COVID-19 Outbreak through Tweeters' Words: Monitoring Italian Social Media Communication about COVID-19 with Text Mining and Word Embeddings

Andrea Sciandra
Department of Communication and Economics
University of Modena and Reggio Emilia
Reggio Emilia, Italy
andrea.sciandra@unimore.it

***Abstract*—In this paper we aim to analyze the Italian social media communication about COVID-19 through a Twitter dataset collected in two months. The text corpus had been studied in terms of sensitivity to the social changes that are affecting people's lives in this crisis. In addition, the results of a sentiment analysis performed by two lexicons were compared and word embedding vectors were created from the available plain texts. Following we tested the informative effectiveness of word embeddings and compared them to a bag-of-words approach in terms of text classification accuracy. First results showed a certain potential of these textual data in the description of the different phases of the outbreak. However, a different strategy is needed for a more reliable sentiment labeling, as the results proposed by the two lexicons were discordant. Finally, although presenting interesting results in terms of semantic similarity, word embeddings did not show a predictive ability higher than the frequency vectors of the terms.**

***Index Terms*—Social Media, COVID-19, Sentiment Analysis, Word Embeddings, Supervised Learning**

I. Introduction

The ongoing outbreak of the novel coronavirus disease (COVID-19) started in December, 2019 in Wuhan, China and it has now spread across the world. COVID-19 was declared as a pandemic by the World Health Organization (WHO) on March, 2020 [39]. Due to the rapid spread of this virus, several countries are/were taking strict measures like nationwide lockdown or by cordoning off the cities that had risks of community spread. Moreover, social distancing measures, travel bans, and business closures are strongly affecting social relationships [1], forcing people out of public spaces. As a result, part of the emergency communication takes place through social media platforms, which can also be useful tools to capture social change. Social media platforms are widely used at the moment to acquire and share various types of information about this crisis. They are playing an important role for disseminating information during the outbreak [21], [27] at a historic and unprecedented scale [18]. They also embody a large and easily accessible by scientists source of data, especially in the case of Twitter [20], [12]. In this way, online social media could allow to identify emerging pattern of discourse about COVID-19 and the way people deal with this crisis. Common analysis practices include text mining and deep learning techniques [29], [31]. In particular, sentiment analysis [19] proved to be a useful method to detect the polarity of social media texts and, in a broader way, to analyze user-generated content produced throughout to a disruptive event. Another powerful tool for discovering relationships in the text corpus as well as analogies between words is the word embedding technique. Word embeddings transform the words of a corpus into meaningful vectors of real numbers that may encode linguistic regularities and patterns [40]. Words close in meaning appear close together in the word embedding space, which has lower dimensionality than the bag-of-words [22], [28].

In this paper we focused on the Italian discussion on Twitter about the COVID-19 crisis. In particular, the collected dataset includes tweets before and throughout the national lockdown (decree of the Presidency of the Council of Ministers of 9th March, 2020) of one of the countries most affected by the outbreak. The main research questions of our analysis can be sum up as follow:

RQ1. Can the dataset capture the differences in language between, before and during lockdown?

RQ2. Is automatic sentiment analysis consistent with the use of two different lexicons?

RQ3. Do word embeddings improve the accuracy of a classification model compared to term frequencies?

The hypothesis will be verified by means of specific indices for the first two points and according to the level of accuracy of two classification models for the third point.

To the best of our knowledge, this is the first attempt that illustrated an analysis of Italian texts from a social media platform regarding the COVID-19 outbreak.

II. Related Work

Since COVID-19 was isolated at the end of 2019 and proved to be a serious pandemic, the scientific literature on the subject cannot be very extensive. Papers appears to be much more extensive in clinical and epidemiological fields [41], [9], while they are quite limited in the social media field. However, online discussion may reveal problems that are a consequence of the epidemic, i.e. anxiety, depression, alcohol use, and lower mental wellbeing [2]. Clearly, the impact on society is diverse and could include, among others, racism, culture (including sports & entertainment), education (school closure), and socio-economic impact [1]. Some panic phenomena occurred due to coronavirus outbreak in countries such as Italy, United States and Hong Kong led to complete shelve clearance, for example for personal hygiene products [1]. Among the negative consequences, in the initial phase discrimination had pervaded Chinese communities, a

978-1-7281-8086-1/20/$31.00 ©2020 IEEE

phenomenon partially attributable to misleading and biased media coverage [38]. This calls into question the importance of understanding risk communication and public's risk perception, an aspect positively linked to the use of social media during infectious disease outbreaks, as in the case of the Middle East Respiratory Syndrome coronavirus (MERS-CoV) outbreak in South Korea, 2015 [25].

Social media could help to disseminate reliable information and to understand the transmission of COVID-19 [21], influencing in this way the response to the pandemic and future public health policies. The most recent studies have shown that during the ongoing outbreak of COVID-19, people use social media to acquire and share various types of information. An example in this sense is that of Sina Weibo [18] (a Chinese microblogging platform very similar to Twitter), whose COVID-19 messages have been classified with natural language processing techniques in order to provide information that helps the concerned authorities or individuals to understand the situation during emergencies.

As soon as we write, the datasets and the first studies on the link between the pandemic and social media begin to rise. A commendable example of this is Chen et al. (2020) [8] who shared a multilingual COVID-19 Twitter dataset collected since January 22, 2020. With people forced out of public spaces, they stress how much conversation about these phenomena now occurs online and the importance of studying online conversation dynamics in the context of a planetary-scale epidemic outbreak of unprecedented proportions and implications.

From a substantive point of view, several studies [11], [17] have shown that people tend to prefer information that confirms their pre-existing attitudes and beliefs. Instead, Pulido et al. [27] report that false information about COVID-19 is tweeted more but retweeted less than science-based or fact-checking tweets, which tend to capture more engagement than mere facts.

In view of this research, it is also important to note the presence of the first works making use of sentiment analysis on messages in social media related to COVID-19. Barkur et al. [4] analyzed the sentiment of 24k tweets from India about COVID-19, finding overall positive and trust sentiments. This result may be an interesting basis for comparison in the following analysis.

III. Method

In this research, we collected tweets in order to monitor Italian social media discussions about COVID-19. We used Twitter REST API (through `rtweet` R package with Oauth authentication), which is known to have a limited rate and data availability not older than a week [36]. Despite these limitations, we continued collecting tweets via the keyword "coronavirus" for two months, from February 14th to April 14th, 2020. In this way, we were able to observe the reactions on this social networking site before and throughout the lockdown (March 9th) and even before the first outbreaks of the virus, which gave rise to the first Italian "red zones" (February 24th). Fig.1 shows the tweets distribution before (570665) and during the national lockdown (1103106). We also tested other keywords, such as COVID or COVID-19, but these generated far fewer tweets (about 9k vs. 160k of "coronavirus") in the first two weeks and were later discarded. The API search was done in multiple sessions in order to retry on limit (usually 18k) and included only tweets in Italian language. No record is duplicated, as retweets were not retrieved a priori.

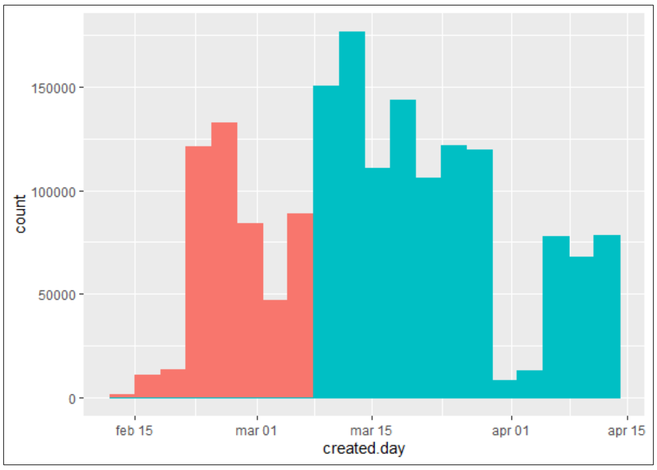

Fig. 1: Tweets distribution before (in red) and after (in blue) the Italian lockdown

A. Dataset

The resulting dataset consists in a corpus of N = 1673761 tweets and 89 other related metadata (i.e.: date, author, retweet count, favorite count, etc.) made available by Twitter. Despite the extent of this dataset, it is known [5] that this data is only a fraction of the Twitter database and the company does not precisely define the representativity of the sample and the methods of extraction. We didn't look at the representativeness issue because, although we are not able to provide a statistically reliable distribution of phenomena such as the polarity of messages, this dataset allows us to (a) analyze the discourses about the epidemic on this social medium and (b) to test empirically methods of simplifying this large amount of information. Moreover, some authors have shown that these limits can be overcome by improving the sample population coverage [30], e.g. by splitting the same keywords across multiple crawlers and, in wider terms, a social media analysis may capture population attitudes and behaviors, even if the characteristics of the users do not reflect the characteristics of the full population [7].

B. Pre-processing

The pre-processing phase involves the following steps:

- Removal of any duplicates identified by the “status_id” field provided by Twitter.
- Text cleaning, in order to normalize text encoding, html, emoticons, punctuation, etc. Therefore, all the URLs appearing in the corpus had been coded in a single word replacing all the links. If an emoticon had been recognized, it had been coded with a word that specify the kind of emoticon (e.g. ";)" had been recoded into "emote_wink").
- Stop words removing in Italian language through the R `TextWiller` package stop words list.

Looking at the most frequent words, apart from the negation "not" and the acronym COVID-19, the most frequent words were: *more*; *emergency*; *home*; *alone*; *cases*; *dead*; and *government*. In this phase, we worked within the bag-of-words framework, as only frequency weighting was applied to the words. Then, we analyzed the main multiword expressions by their frequencies as we tokenized adjacent words into ngrams. As a result of this procedure, we recoded the most frequent bigrams.

The initial corpus had 2124035 type and 25507563 tokens (type/token ratio: 8.3%), while after pre-processing and stop-words removing we had 470478 token and 17474428 type (type/token ratio: 2.7%). Therefore, the corpus appears to be sufficiently extensive for statistical analysis [35], even though the thresholds reported in literature are broad indications and can often be distorted, depending on the context or the extent of the corpus [15], [37].

C. Lockdown

By performing an analysis based on term frequencies, we meant to assess if the dataset can capture the differences in language before and during the lockdown. Effectively, we wanted to find which words are more or less likely to come from each period using the log odds ratio and then assess the result on a descriptive base. Therefore, we count how many times each word was used in the two periods and then we can calculate the log odds ratio for each word [33], using (1):

$$log\ odds\ ratio = \ln\left(\frac{\left[\frac{n+1}{total+1}\right]_{pre-lockdown}}{\left[\frac{n+1}{total+1}\right]_{post-lockdown}}\right) \qquad (1)$$

where *n* is the number of times a given word is used in each period and *total* indicates the total number of words in each period.

In summary, in the period before the lockdown (the words at the top with the red odds ratio in Fig. 2), Italian tweeters posted about:

- a student called Niccolò that was having trouble getting back from China before the Italian lockdown;
- the Diamond Princess ship stopped off the coast of Tokyo for almost a month, due to the spread of the coronavirus on board;
- the first case of a 38-year-old Italian contagious person;
- the idea of an excessive psychosis and alarmism;
- lack of disinfectants;
- the first red zones (Lodi and Venetian area).

Instead, in the second phase (the words at the bottom with the blue odds ratio in Fig. 2) during the lockdown, Italian tweeters posted about:

- an immunosuppressive drug used for the treatment of rheumatoid arthritis, potentially useful against COVID-19;
- some football players tested positive for the virus (Daniele Rugani, Paulo Dybala);
- some international politicians (Boris Johnson, Jair Bolsonaro);
- the self-certification paper to be able to move out of the house during lockdown;
- "coffins", a word that very likely refers to images of the army vehicles that transported the COVID-19 deceased in the Lombardy region.

This comparison answers to RQ1 suggesting that the dataset manages to capture the differences in language between before and throughout the lockdown. In fact, in a descriptive way, the words that best defined each of the two time periods highlighted the topics and the personalities featured in the Italian debate. An important example was the different approach to the virus: before the lockdown the tweeters talked about psychosis and exaggerated alarmism, while after the lockdown communication unfortunately moved on the deaths issue.

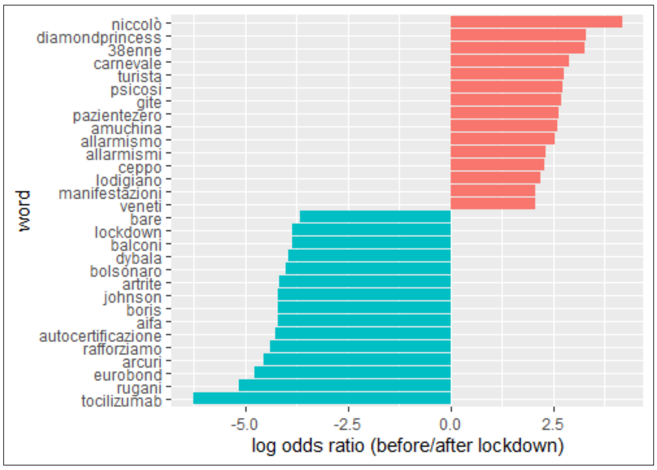

Fig. 2: Comparing the odds ratios of words before (in red) and after (in blue) lockdown

D. Sentiment Analysis

We analyzed the tweets' polarity automatically by using two different lexicons. In this case, we expected a low level of agreement because we used two generic lexicons. We analyzed the sentiment of a text by considering each tweet as a combination of its words and the sentiment content of the whole tweet as the sum of the sentiment content of the words. In order to compare ontological dictionaries that contain a different number of words, we applied the *Sign* function to the sum of the sentiment content of the words, obtaining a classification in three classes: -1 (negative); 0 (neutral); 1 (positive). So, the sentiment of the tweet *i* is calculated according (2):

$$Sentiment_i = sgn\left(\sum_{j=1}^{n_i}(polarity_{ij})\right) \qquad (2)$$

where n_i is the number of words in the tweet *i* and $polarity_{ij}$ is the polarity of each word, assuming values 1 or -1 according to the lexicon used (if a word was not classified as positive or negative, it was not considered). Among the few resources available for the Italian language, we chose the NRC [24] and the TextWiller [34] lexicons, through which we extracted a sentiment score for each tweet, as reported in Table 1.

To assess the level of overlap between the two sentiment distributions we used three measures: the percentage agreement; the Cohen's Kappa for two raters; the Kendall's coefficient of concordance W (with correction for ties). Table 2 summarizes the coefficients values, suggesting a limited level of agreement. In fact, only 48% of tweets had been classified in the same sentiment by the two lexicons and the Cohen's Kappa has a value quite far from the theoretical maximum (equal to 1), although significant. The level of agreement seems to improve, though remaining largely

TABLE I.
SENTIMENT DISTRIBUTION ACCORDING THE TWO LEXICONS

Lexicon	Negative	Neutral	Positive
NRC	41.1%	32.1%	26.8%
TextWiller	26.9%	42.2%	30.9%

TABLE II.
INTERRATER RELIABILITY AND AGREEMENT COEFFICIENTS

Measure	Value	p-value
Percentage agreement	48.3%	-
Cohen's Kappa	0.229	*p < 0.001*
Kendall's coefficient of concordance W	0.66	*p < 0.001*

unsatisfactory, by considering the distributions of sentiment as ordinary variables and therefore calculating Kendall's W coefficient. W coefficient turned out to be significant and equal to 2/3 of the theoretical maximum.

To answer RQ2, we found a low level of agreement between the two lexicons used, probably due to the nature of this specific dataset created during a disruptive event. Other Italian lexicons could show a good level of agreement with one of the two used here. However, we believe that to obtain realistic and reliable results, it would be necessary to use a supervised learning method with human tagging [6] by labelling a sample of tweets or, at least, to use a (COVID-19) context-dependent lexicon [14], [3].

E. *Word Embeddings*

As we pointed out, the analyses included in the previous sections refer to a bag-of-words (BoW) approach. Among the most notable disadvantages of BoW we find that the word order is ignored and BoW generally fails to capture the semantics of words [40]. BoW also suffers from data sparsity and high dimensionality from a computational point of view.

Word embeddings techniques based on neural networks are one of the most common solutions to these problems [40]. These techniques generate dense vectors for word representation, which are generally able to capture semantic relations between the corresponding words. Word embedding vectors could be used to identify analogies such as "Man is to woman as king is to queen", by examining the adjacency of words. This result is achieved by defining a context window, i.e. a string of words before and after a focal word that will be used to train a word embeddings model (each focal word and context words can be represented as a vector of real numbers, with the aim of detecting linguistic regularities and patterns). So, instead of using vectors of word counts, word embeddings can be seen as a particular type of text transformation into features, in which words are represented as coordinates on a latent multidimensional space derived from an underlying deep learning model that considers the contiguous words. One expected result is that words with similar weights should appear contiguous to the same words.

The two most commonly used examples of word embeddings are Word2Vec and GloVe. Word2Vec [22], [23] is a neural network prediction model containing continuous bag-of-words (CBoW) model [22] and Skip-gram (SG) model. The CBoW model predicts a target word from its context words, while the SG model predicts the context words given a target word [40]. Global Vectors (GloVe) [26] approach is very similar to the Word2Vec method, but it is performed on aggregated global word-word co-occurrence matrix derived from a corpus. Both methods have some disadvantages, as they fail to capture polysemy and out-of-vocabulary words from corpus. In comparison with Word2Vec, GloVe seems to show a better performance in identifying sub-linear relationships in the vector space and gives lower weight for highly frequent words, avoiding them to dominate the training process [16]. Word embeddings need a huge corpus of text data sets for training and, for example, GloVe provides pre-trained word vectorizations with different dimensions which are trained over big corpora, including Wikipedia and Twitter content. There are also some Italian word embeddings available [10], but we decided to use this COVID-19 dataset for training, since we are facing a turning point for its impact in human relations that is potentially displayed in linguistic perspective as well.

TABLE III.
EXAMPLES OF WORD SIMILARITY (GLOVE WORD EMBEDDINGS)

Input word	Word1 Cosine sim.	Word2 Cosine sim.	Word3 Cosine sim.	Word4 Cosine sim.	Word5 Cosine sim.
Emergency	Facing 0.740	Health 0.735	Crisis 0.713	Situation 0.654	Confronting 0.654
Doctors	Nurses 0.910	Health care workers 0.729	Frontline 0.665	Hospitals 0.603	Heroes 0.595
Government	Conte 0.746	Decree 0.677	Measures 0.667	Immediately 0.620	Politics 0.605
Lombardy	Veneto 0.749	Fontana 0.688	Region 0.672	Red_zones 0.640	Gallera 0. 624
China	Chinese 0.633	Wuhan 0.607	Epidemic 0. 591	World 0.579	Virus 0.571

First, we chose to limit the dimension of the vocabulary to avoid too uncommon words. So, we took only words which appear at least fifty times, obtaining a 26k terms vocabulary. We set our context window to have a length of 5 words (the number of words to be considered left and right) and set the dimension of the embedding vector to 100. The number of epochs we wanted to train was set to 10. For GloVe algorithm we constructed the term-co-occurrence matrix (TCM). Once we had the vectors for each word, we computed a similarity score among words to help us find regularities in the word vector space. A typical measure of similarity is the cosine similarity, defined by the Euclidean Dot product of two vectors normalized by their magnitude. Table 3 shows some examples of the most similar words to a specific input word, i.e. the closest terms on the latent multidimensional space. These examples seemed quite consistent to describe the main events (e.g.: Wuhan, Red Zones, etc.) and public figures (Prime Minister Conte, Lombardy Region's Governor Fontana, and Health Assessor Gallera) of the COVID-19 emergency. By reducing the matrix to the first two vectors, we can also visualize through a plot the positions of different actors such as heads of state and football players (Fig. 3), which are quite close according to their sphere of work.

F. *Supervised Learning*

Beyond this type of exploratory analysis, word embeddings could be very useful to build features at the word level for a supervised learning classifier. This strategy could be an alternative or a complement to a BoW approach: in order to maximize the performance, it is possible to combine both BoW (words with their term frequencies for each tweet) and embeddings features into a single matrix, and then use a classifier to identify the best set of features.

In this paper we want to compare the accuracy of a model based on terms frequency with a model based on word embeddings (together with some metadata collected with Twitter API, such as favorite and retweet count, in order to obtain a data mashed-up set of features). We decided to use Gradient Boosting classifier [13] (R implementation `xgboost`) for its flexibility and good performances [32]. The Gradient Boosting method results to be particularly suitable for models characterized by sparse features, since it produces

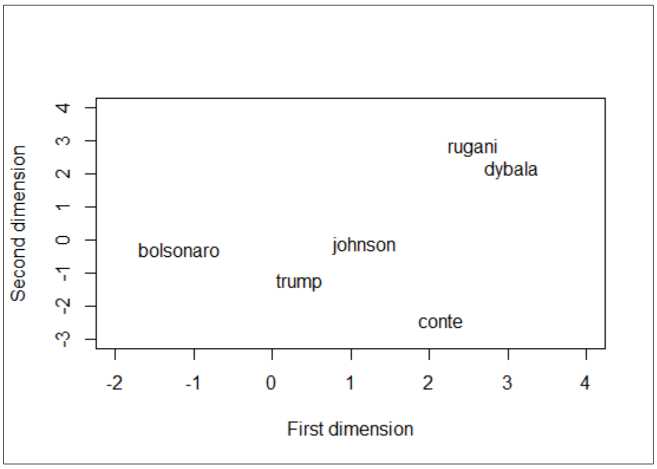

Fig. 3: Plot of the first two vectors related to public figures (Word2Vec)

a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. As stated in RQ3, we expected to find an improvement in estimates using word embeddings compared to term frequencies. After splitting our dataset into training (70%) and validation (30%) sets we tuned Gradient Boosting parameters by cross-validation (learning rate, depth of the tree, etc.). Since we created 100 vectors with both GloVe and Word2Vec, in this analysis we selected the 100 terms with the highest document frequency (above 1%).

We decided to run two experiments: the first one is to estimate if a tweet includes a quote of another tweet (variable "is_quote" available in the metadata downloaded from Twitter), while in the second experiment we built a variable that distinguishes tweets created before and during the lockdown. In response to RQ3, prediction models based on the two word embeddings techniques turned out to be equal or slightly better in terms of accuracy, but not compared to other measures (Tab. 4). Overall, the classification of tweets quoting another tweet shows very high levels of accuracy, while it is more problematic to classify tweets written before and during lockdown. This result could be due to the repetition of the most frequent terms in both phases (political figures, health-related words, etc.), while several of the distinctive terms that emerged in the previous analysis (Fig. 2) were not considered as features. The classification model of the variable "is_quote" revealed generally low levels of all the measures except for the accuracy since the tweets with this trait amount to less than 5% of the total and the classifier had problems in identifying them, while the true negative cases were identified with an accuracy of more than 95%.

TABLE IV.
CLASSIFICATION ACCURACY ON VALIDATION SET

Classified variable	Features	Accuracy	Precision	Recall	F_1 score
Is_quote	GloVe	0.963	0.563	0.018	0.034
	Word2Vec	0.963	0.604	0.018	0.035
	Terms	0.963	0.572	0.042	0.077
Post-lockdown	GloVe	0.763	0.785	0.881	0.830
	Word2Vec	0.765	0.798	0.881	0.838
	Terms	0.739	0.830	0.922	0.874

IV. Conclusions

In this paper we focused on Italian social media communication about COVID-19. The analysis of the conversations going on Twitter, through the odds ratios and the similarities of word embeddings, managed to capture events, topics, and personalities of the COVID-19 emergency. Instead, with respect to the used techniques, the polarity evaluation of the tweets will require further investigation and different methods, while a few experiments showed that the effect of word embeddings on the classification was similar to that of term frequencies.

Our first hypothesis was that we collected a large dataset containing texts able to capture the differences in language between, before and after Italian lockdown decree. Descriptive analysis suggested that the information provided by the data may reveal the terms used differently in various time periods, thereby capturing changes in online discussions as a result of government measures that have a major impact on people's lives. The second hypothesis involved a comparison of the results of a sentiment analysis using two lexicons for the Italian language. In this case, the analysis showed a certain discordance and the need for a reliable classification through human tagging, in order to provide an acceptable estimate of the polarity of the texts. Finally, word embeddings techniques showed at a descriptive level the ability to capture semantic and context similarities, while predictive models based on word embeddings failed to significantly improve the accuracy of the classification of some variables.

As future work, we plan to improve the accuracy of our classification by combining features form word embeddings and from term frequencies (or tf-idf weighting). Besides, as pointed out by other authors [14], [28], [40], the word embeddings approach can also become useful for the sentiment estimation. In the end, we plan to apply other supervised classification techniques, as well as to use this approach on other variables, since we are still collecting data as we write.

Acknowledgment

The authors would like to thank Anna Boaretto (University of Padova) for her support and suggestions.

References

[1] V. Agarwal, and B.K. Sunitha, “COVID –19: Current Pandemic and Its Societal Impact”, *Int. J. of Advanced Science and Technology*, Vol. 29, No. 5s, pp. 432-439, 2020.

[2] M. Ahmed, O. Ahmed, Z. Aibao, S. Hanbin, L. Siyu and A. Ahmad, "Epidemic of COVID-19 in China and associated Psychological Problems", *Asian J.l of Psychiatry*, vol. 51, p. 102092, 2020. Available: 10.1016/j.ajp.2020.102092.

[3] M. Asif, A. Ishtiaq, H. Ahmad, H. Aljuaid and J. Shah, “Sentiment analysis of extremism in social media from textual information”, *Telematics and Informatics*, vol. 48, p. 101345, 2020. Available: 10.1016/j.tele.2020.101345.

[4] G. Barkur, Vibha and G. Kamath, “Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India”, *Asian J. of Psychiatry*, vol. 51, p. 102089, 2020. Available: 10.1016/j.ajp.2020.102089.

[5] D. Boyd and K. Crawford, “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon”, *Information, communication & society*, *15*(5), pp. 662-679, 2012. Available:10.1080/1369118X.2012.678878.

[6] A. Ceron, L. Curini and S. Iacus, “iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content”, *Information Sciences*, vol. 367-368, pp. 105-124, 2016. Available: 10.1016/j.ins.2016.05.052.

[7] A. Ceron, L. Curini and S. Iacus, Politics and Big Data: Nowcasting and Forecasting Elections with Social Media. New York: Routledge, 2017.

[8] E. Chen, K. Lerman, and E. Ferrara, “Covid-19: The first public coronavirus twitter dataset”, *arXiv preprint arXiv:2003.07372*, 2020.

[9] N. Chen et al., "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study", *The Lancet*, vol. 395, no. 10223, pp. 507-513, 2020. Available: 10.1016/s0140-6736(20)30211-7.

[10] A. Cimino, L. De Mattei, F. Dell’Orletta, Multi-task Learning in Deep Neural Networks at EVALITA 2018, in: Proc. of EVALITA ’18, Evaluation of NLP and Speech Tools for Italian, 2018.

[11] M. Del Vicario et al., “The spreading of misinformation online”, *Proc. of the National Academy of Sciences*, vol. 113, no. 3, pp. 554-559, 2016. Available: 10.1073/pnas.1517441113.

[12] M. Felt, “Social media and the social sciences: How researchers employ Big Data analytics”, *Big Data & Society*, vol. 3, no. 1, p. 205395171664582, 2016. Available: 10.1177/2053951716645828.

[13] J. Friedman, “Stochastic gradient boosting”, *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367-378, 2002. Available: 10.1016/s0167-9473(01)00065-2.

[14] M. Huang, H. Xie, Y. Rao, J. Feng and F. Wang, “Sentiment strength detection with a context-dependent lexicon-based convolutional neural network”, *Information Sciences*, vol. 520, pp. 389-399, 2020. Available: 10.1016/j.ins.2020.02.026.

[15] D. Holmes, “The Analysis of Literary Style--A Review”, *J. of the Royal Statistical Society. Series A (General)*, vol. 148, no. 4, p. 328, 1985. Available: 10.2307/2981893.

[16] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes and D. Brown, “Text Classification Algorithms: A Survey”, *Information*, vol. 10, no. 4, p. 150, 2019. Available: 10.3390/info10040150.

[17] D. Lazer et al., “The science of fake news”, *Science*, vol. 359, no. 6380, pp. 1094-1096, 2018. Available: 10.1126/science.aao2998.

[18] L. Li et al., “Characterizing the Propagation of Situational Information in Social Media During COVID-19 Epidemic: A Case Study on Weibo”, *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 556-562, 2020. Available: 10.1109/tcss.2020.2980007.

[19] B. Liu, Sentiment analysis and opinion mining. San Rafael, Calif.: Morgan & Claypool, 2012.

[20] T. McCormick, H. Lee, N. Cesare, A. Shojaie and E. Spiro, “Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing”, *Sociological Methods & Res.*, vol. 46, no. 3, pp. 390-421, 2015. Available: 10.1177/0049124115605339.

[21] R. Merchant and N. Lurie, “Social Media and Emergency Preparedness in Response to Novel Coronavirus”, *JAMA*, 2020. Available: 10.1001/jama.2020.4469

[22] T. Mikolov , K. Chen , G.S. Corrado , J. Dean , Efficient estimation of word representations in vector space, in: Proc. of International Conf. on Learning Representations, 2013.

[23] T. Mikolov , I. Sutskever , K. Chen , G.S. Corrado , J. Dean , Distributed representations of words and phrases and their compositionality, in: Proc. of the Annual Conf. on Neural Information Processing Systems, 2013, pp. 3111–3119 .

[24] S. Mohammad and P. Turney, “Crowdsourcing a word-emotion association lexicon”, *Computational Intelligence*, vol. 29, no. 3, pp. 436-465, 2012. Available: 10.1111/j.1467-8640.2012.00460.x.

[25] S. Oh, S. Lee and C. Han, “The Effects of Social Media Use on Preventive Behaviors during Infectious Disease Outbreaks: The Mediating Role of Self-relevant Emotions and Public Risk Perception”, *Health Communication*, pp. 1-10, 2020. Available: 10.1080/10410236.2020.1724639.

[26] J. Pennington , R. Socher , C.D. Manning , Glove: global vectors for word representation, in: Proc. of the Conf. on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543 .

[27] C. Pulido, B. Villarejo-Carballido, G. Redondo-Sama and A. Gómez, “COVID-19 infodemic: More retweets for science-based information on coronavirus than for false information”, *International Sociology*, 2020. Available: 10.1177/0268580920914755.

[28] S. Rezaeinia, R. Rahmani, A. Ghodsi and H. Veisi, “Sentiment analysis based on improved pre-trained word embeddings”, *Expert Systems with Applications*, vol. 117, pp. 139-147, 2019. Available: 10.1016/j.eswa.2018.08.044.

[29] J. Rout, K. Choo, A. Dash, S. Bakshi, S. Jena and K. Williams, “A model for sentiment and emotion analysis of unstructured social media text”, *Electronic Commerce Res.*, vol. 18, no. 1, pp. 181-199, 2017. Available: 10.1007/s10660-017-9257-8.

[30] J. Sampson, F. Morstatter, R. Maciejewski and H. Liu, Surpassing the limit: Keyword clustering to improve twitter sample coverage, in: Proc. of the 26th ACM Conf. on hypertext & social media, 2015, pp. 237-245.

[31] A. Sapountzi and K. Psannis, “Social networking data analysis tools & challenges”, *Future Generation Computer Systems*, vol. 86, pp. 893-913, 2018. Available: 10.1016/j.future.2016.10.019.

[32] A. Sciandra, A. Surian and L. Finos, “Classifying the Willingness to Act in Social Media Data: Supervised Machine Learning for UN 2030 Agenda”., in: SIS 2019-Smart Statistics for Smart Applications pp. 509-516, Pearson, 2019.

[33] J. Silge and D. Robinson, Text mining with R. Sebastopol, CA: O'Reilly Media, 2017.

[34] D. Solari, A. Sciandra and L. Finos, “TextWiller: Collection of functions for text mining, specially devoted to the Italian language”, *J. of Open Source Software*, vol. 4, no. 41, p. 1256, 2019. Available: 10.21105/joss.01256.

[35] A. Tuzzi, “Analisi statistica del contenuto”, in Percorsi di ricerca sociale, L. Bernardi, Ed. Roma: Carocci, 2005, pp. 237-254.

[36] Twitter, Search tweets, standard search API, developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets (Accessed 8 May 2020).

[37] A. Vermeer, “Coming to Grips with Lexical Richness in Spontaneous Speech Data”, *Language Testing*, vol. 17, no. 1, Jan. 2000, pp. 65–83, Available:10.1177/026553220001700103.

[38] J. Wen, J. Aston, X. Liu and T. Ying, “Effects of misleading media coverage on public health crisis: a case of the 2019 novel coronavirus outbreak in China”, *Anatolia*, vol. 31, no. 2, pp. 331-336, 2020. Available: 10.1080/13032917.2020.1730621.

[39] WHO, WHO Director-General’s Opening Remarks at the Media Briefing on COVID-19, 11 March 2020. Available at: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020 (Accessed 8 May 2020).

[40] L. Zhang, S. Wang and B. Liu, “Deep learning for sentiment analysis: A survey”, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, 2018. Available: 10.1002/widm.1253.

[41] N. Zhu et al. “A novel coronavirus from patients with pneumonia in China, 2019”. *New England J. of Med.*, vol. 382, no. 8, pp. 727-733, 2020. Available: 10.1056/NEJMoa2001017.