Impact of Tweets Pre-processing Techniques on a Dictionary for Environment

Camilla Salvatore[1], Daniele Toninelli[2],
Michela Cameletti[3], Stephan Schlosser[4]

[1] University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126, Milan, Italy,
c.salvatore4@campus.unimib.it
[2] University of Bergamo, via dei Caniana, 2, 24127, Bergamo, Italy,
daniele.toninelli@unibg.it
[3] University of Bergamo, via dei Caniana, 2, 24127, Bergamo, Italy,
michela.cameletti@unibg.it
[4] Universität Göttingen, Goßlerstraße 19, 37073, Göttingen, Germany,
stephan.schlosser@sowi.uni-goettingen.de

Abstract

The availability of unstructured big data, such as the ones produced by social media, highlights the increasing methodological interest on text analysis and on the linked pre-processing phases. Several works have recently studied the impact of different pre-processing treatments on text classification. This aspect has been rarely studied when the target of the research is the definition of a topic-oriented dictionary that could be used to select messages regarding a certain topic among a wide group of unlabelled texts. The latter is a crucial phase: carefully filtering messages is a key aspect to start and to properly develop any type of textual analysis.
In this paper, we aim at setting up a dictionary regarding environment. Starting from a verified list of Twitter Official Social Accounts, we evaluate if and how different pre-processing treatments (and their combination) can affect the final dictionary.

Key Words: Social media, Twitter, Text classification, Text mining

1. Introduction

The large amount of textual digital trace data produced by human interactions on social media (SM) allows, to study new phenomena and to get insights about users' behaviours and attitudes. Differently from probability-based sample surveys, digital SM data are characterized by large volume and fine-grained temporal and spatial resolution, they are relatively cheap to obtain but also unstructured and affected by self-selection. Despite these challenges, SM data can be analysed for different purposes, from different perspectives and in different fields. Applications include, just to cite few examples, the study of voting behaviours (Ceron, Curini, & Iacus, 2016) and of market and citizens' sentiment (Luhmann, 2017; Ranco, Aleksovski, Caldarelli, Grcar, & Mozetic, 2015).

In order to get relevant information from the available unstructured SM data, it is necessary to *clean* them, before the analysis. It is part of the *text pre-processing phase* which is a key step in the study of textual data, because it can affect the final results. The general aim of this procedure is to reduce the dimensionality of data and/or clean them, keeping only the relevant text making the analysis easier and efficient from both a practical and computational point of view. With a wide set of options available, choosing which text pre-processing technique to implement is not an obvious choice; moreover, whether and to which extent there is an impact on the results is still an underdeveloped field of research. After the pre-processing phase, the text is ready to be analysed with further and more specific techniques, including sentiment extraction, topic detection or, more generally, for classification purposes.

Different methodologies can be implemented in order to classify the text and Grimmer & Steward (2013) provide an interesting overview. First, it is worth to distinguish between the case where categories are known or unknown. In the first case, the most common choice is between a dictionary or a supervised learning method. With supervised learning, classification labels are known, and it is necessary to have a *training set* of labelled texts. In the dictionary approach, instead, a list of topic-related *key words* is available and the text class is determined according to the frequency of key words occurring in the text. On the contrary, when categories are unknown, the latent structure of the text is studied. This includes unsupervised learning techniques, such as topic modeling.

In this paper, we focus our attention on the dictionary approach. Differently from the more sophisticated and, usually, black-box machine learning algorithms, dictionary-based classification is a very popular technique to classify the text mainly due to its easy implementation and interpretation. It is considered a very flexible approach, because it is relatively easy and fast to build and apply context-specific dictionaries. This method can be used for different purposes and in different fields. For example, the dictionary approach has been applied in order to study terrorist speeches (Pennebaker, Chung, Krippendorf, & Bock, 2008), emotional contagion on Facebook (Kramer, Guillory, & Hancock, 2014) and policy position (Laver & Garry, 2000), among other topics.

In addition, Official Statistical Institutes are developing experimental statistics through the analysis of textual data and, in particular, through the application of the dictionary approach, such as the Social Tension Indicator[1] by Statistics Netherlands and the Social Mood on Economic Index[2] by the Italian National Institute of Statistics (ISTAT).

In order to apply the dictionary approach, both text and dictionary must be treated with the same text pre-processing techniques. Currently, there are just few extensive studies about the effect of such techniques on the construction of dictionaries and on the performance of analyses based on this approach.

In this framework, our contribution is twofold. First, we propose an automatic method in order to construct domain-specific dictionaries relying on the availability of domain-specific texts from a set of sources that are naturally linked to the type of text we want to classify. Our study aims at finding a flexible and generally reliable method that could be applied in a wide range of research fields. Second, we study the impact of text pre-processing on the construction of the dictionary. We focus on the environmental topic and on UK tweets.

The following part of the paper is organized as follows. Section 2 presents the main text pre-processing techniques and a literature review about their on textual analysis. Section 3 introduces the framework of our analysis and the data we used. In Section 4 we present and discuss our results. Conclusions and ideas for further research are discussed in Section 5.

[1] https://www.cbs.nl/en-gb/over-ons/innovation/project/social-tensions-indicator-gauging-society.

[2] https://www.istat.it/en/archivio/219600.

2. Text pre-processing techniques: a literature review

Generally, the pre-processing phase includes tasks such as the removal of punctuation, symbols, numbers, stop words, infrequently used terms, lowercasing, Part Of Speech (POS) tagging and word normalization. Then, depending upon the type of data, it is also common to remove URLs, user mentions, Unicode strings and replacing slangs, abbreviations and contractions. These last operations are common when working with SM, blogs and review's texts. Even obvious choices, such as the removal of punctuation, symbols and number, should be thoughtfully implemented, because these decisions could potentially affect the data that will be finally obtained and their quality. For example, repeated punctuation symbols such as "!!!" could be seen as an amplifier of the sentiment keyword; numbers can refer to laws in the case of legal texts and they are essential for identifying the underlying topic; currency symbols (i.e., $, €, etc.) could be helpful in identifying economic texts. Lowercasing is also important for distinguish between common and proper names. Stop-words removal is a generally accepted task. It is helpful in reducing the dimensionality dropping non-relevant words. Other two tasks that aim at reducing dimensionality are the removal of infrequent terms and word normalization. Infrequent terms are defined as those words that appear in a number of documents less than a pre-defined threshold. Usually, this arbitrary threshold ranges between 0.5-1% (Denny & Spirling, 2018). An alternative is to select only the most relevant terms using the tf-idf index (Kwartler, 2017), which is a measure of the importance of a word inside a text. Finally, word normalization consists in stemming or lemmatization. These are two exclusive tasks. Stemming keeps only the root of a word (the *stem*), while Lemmatization considers the context of a word (the whole text and POS of words) in order to identify the *lemma*. Stemming is usually preferred because it is less computationally expensive. For more details about such techniques, please refer to Kwartler (2017).

The effect of pre-processing techniques has been studied mainly with reference to machine learning applications and topic modeling. Song, Liu & Yang (2005) found that the impact of stop words removal and stemming was not significant in order to classify English news with supervised learning. A similar result against stop words removal was also obtained by Zhao & Gui (2017) on a ML-based sentiment analysis on Twitter data. On the contrary, stop words removal had a positive impact in the study by Uysal & Gunal (2014). The authors assessed the impact of different text pre-processing techniques in two domains, supervised spam detection and news classification, and two languages, English and Turkish. Based on this analysis, it is recommended to carefully evaluate all possible combinations of techniques. The performance of the classifier changes significantly according to the text pre-processing techniques, also in relationship to the language. A similar study about combining different techniques has been performed by Symeonidis, Effrosynidis & Arampatzis (2018). The objective was to implement different supervised ML sentiment analysis algorithm on two Twitter dataset treated with different text pre-processing techniques. They found that lemmatization, number removal and replacing contractions improved the accuracy, whereas removing punctuation had a negative impact. On the contrary lemmatization was not effective on a similar study about Twitter sentiment analysis performed by Bao et al. (2014).

The order of the applied technique is also important, and these settings should be accurately defined in the research plan. Thus, for example, lowercasing should not be implemented before POS, which is a ML method that assigns each word a part-of-speech label including nouns, proper nouns, verbs, adverbs, adjectives, punctuations and others.

The main message from the literature is that the right mix of pre-processing techniques should be chosen based on the language of the text, on the data type (social media vs standard or formal texts) and, perhaps mostly, on the purpose of the planned method of subsequent analysis, such as supervised classification, sentiment analysis or unsupervised classification.

3. Methods and data

3.1 Methods

Our first contribution is to propose a method that allows to build a high-quality context-specific dictionary. We propose to follow three steps. First, we argue that it is necessary to identify a list of SM accounts linked to the topic or to the area of interest. This is because it is reasonable to assume that the language used by such users is similar to the one used in the tweets we aim at classifying. In addition to the keywords that can be identified by experts, this approach allows to include source-specific words. For example, Twitter users tend to use hashtags (#) in order to define the topic of their message and to associate their tweet to the others about the same topic. This approach allows also to account for the features of the written language on social media and thus, for specific slangs. The second step is to build the dictionary using text mining techniques. In this step, the text pre-processing techniques are implemented and the metric to extract the more relevant keywords is set. Our second contribution relates this step. We perform an experiment combining different text pre-processing techniques and we study the effect on the construction of the dictionary. The last step relates the human judgment and evaluation to fine-tuning and validating the dictionary. In this phase, experts assess the list of words and decide which keywords to include in the final dictionary.
As source-specific account we consider different categories of Official Social Accounts (OSA), i.e., institutional or recognized associations accounts related to the environment. The selection of such accounts is addressed more specifically in Section 3.2 together with the presentation of our data. In this part we focus on the experimental plan. In order to measure the impact of different pre-processing techniques, we apply some very common choices, such as: removal of punctuation, numbers, symbols, URLs and non-ASCII characters. We also evaluate some optional exclusive choices, such as stemming versus lemmatization and removal of stop words versus part of speech tagging. We also combine these techniques in order to assess their main impact and their interaction effect. For this purpose, we create six experimental group defined in Table 1.

Table 1: Experimental Plan

Group	Name	Text pre-processing techniques
Group 1	SS	Lowercasing > Stem (S) > Remove stop words (S)
Group 2	SP	Lowercasing > Stem (S) > Part Of Speech (P)
Group 3	LP	Lemmatization (L) > Lowercasing > Part Of Speech (P)
Group 4	LS	Lemmatization (L) > Lowercasing > Remove stop words (S)
Group 5	WS	Word (W) > Lowercasing > Remove stop words (S)
Group 6	WP	Word (W) > Lowercasing > Part Of Speech (P)

As Part of Speech, we only selected nouns and adjectives. In order to pre-process the text, we use *udpipe* (Wijffels, 2021) for lemmatization and POS tagging and *quanteda* (Kenneth, et al., 2018) for the rest of pre-processing techniques. We focus on unigrams, i.e. on single words. In order to select the most relevant words by OSA category, we consider the tf-idf (term frequency-inverse document frequency) measure. It is a measure of the importance of a word in a text. The mathematical definition for each term is the following (Kwartler, 2017).

$$TFIDF = TF \cdot IDF$$

Where,

$$TF = \frac{term\ occurences\ in\ a\ document}{total\ unique\ terms\ in\ the\ document}$$

and

$$IDF = \ log \frac{total\ number\ of\ documents}{number\ of\ documents\ with\ that\ term}$$

The intuition is that if a word is common (in terms of frequency), it should play a relevant role, but if the word appears in all documents, then, it is not informational.

Given the tf-idf set of values for each word and category, we select the most relevant terms based on the deciles of the tf-idf distribution. We eliminate the terms with tf-idf equal to zero and we select only terms with if-idf greater than the 99th percentile.

Finally, in order to assess the impact of the text-preprocessing techniques, we compare the resulting dictionaries considering the number of terms, of relevant terms (those related to the environment) and the percentage of overlap between dictionaries.

3.2 Data

The key aspect of the proposed methodology is to select a list of accounts linked to the topic of interest. Our aim is to build a context-specific dictionary about the environment. Our dictionary is UK-oriented, since the tweets to be analysed in a future work are geo-tagged tweets collected for UK (Schlosser, Toninelli, & Cameletti, 2021). We focus on OSA, i.e. on institutional or recognized associations accounts. In order to select the OSA, we use the list of organizations and institutions listed in the UK governmental website[3]. In total, we consider 38 OSA classified into five categories, namely: environment in general, agriculture/farm, forests, parks and water. The aim is to get the most relevel words by category. As reference period, we consider tweets posted between February 16th, 2019, and February 15th, 2020 on the OSA timelines. This base is set with the objective of analysing, in a future work, all geotagged UK tweets posted starting from February 16th, 2020. Data were retrieved using the Twitter Academic API[4] and by means of the R statistical software. The total number of OSA tweets retrieved is 26,345. Table 2 shows the number of OSA by category and the number of Tweets by OSA category.

The majority of OSA is about the general topic environment. Then, there are Parks (National parks or gardens) and Agriculture/Farm accounts. Finally, there are Forests and accounts about Water management and related issues. Looking at the number of tweets, Environmental and Parks account were more active. Then, there is Agriculture/Farm, Water and, finally, Forests.

[3] https://www.gov.uk/government/organisations

[4] https://developer.twitter.com/en/products/twitter-api/academic-research

Table 2: OSA categories and number of tweets

OSA Category	Number of Accounts (perc.)	Number of Tweets (perc.)
Environment in general	12 (31.6)	9,359 (35.5)
Agricolture/Farm	8 (21.1)	4,872 (18.5)
Forests	4 (10.5)	1,437 (5.5)
Parks	10 (26.3)	6,890 (26.1)
Water	4 (10.5)	3,787 (14.4)
Total	**38**	**26,345**

4. Results

In order to evaluate the effect of the text pre-processing techniques on the construction of dictionaries we compare the following metrics. Human validation is essential in this phase of the analysis. Each dictionary is analysed by expert and the *relevant terms* are identified. These are the words that relate the environment.

Table 3 shows the number of selected terms and the number of relevant terms. It is possible to notice that dictionaries are generally bigger, if they are based on words, rather then on lemmas and stems. Indeed, the objective of such methods is the dimensionality reduction. Then, clearly the size is smaller, when considering POS tagging, instead of stop words removal. The percentage of relevant terms is always around 50% in each experimental group.

Table 3: Dictionaries Features

Dictionary (Experimental Group)	Number of Terms	Number (perc.) of relevant terms
Group 1 – SS	306	158 (51.63)
Group 2 – SP	194	96 (49.5)
Group 3 - LP	140	76 (54.3)
Group 4 - LS	361	179 (49.6)
Group 5 – WS	405	197 (48.6)
Group 6 - WP	254	121 (47.6)
Mean	*277*	*138*

Another key point worth to be studied, is to compare the overlap rate between dictionaries considering only relevant terms in order to assess their similarity (Table 4). Due to the different structure, stems are not comparable with words or lemmas. Thus, comparisons are possible between Group 1 and Group 2, and then between all other groups. We highlight that the experimental Group 5, which relates the use of words and removal of stop words, and Group 4, about Lemmas and removal of stop words, share the higher number of terms. On the opposite, Group 5 and Group

3 (the one about lemmatization and POS tagging) show the lowest rate. The remaining ones range between 35% and 48%. In general, these percentages are not-so-high, indicating a different composition of the dictionaries.

Table 4: Overlapping Between Dictionaries (percentages)

	Group 1 **SS**	**Group 2** **SP**	**Group 3** **LP**	**Group 4** **LS**	**Group 5** **WS**	**Group 6** **WP**
Group 1 **SS**	100					
Group 2 **SP**	45	100				
Group 3 **LP**	-	-	100			
Group 4 **LS**	-	-	36	100		
Group 5 **WS**	-	-	26	74	100	
Group 6 **WP**	-	-	40	35	48	100

5. Conclusion

From this analysis it is clear that the composition of the dictionary is strongly affected by the mix of text pre-processing techniques implemented. Thus, the text pre-processing phase is crucial and should be carefully designed. A generally valid mix of methods cannot be suggested, but it strongly depends upon the type of data to be analysed and the scope of the analysis. Moreover, human evaluation is fundamental for validating and building the final dictionary. The next possible steps for this research can include an analysis including *n*-grams in the construction of dictionaries, the inclusion of an experiment additional text pre-processing techniques and the test of the dictionary effect on real data and, finally, the dictionary stability over time.

In conclusion, we argue that the dictionary approach is easy and fast to apply, in comparison to other techniques. Moreover, it has the advantage of being source and context specific. However, there are relevant choices to make before the analysis, including the selection of OSA and the choice of the text pre-processing techniques. For these reasons, it is necessary to investigate more deeply these topics.

Acknowledgements

This research project has been developed with the support of the 60% University funds (University of Bergamo).

References

Bao, Y., Quan, C., Wang, L., & Ren, F. (2014). The role of pre-processing in twitter sentiment analysis. *In International conference on intelligent computing* (p. 615-624). Springer, Cham.

Ceron, A., Curini, L., & Iacus, S. M. (2016). *Politics and Big Data: Nowcasting and forecasting elections.* London: Routledge.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2), 168-189.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3), 267-297.

Kenneth, B., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo., A. (2018). "quanteda: An R package for the quantitative analysis of textual data". *Journal of Open Source Software*, 3(30), 774.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24), 8788-8790.

Kwartler, T. (2017). *Text mining in practice with R.* John Wiley & Sons.

Laver, M., & Garry, J. (2000). Estimating policy positions from political texts. *American Journal of Political Science*, 619-634.

Luhmann, M. (2017). Using Big Data to study subjective well-being. *Current Opinion in Behavioral Sciences*, 18, 28–33.

Pennebaker, J. W., Chung, C. K., Krippendorf, K., & Bock, M. A. (2008). Computerized text analysis of Al-Qaeda transcripts. *A content analysis reader. Thousand Oaks, CA: Sage.*, 453-465.

Ranco, G., Aleksovski, D., Caldarelli, G., Grcar, M., & Mozetic, I. (2015). Price effects of Twitter sentiment. *PLOS One*, 10(9).

Schlosser, S., Toninelli, D., & Cameletti, M. (2021). Comparing Methods to Collect and Geolocate Tweets in Great Britain. *Journal of Open Innovation: Technology, Market, and Complexity*, 7(1), 44.

Song, F., Liu, S., & Yang, J. (2005). A comparative study on text representation schemes in text categorization. *Pattern analysis and applications*, 8(1), 199-209.

Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, 110, 298-310.

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information processing & management*, 50(1), 104-112.

Wijffels, J. (2021). udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the 'UDPipe''NLP'toolkit. [Computer software].

Zhao, J., & Gui, X. (2017). Comparison research on text pre-processing methods on. *IEEE Access*, 5, 2870–2879.