBERTo and UmBERTo (both based on RoBERTa).

AlBERTo is a BERT language understanding model for the Italian language focused on the language used in social networks, specifically on Twitter.

The results achieved by testing this model are shown in Table 3.33

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.91	0.86	0.88	209.0
1_teenagers	0.68	0.78	0.71	179.0
2_adults	0.73	0.59	0.63	126.0
Accuracy	0.76			
Macro avg	0.77	0.74	0.74	514
Weighted avg	0.79	0.76	0.76	514

Table 3.33: AlBERTo classification report.

AlBERTo achieved the same accuracy as Electra (0.76). Like the other models, it performs best in the children's category, achieving the highest precision so far among all models tested on Italian, with a precision of 0.91. In the adults' category, where all models faced

challenges, AlBERTo also attained a higher recall of 0.59, the best result among all models tested on Italian in this category.

The final two models tested are both based on RoBERTa. GilBERTo is an Italian pre-trained language model built on Facebook's RoBERTa architecture with a CamemBERT text tokenization approach. UmBERTo is another RoBERTa-based model, trained on extensive Italian corpora, utilizing two innovative methods: SentencePiece and Whole Word Masking.

The results for GilBERTo are presented in Table 3.34, and the results for UmBERTo in Table 3.35.

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.94	0.94	0.94	209.0
1_teenagers	0.79	0.89	0.84	179.0
2_adults	0.91	0.73	0.80	126.0
Accuracy	0.87			
Macro avg	0.88	0.86	0.86	514
Weighted avg	0.88	0.87	0.87	514

Table 3.34: GilBERTo classification report.

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.93	0.97	0.95	209.0
1_teenagers	0.85	0.91	0.88	179.0
2_adults	0.97	0.78	0.85	126.0
Accuracy	0.90			
Macro avg	0.92	0.89	0.89	514
Weighted avg	0.91	0.90	0.90	514

Table 3.35: UmBERTo classification report.

With the RoBERTa-based models trained specifically for Italian, performance improved significantly. In the adults category, GilBERTo achieved an F1-score of 0.94, while UmBERTo reached 0.95. Across other categories, the overall trend observed with previous models persisted, namely, lower precision and high recall in the teenagers category and high precision with lower recall in the adults category, but with noticeably stronger results. For example, GilBERTo achieved a precision of 0.79 and a recall of 0.89 in the teenagers category, and in the adults category, it achieved a precision of 0.91 and a recall of 0.73, resulting in an overall accuracy of 0.87. UmBERTo yielded the highest performance, achieving an accuracy of 0.90. Specifically, it reached an F1-score of 0.95 in the children's category, a precision of 0.85 with a recall of 0.89 in the teenagers category, and a precision of 0.97 with

a recall of 0.78 in the adults category.

3.3.1 Comprehensive Evaluation of Transformer-Based Models

In conclusion, the results across all observed models reveal some common trends. All models perform exceptionally well in the children category, with F1-scores starting from 0.85. In English, models like mBERT and RoBERTa achieved a perfect F1-score of 1.00, while similar results were obtained with GilBERTo and UmBERTo for Italian, reaching F1-scores between 0.94 and 0.95.

In the teenagers category, there is a general pattern of lower precision and higher recall across all models, suggesting that while the models successfully identify most teenage-targeted content, they tend to misclassify other content as belonging to this category. RoBERTa-based models perform better here, with GilBERTo achieving a precision of 0.79 and recall of 0.89, while UmBERTo reaches a precision of 0.85 and recall of 0.89.

The adults category remains the most challenging for all models, characterized by high precision but often low recall, indicating that while adult content is identified correctly, not all adult texts are detected. However, UmBERTo and GilBERTo (both RoBERTa-based) show notable improvements in this category, with UmBERTo achieving an F1-score of 0.85 and GilBERTo reaching 0.80.

Overall, the RoBERTa-based models were the best performers across both English and Italian, particularly UmBERTo for Italian, with an accuracy of 0.90, and RoBERTa for English, with an accuracy of 0.89.

3.4 Large Language Models

As described in Chapter 2, this experimental phase also evaluates the performance of several LLMs on the core classification task addressed in this thesis, for both English and Italian datasets. Also in this case, each model was fine-tuned on the built corpora with standardized parameters and evaluated through a 5-fold cross-validation for consistent results.

Additionally, the bad-word filter was applied post-prediction to reclassify content as necessary, following a preliminary test that demonstrated its effectiveness in improving classification accuracy. The results indicated a noticeable improvement in accuracy, particularly for the Italian dataset. The preliminary test was performed using Flan-T5. For English, accuracy increased slightly from 0.87 to 0.88 with the rule in place. However, for Italian, accuracy saw a more significant improvement, rising from 0.71 to 0.76. This preliminary result, reported in Table 3.36, highlighted the value of the bad words rule, particularly in handling Italian content, justifying its inclusion as part of the classification process in the experiment.

Classifier	ITA	BW_ITA	ENG	BW_ENG
Flan-T5	0.71	**0.76**	0.87	**0.88**
GPT-2	-	0.75	-	0.83
Minerva	0.72	**0.75**	-	0.83

Table 3.36: Accuracies of tested LLMs before and after applying the bad words rule in both Italian and English.

Among the various LLMs, Flan-T5, GPT-2, and Minerva were selected for testing. The results of Flan-T5 are shown in Talbe 3.37 for English and Table 3.38 for Italian.

	Precision	Recall	F1-Score	Support
0_kids	0.99	1.00	1.00	200.0
1_teenagers	0.73	0.99	0.84	146.0
2_adults	0.99	0.46	0.62	100.0
Accuracy	0.88			
Macro avg	0.90	0.82	0.82	446
Weighted avg	0.91	0.88	0.87	446

Table 3.37: Flan-T5 classification report for the English language.

	Precision	Recall	F1-Score	Support
0_kids	0.77	0.97	0.86	209.0
1_teenagers	0.72	0.80	0.76	179.0
2_adults	0.90	0.34	0.48	126.0
Accuracy	0.76			
Macro avg	0.80	0.71	0.70	514
Weighted avg	0.79	0.76	0.73	514

Table 3.38: Flan-T5 classification report for the Italian language.

The Flan-T5 model performed exceptionally well in the Kids category for English, achieving an F1-score of 1.0. However, as with previously tested models, it encountered more challenges in the "Teenagers" and "Adults" categories. Specifically, in the "Teenagers" category, it reached a high recall of 0.99 but had a lower precision of 0.73, indicating some misclassification. For the "Adults" category, the model attained a high precision of 0.99 but struggled with recall, achieving only 0.46. This overall pattern resulted in a final accuracy of 0.88 for English, consistent with other transformers tested.

Observing the results for Italian, we notice that the model encountered some difficulties

even in the children's category, achieving an F1-score of 0.86 but with a precision of only 0.77. In the "Teenagers" category, the F1-score was also 0.86, while the "Adults" category showed a significantly lower F1-score, largely due to a recall of only 0.34. Overall, the model reached an accuracy of 0.76 for Italian, which aligned with the results of other transformers that were not specifically specialized for Italian.

The OpenAI GPT-2 model, a causal transformer developed by Radford et al. (2019), was also evaluated for this classification task. As previously outlined in the methodology, this model was fine-tuned with a sequence classification objective, leveraging a 5-fold cross-validation approach to ensure reliable performance metrics.

Table 3.39 (English) and Table 3.40 (Italian) illustrate GPT-2's performance.

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.99	1.00	0.99	200.0
1_teenagers	0.79	0.66	0.70	146.0
2_adults	0.61	0.68	0.63	100.0
Accuracy	0.83			
Macro avg	0.80	0.78	0.77	446
Weighted avg	0.84	0.83	0.82	446

Table 3.39: GPT-2 classification report for the English language.

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.88	0.87	0.87	209.0
1_teenagers	0.73	0.68	0.70	179.0
2_adults	0.64	0.67	0.64	126.0
Accuracy	0.75			
Macro avg	0.75	0.74	0.74	514
Weighted avg	0.77	0.75	0.75	514

Table 3.40: GPT-2 classification report for the Italian language.

The results show that the model tested on English data achieved a lower precision than Flan-T5 and recorded the lowest overall accuracy among all models tested up to that point. Consistent with previous models, its best performance was observed in the Kids category, where it reached an F1-score of 0.99. However, the model struggled more in other categories, with an F1-score of only 0.70 in the Teenagers category and an even lower score of 0.63 in the Adults category.

For Italian, while the overall performance remained modest, it aligned with results seen in other transformers. The model achieved an accuracy of 0.75, performing best in the

Kids category with an F1-score of 0.87. In the Teenagers category, it scored 0.70, and in the Adults category, it reached 0.64, slightly outperforming its English counterpart in this specific category.

The last model evaluated in this experiment was Minerva, specifically the `Minerva-350M-base-v1.0` variant. Developed by Sapienza NLP, Minerva was the first family of LLMs pre-trained from scratch on English and Italian.

Table 3.41 presents the results for English, while Table 3.42 displays the results for Italian.

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.99	0.99	0.99	200.0
1_teenagers	0.71	0.83	0.76	146.0
2_adults	0.68	0.50	0.57	100.0
Accuracy	0.83			
Macro avg	0.79	0.77	0.77	446
Weighted avg	0.83	0.83	0.83	446

Table 3.41: Minerva report for the English language.

	Precision	**Recall**	**F1-Score**	**Support**
0_kids	0.85	0.91	0.88	209.0
1_teenagers	0.68	0.67	0.67	179.0
2_adults	0.72	0.59	0.62	126.0
Accuracy	0.75			
Macro avg	0.75	0.73	0.72	514
Weighted avg	0.76	0.75	0.75	514

Table 3.42: Minerva report for the Italian language.

Examining the results, we observe that Minerva achieved an accuracy of 0.83 on English data, performing best in the Kids category with an F1-score of 0.99. In the Teenagers category, Minerva reached an F1-score of 0.76, while in the Adults category, its performance declined, achieving an F1-score of only 0.57.

For Italian data, despite being trained on Italian sources, Minerva's performance was lower overall, with an accuracy of 0.75. Similar to its results in English, the best performance was observed in the Kids category with an F1-score of 0.88. However, in the Teenagers and Adults categories, the model struggled, yielding F1-scores of 0.67 and 0.62, respectively.

In summary, each LLM performed best in the Kids category, reflecting the models' ability to accurately identify less complex content. Challenges arose in classifying the Teenagers and

Adults categories, where models showed high precision but struggled with recall. Models tested on English typically showed slightly better accuracy than those tested on Italian, including Minerva, despite its Italian-specific training.

3.4.1 Final considerations on the experiment

The detailed results for each model in English and Italian are summarized in Table 3.43 and Table 3.44, respectively, with Table 3.43 presenting the performance metrics for English models and Table 3.44 illustrating the results for Italian models.

Classifier	**Representation**	**Precision**	**Recall**	**F1-score**	**Accuracy**
SVM	TF-IDF	0.88	0.80	0.80	0.86
Naive Bayes	TF-IDF	0.90	0.82	0.83	**0.88**
Decision Tree	TF-IDF	0.82	0.81	0.81	0.85
Random Forest	TF-IDF	0.90	0.81	0.82	0.87
K-Nearest Neighbors	TF-IDF	0.85	0.82	0.82	0.87
CNN	Word Embedding	0.84	0.81	0.81	**0.86**
LSTM	Word Embedding	0.81	0.80	0.80	0.84
BiLSTM	Word Embedding	0.82	0.81	0.81	0.85
BERT	Contextual Embeddings	0.87	0.80	0.80	0.87
mBERT	Contextual Embeddings	0.88	0.80	0.80	0.87
DistilBERT	Contextual Embeddings	0.85	0.81	0.81	0.86
roBERTa	Contextual Embeddings	0.91	0.83	0.83	**0.89**
Flan T5	Contextual Embeddings	0.90	0.82	0.82	**0.88**
Minerva (350M-base)	Contextual Embeddings	0.79	0.77	0.77	0.83
GPT-2	Contextual Embeddings	0.80	0.78	0.77	0.83

Table 3.43: Classification Results for English.

Among the models tested, transformer-based architectures proved the most effective, with roBERTa leading in English (accuracy: 0.89) and UmBERTo for Italian (accuracy: 0.90). These models demonstrated a superior capacity to handle age-based text classification, indicating that transformers, with their advanced contextual embeddings, are particularly adept at capturing nuanced linguistic markers essential for this task. In particular, all models achieved satisfactory accuracy for English, surpassing 0.80, while for Italian, all models scored above 0.70, reflecting strong but relatively lower results for Italian. This gap may arise from intrinsic differences in language structure; Italian morphological complexity and syntactic variability often present a greater challenge for classification than English, where structural simplicity aids model accuracy. Furthermore, the exceptional performance

Classifier	Representation	Precision	Recall	F1-score	Accuracy
SVM	TF-IDF	0.80	0.75	0.76	0.78
Naive Bayes	TF-IDF	0.81	0.75	0.76	0.78
Decision Tree	TF-IDF	0.72	0.69	0.70	0.72
Random Forest	TF-IDF	0.81	0.75	0.76	**0.79**
K-Nearest Neighbors	TF-IDF	0.71	0.69	0.69	0.70
CNN	Word Embedding	0.79	0.74	0.75	**0.78**
LSTM	Word Embedding	0.75	0.75	0.75	0.77
BiLSTM	Word Embedding	0.73	0.73	0.72	0.74
BERT	Contextual Embeddings	0.73	0.70	0.68	0.72
mBERT	Contextual Embeddings	0.78	0.70	0.70	0.74
DistilBERT	Contextual Embeddings	0.79	0.71	0.70	0.75
roBERTa	Contextual Embeddings	0.84	0.76	0.77	0.80
dbmdz/bert-italian	Contextual Embeddings	0.76	0.70	0.69	0.74
ELECTRA	Contextual Embeddings	0.82	0.72	0.71	0.76
AlBERTo	Contextual Embeddings	0.77	0.74	0.74	0.76
GilBERTo	Contextual Embeddings	0.88	0.86	0.86	0.87
UmBERTo	Contextual Embeddings	0.92	0.89	0.89	**0.90**
Flan T5	Contextual Embeddings	0.80	0.71	0.70	**0.76**
Minerva (350M-base)	Contextual Embeddings	0.75	0.73	0.72	0.75
GPT-2	Contextual Embeddings	0.75	0.74	0.74	0.75

Table 3.44: Classification Results for Italian.

of UmBERTo, a model specifically trained in Italian, reinforces the importance of language-specific training for accurate classification.

In comparison, simpler models like NB performed competitively in English but struggled more with Italian, potentially due to the inherent complexity of Italian morphology and syntax. Simpler models like NB are less capable of managing the linguistic intricacies and rich morphology present in Italian, which probably explains their lower performance in this context. The underperformance of k-NN in Italian further underscores these difficulties, as simpler ML models generally struggle with the linguistic variability that characterizes Italian.

Across the models, LLMs like GPT-2 and Minerva showed the lowest performance in English, confirming that the layer-based embeddings of the transformers capture contextual dependencies more effectively than the more specialized, broader embeddings found in the LLMs. This advantage of transformers lies in their ability to integrate sentence-level context

directly into their embeddings, allowing them to detect subtle, age-specific language features. LLMs, while powerful in general text generation, lack this fine-tuned context differentiation, making them less effective for nuanced tasks such as age-based content classification.

The introduction of a rule-based filter for "bad words" significantly improved the performance of the model, particularly in the category "Adults". This filter reclassifies texts containing explicit language as "adult" content, effectively compensating for the models' struggle to identify mature themes based solely on more subtle linguistic indicators. The filter also demonstrated particular effectiveness in YouTube Poop (YTP) content, where child-appropriate formats like cartoons are overlaid with inappropriate language. This finding highlights a key limitation in the models: Without explicit markers, they have difficulty detecting shifts in tone that result from content repurposing. This contextually nuanced content, where form (e.g. cartoon characters) contrasts with function (e.g. inappropriate language), requires explicit cues, such as bad words, for effective classification.

Performance on children's content, by contrast, was consistently high across all models. The simpler structure and direct language typical of texts intended for young audiences provide clear, unambiguous cues that facilitate classification. This trend aligns with the observation that straightforward language is more readily captured by models, enabling them to achieve high precision and recall scores in the children's category.

However, distinguishing between "teenagers" and "adults" proved challenging. Many models displayed high recall but low precision for the "teenagers" category, and the opposite for "adults", revealing a common misclassification tendency. This difficulty likely reflects the nature of the corpus itself; texts for both groups, including TED Talks, share similarities in tone, vocabulary, and structure, creating an ambiguity that even human annotators found challenging to resolve. This confusion is reflected in the models, which tend to misclassify adult-targeted texts as suitable for teenagers, leading to lower recall for adults and high recall for teenagers. Such results indicate that while the models handle clear-cut categories like children's content effectively, they struggle with more nuanced distinctions, particularly when texts share stylistic similarities but differ in intended audience.

In summary, UmBERTo for Italian and roBERTa for English demonstrated the highest effectiveness across the board, confirming the superiority of transformers in this domain. Traditional models, particularly NB, still show competitive results in English, suggesting that classical classifiers may be viable for certain languages where simpler syntactic structures prevail. For Italian, however, the superiority of transformers is more pronounced, indicating their value for more complex linguistic tasks.

Given these findings, future research could focus on developing refined age-detection strategies that go beyond explicit markers like bad words, integrating features that capture nuanced language shifts more accurately. Additionally, exploring hybrid models that

combine traditional ML classifiers with transformer embeddings could further enhance classification performance in complex cases. Expanding language-specific training resources, especially for languages like Italian, would be beneficial in strengthening the models' capacity to manage linguistic diversity, ultimately supporting a robust foundation for safe and accurate multimedia content classification across languages.

3.5 Semantic Analysis

As described in Chapter 2, the steps undertaken in this second experiment to achieve semantic classification included:

1. Constructing specialized dictionaries (covering categories such as violence, drugs, offensive language, and emotions, as described in Section 2.4).
2. Building the corpora (see Section 2.5).
3. Extracting the most frequent terms from each text.
4. Expanding these lists automatically using semantic similarity algorithms.
5. Generating a network of semantically similar texts with cosine similarity.
6. Representing the networks using *Gephi* (Bastian et al. 2009) and measuring their strength with the Modularity Class.
7. Labeling the networks through a dedicated labeling algorithm.

The first step in the experimental process involved using a distributional semantic matrix to compute similarity values between words and identify the nearest neighbors for each frequently occurring term. This similarity information was used to build networks of related terms and categories within the texts, representing nuanced relationships among content items in both languages.

In the subsequent steps, semantic networks were generated using cosine similarity to define edges between similar texts. These networks were visualized in *Gephi*, where the Modularity Class algorithm was applied to measure community strength, enabling the identification of distinct text clusters based on semantic similarity. The resulting networks were labeled with a majority-based algorithm, which was implemented using the Python library pandas, allowing for the aggregation and assignment of the most frequent label for each identified cluster. Smaller clusters with fewer than ten texts were filtered out to maintain result accuracy.

To implement the Word2Vec model, the `Gensim` library was selected, and preliminary tests were conducted to determine the optimal parameter combination. Table 3.45 presents the two configurations tested for each language along with the results obtained.

	Vector Size	Skip-gram	English Results	Italian Results
Test 1	100	0	0.41	0.22
Test 2	300	1	0.29	0.22

Table 3.45: Tested configurations for each language and obtained results.

To evaluate the performance of the models, reference words (ground truth) were selected, and accuracy was calculated as follows: for each target word, a list of similar words was generated using Gensim's `most_similar` method. These lists were then compared to predefined ground-truth lists, with accuracy determined by the number of matching words between the model's output and the ground-truth list. Specifically, the proportion of words from the ground truth that appeared in the model's list was calculated, with precision defined as the ratio of matching words to the total words in the ground-truth list. This process was repeated for each target word, and the overall accuracy was reported for each model.

Despite the lower accuracy values, which are expected in this type of evaluation since the goal is embedding optimization, these preliminary tests were essential for selecting the most suitable Word2Vec configuration to ensure high-quality word representations for the subsequent analysis.

The tests showed that the English models generally achieved higher accuracy than the Italian models. While the configurations tested for Italian showed no significant differences in accuracy, those for English did. The configuration with a lower vector count and the use of `CBOW` instead of `Skip-gram` performed best for English. Consequently, Test 1 was selected for model training. This choice aligns with the literature, which suggests that `CBOW` is particularly suited for capturing common words rather than rare ones, making it ideal for this study's objectives[3].

The matrix was used to perform a semantic expansion of the most significant words in the transcripts: the 100 highest-frequency terms from each text were extracted, and the 10 nearest neighbors for each were identified. This produced a vector where each text is represented by its list of high-frequency terms along with their semantically similar words, each associated with respective frequency and similarity values.

The subtitles corpus underwent comprehensive preprocessing: each text was converted to lowercase, punctuation was removed, English and Italian stopwords were eliminated using the `NLTK` library, and the text was tokenized and lemmatized using `SpaCy`. The preprocessed corpus was then applied to the previously trained Word2Vec model (see Section 2.3) to extract the 10 most semantically similar words to each of the 100 most frequent terms,

[3] *Word Embeddings in NLP: Comparison Between CBOW and Skip-Gram Models* in https://www.geeksforgeeks.org/word-embeddings-in-nlp-comparison-between-cbow-and-skip-gram-models/, last accessed: 2024/07/26.

generating a representative list of words for each text in the corpus.

Table 3.46 provides examples of the extracted words for English, while Table 3.47 does so for Italian. The selection of words was carried out manually, choosing standard terms from different texts to clearly illustrate how the algorithm operates in practice. Standard words were intentionally reported in the table, as there are no strictly representative words for a specific age group. While some words may appear frequently across different texts, their relevance is determined by the words they co-occur with. A term like *palla* (ball), for example, may frequently appear in both a children's story and an adult-themed text, but in the former, it will co-occur with words like *gioca* (play), while in the latter, it may appear alongside offensive or inappropriate terms. This highlights how co-occurrence patterns, rather than individual words, are key to capturing contextual differences in semantic analysis.

Word	**Semantically related words**
winter	summer, spring, autumn, summers, harvest, weather, sunshine, rains, rain, frost
okay	Okay, alright, Right, yeah, yes, Yeah, Yes, Alright, Mhm, OK
day	night, evening, week, afternoon, morning, weekend, half-hour, Friday, month, Saturday
friend	colleague, neighbour, cousin, girlfriend, uncle, brother, companion, boyfriend, lover, son
tea	coffee, sandwiches, breakfast, dinner, cup, supper, lunch, toast, drink, brandy

Table 3.46: Extract of the output showing the 10 semantically related words to each target word in English.

The next step involved creating a vector for each script, indicating the presence (1) or absence (0) of each word from the common vocabulary in the script text. These vectors were stored in a DataFrame to calculate script similarity. The process included:

- Building the common vocabulary: All files in the lists folder were read, and words from each file were converted to lowercase and added to a set of unique words. The vocabulary was then created by sorting this set alphabetically.
- Creating a DataFrame of presence vectors: A DataFrame was initialized with columns representing the vocabulary words. For each file, the script content was loaded, and a presence vector was created to denote each word's presence (1) or absence (0) in the text, then added to the DataFrame.

Words	Semantically related words
primo	terzo, successivo, quinto, secondo, sesto, seguente, precedente, prossimo, quarto, nono
decina	ventina, quindicina, manciata, trentina, cinquantina, quarantina, sessantina, settantina, miriade, dieci
rimanere	restare, trovarsi, esserci, muoversi, stare, venire, esserlo, rientrare, starci, esser
applicare	adottare, eseguire, impostare, attuare, estendersi, emettere, influenzare, indurre, effettuare, operare
superare	misurare, sopportare, sfruttare, raggiungere, recuperare, ridurre, mantenere, soddisfare, fronteggiare, determinare

Table 3.47: Extract of the output showing the 10 semantically related words to the target word in Italian.

- Calculating cosine similarity: The `cosine_similarity` function from scikit-learn was applied to calculate the similarity between all script presence vectors.

The two large similarity matrices obtained, one for English and one for Italian, were used to represent a network of texts in *Gephi* to calculate the classes. In *Gephi*, the *Fruchterman-Reingold* algorithm was chosen for representation. Force-directed graph drawing algorithms, like *Fruchterman-Reingold*, are designed to create graph layouts by positioning nodes in two- or three-dimensional space such that edges are approximately equal in length and the number of crossing edges is minimized. This is achieved by assigning forces to edges and nodes based on their relative positions and then using these forces to simulate the movement of edges and nodes or to minimize the system's energy (Kobourov 2012).

These algorithms balance attractive and repulsive forces: attractive forces pull connected nodes closer together to reflect their linkage, while repulsive forces push nodes apart, reducing overlap. The algorithm iterates until these forces reach equilibrium.

3.5.1 The English case

The document similarity values were imported as an undirected edge graph, with all edges below a weight of 0.76 filtered out to improve readability. The *Modularity Class* algorithm was then applied to calculate sub-communities of nodes, facilitating the identification of specific word classes. A resolution parameter of 1.2 was used for the Modularity Class to reduce the number of communities in the network.

In the case of the English language, the unsupervised classification step produced seven main classes identified in Figure 3.8 with different colors. Table 3.48 shows the distribution

of each class.

Classes	Color	No. Documents	Percentage
13	Purple	171	38,4%
50	Green	66	14,8%
31	Light Blue	57	12,78%
49	Black	33	7,4%
51	Orange	29	6,5%
33	Turquoise	13	2,91%
32	Magenta	13	2,91%
Residual	Grey	64	14,35%

Table 3.48: Classes composition for English.

In the representation, it is evident that the most populous class is the purple one, predominantly composed of TED Talks, suggesting that the model successfully identified this specific type of content as similar. However, as shown in Figure 3.9, where each original label is color-coded (green for the "Kids" category, purple for "Adults", and orange for "Teenagers"), this group also includes nodes from both the "Teenagers" and "Adults" categories. This highlights the model's limited ability to distinguish between texts for adults and teenagers within this family. Nevertheless, the model effectively identifies texts exclusively for adults. For instance, Class 49 consists of subtitles from adult films and YouTube Poop videos, texts that statistical models have historically struggled to differentiate from children's content.

The model demonstrates strong performance in distinguishing texts for children, as evidenced by classes 50, 31, 51, 33, and 32, which are composed exclusively of texts for children. Additionally, the model successfully differentiates between texts intended for children under the age of 6 (referred to as "preschool" in the corpus) and those for children aged 7 to 12 (referred to as "schoolage" in the corpus). A closer examination of the corpus reveals that classes 32 and 51 are distinctly separated because they consist of texts from the same series or cartoon. For example, the texts in class 51 belong to the *Teeny Tiny Stories* series.

The residual class, on the other hand, includes all texts that are not connected to other nodes or have a connection strength below 0.76 and have therefore been filtered out.

The final step of the analysis involved assigning labels to the identified classes using a labeling algorithm. A majority-based matching algorithm was selected for this purpose. Specifically, a mapping was created using the Python library `pandas` by aggregating data for each class identified by the Modularity Class and determining which predicted class was the most frequent. This label was then assigned to the entire Modularity Class. To prevent residual classes from influencing the results (as these would inevitably be assigned the correct

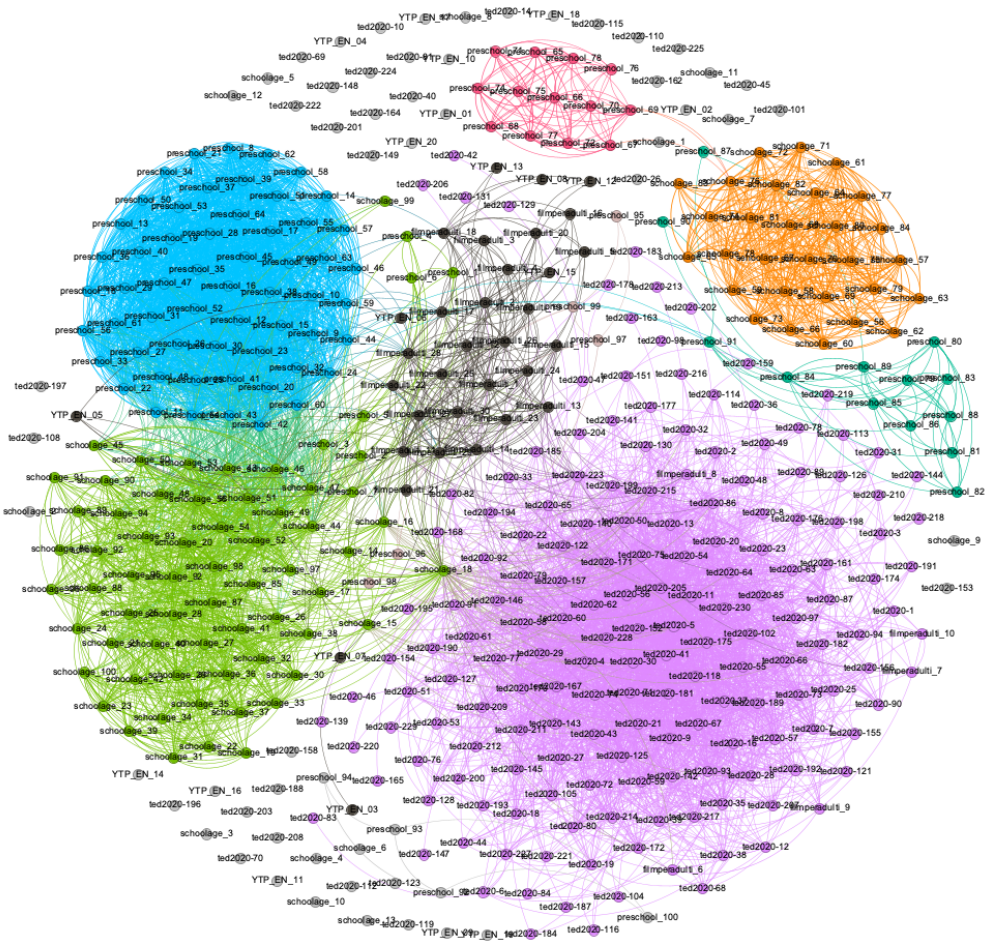


Figure 3.8: Representation with Gephi of the families identified by the Modularity Class for English.

label due to containing only one text), classes with fewer than ten texts were filtered out.

After labeling all modularity classes, a comparison with the original dataset labels was performed, allowing for the calculation of Accuracy, Recall, and F1-scores. The results are shown in Table 3.49

As observed from the graph obtained with Gephi, the model indeed successfully predicted all the texts for children, achieving a precision and recall of 1.0. The model also correctly predicted all suitable texts for adults, with a precision of 1.00, but achieved a low recall because it failed to identify all suitable texts for adults, recognizing only a small portion. As we saw in the representation, many of them were confused with those in the Teenager category. Observing the results for the latter, we note that, on the contrary, the recall is high, meaning that all the examples of teenagers were correctly identified as such, but this label did not always correspond to the true label, confirming that some texts for adults were

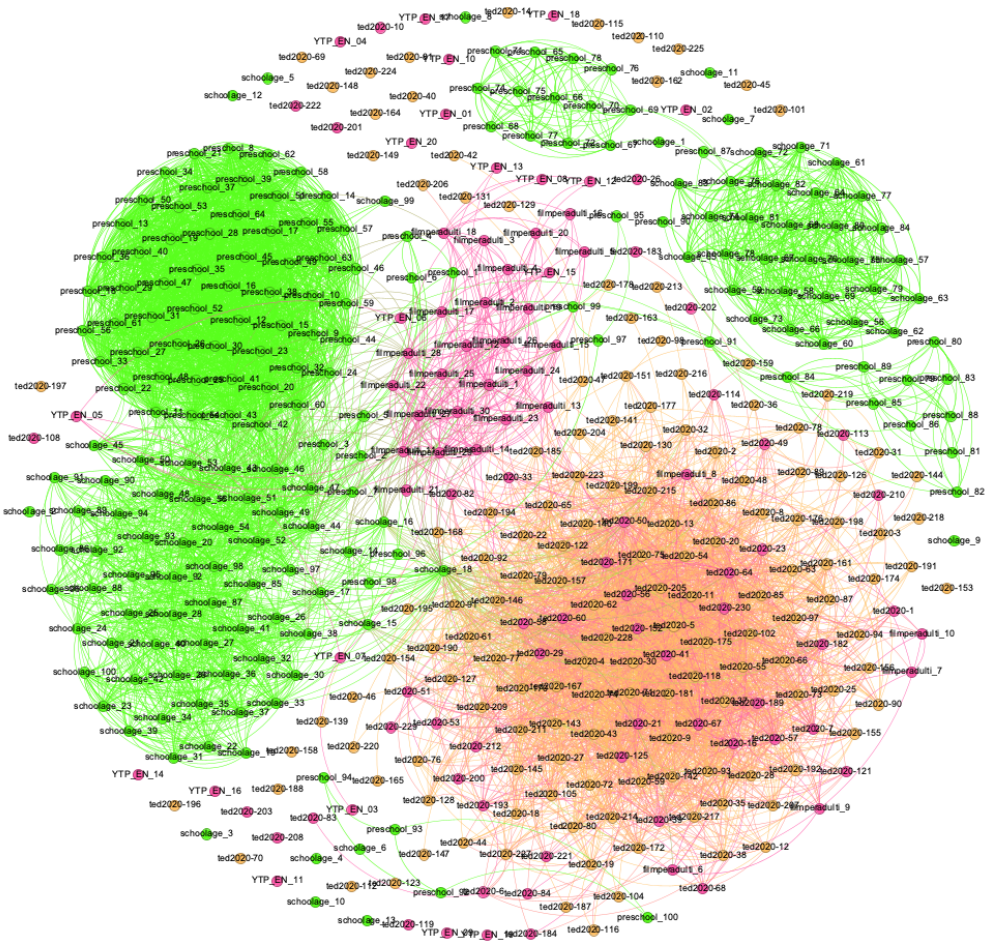


Figure 3.9: Representation with Gephi of the original labels for English.

	Precision	Recall	F1-Score	Support
Kids	1.00	1.00	1.00	178
Teenagers	0.73	1.00	0.84	124
Adults	1.00	0.41	0.58	80
Accuracy			0.88	382
Macro Avg	0.91	0.80	0.81	382
Weighted Avg	0.91	0.88	0.86	382

Table 3.49: Semantic Classification report results for English.

labeled as teenagers. However, there is a good balance between the two metrics for this category, resulting in an F1-score of 0.84. Overall, the model accuracy is 0.88.

3.5.2 The Italian case

The case of Italian differs significantly from that of English. In this instance, the document containing similarity values was also imported as an undirected graph; however, the filter applied to make the representation readable was set to 0.73. To calculate the subcommunities, the Modularity Class was used with a resolution parameter of 1, in order to obtain the appropriate number of communities. For the Italian language, the model generated eight significant communities, which are visualized in Figure 3.10 with different colors. Table 3.50 illustrates the distribution of each class.

Classes	**Color**	**No. Documents**	**Percentage**
14	Pink	67	13,04%
285	Green	48	9,34%
63	Yellow	42	8,17%
33	Light Blue	30	5,84%
82	Purple	8	1,56%
286	Orange	5	0,97%
166	Lime	4	0,78%
266	Turquoise	4	0,78%
Residual	Grey	306	59,53%

Table 3.50: Classes composition for Italian.

In the representation, the group consisting of three classes identified by the Modularity Class (14, 63, and 33) immediately stands out, as it is primarily composed of TED Talks. Another clearly visible class is 285, which mainly consists of films for adults, along with a smaller class, 286, that also contains adult content. Unlike the English case, where texts for children were well-distinguished even by age group, the model for Italian faced greater difficulty in achieving this distinction.

The remaining classes, 82, 166, and 266, are very small and consist of texts from YouTube Kids. An interesting observation from the figure is that the TED Talks are divided into three distinct groups; this suggests that the model may have failed to identify the viewer's age and instead captured the topic of the texts. Future analyses could further investigate this aspect.

Figure 3.11 shows the nodes colored according to the original labels of the texts: green for children's texts, orange for teenagers, and purple for adults. Observing this representation, it is evident that the model struggled to clearly distinguish between TED Talks for adults and those for teenagers, as the nodes are intermixed. However, it effectively identified texts specifically for adults, and, as observed in the classes obtained through the Modularity Class,

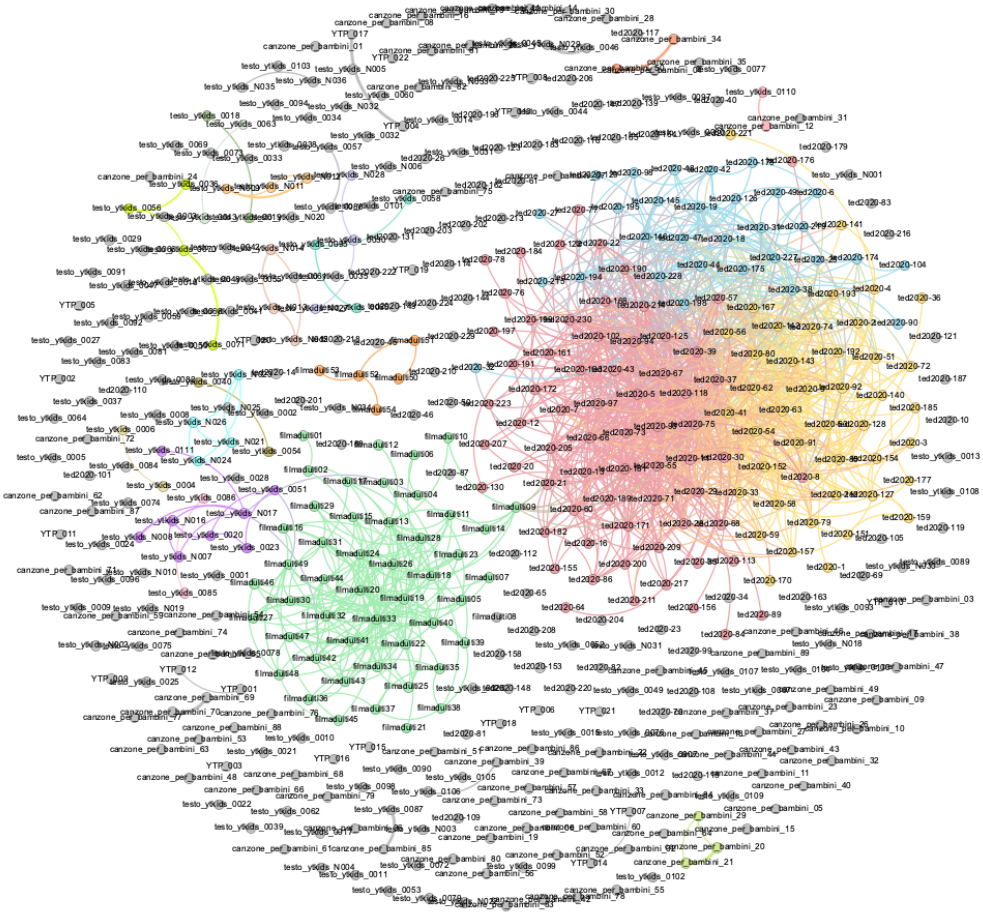


Figure 3.10: Representation with Gephi of the families identified by the Modularity Class for Italian.

texts for children are distinct and positioned on the left side of the representation, though most lack strong semantic connections.

The residual class, which consists of all texts not connected to other nodes or with a connection strength below 0.73, is notably large in this case.

As a final step in the analysis, a majority-based algorithm was once again used to assign labels to the classes by aggregating data for each class identified by the Modularity Class and determining the most frequent predicted class. Similar to the English dataset, smaller classes in the Italian dataset were filtered out to avoid influencing the metrics; specifically, a filter was applied to exclude classes with three or fewer texts. By comparing the assigned labels with the original labels, Accuracy, Recall, and F1-score were calculated, as shown in Table 3.51.

Examining the metrics, it is evident that both precision and recall for the children's

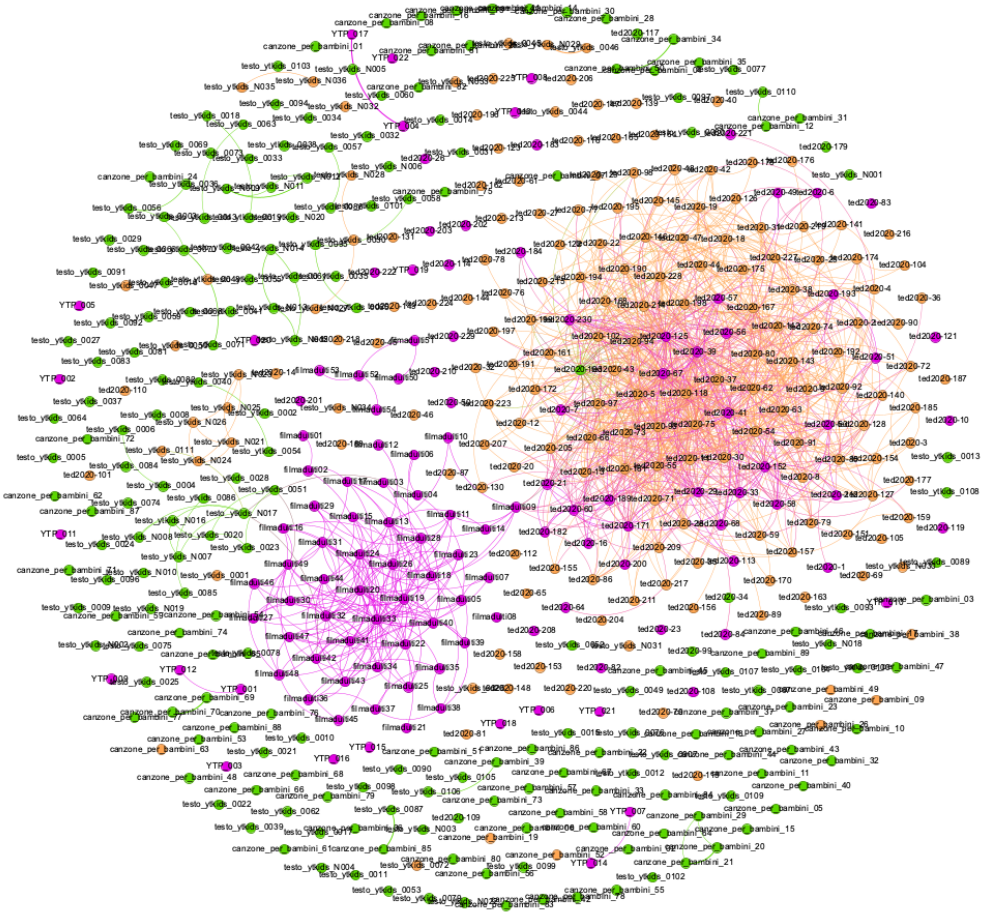


Figure 3.11: Representation with Gephi of the original labels for Italian.

	Precision	Recall	F1-Score	Support
Kids	0.96	0.96	0.96	24
Teenagers	0.77	0.99	0.87	114
Adults	1.00	0.62	0.77	85
Accuracy			0.85	223
Macro Avg	0.91	0.86	0.87	223
Weighted Avg	0.88	0.85	0.84	223

Table 3.51: Semantic Classification report results for Italian.

category are very high, although they are based on only twenty-four texts with strong relationships to one another. For the Teenagers category, however, precision is lower at 0.77, indicating that not all texts classified as teenage content actually belonged to this

category. Recall, on the other hand, is very high, meaning that nearly all examples in the Teenagers category were correctly identified. The F1-score for this category is 0.87.

In the Adults category, precision is high, reaching 1.00, meaning that all adult texts were correctly classified. However, recall is lower at 0.62, indicating that only 62% of the texts in this class were correctly identified. This is likely due to some adult texts being misclassified as part of the Teenagers category, as observed in the English analysis. The F1-score for the Adults category is 0.77. Overall, the model developed for analyzing Italian achieves an accuracy of 0.85.

3.5.3 Content analysis with high-specialized lexicons

The final experiment in this study involved the automatic extraction of specific feature frequencies within each text. Indicators were identified to infer the themes addressed in the texts and to evaluate their suitability for minors based on the presence of certain topics. The focus was specifically on identifying the following indicators: *Bad Language*, *Violence*, and *Drugs*. The experiment was conducted using the dictionaries described in Section 2.4. This approach also proved valuable for testing and refining the resources created, resulting in more accurate research outcomes that could be applied to other fields. The test was performed by comparing the results of the automatic extraction with the annotations made during the corpus annotation phase, as described in Section 2.6.

To extract the indices, the following procedure was adopted: Each text was pre-processed, then lemmatized, and the stop words were removed. The presence of words related to each indicator was then analyzed in terms of frequency and relative frequency within each text. The indicators corresponded to the custom dictionaries created and the NRC Lexicon dictionaries for each negative emotion (namely fear, disgust, anger, and sadness). This process was conducted in parallel on both the Italian and English corpora.

The indices for each feature were calculated as follows:

- Bad Language was calculated as the sum of the total frequency of words from the Bad Language and Discrimination dictionaries.
- Violence included the combined frequency of words from the Violence, Anger, Fear, and Disgust dictionaries.
- The Drugs indicator only included the total frequency of words from the Drugs dictionary.

Table 3.52 presents the minimum, maximum and average values for the three indices for English, while Table 3.53 presents the same values for Italian.

In Tables 3.54 and 3.55, the five titles of multimedia content with the highest values and the ten titles with the lowest values for each indicator, respectively, for English and

Indicator	Lower Value	Higher Value	Average Value
Bad Language	0.0	7.927	1.159
Violence	0.0	9.411	1.300
Drugs	0.0	1.074	0.039

Table 3.52: Lower, higher and average values for the three indicators, for English language.

Indicator	Lower Value	Higher Value	Average Value
Bad Language	0.0	8.0	0.219
Violence	0.0	10.2	2.636
Drugs	0.0	0.847	0.012

Table 3.53: Lower, higher and average values for the three indicators, for Italian language.

Italian, are reported, together with the age rating assigned to that specific content. These elements allow for an initial assessment of the effectiveness of the methodology, and at the same time allow us to conduct an analysis to understand the origin of any issues related to the constructed resources.

By analyzing the results obtained for English, we observe that the five titles with the lowest values across all three indicators are texts suitable for children aged three (with the exception of one, which is suitable for individuals aged twelve). These are predominantly videos categorized by the source website as preschool videos, along with a few TED Talks that were still annotated as appropriate for younger audiences. On the other hand, when we look at the highest values, we notice that almost all texts with a high Bad Language index are suitable for people aged 18 or older. These consist of YouTube Poops and films categorized as adult content by the source website. The only exception is a text classified as school-age content and annotated as suitable for children aged six years, which likely represents an error. This case, as mentioned earlier, allows us to validate the resources built and identify potential issues or challenges to be addressed. Analyzing the specific text (*schoolage_10*), we find that the high score for Bad Language is due to a combination of factors. First, like most children's texts, it is short, so the presence of certain terms, often repeated due to the nature of songs or nursery rhymes, causes the relative frequency of these terms to be high. The topic of the text is lice, and consequently, the term *louse* appears frequently. This term also matches in the profanity dictionary, as it is a common insult. In addition to this, the text contains many words associated with negative emotions (e.g., *resistance*, *choke*, *remove*, *wan*). All these elements combined result in the text having one of the highest scores for Bad Language, even though there are no actual insults.

The issue of term ambiguity is, unfortunately, one of the most common problems in

Indicator	Videos with the lowest score (*Title*, Age, Value)	Videos with the highest score (*Title*, Age, Value)
Bad Language	*preschool_9*, **3**, 0.0 *preschool_50*, **3**, 0.0 *preschool_8*, **3**, 0.0 *preschool_65*, **3**, 0.0 *preschool_66*, **3**, 0.0	*ytp_en_09*, **18**, 7.927 *ytp_en_01*, **18**, 7.609 *schoolage_10*, **6**, 7.5 *ytp_en_03*, **18**, 6.593 *filmadulti12*, **18**, 6.489
Violence	*ted2020-179*, **3**, 0.0 *ted2020-115*, **12**, 0.0 *preschool_99*, **3**, 0.0036 *preschool_44*, **3**, 0.0037 *preschool_74*, **3**, 0.0041	*ted2020-33*, **18**, 9.413 *ted2020-225*, **15**, 8.663 *ted2020-84*, **18**, 7.906 *ted2020-29*, **18**, 7.612 *ted2020-12*, **15**, 7.223
Drugs	*preschool_24*, **3**, 0.0 *preschool_23*, **3**, 0.0 *preschool_22*, **3**, 0.0 *preschool_21*, **3**, 0.0 *preschool_20*, **3**, 0.0	*ted2020-151*, **15**, 1.074 *schoolage_63*, **6**, 0.758 *schoolage_6*, **3**, 0.696 *schoolage_79*, **6**, 0.671 *ted2020-117*, **3**, 0.556

Table 3.54: Videos with the lowest and highest scores for each indicator, for English language.

NLP (Maisto et al. 2021c). Ambiguity involves a single lexical unit representing multiple meanings (polysemy) or the expression of the same meaning through different lexical units (synonymy). In this case, a preliminary evaluation is recommended to determine whether a term with multiple meanings is frequently used as an insult,justifying its inclusion in the dictionary, or if it is rarely used as such in its common form, in which case it could be removed to enhance the system's precision.

The Violence indicator, on the other hand, demonstrates the system's efficiency, as all texts with the lowest values are suitable for children aged three or twelve, while those with the highest values are suitable for individuals aged fifteen or eighteen.

The Drugs indicator highlights another aspect of the system: all texts with the lowest values are indeed suitable for minors, specifically children aged three. However, among the texts with the highest values, which would typically include adult content, there are also children's texts. It is important to note that these values are very low, even below 1, meaning that the repeated occurrence of a single word in shorter texts can result in those texts being classified as having a high value for drugs. Analyzing the children's texts that displayed a high Drugs indicator, I found that the words flagged by the system were again ambiguous but also common in drug-related slang, making it impossible to remove them without risking the omission of important elements in other texts. For example, flagged

Indicator	Videos with the lowest score (*Title*, Age, Value)	Videos with the highest score (*Title*, Age, Value)
Bad Language	*testo_ytkids_0108*, **3**, 0.0	*ytp_004*, **18**, 8.0
	testo_ytkids_0107, **3**, 0.0	*ytp_011*, **18**, 6.25
	testo_ytkids_0106, **3**, 0.0	*ytp_010*, **18**, 5.0
	testo_ytkids_0105, **3**, 0.0	*ytp_017*, **18**, 4.839
	testo_ytkids_0104, **3**, 0.0	*ytp_012*, **18**, 4.237
Violence	*canz._per_bambini_45*, **3**, 0.0	*canz._per_bambini_59*, **6**, 10.216
	testo_ytkids_n001, **3**, 0.0	*ted2020-84*, **18**, 10.070
	canz._per_bambini_25, **3**, 0.0	*testo_ytkids_0050*, **12**, 9.738
	ted2020-179, **3**, 0.0	*ted2020-163*, **15**, 8.705
	canz._per_bambini_11, **3**, 0.0	*ted2020-33*, **18**, 8.547
Drugs	*canz._per_bambini_34*, **6**, 0.0	*ytp_012*, **18**, 0.847
	canz._per_bambini_33, **3**, 0.0	*ytp_017*, **18**, 0.806
	canz._per_bambini_32, **6**, 0.0	*ytp_004*, **18**, 0.800
	canz._per_bambini_31, **3**, 0.0	*ted2020-117*, **3**, 0.568
	canz._per_bambini_30, **6**, 0.0	*filmadulti19*, **18**, 0.476

Table 3.55: Videos with the lowest and highest scores for each indicator, for Italian language.

words included *pot*[4], which can refer to a flower pot or cooking pot in a children's text but is also slang for marijuana; *acid*[5], which is slang for a hallucinogenic drug but refers to the simple chemical acid in children's texts; and *sniff*[6], which is slang for cocaine or inhaling powdered drugs but is often used in children's texts to describe smelling a flower or perfume.

If we look at the Italian language, we notice that among the five texts with the lowest Bad Language indicator values, all are texts for three-year-old children, each with a consistent score of 0.0. In contrast, the five texts with the highest values are all suitable for an adult audience. Similarly, among the five least violent texts, all are appropriate for three-year-old children. However, within the five texts with the highest Bad Language values, two children's texts appear with notably high scores.

Upon closer examination of these specific texts, we find that *Canzone_per_bambini_59* is a song centered on the theme of lice, containing terms such as *disgrazia* (misfortune), *invadere* (invade), *paura* (fear), *dispettoso* (naughty), and *male* (harm), which convey anger. This song is also very short (189 tokens), with these terms repeated multiple times.

Turning to the Drugs indicator, we observe that among the five texts with the lowest

[4]Pot in https://www.urbandictionary.com/define.php?term=pot, last accessed: 2024/09/16.
[5]Acid in https://www.urbandictionary.com/define.php?term=acid, last accessed: 2024/09/16.
[6]Sniff in https://www.urbandictionary.com/define.php?term=sniff, last accessed: 2024/09/16.

values, all are suitable for children aged three or six, with values consistently at 0. However, among the five texts with the highest Drug indicator values, there is one text for children aged three. As described in Chapter 2, the TED Talks portion of the English corpus corresponds exactly to that of the Italian corpus, and we find *ted2020-117* here as well, which, for similar reasons, scores higher than other texts. It is important to note that once again, all scores are very low, below one, as was the case for English texts.

In summary, the analysis of both the English and Italian corpora demonstrates the system's effectiveness in determining the suitability of texts for minors. However, the presence of ambiguous terms, especially in shorter texts, can lead to inflated scores for certain indicators, such as Bad Language or Drugs, even in the absence of actual insults or references to drugs. The results of this experiment provide a basis for future improvements to the system by refining lexicons and implementing context-aware mechanisms to minimize false positives.

An additional experiment was conducted to compare the results automatically detected by the system, using specialized lexical dictionaries, with those annotated by human annotators. The aim was to assess whether the specific themes identified by the system were significant enough to also be recognized by a human. The annotation phase is described in Section 2.6.

The first step of the experiment was to define the transformation logic:

- If the annotation indicates the presence of drugs, violence, or discrimination, the value is set to 1;

- If these terms are not annotated, the value is set to 0.

The process of comparing the results obtained from automated analysis with manual annotations involved an initial comparison phase for each individual indicator (bad language, violence, drugs).

The metrics calculation was performed by simply checking whether, for each text with an automatically assigned score greater than 0, there was a corresponding manual annotation (where 1 indicates presence, following the logic described above). As shown in Table 3.56 for English and Table 3.57 for Italian, this system proved to be imprecise due to the presence of many noisy elements.

Comparison with the annotations highlights the high sensitivity of this approach, which relies on highly specialized dictionaries. Although the system exhibits low precision due to numerous false positives, it effectively captures all instances where a specific indicator is indeed present, resulting in high recall. Thus, the dictionaries demonstrate their comprehensiveness and usefulness in detecting specific topics within texts. However, the results may require calibration based on the specific goals of the analysis. For example, setting a

Indicator	Precision	Recall	F1-score	Accuracy
Bad Language	0.14	0.98	0.24	0.25
Violenza	0.10	1.00	0.19	0.11
Droga	0.06	0.50	0.11	0.79
Overall	0.10	0.83	0.18	0.38

Table 3.56: Metrics for Presence and Absence for the English Language.

Indicator	Precision	Recall	F1-score	Accuracy
Bad Language	0.65	0.82	0.73	0.90
Violenza	0.11	1.00	0.19	0.13
Droga	0.21	0.32	0.26	0.93
Overall	0.32	0.71	0.39	0.65

Table 3.57: Metrics for Presence and Absence for the Italian Language.

threshold above which a text is deemed "violent" could refine the results based on specific objectives.

Reducing noise involves filtering out cases detected by the system despite having very low scores. By applying a threshold of 3% and comparing the results with the manual annotations, we observe improved overall metrics, which better align with the lower sensitivity typical of human annotators.

Most cases with scores below 3% were not annotated by human annotators, suggesting that they were not considered significant. However, the system also allows for cases where human annotators may flag a case as relevant even with a score below 3%, capturing subtleties that the model may not fully recognize. This threshold helped reduce noise and false positives from low-frequency words that do not reflect meaningful indicators, while still respecting the judgment of human annotators in borderline cases.

The results of this second calculation are presented in Table 3.58 for English and Table 3.59 for Italian.

Indicator	Precision	Recall	F1-score	Accuracy
Bad Language	0.86	0.98	0.92	0.98
Violenza	0.63	1.00	0.77	0.94
Droga	0.12	0.50	0.19	0.89
Overall	0.54	0.83	0.63	0.94

Table 3.58: Evaluation metrics after applying a threshold for English.

Indicator	**Precision**	**Recall**	**F1-score**	**Accuracy**
Bad Language	1.00	0.82	0.90	0.97
Violenza	0.27	1.00	0.42	0.72
Droga	1.00	0.32	0.48	0.97
Overall	0.76	0.71	0.60	0.89

Table 3.59: Evaluation metrics after applying the threshold for Italian.

We can observe that the precision has improved significantly.

Not all metrics improved equally, as the values assigned to each indicator are distributed on different scales (as shown in Tables 3.52 and 3.53). However, adjusting these values according to specific objectives and calibrating them as independent variables can yield good results.

These findings suggest that adjusting thresholds based on individual values, rather than applying a uniform empirical threshold across all indicators, may further enhance the results.

Therefore, further analysis and potential refinement of the specialized dictionaries, which currently include ambiguous terms, may prove useful.

Chapter 4

Conclusions and Future Works

In this thesis, various methods for classifying online multimedia content have been explored by leveraging the potential of textual analysis. Resources such as highly specialized lexical dictionaries and annotated corpora were created, which could be valuable for future NLP projects. A range of algorithms, from SVM and NB to NNs, transformers, and LLMs, were tested on this specific task, evaluating methods such as rule augmentation to enhance performance. Additionally, a semantic and unsupervised approach was investigated, and a lexical analysis was conducted to assess the effectiveness of the resources created.

In order to reach this goal, the research questions investigated aimed at:

1. Understand what has already been done in the field of computational linguistics in order to solve this problem.

2. Outline a methodology for testing supervised classification and semantic analysis.

3. Create specific resources for the task of classifying multimedia content through transcriptions, such as domain-specific dictionaries and annotated corpora, which can also be used for additional tasks in the future.

4. Implement an experimental phase to test various algorithms on the specific task, conduct a semantic analysis and a lexical analysis to evaluate the best approach, and test the lexical resources created.

Chapter 1 tried to answer the research question: "What has already been done in the field of computational linguistics on this problem and what elements need to be developed for the realization of such a system?" by framing the state-of-the-art. The studies conducted on this topic, which form the basis for the motivation behind this research, were thoroughly investigated, specifically, the various efforts aimed at addressing the issue of exposure of minors to content inappropriate for their age. Legislative proposals were examined (with a

particular focus on the AGCOM guidelines, which also underpin the corpus annotation process developed in this study), as well as partial solutions offered by video-sharing platforms. Special attention was given to the Elsagate phenomenon, which led to the development of a line of research well summarized in Alqahtani et al. (2023)'s work. This work conducted a systematic review of ML, DL, and NLP techniques to provide an overview of existing methods, tools, and approaches proposed or developed by scholars to protect, prevent, and safeguard children from inappropriate or illegal video content.

Despite these advances, there remains a need for further research, particularly in approaches that focus on analyzing the content itself rather than user comments or metadata. Based on the literature reviewed in the first chapter of this thesis, it emerges that most of the works employed text as a key element for identifying inappropriate content, primarily focusing on specific text sources, such as user comments or metadata-based filtering on web pages.

From the literature examined, it emerged that even when transcriptions were analyzed, the focus was often on general appropriateness, typically involving binary classifications (appropriate vs. inappropriate), rather than a detailed age-based categorization of texts related to the content. Many studies focused on related issues such as discrimination, cyberbullying, or hate speech, but lacked a specific emphasis on precise age-based distinctions. This work provided a unique focus on age categorization through a comprehensive analysis of subtitles and text, which directly reflect the language used within the media.

Additionally, while the majority of studies focused on other aspects for classifying content, such as images, video frames, or audio, some researchers attempted a multimodal analysis. As highlighted in Balat et al. (2024), their future work emphasizes the need for more language-based research to achieve a comprehensive understanding of video content. This thesis provided a detailed analysis of the performance of different ML models and explored whether a semantic approach could yield better results for this specific task.

LLMs, a relatively new area compared to the project's beginning, have also been explored for their potential. Although they did not achieve optimal results in this specific research, they remain a promising field that warrants further investigation. As discussed in the theoretical background, the fine-tuning conducted in this thesis provides a solid foundation for future work that might incorporate prompting strategies, thereby more fully leveraging the capabilities of generative models. This combined approach could help achieve better performance and broaden the effectiveness of LLMs in increasingly complex application contexts.

Furthermore, unlike many of the studies analyzed that predominantly addressed English-language content, this work also focused on creating resources for Italian. This is particularly important because it provides resources, such as annotated corpora and specialized lexicons for Italian, which can be used in future research. Additionally, it offered a comparative

study, alongside English, of various ML models for age-appropriate content classification, highlighting their strengths and suitability for this specific task.

The comparative analysis carried out in the thesis also simplifies future research efforts by providing a study that can be readily leveraged by other scholars. This work can serve as a useful resource for understanding the effectiveness of different approaches, helping streamline future efforts in developing content moderation systems. Further studies could focus on different methods of text representation in vectors, potentially combining these methods to expand the comparative aspects of this research. Such work could integrate both semantic approaches and more traditional models. Additionally, testing a combination of models on this specific task could provide valuable insights, as has been done in other works explored in the theoretical background for different types of classification.

Chapter 2, and in particular, sections 2.4, 2.5 and 2.6 aimed at describing the methodological approach employed to answer the research question: "What resources can be exploited and how existing frameworks can be improved?".

Specifically, in Section 2.4 the construction methods for the domain-specific electronic dictionaries are described, which were later tested in the experimental part of the thesis (see Section 3.5.3). The dictionaries were created manually by extracting resources from various websites and thesauri. For this particular thesis project, I refined the English dictionaries that I had previously developed in earlier work, while the Italian dictionaries were created from scratch. The English dictionaries had been automatically expanded, necessitating a refinement process to exclude generic terms. In contrast, for the Italian dictionaries, automatic expansion was intentionally avoided. The tests presented in Section 3.5.3 demonstrate that, although the English dictionaries show broader term coverage, they suffer from lower precision due to false positives. In contrast, the Italian dictionaries, being more streamlined, demonstrate higher precision but somewhat lower recall. This experimental result underscores the importance of balancing coverage and accuracy when creating specialized lexicons, suggesting that future work could focus on refining both sets of resources to better manage ambiguity and improve overall reliability.

This research also focused on adapting emotion lexicons to identify whether specific negative emotions, such as anger, sadness, disgust, and fear, are expressed within the language of the analyzed content. To analyze these emotions, the NRC Emotion Intensity Lexicon (NRC-EIL) was used. This lexicon assigns real-valued intensity scores for eight primary emotions in English. A specific focus was placed on the Italian dictionaries, which were expanded through MultiWordNet and manually revised in order to avoid that terms that, having originally been translated via automatic translation, did not accurately convey that specific emotion in Italian were reviewed and adjusted based on my expertise as a native Italian speaker. Addictionally, the inflected forms of each lemma were added to the dic-

tionary, capturing different verb tenses, moods, and noun forms. The updated dictionary can be useful for projects specifically focused on Italian, with an emphasis on identifying emotions within a text. However, in the lexical analysis task, described in Section 3.5.3, another resource was chosen from the NRC-EIL: the individual emotion files. The objective was not to understand which emotion was generally present in the text, but rather to determine the presence and intensity of specific negative emotions, anger, fear, and disgust, which, when combined with other dictionaries, can generate values indexing the presence or absence of particular features within texts. Moreover the texts used for the experiment were lemmatized. However, the emotion dictionary only exists in English, so for Italian, an automatic translation using DeepL was needed. Nevertheless, as outlined above, it would be beneficial in the future to refine this resource by incorporating native Italian speaker expertise and inflecting the lemmas.

In Section 2.5 the building process of the corpora used for the experiment has been described. Since the study focuses on both English and Italian, two separate corpora were built. One part consists of TED Talks, which are the same for both Italian and English as they were taken from an existing corpus. Other texts were taken from specific websites, while the remaining content was gathered through a semi-automatic procedure: an initial automated extraction of subtitles from YouTube was performed using a scraping code, followed by a careful manual review to correct any errors in automatically generated subtitles and to ensure that essential "bad words" were not censored, which is crucial in a study of this nature. Although the corpora are not parallel, efforts were made to balance them. They consist of TED Talks, content specifically aimed at children, and content specifically aimed at adults.

In Section 2.6 the annotation procedure of the corpora was described. In this study, annotation guidelines were developed based on AGCOM's criteria for web-based content, tested with three annotators on a sample of 30 Italian texts to calculate the IAA. Although the IAA score was relatively low (0.25 on Fleiss' Kappa), largely due to subjective interpretations in age-appropriateness ratings, the annotations on specific thematic elements like discrimination, nudity, and violence achieved high consistency. These results highlighted challenges in subjective age classification, as annotators sometimes relied on personal judgment beyond the objective AGCOM guidelines.

Given resource constraints, I proceeded to annotate the remainder of the corpus independently, ensuring a consistent application of these guidelines across the dataset. This approach balanced the initial limitations by providing uniformity in annotation while also setting a foundation for potential future refinements. Future developments could include integrating a perspectivist approach to account for annotator disagreements and involving a greater number of annotators to further enhance reliability and reduce individual biases

in classification.

Following preliminary tests and discussions during the annotation phase, it was decided to reduce the labels from five to three categories: Kids (ages 0 to 11), Teenagers (ages 12 to 17), and Adults (ages 18 and up).

The sections 2.2 and 2.3 outlined the methodology underlying the experiments, which are subsequently conducted in Chapter 3, that directly address the research questions "Which of the currently existing algorithms commonly used for classification tasks works best for this specific task?" and "How could a classification approach based on semantics work instead?".

Specifically, to address the first research question, various algorithms were tested on the built corpora with the aim of identifying the best method, for classifying content according to audience age. To compare results consistently, a common set of metrics was used for all models tested. For model evaluation, the k-fold cross-validation was chosen with k set to 5. The models selected for testing included SVM, NB, DT, RF, and k-NN, as well as NNs (CNN, LSTM, and BiLSTM), various Transformers, and some LLMs.

To improve classification accuracy and adhere to AGCOM guidelines, a lexicon-based rule was implemented. Specifically, if a Bad Word was detected in a text labeled as "for kids," the label was revised to "for adults". The Bad Words lists for both English and Italian were derived from Discriminative-terms dictionaries, applying a selection methodology that excluded low-impact expressions as well as more generic insults. This rule was consistently applied across all tested models to ensure accurate categorization. As shown in the experimental results in Chapter 3, Table 3.1, this approach yielded improved performance across all classifiers.

To address the second research question, a semantic classification approach was implemented. Specifically, the Word2Vec word embedding model, trained on the PAISÀ corpus for Italian and on BNC for English, was employed to identify the words most semantically similar. A distributional semantic matrix was used for extracting similarity values between words and locating the nearest neighbors of the most frequent terms in each transcript was made to facilitate an analysis free from any predefined context.

The final research question, "What are the results, and what could be the future developments of the project?" is addressed in Chapter 3.

The results show that among the models tested, the best performers were roBERTa-based architectures, with roBERTa leading for English, and UmBERTo, a roBERTa-based model trained on Italian corpora, performing best for Italian. roBERTa achieved an accuracy of 0.89, while UmBERTo reached 0.90. All models achieved satisfactory accuracy for English, surpassing 0.80, while the Italian models generally achieved slightly lower results, though all exceeded 0.70. Several factors may explain these results, including the increased morphological complexity of Italian. Many of the models, only fine-tuned for this task, were

initially trained on English texts, which likely contributed to their reduced performance on Italian. This is evidenced by the fact that, when using a classifier like UmBERTo, specifically trained in Italian, the results not only improve but even slightly surpass those obtained for English.

Simpler models such as Random Forest for Italian and Naive Bayes for English achieved excellent results, outperforming more complex models like LLMs. This may be due to the risk of overfitting when using a small corpus for both fine-tuning and testing, which tends to favor simpler models. This is especially true for a classification task, as LLMs are primarily designed for generation rather than classification.

The experiment confirmed that the introduction of a rule-based filter for "bad words" significantly improved the model's performance, particularly in the "adults" category. In general, the models performed very well in identifying content for children but encountered greater difficulty distinguishing between content for adults and teenagers. This challenge could be related to subtle differences in meaning, especially in TED Talks, which have a tone, vocabulary, and structure that are less intense, as noted during the annotation phase, making it harder for even humans to distinguish between these categories. However, this could be beneficial when the goal of classification is to ensure that only appropriate content is shown to children.

The limitations highlighted by these results allow for some reflections on potential future developments in this area. While the bad words rule improved the results, it could be refined, as not all "bad words" are necessarily inappropriate for teenagers. The rule could be enhanced by considering factors such as the proportion of bad words relative to the total text or the specific types of bad words. Further annotation could also improve this filter. Although the results are satisfactory even with a corpus of this size, expanding the corpus could allow for further assessment of whether performance improvements are achievable with LLMs.

The aspect of YTP should also be further explored, as these videos are the ones most likely to blend in with content intended for children due to their similar structure, use of familiar names, and the potential to attract children while misleading parents with child-friendly thumbnails. This research revealed that, without the bad words rule, YTPs are particularly challenging for classifiers, as they often struggle to accurately distinguish these videos. While the bad words rule partially addresses this issue, further assessments would be beneficial to enhance classification accuracy for this type of content.

In Section 3.5, the experiment on semantic analysis was conducted, yielding satisfactory results. For English, the classification achieved an accuracy of 0.88, very close to the performance of the best classifier. For Italian, the achieved accuracy was 0.85, higher than the average of the tested classifiers but slightly lower than the results obtained when testing

UmBERTo.

Conducting a semantic-based analysis allowed us to fully exploit the potential of the text. The graphical representation enabled us to make several observations about classification; we observed how this type of analysis, focusing solely on the meaning of words without considering text structure, was able to classify texts belonging to the same content category into the same network group. In particular, this approach was successful in correctly classifying, especially in English, texts that were challenging for statistical models, such as YTPs. The same difficulties in classifying the age category of TED Talks were also encountered with this approach.

The results are certainly encouraging, and the semantic approach warrants further exploration in the context of content classification systems. Future studies could investigate its application as a preliminary filter, given its strong ability to recognize content belonging to certain classes. Additionally, this is an unsupervised approach, which is beneficial in contexts where annotated texts are not available.

In Section 3.5.3, an experiment was conducted using a lexical analysis to assess the presence of specific indicators unsuitable for minors within the text. As we have seen, this type of experiment was also useful for evaluating the resources created and understanding their limitations, as well as providing insights for future improvements. Specifically, the presence of bad language, violence, and drugs was calculated, including negative emotions within the calculation for violence. By examining the top 5 results for each indicator, the system's effectiveness in determining the suitability of texts for minors was demonstrated for both Italian and English. However, the presence of ambiguous terms, especially in shorter texts, can lead to inflated scores for certain indicators, such as Bad Language or Drugs, even in the absence of actual insults or references to drugs. The results of this experiment provide a basis for future improvements to the system by refining lexicons and implementing context-aware mechanisms to minimize false positives. In particular, this methodology could be advantageous for automatically extracting keywords that can serve as reliable indicators of specific features.

An additional experiment was conducted to compare the results automatically detected by the system, using specialized lexical dictionaries, with those annotated by human annotators. It was observed that the unexpanded Italian dictionaries, particularly those not derived from translations of English resources, achieved higher and more satisfactory results compared to those for English, which demonstrated very high recall but notably low precision. Thus, dictionaries demonstrate their comprehensiveness and usefulness in detecting specific topics within texts. However, the results may require calibration based on the specific goals of the analysis. For instance, setting a threshold above which a text is deemed "violent" could refine the results according to specific objectives. Noise reduction involves

filtering out cases detected by the system with very low scores. By applying a 3% threshold and comparing the results with manual annotations, we observed an improvement in overall metrics, which better align with the lower sensitivity typical of human annotators. Therefore, further analysis and potential refinement of the specialized dictionaries, which currently include ambiguous terms, may prove beneficial.

In the future, it may be beneficial to explore research that incorporates user comments or integrates metadata to enhance the classification process with additional contextual elements. Developing a multimodal classification system that leverages images or video frames, based on this research as a foundation for text analysis, would be particularly valuable. The linguistic features related to violence, for example, are often inadequate to describe the presence of this feature; on the contrary, they are highly useful for detecting bad language. For this reason, exploring multimodal analyses could be particularly useful.

In conclusion, this dissertation presents a comprehensive study, supported by various experiments, on the automatic classification of text associated with online content. Both supervised and unsupervised approaches were explored, along with a lexical-based analysis. The primary aim was to provide findings that could be valuable for future applications in multimedia content supervision, specifically to protect minors from inappropriate material by fully leveraging the potential of textual content.

The main outcomes of this thesis include specialized high-lexicon dictionaries and age-annotated corpora of multimedia content transcriptions in both English and Italian, as well as a comparative study of various approaches for the specific task of age-based classification. These resources and insights lay a foundation for future research and applications in safeguarding younger audiences from unsuitable online content.

Appendix

Age	Discrimination	Drugs	Dangerous behaviour	Bad language
3	Absent (unless in the context of social initiatives)	Absent	Absent or some mentions	No offensive language
6	Absent (unless for educational purposes or clearly disapproved)	Admissible only for educational purposes	Occasional, with disapproval	Allowed only if sporadic
12	Present but justified by the narrative context and condemned, attributable to negative characters	Admissible only for educational purposes	Allowed if not encouraged	Allowed only in comedic contexts
15	Present but not encouraged	Allowed if not glorified	Frequent if discouraged	Allowed only if not portrayed as positive
18	Free presence	Free presence	Free presence	Allowed, blasphemy excluded

Table 1: Summary table of evaluation criteria based on AGCOM guidelines. Part 1.

Age	Nudity	Sex	Threats	Violence
3	Absent, unless related to scientific contexts	No representation	Absent	Absent or related to unrealistic contexts
6	Partial and only if not related to a sexual context	No representation but allowed displays of affection	Occasional	Allowed short duration scenes related to fiction
12	Complete, only if not in a sexual context, partial also in sexual contexts	Infrequent, devoid of explicit language	Allowed if of short duration and with justifiable outcome	Scenes of mild violence allowed only if contextualized, morally sanctioned, and devoid of insistence on graphic or brutal details
15	Allowed, including in sexual contexts	Infrequent, allowed even with explicit language	Allowed if of short duration and devoid of graphic violence	Allowed, even if portrayed insistently, as long as it is not presented in a positive light and devoid of graphic or brutal details
18	Allowed, including in sexual contexts	Allowed even with explicit language	Free presence	Allowed if not glorified, excluding sadistic representation

Table 2: Summary table of evaluation criteria based on AGCOM guidelines. Part 2.

Bibliography

Abbasi, Mohsin Manshad and AP Beltiukov (2019), "Summarizing emotions from text using Plutchik's wheel of emotions", in *Scientific Conference on Information Technologies for Intelligent Decision Making Support*.

Adewumi, Tosin, Foteini Liwicki, and Marcus Liwicki (2022), "Word2Vec: Optimal hyper-parameters and their impact on natural language processing downstream tasks", *Open Computer Science*, 12, 1, pp. 134-141.

AGCOM, Delibera (2017), "Regolamento sulla classificazione delle opere audiovisive destinate al Web e dei videogiochi di cui all'Art. 10, commi 1 e 2, del Decreto Legislativo 7 Dicembre 2017, N. 203 recante "Riforma delle disposizioni legislative in materia di tutela dei minori nel settore cinematografico e audiovisivo, a norma dell' Art. 33 della Legge 14 Novembre 2016, N. 220"".

Aggarwal, Anubhav, Jasmeet Singh, and Dr Kapil Gupta (2018), "A review of different text categorization techniques", *International Journal of Engineering & Technology*, 7, 3.8, pp. 11-15.

Ahmed, Md Tofael, Maqsudur Rahman, Shafayet Nur, Abu Zafor Muhammad Touhidul Islam, and Dipankar Das (2021), "Natural language processing and machine learning based cyberbullying detection for Bangla and Romanized Bangla texts", *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 20, 1, pp. 89-97.

Ahmed, Syed Hammad, Shengnan Hu, and Gita Sukthankar (2023), "The Potential of Vision-Language Models for Content Moderation of Children's Videos", in *2023 International Conference on Machine Learning and Applications (ICMLA)*, IEEE, pp. 1237-1241.

Alakrot, Azalden, Liam Murray, and Nikola S Nikolov (2018), "Dataset construction for the detection of anti-social behaviour in online communication in Arabic", *Procedia Computer Science*, 142, pp. 174-181.

Alba, Davey (2015), "Google Launches 'YouTube Kids,' a New Family-Friendly App", *Wired*, https://www.wired.com/2015/02/youtube-kids/.

Alcântara, Cleber, Viviane Moreira, and Diego Feijo (2020), "Offensive video detection: dataset and baseline results", in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4309-4319.

Alghowinem, Sharifa (2019), "A safer youtube kids: An extra layer of content filtering using automated multimodal analysis", in *Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 1*, Springer, pp. 294-308.

Alqahtani, Saeed Ibrahim, Wael MS Yafooz, Abdullah Alsaeedi, Liyakathunisa Syed, and Reyadh Alluhaibi (2023), "Children's Safety on YouTube: A Systematic Review", *Applied Sciences*, 13, 6, p. 4044.

Alshamrani, Sultan (2020), "Detecting and measuring the exposure of children and adolescents to inappropriate comments in YouTube", in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 3213-3216.

Alshamrani, Sultan, Mohammed Abuhamad, Ahmed Abusnaina, and David Mohaisen (2020), "Investigating Online Toxicity in Users Interactions with the Mainstream Media Channels on YouTube.", in *CIKM (Workshops)*.

Alshamrani, Sultan, Ahmed Abusnaina, Mohammed Abuhamad, Daehun Nyang, and David Mohaisen (2021), "Hate, obscenity, and insults: Measuring the exposure of children to inappropriate comments in youtube", in *Companion Proceedings of the Web Conference 2021*, pp. 508-515.

Alshamrani, Sultan, Ahmed Abusnaina, and David Mohaisen (2020), "Hiding in plain sight: A measurement and analysis of kids' exposure to malicious urls on youtube", in *2020 IEEE/ACM Symposium on Edge Computing (SEC)*, IEEE, pp. 321-326.

Alsubait, Tahani and Danyah Alfageh (2021), "Comparison of machine learning techniques for cyberbullying detection on youtube arabic comments", *International Journal of Computer Science & Network Security*, 21, 1, pp. 1-5.

Altınel, Berna and Murat Can Ganiz (2016), "A new hybrid semi-supervised algorithm for text classification with class-based semantics", *Knowledge-Based Systems*, 108, pp. 50-64.

— (2018), "Semantic text classification: A survey of past and recent advances", *Information Processing & Management*, 54, 6, pp. 1129-1153.

Anand, Vishal, Ravi Shukla, Ashwani Gupta, and Abhishek Kumar (2019), "Customized video filtering on YouTube", *arXiv preprint arXiv:1911.04013*.

Arslan, Yusuf, Kevin Allix, Lisa Veiber, Cedric Lothritz, Tegawendé F Bissyandé, Jacques Klein, and Anne Goujon (2021), "A comparison of pre-trained language models for multi-class text classification in the financial domain", in *Companion Proceedings of the Web Conference 2021*, pp. 260-268.

Ashraf, Noman, Arkaitz Zubiaga, and Alexander Gelbukh (2021), "Abusive language detection in youtube comments leveraging replies as conversational context", *PeerJ Computer Science*, 7, e742.

Awal, Md Abdul, Md Shamimur Rahman, and Jakaria Rabbi (2018), "Detecting abusive comments in discussion threads using Naıve Bayes", in *2018 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, IEEE, pp. 163-167.

Baccianella, Stefano, Andrea Esuli, Fabrizio Sebastiani, et al. (2010), "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.", in *Lrec*, 2010, vol. 10, pp. 2200-2204.

Bafna, Prafulla, Dhanya Pramod, and Anagha Vaidya (2016), "Document clustering: TF-IDF approach", in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, IEEE, pp. 61-66.

Balat, Mazen, Mahmoud Essam Gabr, Hend Bakr, and Ahmed B Zaky (2024), "TikGuard: A Deep Learning Transformer-Based Solution for Detecting Unsuitable TikTok Content for Kids", *arXiv preprint arXiv:2410.00403*.

Barberá, Pablo, Amber E Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler (2021), "Automated text classification of news articles: A practical guide", *Political Analysis*, 29, 1, pp. 19-42.

Barrett, Lisa Feldman (2017), "The theory of constructed emotion: an active inference account of interoception and categorization", *Social cognitive and affective neuroscience*, 12, 1, pp. 1-23.

Basile, Valerio, Federico Cabitza, Andrea Campagner, and Michael Fell (2021), "Toward a perspectivist turn in ground truthing for predictive computing", *arXiv preprint arXiv:2109.04270*.

Bassignana, Elisa, Valerio Basile, Viviana Patti, et al. (2018), "Hurtlex: A multilingual lexicon of words to hurt", in *CEUR Workshop proceedings*, CEUR-WS, vol. 2253, pp. 1-6.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy (2009), "Gephi: an open source software for exploring and manipulating networks", in *Proceedings of the international AAAI conference on web and social media*, 1, vol. 3, pp. 361-362.

Belainine, Billal, Fatiha Sadat, Mounir Boukadoum, and Hakim Lounis (2020), "Towards a multi-dataset for complex emotions learning based on deep neural networks", in *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pp. 50-58.

Bender, Emily M and Alexander Koller (2020), "Climbing towards NLU: On meaning, form, and understanding in the age of data", in *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5185-5198.

Bhuiyan, Hanif, Jinat Ara, Rajon Bardhan, and Md Rashedul Islam (2017), "Retrieving YouTube video by sentiment analysis on user comment", in *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, IEEE, pp. 474-478.

Binh, Le, Rajat Tandon, Chingis Oinar, Jeffrey Liu, Uma Durairaj, Jiani Guo, Spencer Zahabizadeh, Sanjana Ilango, Jeremy Tang, Fred Morstatter, et al. (2022), "Samba: Identifying inappropriate videos for young children on YouTube", in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 88-97.

Bisk, Yonatan, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. (2020), "Experience grounds language", *arXiv preprint arXiv:2004.10151*.

Blei, David M (2012), "Probabilistic topic models", *Communications of the ACM*, 55, 4, pp. 77-84.

Blei, Andrew Y. Ng, and Michael I. Jordan (2002), "Latent dirichlet allocation", in *Advances in neural information processing systems*, pp. 601-608.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre (2008), "Fast unfolding of communities in large networks", *Journal of statistical mechanics: theory and experiment*, 2008, 10, P10008.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017), "Enriching word vectors with subword information", *Transactions of the association for computational linguistics*, 5, pp. 135-146.

Bolasco, Sergio and Tullio De Mauro (2021), *"L'" analisi automatica dei testi: fare ricerca con il text mining*, Carocci.

Bonifazi, Gianluca, Silvia Cecchini, Enrico Corradini, Lorenzo Giuliani, Domenico Ursino, and Luca Virgili (2022), "Investigating community evolutions in TikTok dangerous and non-dangerous challenges", *Journal of Information Science*, p. 01655515221116519.

Boucouvalas, Anthony C (2002), "Real time text-to-emotion engine for expressive internet communications", in *Proceedings of International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP-2002)*.

Bowman, Samuel R (2023), "Eight things to know about large language models", *arXiv preprint arXiv:2304.00612*.

Brandom, Russell (2017), "Inside Elsagate, the conspiracy-fueled war on creepy YouTube kids videos", *The Verge*.

Breiman, Leo (2017), *Classification and regression trees*, Routledge.

Brezeale, Darin and Diane J Cook (2008), "Automatic video classification: A survey of the literature", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38, 3, pp. 416-430.

Bridle, James (2017), "Something is wrong on the internet", *Medium*, https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al.

(2020), "Language models are few-shot learners", *Advances in neural information processing systems*, 33, pp. 1877-1901.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. (2023), "A survey on evaluation of large language models", *arXiv preprint arXiv:2307.03109*.

Chen, Jingxiang, Kai Wei, and Xiang Hao (2021), "Detect profane language in streaming services to protect young audiences", in *Proceedings of the 4th Workshop on e-Commerce and NLP*, pp. 123-131.

Chomsky, Noam (2014), *Aspects of the Theory of Syntax*, 11, MIT press.

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. (2024), "Scaling instruction-finetuned language models", *Journal of Machine Learning Research*, 25, 70, pp. 1-53.

Chuttur, MY and A Nazurally (2022), "A multi-modal approach to detect inappropriate cartoon video contents using deep learning networks", *Multimedia Tools and Applications*, 81, 12, pp. 16881-16900.

Creators, YouTube (2017), "L'algoritmo" - Come funziona il sistema di ricerca e scoperta di YouTube", https://www.youtube.com/watch?v=hPxnIix5ExI#strategies-zippy-link-3.

Cunha, Alexandre Ashade Lassance, Melissa Carvalho Costa, and Marco Aurélio C Pacheco (2019), "Sentiment analysis of youtube video comments using deep neural networks", in *Artificial Intelligence and Soft Computing: 18th International Conference, ICAISC 2019, Zakopane, Poland, June 16–20, 2019, Proceedings, Part I 18*, Springer, pp. 561-570.

Da Costa, Liliane Soares, Italo L Oliveira, and Renato Fileto (2023), "Text classification using embeddings: a survey", *Knowledge and Information Systems*, 65, 7, pp. 2761-2803.

Dadvar, Maral and Kai Eckert (2020), "Cyberbullying detection in social networks using deep learning based models", in *Big Data Analytics and Knowledge Discovery: 22nd International Conference, DaWaK 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings 22*, Springer, pp. 245-255.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber (2017), "Automated hate speech detection and the problem of offensive language", in *Proceedings of the international AAAI conference on web and social media*, 1, vol. 11, pp. 512-515.

Dell'Orletta, Felice, Simonetta Montemagni, and Giulia Venturi (2011), "READ–IT: Assessing readability of Italian texts with a view to text simplification", in *Proceedings of the second workshop on speech and language processing for assistive technologies*, pp. 73-83.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018), "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*.

Di Gennaro, Giovanni, Amedeo Buonanno, and Francesco AN Palmieri (2021), "Considerations about learning Word2Vec", *The Journal of Supercomputing*, pp. 1-16.

Eickhoff, Carsten and Arjen P de Vries (2010), "Identifying suitable YouTube videos for children", *3rd Networked and electronic media summit (NEM)*.

Eickhoff, Carsten, Pavel Serdyukov, and Arjen P De Vries (2011), "A combined topical/non-topical approach to identifying web sites for children", in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 505-514.

Eickhoff, Carsten, Pavel Serdyukov, and Arjen P de Vries (2010), "Web page classification on child suitability", in *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 1425-1428.

Ekman, Paul (1992), "An argument for basic emotions", *Cognition & emotion*, 6, 3-4, pp. 169-200.

Elhoseiny, Mohamed, Jingen Liu, Hui Cheng, Harpreet Sawhney, and Ahmed Elgammal (2016), "Zero-shot event detection by multimodal distributional semantic embedding of videos", in *Proceedings of the AAAI conference on artificial intelligence*, 1, vol. 30.

Erikson, E. (1963), *Children and society*, New York: Narton.

Fleiss, Joseph L (1971), "Measuring nominal scale agreement among many raters." *Psychological bulletin*, 76, 5, p. 378.

Franks, Jason (2022), "Text Classification for Records Management", *Journal on Computing and Cultural Heritage (JOCCH)*.

Gagliardi, Gloria et al. (2018), "Inter-Annotator Agreement in linguistica: una rassegna critica", in *CEUR WORKSHOP PROCEEDINGS*, CEUR-WS, vol. 2253, pp. 206-212.

Gasparetto, Andrea, Matteo Marcuzzo, Alessandro Zangari, and Andrea Albarelli (2022a), "A Survey on Text Classification Algorithms: From Text to Predictions", *Information*, 13, 2, https://www.mdpi.com/2078-2489/13/2/83.

— (2022b), "A survey on text classification algorithms: From text to predictions", *Information*, 13, 2, p. 83.

Gautam, Anmol, Sohini Hazra, Rishabh Verma, Pallab Maji, and Bunil Kumar Balabantaray (2023), "ED-NET: Educational Teaching Video Classification Network", in *Computer Vision and Machine Intelligence: Proceedings of CVMI 2022*, Springer, pp. 141-151.

Gkolemi, Myrsini, Panagiotis Papadopoulos, Evangelos Markatos, and Nicolas Kourtellis (2022), "YouTubers Not madeForKids: Detecting Channels Sharing Inappropriate Videos Targeting Children", in *Proceedings of the 14th ACM Web Science Conference 2022*, pp. 370-381.

Goodrow, Cristos (2021), "On YouTube's recommendation system", *YouTube Official Blog*.

Gossen, Tatiana and Andreas Nürnberger (2013), "Specifics of information retrieval for young users: A survey", *Information Processing & Management*, 49, 4, pp. 739-756.

Haloi, Pranabjyoti, Arnab Kisor Bordoloi, and MK Bhuyan (2019), "Unsupervised broadcast news video shot segmentation and classification", in *2019 2nd International Conference on Innovations in Electronics, Signal Processing and Communication (IESC)*, IEEE, pp. 243-251.

Hammami, Mohamed, Youssef Chahir, and Liming Chen (2003), "WebGuard: Web based adult content detection and filtering system", in *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, IEEE, pp. 574-578.

Hani, John, Nashaat Mohamed, Mostafa Ahmed, Zeyad Emad, Eslam Amer, and Mohammed Ammar (2019), "Social media cyberbullying detection using machine learning", *International Journal of Advanced Computer Science and Applications*, 10, 5.

Harris, Zellig S (1954), "Distributional structure", *Word*, 10, 2-3, pp. 146-162.

Hauptmann, Alexander, Rong Yan, Yanjun Qi, Rong Jin, Michael G Christel, Mark Derthick, Ming-yu Chen, Robert Baron, W-H Lin, and Tobun D Ng (2002), "Video classification and retrieval with the informedia digital video library system".

HB, Barathi Ganesh, M Anand Kumar, and KP Soman (2016), "Distributional Semantic Representation for Text Classification and Information Retrieval.", in *FIRE (Working Notes)*, pp. 126-130.

Hochreiter, S (1997), "Long Short-term Memory", *Neural Computation MIT-Press*.

Hovy, Dirk and Diyi Yang (2021), "The importance of modeling social factors of language: Theory and practice", in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pp. 588-602.

Hu, Yongjun, Jia Ding, Zixin Dou, and Huiyou Chang (2022), "Short-Text Classification Detector: A Bert-Based Mental Approach", *Computational Intelligence and Neuroscience*, 2022.

Hu, Zeyu, Jintao Cui, Wei-Hua Wang, Feng Lu, and Binhui Wang (2022), "Video Content Classification Using Time-Sync Comments and Titles", in *2022 7th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, pp. 252-258, DOI: 10.1109/ICCCBDA55098.2022.9778285.

Huang, Chunneng, Tianjun Fu, and Hsinchun Chen (2010), "Text-based video content classification for online video-sharing sites", *Journal of the American Society for Information Science and Technology*, 61, 5, pp. 891-906.

Ishikawa, Akari, Edson Bollis, and Sandra Avila (2019), "Combating the elsagate phenomenon: Deep learning architectures for disturbing cartoons", in *2019 7th International Workshop on Biometrics and Forensics (IWBF)*, IEEE, pp. 1-6.

Jagtap, Raj, Abhinav Kumar, Rahul Goel, Shakshi Sharma, Rajesh Sharma, and Clint P George (2021), "Misinformation Detection on YouTube Using Video Captions", *arXiv preprint arXiv:2107.00941*.

Ježek, E. and R. Sprugnoli (2023), *Linguistica computazionale: introduzione all'analisi automatica dei testi*, Itinerari. Linguistica, Il mulino., ISBN: 9788815290359, https://books.google.it/books?id=_UqmzQEACAAJ.

Jiang, Bo, Lei Zhou, Li Lin, Binbin Xu, Jiahong Yu, Xuping Zheng, and Kailin Wu (2019), "A Real-Time Multi-Label Classification System for Short Videos", in *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, pp. 534-538.

Jurgens, David and Keith Stevens (2010), "The S-Space package: an open source package for word space models", in *Proceedings of the ACL 2010 system demonstrations*, pp. 30-35.

Kadhim, Ammar Ismael (2019), "Survey on supervised machine learning techniques for automatic text classification", *Artificial Intelligence Review*, 52, 1, pp. 273-292.

Kant, Neel, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro (2018), "Practical text classification with large pre-trained language models", *arXiv preprint arXiv:1812.01207*.

Kaushal, Rishabh, Srishty Saha, Payal Bajaj, and Ponnurangam Kumaraguru (2016), "KidsTube: Detection, characterization and analysis of child unsafe content & promoters on YouTube", in *2016 14th Annual Conference on Privacy, Security and Trust (PST)*, IEEE, pp. 157-164.

Kerz, Elma, Yu Qiao, and Daniel Wiechmann (2021), "Language that captivates the audience: predicting affective ratings of ted talks in a multi-label classification task", in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 13-24.

Khatua, Aparup, Apalak Khatua, and Erik Cambria (2019), "A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks", *Information Processing & Management*, 56, 1, pp. 247-257.

Kobla, Vikrant, Daniel DeMenthon, and David Scott Doermann (1999), "Identifying sports videos using replay, text, and camera motion features", in *Storage and Retrieval for Media Databases 2000*, SPIE, vol. 3972, pp. 332-343.

Kobourov, Stephen G (2012), "Spring embedders and force directed graph drawing algorithms", *arXiv preprint arXiv:1201.3011*.

Kolluri, Johnson, Dr Shaik Razia, and Soumya Ranjan Nayak (2019), "Text classification using machine learning and deep learning models", in *International Conference on Artificial Intelligence in Manufacturing & Renewable Energy (ICAIMRE)*.

Kontostathis, April, Lynne Edwards, Jen Bayzick, Amanda Leatherman, and Kristina Moore (2009), "Comparison of rule-based to human analysis of chat logs", *communication theory*, 8, 2, p. 12.

Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown (2019), "Text classification algorithms: A survey", *Information*, 10, 4, p. 150.

Landis, J Richard and Gary G Koch (1977), "The measurement of observer agreement for categorical data", *biometrics*, pp. 159-174.

Leech, Geoffrey Neil (1992), "100 Million Words of English: The British National Corpus (BNC)".

Lekach, Sasha (2018), "'The Cleaners' shows the terrors human content moderators face at work", *Mashable*, https://mashable.com/article/the-cleaners-content-moderators-facebook-twitter-google#5um_Awr_3PqM.

Lenc, Ladislav and Pavel Král (2017), "Word embeddings for multi-label document classification.", in *RANLP*, pp. 431-437.

Lenci, Alessandro (2008), "Distributional semantics in linguistic and cognitive research", *Italian journal of linguistics*, 20, 1, pp. 1-31.

Lenci, Alessandro and Magnus Sahlgren (2023), *Distributional semantics*, Cambridge University Press.

Leonardelli, Elisa, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli (2021), "Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement", *arXiv preprint arXiv:2109.13563*.

Li, Zhuoyan, Hangxiao Zhu, Zhuoran Lu, and Ming Yin (2023), "Synthetic data generation with large language models for text classification: Potential and limitations", *arXiv preprint arXiv:2310.07849*.

Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. (2022), "Holistic evaluation of language models", *arXiv preprint arXiv:2211.09110*.

Lieto, Antonio (2021), *Cognitive design for artificial minds*, Routledge.

Lieto, Antonio, Gian Luca Pozzato, Stefano Zoia, Viviana Patti, and Rossana Damiano (2021), "A commonsense reasoning framework for explanatory emotion attribution, generation and re-classification", *Knowledge-Based Systems*, 227, p. 107166.

Liu, Yinhan (2019), "Roberta: A robustly optimized bert pretraining approach", *arXiv preprint arXiv:1907.11692*.

Luccioni, Alexandra Sasha and Anna Rogers (2023), "Mind your Language (Model): Fact-Checking LLMs and their Role in NLP Research and Practice", *arXiv preprint arXiv:2308.07120*.

Lund, Kevin and Curt Burgess (1996), "Producing high-dimensional semantic spaces from lexical co-occurrence", *Behavior research methods, instruments, & computers*, 28, 2, pp. 203-208.

Lyding, Verena, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, Vito Pirrelli, et al. (2014), "The paisa'corpus of italian web texts", in *Proceedings of the 9th web as corpus workshop (WaC-9)@ EACL 2014*, EACL (European chapter of the Association for Computational Linguistics), pp. 36-43.

Ma, Huan, Changqing Zhang, Huazhu Fu, Peilin Zhao, and Bingzhe Wu (2023), "Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning", *arXiv preprint arXiv:2310.03400*.

Maisto, Alessandro, Giandomenico Martorelli, Antonietta Paone, and Serena Pelosi (2021a), "Automatic Classification and Rating of Videogames Based on Dialogues Transcript Files", in *International Conference on Emerging Internetworking, Data & Web Technologies*, Springer, pp. 301-312.

— (2021b), "Extracting video games rating labels from transcript files", *Internet of Things*, 16, p. 100439.

— (2021c), "Semi-automatic Knowledge Base Expansion for Question Answering", in *Advances in Intelligent Networking and Collaborative Systems: The 12th International Conference on Intelligent Networking and Collaborative Systems (INCoS-2020) 12*, Springer, pp. 173-182.

Manning, Christopher D (2022), "Human language understanding & reasoning", *Daedalus*, 151, 2, pp. 127-138.

Martinez, Victor R, Krishna Somandepalli, Karan Singla, Anil Ramakrishna, Yalda T Uhls, and Shrikanth Narayanan (2019), "Violence rating prediction from movie scripts", in *Proceedings of the AAAI conference on artificial intelligence*, 01, vol. 33, pp. 671-678.

Meng, Yu, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han (2020), "Text classification using label names only: A language model self-training approach", *arXiv preprint arXiv:2010.07245*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013), "Distributed representations of words and phrases and their compositionality", *Advances in neural information processing systems*, 26.

Miller, George A and Walter Charles (1991), "Contextual correlates of semantic similarity", *Language and cognitive processes*, 6, 1, pp. 1-28.

Mohammad, Saif, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko (2018), "Semeval-2018 task 1: Affect in tweets", in *Proceedings of the 12th international workshop on semantic evaluation*, pp. 1-17.

Mohammad, Saif M (2012), "From once upon a time to happily ever after: Tracking emotions in mail and books", *Decision Support Systems*, 53, 4, pp. 730-741.

Mohammad, Saif M et al. (2013), "Tracking sentiment in mail: How genders differ on emotional axes", *arXiv preprint arXiv:1309.6347*.

Mohammad, Saif M (2017), "Word affect intensities", *arXiv preprint arXiv:1704.08798*.

— (2020), "Practical and ethical considerations in the effective use of emotion and sentiment lexicons", *arXiv preprint arXiv:2011.03492*.

— (2021), "Ethics sheet for automatic emotion recognition and sentiment analysis", *arXiv preprint arXiv:2109.08256*.

Mohammad, Saif M and Peter D Turney (2013a), "Crowdsourcing a word–emotion association lexicon", *Computational intelligence*, 29, 3, pp. 436-465.

— (2013b), "Nrc emotion lexicon", *National Research Council, Canada*, 2.

Mohammad, Saif M. (2018), "Word Affect Intensities", in *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*.

Mohammadi, Sogand Mehrpour, Meysam Gouran Orimi, and Hamidreza Rabiee (2023), "Enhancing the Prediction of Emotional Experience in Movies using Deep Neural Networks: The Significance of Audio and Language", *arXiv preprint arXiv:2306.10397*.

Mohaouchane, Hanane, Asmaa Mourhir, and Nikola S Nikolov (2019), "Detecting offensive language on arabic social media using deep learning", in *2019 sixth international conference on social networks analysis, management and security (SNAMS)*, IEEE, pp. 466-471.

Mollas, Ioannis, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas (2022), "ETHOS: a multi-label hate speech detection dataset", *Complex & Intelligent Systems*, 8, 6, pp. 4663-4678.

Moore, Timothy E (2014), *Cognitive development and acquisition of language*, Elsevier.

Mouheb, Djedjiga, Raghad Albarghash, Mohamad Fouzi Mowakeh, Zaher Al Aghbari, and Ibrahim Kamel (2019), "Detection of Arabic cyberbullying on social networks using machine learning", in *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, pp. 1-5.

Navigli, Roberto (2009), "Word sense disambiguation: A survey", *ACM computing surveys (CSUR)*, 41, 2, pp. 1-69.

Neumann, Michelle M and Christothea Herodotou (2020), "Evaluating YouTube videos for young children", *Education and Information Technologies*, 25, pp. 4459-4475.

Obadimu, Adewale, Esther Mead, Muhammad Nihal Hussain, and Nitin Agarwal (2019), "Identifying toxicity within youtube video comment", in *Social, Cultural, and Behavioral Modeling: 12th International Conference, SBP-BRiMS 2019, Washington, DC, USA, July 9–12, 2019, Proceedings 12*, Springer, pp. 214-223.

Of America, Motion Picture Association (2010), "MPAA, Motion Picture Association of America. Classification and rating rules".

Oliveira, Italo L, Renato Fileto, René Speck, Luis PF Garcia, Diego Moussallem, and Jens Lehmann (2021), "Towards holistic entity linking: Survey and directions", *Information Systems*, 95, p. 101624.

Ormord, J. and K. Davis (1999), *Human Learning*, Merrill.

Pang, Bo and Lillian Lee (2004), "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts", *arXiv preprint cs/0409058*.

Pang, Bo, Lillian Lee, et al. (2008), "Opinion mining and sentiment analysis", *Foundations and Trends® in information retrieval*, 2, 1–2, pp. 1-135.

Pang, Nuo, Songlin Guo, Ming Yan, and Chien Aun Chan (2023), "A Short Video Classification Framework Based on Cross-Modal Fusion", *Sensors*, 23, 20, p. 8425.

Papadamou, Kostantinos, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos (2019), "Disturbed youtube for kids: Characterizing and detecting disturbing content on youtube", *arXiv preprint arXiv:1901.07046*.

— (2020), "Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children", in *Proceedings of the international AAAI conference on web and social media*, vol. 14, pp. 522-533.

Parisi, Loreto, Simone Francia, and Paolo Magnani (2020), *UmBERTo: an Italian Language Model trained with Whole Word Masking*, https://github.com/musixmatchresearch/umberto.

Park, Eunjeong L and Sungzoon Cho (2014), "KoNLPy: Korean natural language processing in Python", in *Proceedings of the 26th annual conference on human & cognitive language technology*, Chuncheon Korea, vol. 6, pp. 133-136.

Passaro, Lucia, Laura Pollacci, Alessandro Lenci, et al. (2015), "Item: A vector space model to bootstrap an italian emotive lexicon", in *Proceedings of the second Italian Conference on Computational Linguistics CLiC-it 2015*, Academia University Press, pp. 215-220.

Pavlinek, Miha and Vili Podgorelec (2017), "Text classification method based on self-training and LDA topic models", *Expert Systems with Applications*, 80, pp. 83-93.

Pearl, Lisa and Mark Steyvers (2010), "Identifying emotions, intentions, and attitudes in text using a game with a purpose", in *Proceedings of the naacl hlt 2010 workshop on computational approaches to analysis and generation of emotion in text*, pp. 71-79.

Peña, Alejandro, Aythami Morales, Julian Fierrez, Ignacio Serna, Javier Ortega-Garcia, Iñigo Puente, Jorge Cordova, and Gonzalo Cordova (2023), "Leveraging Large Language Models for Topic Classification in the Domain of Public Affairs", *arXiv preprint arXiv:2306.02864*.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014), "Glove: Global vectors for word representation", in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543.

Piaget, J. and B. Inhelder (1972), *The Psychology Of The Child*, Basic Books, ISBN: 9780465095001, https://books.google.it/books?id=VR1btAEACAAJ.

Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi (2002), "MultiWordNet: developing an aligned multilingual database", in *First international conference on global WordNet*, pp. 293-302.

Pittaras, Nikiforos, George Giannakopoulos, Georgios Papadakis, and Vangelis Karkaletsis (2021), "Text classification with semantically enriched word embeddings", *Natural Language Engineering*, 27, 4, pp. 391-425.

Plutchik, Robert (1984), "Emotions: A general psychoevolutionary theory", *Approaches to emotion*, 1984, pp. 197-219.

Polignano, Marco, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, Valerio Basile, et al. (2019), "Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets", in *CEUR workshop proceedings*, CEUR, vol. 2481, pp. 1-6.

Pollacci, Laura and Alessandro Lenci (n.d.), "Italian Emotive Lexicon" ().

Potha, Nektaria and Efstathios Stamatatos (2019), "Improving author verification based on topic modeling", *Journal of the Association for Information Science and Technology*, 70, 10, pp. 1074-1088.

Pujar, Manjunath and Monica R Mundada (2021), "A Systematic Review Web Content Mining Tools and its Applications", *International Journal of Advanced Computer Science and Applications*, 12, 8.

Purdie-Vaughns, Valerie and Richard P Eibach (2008), "Intersectional invisibility: The distinctive advantages and disadvantages of multiple subordinate-group identities", *Sex roles*, 59, pp. 377-391.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. (2019), "Language models are unsupervised multitask learners", *OpenAI blog*, 1, 8, p. 9.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu (2020), "Exploring the limits of transfer learning with a unified text-to-text transformer", *The Journal of Machine Learning Research*, 21, 1, pp. 5485-5551.

Ramesh, M and K Mahesh (2020), "A Performance Analysis of Pre-trained Neural Network and Design of CNN for Sports Video Classification", in *2020 International Conference on Communication and Signal Processing (ICCSP)*, IEEE, pp. 0213-0216.

Raschka, S. and V. Mirjalili (2020), *Machine learning con Python. Costruire algoritmi per generare conoscenza*, Guida completa, Apogeo, ISBN: 9788850335244.

Rathi, RN and A Mustafi (2022), "The importance of Term Weighting in semantic understanding of text: A review of techniques", *Multimedia Tools and Applications*, pp. 1-23.

Reddy, Sanjana, Nikitha Srikanth, and GS Sharvani (2021), "Development of kid-friendly youtube access model using deep learning", in *Data Science and Security: Proceedings of IDSCS 2020*, Springer, pp. 243-250.

Reimers, Nils and Iryna Gurevych (2020), "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation", in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, https://arxiv.org/abs/2004.09813.

Rohde, Douglas LT, Laura M Gonnerman, and David C Plaut (2006), "An improved model of semantic similarity based on lexical co-occurrence", *Communications of the ACM*, 8, 627-633, p. 116.

Rong, Xin (2014), "word2vec parameter learning explained", *arXiv preprint arXiv:1411.2738*.

Sanh, V (2019), "DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter", *arXiv preprint arXiv:1910.01108*.

Schmidt, Anna and Michael Wiegand (2017), "A survey on hate speech detection using natural language processing", in *Proceedings of the fifth international workshop on natural language processing for social media*, pp. 1-10.

Schwartz, Matthew S. (2019), "Advertisers Abandon YouTube Over Concerns That Pedophiles Lurk In Comments Section", *npr*, https://www.npr.org/2019/02/22/696949013/advertisers-abandon-youtube-over-concerns-that-pedophiles-lurk-in-comments-secti#:~:text=Big%20brands%20are%20pulling%20their,purge%20the%20system%20of%20pedophiles.

Sebastiani, Fabrizio (2002), "Machine learning in automated text categorization", *ACM computing surveys (CSUR)*, 34, 1, pp. 1-47.

Sebastiani, Fabrizio and Andrea Esuli (2006), "Sentiwordnet: A publicly available lexical resource for opinion mining", in *Proceedings of the 5th international conference on language resources and evaluation*, European Language Resources Association (ELRA) Genoa, Italy, pp. 417-422.

Shafaei, Mahsa, Niloofar Safi Samghabadi, Sudipta Kar, and Thamar Solorio (2019), "Rating for parents: Predicting children suitability rating for movies based on language of the movies", *arXiv preprint arXiv:1908.07819*.

Shah, Foram P and Vibha Patel (2016), "A review on feature selection and feature extraction for text classification", in *2016 international conference on wireless communications, signal processing and networking (WiSPNET)*, IEEE, pp. 2264-2268.

Shekhar, Shubhanshu, Aman Saini, et al. (2021), "Utilizing Topic Modelling to Identify Abusive Comments on YouTube", in *2021 International Conference on Intelligent Technologies (CONIT)*, IEEE, pp. 1-5.

Shen, Wei, Jianyong Wang, and Jiawei Han (2014), "Entity linking with a knowledge base: Issues, techniques, and solutions", *IEEE Transactions on Knowledge and Data Engineering*, 27, 2, pp. 443-460.

Sprugnoli, Rachele et al. (2020), "Multiemotions-it: a new dataset for opinion polarity and emotion analysis for italian", in *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, Accademia University Press, pp. 402-408.

Stein, Roger Alan, Patricia A Jaques, and Joao Francisco Valiati (2019), "An analysis of hierarchical text classification using word embeddings", *Information Sciences*, 471, pp. 216-232.

Stoica, Alexandru Stefan, Stella Heras, Javier Palanca, Vicente Julián, and Marian Cristian Mihaescu (2021), "Classification of educational videos by using a semi-supervised learning method on transcripts and keywords", *Neurocomputing*, 456, pp. 637-647.

Strapparava, Carlo, Alessandro Valitutti, et al. (2004), "Wordnet affect: an affective extension of wordnet.", in *Lrec*, 1083-1086, Lisbon, Portugal, vol. 4, p. 40.

Sun, Xiaofei, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang (2023), "Text Classification via Large Language Models", *arXiv preprint arXiv:2305.08377*.

Tahir, Rashid, Faizan Ahmed, Hammas Saeed, Shiza Ali, Fareed Zaffar, and Christo Wilson (2019), "Bringing the kid back into youtube kids: Detecting inappropriate content on video streaming platforms", in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 464-469.

Tang, Zhong, Wenqiang Li, and Yan Li (2022), "An improved supervised term weighting scheme for text representation and classification", *Expert Systems with Applications*, 189, p. 115985.

Thangaraj, M and M Sivakami (2018), "Text classification techniques: a literature review", *Interdisciplinary Journal of Information, Knowledge, and Management*, 13, p. 117.

Tolles, Juliana and William J Meurer (2016), "Logistic regression: relating patient characteristics to outcomes", *Jama*, 316, 5, pp. 533-534.

Turian, Joseph, Lev Ratinov, and Yoshua Bengio (2010), "Word representations: a simple and general method for semi-supervised learning", in *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 384-394.

Turner, Clayton A, Alexander D Jacobs, Cassios K Marques, James C Oates, Diane L Kamen, Paul E Anderson, and Jihad S Obeid (2017), "Word2Vec inversion and traditional

text classifiers for phenotyping lupus", *BMC medical informatics and decision making*, 17, 1, pp. 1-11.

Vasantharajan, Charangan and Uthayasanker Thayasivam (2022), "Towards offensive language identification for Tamil code-mixed YouTube comments and posts", *SN Computer Science*, 3, 1, p. 94.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin, et al. (2017), "Advances in neural information processing systems", *Attention is All you Need*.

Vichi de Marchi Diletta Pistono, Cristiana Pulcinelli (2023), *Tempi Digitali, Atlante dell'infanzia (a rischio) dell'Italia*.

Vietri, Simonetta (2004), *Lessico-grammatica dell'italiano: metodi, descrizioni e applicazioni*, UTET libreria.

Vitale, Pierluigi, Serena Pelosi, and Mariacristina Falco (2020), "# AndràTuttobene: Images, texts, emojis and geodata in a sentiment analysis pipeline", in *CEUR Workshop Proc.*

Wang, Kai, Yuqi Liu, Bin Cao, and Jing Fan (2023), "T k TC: A framework for top-k text classification of multimedia computing in wireless networks", *Wireless Networks*, 29, 4, pp. 1523-1534.

Wang, Peng, Rui Cai, and Shi-Qiang Yang (2003), "A hybrid approach to news video classification multimodal features", in *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, IEEE, vol. 2, pp. 787-791.

Wartella, Ellen A, Elizabeth A Vandewater, and Victoria J Rideout (2005), *Introduction: electronic media use in the lives of infants, toddlers, and preschoolers*.

Wilson, Heather (2020), "Youtube is unsafe for children: Youtube's safeguards and the current legal framework are inadequate to protect children from disturbing content", *Seattle J. Tech. Envtl. & Innovation L.*, 10, p. 237.

Yafooz, Wael MS, Abdullah Alsaeedi, Reyadh Alluhaibi, and Abdel-Hamid Mohamed Emara (2022), "Enhancing multi-class web video categorization model using machine and deep learning approaches".

Yang, Kai-Cheng and Filippo Menczer (2023), "Large language models can rate news outlet credibility", *arXiv preprint arXiv:2304.00228*.

Yenala, Harish, Ashish Jhanwar, Manoj K Chinnakotla, and Jay Goyal (2018), "Deep learning for detecting inappropriate content in text", *International Journal of Data Science and Analytics*, 6, pp. 273-286.

Yousaf, Kanwal and Tabassam Nawaz (2022), "A deep learning-based approach for inappropriate content detection and classification of youtube videos", *IEEE Access*, 10, pp. 16283-16298.

Yu, Hyerim and Eunil Park (2023), "A harmless webtoon for all: An automatic age-restriction prediction system for webtoon contents", *Telematics and Informatics*, 76, p. 101906.

Zhang, Harry (2004), "The optimality of naive Bayes", *Aa*, 1, 2, p. 3.

Zhang, Jingqing, Piyawat Lertvittayakumjorn, and Yike Guo (2019), "Integrating semantic knowledge to tackle zero-shot text classification", *arXiv preprint arXiv:1903.12626*.

Zhu, Haoran and Lei Lei (2022), "The Research Trends of Text Classification Studies (2000–2020): A Bibliometric Analysis", *SAGE Open*, 12, 2, p. 21582440221089963.

La borsa di dottorato cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005),
risorse Fondo Sociale Europeo REACT-EU, Azione I.1 “Dottorati Innovativi con caratterizzazione industriale”, Azione IV.4 “Dottorati e contratti di ricerca su tematiche dell’innovazione” e Azione IV.5 “Dottorati su tematiche Green”