Linguistic complexity and engagement in social media communication for cultural heritage

Massimo Guarino[1], Violetta Simonacci[2], Michele Gallo[1]

[1]University of Naples "L'Orientale" – mguarino@unior.it
[2]University of Naples Federico II – violetta.simonacci@unior.it

Abstract
This contribution has the dual goal of testing the usage of readability and lexical complexity tools in social media writing and analyzing the relationship between linguistic complexity, cultural topics, and social engagement. A corpus of tweets from the official Twitter page of the Italian Ministry of Culture has been stored together with available information on social engagement. Once the tweets have been properly processed with standard Natural Language Processing tasks, linguistic complexity is measured with eight readability and lexical diversity indicators developed outside the social media context. Topic modeling is then carried out through the LDA technique. An explorative study of cause-effect relationships between lexical complexity, topic grouping, and engagement measures is performed by using a PLS-PM mixed approach.

Keywords: corpus, NLP, readability indicators, tokenization, tweets, vocabulary

1. Introduction

The official Twitter page of the Italian Ministry of Culture (MiC) is a powerful resource of social media textual data on front-line cultural topics. A corpus, consisting of a large body of tweets from this page, and some additional measures such as likes and retweets were collected with the purpose of conducting a detailed study on readability, engagement, and cultural topics.
The analysis has been carried out in three working moments. The first phase includes a preprocessing stage and the implementation of a series of Natural Language Processing (NLP) tasks on all the documents of the corpus to clean, organize and lemmatize the data, preparing it for statistical analysis.
In the second working phase, topic and language complexity measures were computed. In detail, a topic analysis was conducted with the Latent Dirichlet Allocation (LDA) approach (Blei, 2012) to obtain indicators that differentiate documents based on dominant topic and common tokens incidence. Moreover,

a set of eight readability[1] indicators was calculated for each cleaned tweet. Some of these measures focus mainly on vocabulary choices, while others are devised to give a better grasp of language richness[2]. Testing the readability of texts in the social media context can help further explain how the use of the language on mass media communication platforms can impact organizations' success with respect to their social engagement policies.

In the final working phase of this project, the relationships among the considered measures, both directly observed and computed, are analyzed via PLS-PM. In detail, the following questions were considered when building the model: 1) What is the relationship between the different computed measures, and can they be synthesized in meaningful compound constructs? 2) Are the readability indicators a good measure of language structure and richness in a social media setting? 3) What role do language complexity and topic indicators play in social engagement with respect to cultural communication?

It is important to note that the PLS-PM model is used for theory building rather than for confirming hypotheses. In this sense, results are viewed from an exploratory standpoint as a first step in the search for possible theories and connections of interest without assumptions (Henseler, 2008). For this reason, the well-known limits of the procedure were not addressed in detail here (see Rönkkö et al., 2016 for an in-depth discussion on PLS shortcomings).

The constructed dataset is quite rich and complex. Details on preprocessing and indicators construction are provided in Section 2. A quick overview of the PLS-PM modeling tool is presented in Section 3. Preliminary results of the model are presented in Section 4.

2. Data and readability indicators

We built the textual dataset based on the tweets of the MiC collected on its Twitter official page (about 2,300 tweets). For each tweet, we also collected the number of likes and retweets and its creation date. Furthermore, we performed a series of NLP tasks in order to prepare all the documents for the computation of indicators.

Preprocessing steps were necessary to remove textual noises and extra whitespaces to perform better tokenization (i.e., splitting each text into a set of sentences and sentences into a set of words, the "word-tokens"). Once preprocessed, word-tokens were lowercased to put all the texts on a level playing

[1] Readability can be defined as "what makes some texts easier to read than others" (DuBay, 2004).

[2] Vocabulary knowledge, including lexical diversity and richness, are principal aspets in reading comprehension (Heilman et al., 2008; Gooding et al., 2021). Moreover, the importance of lexical richness has been investigated also in readability systems (Vajjala and Meurers, 2012; Xia, 2019)

field and two other relevant NLP tasks were performed using Stanza[3]: Part of Speech (PoS) tagging and lemmatization.

Pre-processed tweets were clustered into 17 groups via LDA, an unsupervised machine learning technique providing highly interpretable topics based on Bayesian algorithms and Dirichlet probability distributions (Blei, 2012). Given the nature and length of the tweets, we chose to focus only on some content words (adjectives, nouns, and verbs) for topic detection. The number of groups was chosen by maximizing topic coherence, a reliable measure used to determine the readability/interpretability of topics (Tijare and Rani, 2020). Two topic indexes were built: Top1, which groups documents based on the incidence of common tokens (a higher value is associated with a smaller percentage of such tokens), and Top2 which expresses the percentage of dominant topic.

Successively, eight readability/lexical richness indicators were computed for each tweet. To start, we define: TTR[4], the type/token ratio, namely the ratio between the number of different words in a text and the total amount of its word-tokens; HAP, the percentage of hapaxes (words which occur only once in a text); and CONT the percentage of content words[5].

A fourth fundamental indicator is given by the *Gulpease* index (Gulp), which is of particular interest because it was developed in reference to the Italian language and educational system (Lucisano and Piemontese, 1988). It can be described by the following formulation:

$$Gulpease = 89 + \frac{(300 \times sentences - 10 \times letters)}{number\ of\ words} \quad (1)$$

The index ranges between 0 (un-intelligible) and 100 (maximum readability) and it is inversely correlated with the degree of educational level. Documents with a

[3] While for sentence recognition and tokenization we used Stanza, the NLP tool developed within the Stanford NLP research group (Manning et al., 2014; Peng et al., 2020), for words and letters tokenization we used regular expressions, which are sequences of characters that define a search pattern (Aho, 1991). Word tokenization refers only to words, thus excluding symbols and punctuations from the computation.

[4] Given that tweets do not greatly differ in terms of texts length, the ratio between types and tokens can be considered stable.

[5] There are 3 main class words in the universal part of speech tagging framework (i.e., the standard part of speech tagset for many languages) (Murphy 2010, Petrov et al., 2011): open class words (the most important class comprising the so called content words, that is to say words giving meaning to a text as adjectives, nouns, verbs (other than auxiliary verbs), proper nouns, adverbs and interjections), closed class words (this is the set comprising the so called function words, that is to say prepositions, auxiliary verbs, conjunctions, pronoun, etc..), and other words (everything not comprised within the previous classes such as punctuation, symbols and other misclassified elements).

low level of readability (i.e. with an index lower than 60) are hardly understood by native speakers with junior high school education; a value lower than 40 means that documents are not easily understandable by those with high school education.

The last four indicators are connected to the use of the Italian *Base* Vocabulary (De Mauro, 1980; De Mauro and Chiari, 2016), namely a collection of the first high-frequency words (about 7,000 lemmas) of the written Italian language which are greatly understandable also by those who are less educated. The very first 2,000 lemmas of this collection constitute the Italian *Fundamental* Vocabulary. In this perspective, we computed the share of content words present in the base vocabulary (Dic0) and fundamental vocabulary (Dic1). To avoid biases due to the tokenization of the tweets, we also computed two more indicators with respect to the complete text, and not just the content words (labelled respectively Dic2 for the base vocabulary, and Dic3 for the fundamental one).

In addition to the described indicators, two simple volume measures were also computed to give general size information: the total number of words (Vol1) and the total number of words without repetitions (Vol2). Lastly, the Engagement measures considered in the model were extracted directly from the tweets and include the number of likes (Hot1) and the number of retweets (Hot2)

3. PLS-PM method

To investigate relationships among the different types of measures discussed, a multivariate procedure capable of unveiling latent constructs and probing linear structures can be very helpful. An exploratory PLS-PM approach appears to be a suitable choice as no specific assumptions were made. The procedure identifies latent constructs for the manifest variables, represented in our data by readability indicators, volume, topic, and engagement measures, while simultaneously verifying the relationships between the constructs. PLS-PM can be seen as a parallel combination within an iterative scheme of path analysis to estimate the relationship among latent variables (structural or inner part of the model) and a factor analysis to obtain these latent measures from the observed variables (measurement or outer part of the model).

The structural or inner model can be described in brief by the following formulation where the path coefficient β_{mi} expresses the causal effect between the generic endogenous latent variables ξ_j and the $[1, \dots, m, \dots, M]$ exogenous latent variables ξ_m:

$$\xi_j = \sum_{m=1}^{M} \beta_{mj}\, \xi_m + \tau_j \qquad (2)$$

The value τ_j denotes the structural error. The measurement or outer model for linking the manifest variables to the latent constructs can be either reflective, formative, or mixed, on the basis of theoretical or empirical reasons. In the reflective blocks, the manifest variables are intended as the manifestations of the latent construct. Let us consider the q-th block referring to the $[1, \dots, p, \dots, P]$ manifest variables x_{p_q}. The relationship between the manifest variables and their corresponding latent construct ξ_q can be formulated as:

$$x_{pq} = \lambda_{pq}\xi_{pq} + \varepsilon_q \tag{3}$$

Here $\lambda_{pq}$is the simple regression coefficient between the manifest variable and the latent variable which is denoted as loading.

In formative blocks, the latent variables are seen as the best predictors of the manifest variables under the constraint of explaining the relationships between latent adjacent measures. In this case, the latent construct is expressed as the linear combination of the manifest variables with associated weights w_{p_q}in the following manner:

$$\xi_q = \sum_{p=1}^{p_q} w_{p_q} x_{p_q} + \delta_q \tag{4}$$

Here a mixed scheme is used for the measurement model. Parameter estimation is carried out with an iterative process that allows the simultaneous estimation of latent variables, outer and inner weights. After initializing of outer weights, the procedure conducts alternating inner and outer estimating steps through a system of linear equations by differentiating outer weights computation for reflective and formative blocks, until convergence is reached. For a detailed illustration of the PLS-PM procedure please refer to Tenenhaus et al., 2005.

4. Preliminary findings and discussion

The model input includes 14 manifest variables arranged in $q=6$ blocks, consisting of 8 language complexity indicators, 2 volume, 2 topic, and 2 engagement measures. Three blocks have reflective relations and include: 1) Volume {Vo1; Vol2}; 2) Dictionary {CONT, Dict0, Dict1, Dict2, Dict3}; and 3) Languange Richness {TTR, HAP}. Here, the reliability of the constructs is demonstrated by the high Cronbach's alpha values, equal to 0.976, 0.895, and 0.961 respectively. All loadings are acceptable except for Dict3, which records a value of 0.197. The remaining blocks have formative ties and include: 4) Topic {Top1, Top2}; 5) Structure {Gulp}; and 6) Engagement {Hot1; Hot2}.

The structural model estimates 1) the relationships between the exogenous latent variables Volume, Dictionary, and Richness with Structure and Topic and 2) the relations between the variables Structure and Topic with Engagement. To assess results, beta coefficients are reported in Table 1:

Table 1. Structural model Path Coefficients

	Structure	Topic	Engagement
Volume	0.008	-0.117	-
Dictionary	0.163***	0.597	-
Richness	-0.073***	0.619	-
Structure	-	-	0.344
Topic	-	-	-0.49

R^2 determination coefficients are quite low with the exception of Topic, where it corresponds to 0.774. This is, however, expected as many other aspects impact engagement on social media which are not the focus of this analysis. In addition, very few path coefficients are statistically significant. This is also not surprising nor problematic at this stage as the model was carried out with a heuristic approach to explore rather than to infer.

$$\xi_q = \sum_{p=1}^{p_q} w_{p_q} x_{p_q} + \delta_q \quad (3)$$

These preliminary results, which need additional evidence by extending the corpus and refining indicators, show how an Engagement increase could be brought by a higher value of Structure, which stems from the *Gulpease* readability index. On the other hand, as the topic scores increase (the topic becomes more specific), Engagement may be reduced. A positive connection also links Dictionary with Structure and with Topic, and Richness with Topic.

References

Aho A. V. (1991). Algorithms for finding patterns in strings. *Handbook of theoretical computer science (vol. A): algorithms and complexity*. MIT Press, Cambridge, MA.

Blei D. M (2012). Probabilistic topic models. *Communications of the ACM,* 55(4): 77–84

Manning C. D., Surdeanu M., Bauer J., Finkel J. R., Bethard S. and McClosky D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*.

De Mauro T. and Chiari I. (2016). Il Nuovo vocabolario di base della lingua italiana. https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana.

De Mauro T. (1980). *Guida all'uso delle parole: Come parlare e scrivere semplice e preciso: Uno stile italiano per capire e farsi capire*. Editori Riuniti.

DuBay W. H. (2004). *The Principles of Readability*. *Costa Mesa: Impact Information*, vol. (76).

Gooding S., Berzak Y., Mak T. and Sharifi, M. (2021). Predicting Text Readability from Scrolling Interactions*. arXiv preprint arXiv:2105.06354.*