ConteCorpus: An Analysis of People Response to Institutional Communications During the Pandemic

Viviana Ventura, Elisabetta Jezek
Department of Humanities
University of Pavia, Pavia, Italy
viviana.ventura01@universitadipavia.it, jezek@unipv.it

Abstract

The study of institutional communication related to the pandemic, and to the population's response to it, is of great relevance today. The Italian spokesperson for communication regarding the pandemic has been, during the year 2020, the former Prime Minister Giuseppe Conte. We retrieved 4,860,395 comments from his Facebook official page and built the ConteCorpus, a new Italian resource annotated in CoNLL-U format. A first aim of the research was to evaluate the performance of the model used to annotate the corpus. Models trained on social media texts are usually not very generalizable. Nevertheless, the results of the evaluation were good, especially in parsing metrics, and showed that a parser trained on Twitter data can be successfully applied to Facebook data. A second aim of the research was to provide an overall view of the content of such a large corpus; for this purpose, topic modeling was conducted, training an LDA model. The model generated 5 topics that cover different aspects linked to the pandemic emergency, from economic to political issues. Through the topic modeling we investigated which topics are prevalent on particular days.

1 Introduction

During the year 2020, the Prime Minister Giuseppe Conte has played a major role in institutional communication, particularly in communication regarding the policies undertaken to manage the health emergency. We assumed that interesting content from the point of view of the response of the population to institutional communications regarding the pandemic would have been found on his social media profiles. Therefore, we created ConteCorpus,[1] retrieving more than 4 million comments from his Facebook page[2] starting from January 2020 until December 2020, and we annotated it in CoNLL-U format[3].

A first aim of the research was to evaluate the performance of the model used to annotate the dataset. Models trained on social media texts usually are poorly generalizable even on text retrieved from the same social media, therefore we wanted to test the performance on Facebook texts of a model trained on Twitter texts. In order to evaluate the model, we created a gold standard by extracting 1,000 sentences from the ConteCorpus and manually revising them.

A second aim of the research was to provide an overall view of this large corpus. For this purpose we performed a Topic Modeling. We trained a LDA model sampling 10% of the ConteCorpus. The LDA model generated 5 topics related to different aspects of the pandemic emergency. The model was used to see which topics were the most relevant before and after the announcement of the first and the second period of restrictions adopted to fight the pandemic in Italy.

The paper is structured as follows: we first review the relevant literature for our research (section 2), then we describe the data collection and the creation of the corpus (section 3). In section 4, we describe the evaluation we performed of the model we used to annotate the corpus in CoNLL-U format, and in section 5 we report the results of the topic modeling experiment. In section 6 we provide some concluding observations.

[1] https://github.com/Viviana-dev/Conte_Corpus
[2] https://www.facebook.com/GiuseppeConte64/
[3] https://universaldependencies.org/format.html

	January	February	March	April	May	June	July	August	September	October	November	December	Tot
Post	48	59	48	45	26	44	61	24	43	75	33	28	534
Comment	115,971	154,266	681,221	775,972	361,179	335,772	449,913	190,777	260,237	666,126	441,822	427,139	4,860,395

Table 1. Number of posts and comments retrieved for each month.

2 State of the Art

Since the beginning of the health emergency, there has been a proliferation of computational analyses that exploit data extracted from social media. These data are considered relevant as they allow us to generalize about human social and linguistic behavior, especially regarding the pandemic event. Among the tasks that have been conducted on data drawn from social media in this period, sentiment analysis, emotion profiling and topic modeling are the most common (Gagliardi et al., 2020; Tamburini, 2020; Vitale et al., 2020; Stella et al., 2020a; Stella et al., 2020b; Stella et al., 2021; De Santis et al., 2020; Sciandra, 2020; Trevisan et al., 2021; Gozzi et al., 2020; Kruspe et al., 2020; Hussain et al., 2021; Chakraborty et al., 2020; Nemes e Kiss, 2020; Jelodar et al., 2021; Lamsal, 2020; Duong et al., 2021; Gupta et al., 2021; Sullivan et al., 2021; Su et al., 2020; Garcia et Berton, 2021; Ahmed et al., 2020).

In particular, Topic Modeling aims at finding hidden semantic structures within the texts and to model them into concepts. The unsupervised clustering technique LDA (Latent Dirichlet Allocation), developed by Blei (2003), has been used extensively in analyses conducted on social media data during the pandemic (Dashtian et Murthy, 2021; Feng et Zhou, 2020; Ordun et al., 2020; Wang et al., 2020; Kabir et Mandria, 2020; Amara et al., 2020; Abd-Alzaraq et al., 2020; Naseem et al., 2021; Low et al. 2020, Andreadis et al., 2021). LDA is a statistical model that represents each document in a corpus as a probabilistic distribution over latent topics and each topic as a probabilistic distribution over words. A topic has a probability of generating various words, where the words are all the observed words in the corpus. Thus, the terms in the set of documents are used to discover hidden topics in a large corpus.

As is well known, the language of the web is characterized by deviation from the standard language that challenges the use of NLP tools. Several classifications have been proposed to label the nature of web and social media language. In general, the labels aim to define a variety of language that is diaphasically low and at an indefinite point on the diamesic axis, e.g., “netspeak” (Crystal, 2001). Web and social media language is characterized by little planning in text structure and a greater propensity for parataxis, absence of revision and punctuation, abrupt interruption of periods, and an imitation of the continuous flow of speech (Fiorentino, 2013). Although some persistent traits of web and social media language can be described, it does not constitute a single variety of language from a sociolinguistic perspective (Fiorentino, 2013). This poses a double challenge in the use of NLP tools. First, because the tools are calibrated to standard language variety resources. Secondly, even if we created models that are better suited to web and social media languages, they would not be generalizable to every language variety on the web (Sanguinetti et al., 2018).

3 ConteCorpus Construction

3.1 Data Collection

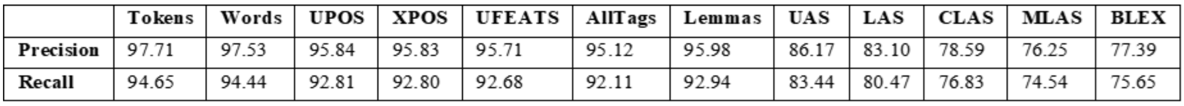
We have downloaded 4,860,395 comments and 534 posts published during the year 2020 on Giuseppe Conte’s Facebook official profile. We made call to any 2020 post ID of Giuseppe Conte’s official page to retrieve text, object id, and created time of comments. The calls to the Facebook API Graph[4] were made month to month in the same fashion. Nevertheless, as Table 1 shows, a larger amount of data has been retrieved in the month of March, April, and October. In the same period in Italy the more restrictive measures to fight pandemic were taken by the government.

3.2 Processing with the Neural Pipeline Stanza

After the data collection, we processed the data with the Neural Pipeline Stanza[5] to enrich the texts with some annotations. Stanza is an open-source Python NLP toolkit, which “features a language-agnostic fully neural pipeline for text analysis, including tokenization, multiword token expansion, lemmatization, part-of-speech and morphological feature tagging, dependency parsing, and named entity” (Qi et al., 2020). The kit supports more than 77 human languages and uses the

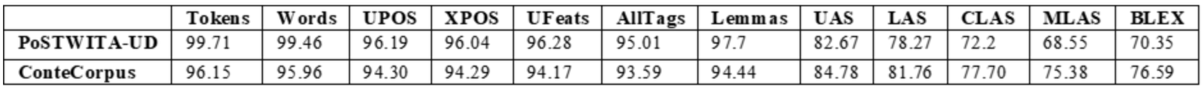
[4] https://developers.facebook.com/docs/graph-api?locale=it_IT

[5] https://stanfordnlp.github.io/stanza/



	Tokens	Words	UPOS	XPOS	UFEATS	AllTags	Lemmas	UAS	LAS	CLAS	MLAS	BLEX
Precision	97.71	97.53	95.84	95.83	95.71	95.12	95.98	86.17	83.10	78.59	76.25	77.39
Recall	94.65	94.44	92.81	92.80	92.68	92.11	92.94	83.44	80.47	76.83	74.54	75.65

Table 2. Performance of Stanza's UD pre-trained model tested on the test set of ConteCorpus.



	Tokens	Words	UPOS	XPOS	UFeats	AllTags	Lemmas	UAS	LAS	CLAS	MLAS	BLEX
PoSTWITA-UD	99.71	99.46	96.19	96.04	96.28	95.01	97.7	82.67	78.27	72.2	68.55	70.35
ConteCorpus	96.15	95.96	94.30	94.29	94.17	93.59	94.44	84.78	81.76	77.70	75.38	76.59

Table 3. Performance of Stanza's UD pre-trained model tested on official test set of PoSTWITA-UD and on test set of ConteCorpus. The scores shown are calculated using the F-measure.

formalism Universal Dependencies[6] Knowing the difficulties of annotating non standard texts such as those derived from social media, we chose to use this pipeline because the evaluation of its models found that Stanza neural language agnostic architecture "adapts well to text of different genres […] achieving state-of-the-art or competitive performance at each step of the pipeline" (Qi et al., 2020). Moreover, models that can be downloaded from Stanza have been trained each on a single language and on a specific text genre dataset. We chose to download the model trained on PoSTWITA-UD.[7] PoSTWITA-UD is an Italian Twitter treebank in Universal Dependencies (Sanguinetti et al., 2018). Although the language of social media is very peculiar and changes from one social media to another and from groups to groups (Fiorentino, 2013), we thought that the model downloadable from Stanza - trained on this dataset - could be generalizable to our data, being in-domain. Moreover, Sanguinetti et al. (2018) have added customized tags to the UD scheme to deal with some social media peculiar phenomena: "discourse:emo" for emojis and emoticons, and "parataxis:hashtag" for hashtags. They tagged the link found in some sentences as "dep" (unspecified relation) and used the "upos" (universal part-of-speech) tag "SYM" (symbol) for hashtags and emojis. Additionally, they manually inserted the lemma of non-standard word forms not recognized by the lemmatizer (Sanguinetti et al., 2018).

We processed the data divided in 12 packages; each correspond to one month data. We used every processor of the pipeline, besides the Named Entity Recognition module (TokenizeProcessor, POSProcessor, LemmaProcessor, DepparseProcessor). We personalized the model in order not to split the sentences,[8] forcing the TokenizeProcessor to consider each comment as a sentence. Furthermore, we added two metadata to each sentence: one refers to the id of the post from which the comment was retrieved, and the other is the creation time of the comment. The aim is to make it easier to retrieve the comments from the corpus by their created time or post id if one needs to analyze a particular period of time or a particular post.

4 End-to-End Evaluation

4.1 Construction of the Gold Standard

We built a gold standard with a dual purpose: to evaluate the performance of the model on this new collection of social media texts, and to create a standard that can be used for future training and testing. We randomly selected 83 sentences from each file of the corpus annotated automatically (one file is composed of one-month comments), and manually revised the 1,000 sentences collected. The manual revision has followed the principle that what is understandable by a human would be correct.

4.2 Evaluation with CoNLL 2018 UD Shared Task Official Evaluation Script

To perform the evaluation, we used CoNLL 2018 UD shared task official evaluation script.[9] Table 2 shows the scores of evaluation metrics resulting from the performance of Stanza model on the test set of the ConteCorpus. Table 3 compares the scores of evaluation metrics resulting from the performance of Stanza model on the test set of PoSTWITA-UD and the ConteCorpus. The first two columns are the scores on metrics that evaluate segmentation. The row called UPOS shows the

[6] Universal Dependencies (UD) is a "framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages" (https://universaldependencies.org/).

[7] https://universaldependencies.org/treebanks/it_postwita/index.html

[8] Sentence segmentation and tokenization are jointly performed by the TokenizeProcessor (Qi et al., 2020).

[9] https://universaldependencies.org/conll18/evaluation.html

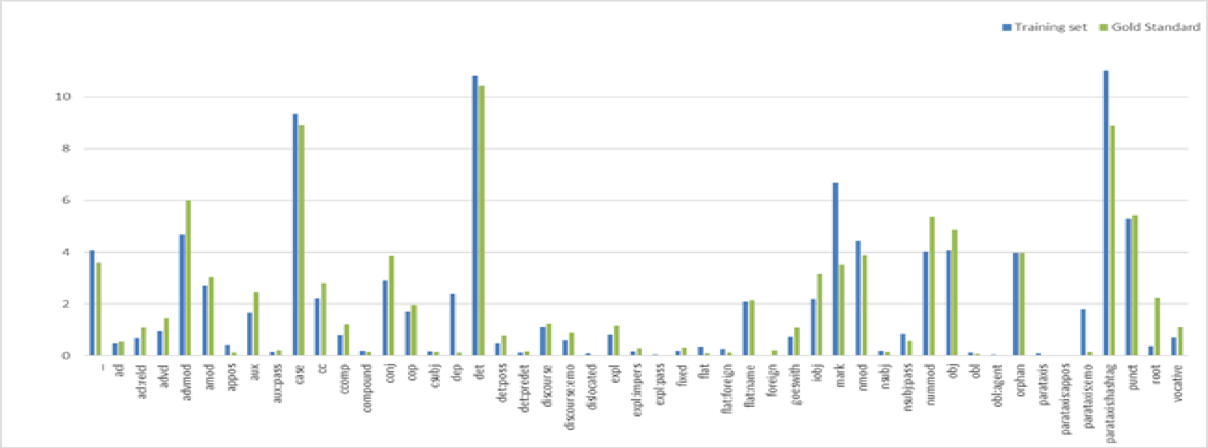


Figure 1. Frequency distribution of syntactic relation tags in the training set and the gold standard.

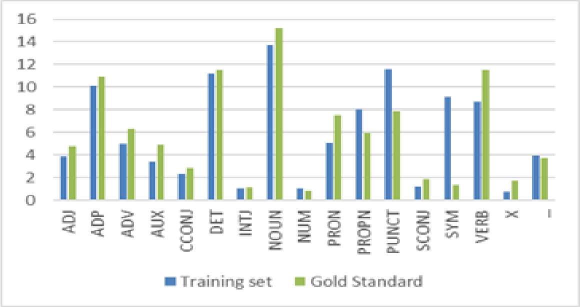


Figure 2. Frequency distribution of part-of-speech tags in the training set and the gold standard.

resulting scores on Universal part-of-speech tagging metric, XPOS on language-specific part-of-speech tagging metric, and UFeats on morphological features tagging metric. The last 5 rows show scores in five different parsing metrics.

What we found most challenging during the manual revision of the 1,000 sentences annotated automatically was correcting the errors in tokenization: many words that the tokenizer should have splitted were joined together. This type of tokenization error is often found when punctuation is used with non standard function. For example: we found that the token "oneste…volevo" ("honest…I wanted to") - an adjective, a punctuation mark and a verb - are conflated in a single token. In the manual revision, tokens like this have been splitted in three different tokens and other missing tags were added. The presence of such conflated words mayhave caused a worse score in the metric that evaluates the performance of segmentation, and consequentially in the other scores. The evaluation on the parser starts with aligning system nodes and gold nodes; their respective parent nodes are also considered; if the system parent is not aligned with the gold parent or if the relation label differs, the word is not counted as correctly attached. Despite errors in segmentation seem frequent in the corpus, this did not cause an excessive lowering of the scores on the various metrics reported in Table 2 and 3. Another error that appears frequently regards the lemma assigned to the abbreviations that are not present in PoSTTWITA-UD. Canonical abbreviations are tagged correctly, for example "cmq" for "comunque" ("however"). The abbreviations tagged incorrectly are those which appeared few times: such as "ql" that stands for "quelli" (those). An unexpected good result has been achieved on parsing metrics. This result could be due to the "preference of UD scheme in assigning headedness to content words" (Sanguinetti et al., 2018); therefore, the tendency of the social media languages to eliminate function words does not affect the performance of the parser. Another explanation can be found in the very similar frequencies distribution of part-of-speeches and syntactic relations in the training set and the gold standard, as shown in Figure 1 and 2.

Overall, the model trained on PoSTWITA-UD turned out to perform well on the test set of the ConteCorpus because PoSTWITA-UD tagset has been adapted with attention to some recurrent features of social media languages. Our evaluation showed that a model trained on texts retrieved by social media can adapt well to other social media texts if one pays attention to the neural architecture of the model and the annotation format being used.

5 Topic Modeling

To provide an overall view of the content of this large corpus we performed a Topic Modeling training and testing an LDA model on the ConteCorpus.

5.1 Methodology

Topic	1: Economics	2: Prime Minister	3: Politics	4: Pandemic	5: Home
Terms	pagare, soldo, italia, euro, chiudere, mese, debito, azienda, prestito, lavorare	presidente, grazie, Conte, lavoro, bravo, italia, italiano, signore, giuseppe, caro	italiano, europa, italia, paese, banca, popolo, governo, chiedere, germania, storia	uscire, miliardo, firmare, virus, decreto, Salvini, maria, pandemia, chiedere, italy	sperare, casa, aspettare, perdere, impresa, tedesco, subito, tempo, fondo, stipendio
English Translation	to pay, money, italy, euro, to close, month, loan, company, to work	prime minister, thank you, Conte, work, bravo, italy, italian, sir, giuseppe, dear	italian, europe, italy, country, bank, people, government, to ask, germany, story	to go out, billion, to sign, virus, decree, Salvini, maria, pandemic, to ask, italy	to hope, home, to wait, to lose, business, german, immediately, time, capital, salary

Table 4. Topic generated from the LDA model and the ten most frequent terms.

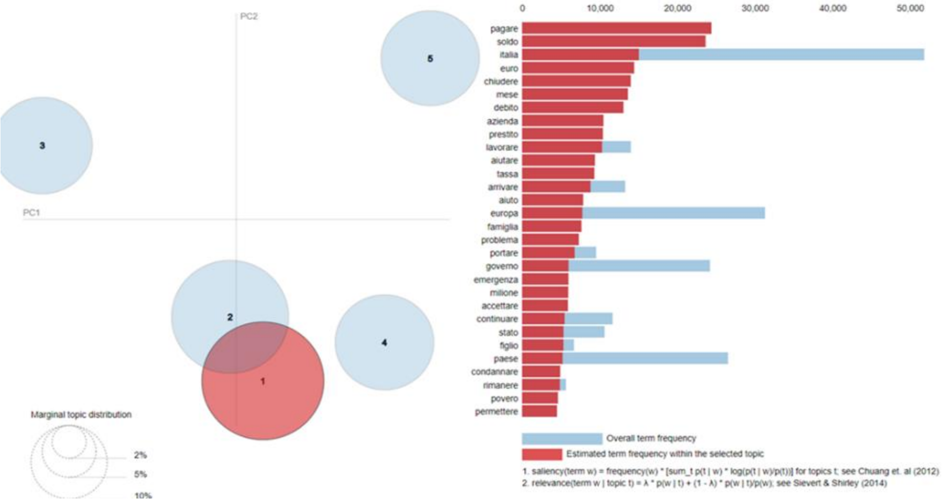


Figure 3. Intertopic distance Map and Top-30 most relevant terms for Topic 1. For a better view visit: https://sites.google.com/view/ldavisualizationcontecorpus/home-page.

To perform topic modelling, we sampled 10% of the sentences in our dataset and trained a LDA model. We treated each sentence as a document. We pre-processed lemmas removing stopwords, downloading Italian stopwords list from the NLTK (Natural Language Toolkit) library[10] and manually inserting missing stopwords. We filtered out tokens that appear in less than 15 documents and tokens with less than three letters; additionally, we kept only the 100,000 most frequent words. We transformed the documents into vectors creating a bag-of-words representation of each document. Then, we performed the term frequency-inverse document frequency (TF-IDF) on the whole corpus to assign higher weights to the most important words. Gensim LDA model[11] was applied first to the bags-of-words and secondly on the TF-IDF corpus to extract latent topics. Better performances were achieved with the LDA model applied to bags-of-words. We determined the optimal number of topics in LDA using the Coherence Value metric.[12] The underlying idea is that a good model will generate topics with high topic Coherence Value score. We ran different LDA experiments varying the number of topics and selected the model with the highest medium topic Coherence Value score. Our final model generated 5 topics and has a topic medium Coherence Value score of 0.5. Table 4 illustrates the top ten most representative terms associated with each detected topic.

5.2 Results

As expected, all the topics extracted from the corpus are related to the concerns about the emergency. The focus is on the economic aspect of the emergency. The first ten most frequent words in *Economics* topic (Table 4 and Figure 3) are economic terms: “loan”, “company”, “to pay” “money” etc. In all the other topics at least one of the 10 most frequent words comes from the economic sphere. Among the ten most frequent words of each topic there are only two words regarding the pandemic, found in *Pandemic* topic: "virus" and "pandemic". It is no coincidence that the most frequent word in this topic is “to go out”. The need to face the emergency through the intervention of the institutions is evident. This is shown espe

[10] https://www.nltk.org/.

[11] https://radimrehurek.com/gensim/models/ldamodel.html.

[12] Coherence Value metric is developed by Roder (2015). It evaluates a single topic by measuring the degree of semantic similarity between high scoring words in the topic.

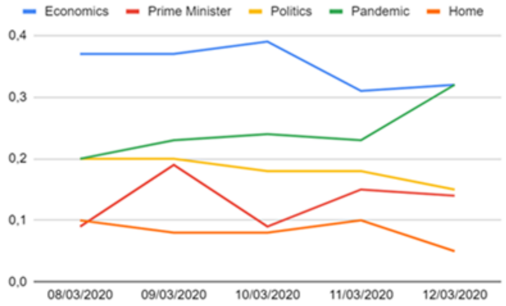


Figure 4. Prevalence of topics during the days 8-12 March 2020.

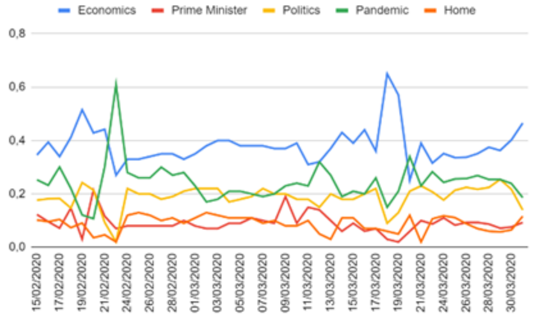


Figure 5. Prevalence of topics during the days 15 February-30 March 2020.

cially by *Prime Minister* and *Politics* topics (Table 4). *Prime Minister* topic most frequent words are related to the Prime Minister. Perhaps words like "bravo" and "thank you" and "dear" show a positive judgement towards him. In *Politics* topic one finds words of the institutional sphere such as: "country", "government", "people", "bank". *Home* topic is related to the private sphere with words like "to hope", "home", "to wait", "to lose", although there is no shortage of words from the economic sphere. In Figure 3 the distance between the centre of the circles indicates the similarity between the topics. Here you can see that only *Economics* topic and *Prime Minister* topic overlap; this indicates that the two topics are more similar with respect to the other topics. Moreover, the size of the area of each circle represents the importance of the topic relative to the corpus. *Economics* topic is the most important topic in the corpus. Finally, we tested our model on unseen documents: the comments published between 15 February and 30 March 2020, before and after the announcement of the first period of restrictions to combat the pandemic, and between 1 October and 14 November 2020, before and after the announcement of the second period of restrictions. Figures 4, 5 and 6 show trends in topics over time. Each line represents a topic and the x-axis shows the time progression. On 23 February, the first restrictive policies were announced for some Italian cities: Figure 5 shows a peak in the pandemic topic on that day. Figure 4 shows how the prevalence of the five topics changes on 8-12 March 2020. The Figure shows a peak on 9 March in *Prime Minister* topic: on that day he announced the first national restrictions period to combat the pandemic. Overall, the prevalent topics on those days are economics and pandemic. On 13 October, after a summer without major restrictions, with a new exponential increase in the curve of contagions, the Italian Parliament passed a decree limiting the possibility of aggregation. That day we have a new peak in the *Pandemic* theme (Figure 6). In the days that followed, the prevailing topic is *Economics*: on 28 October, the "ristoro" decree was approved to financially support commercial activities. A peak in the topic of *Economics* occurred on 18 March: on those days, discussions were taking place on whether to ask the European Union for financial aid to overcome the pandemic. The prevailing topics are therefore usually related to current events.

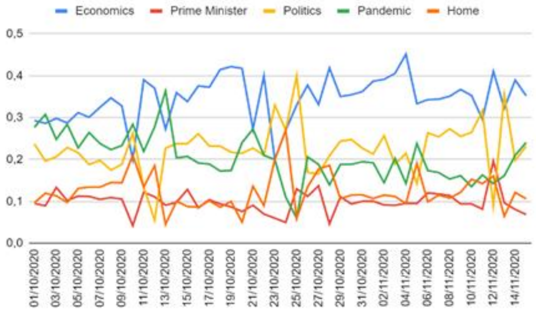


Figure 6. Prevalence of topics during the days 1 October-15 November 2020.

6 Concluding Observations

As mentioned before, models trained with data from social media are hardly generalizable. This stems from the fact that from a sociolinguistic perspective, the language of social media does not constitute a single variety. So, we expected that the results in the various evaluation metrics we performed would be worse than the results in the evaluation conducted on the PoSTWITA-UD test set. Surprisingly, in some metrics the results on evaluating the ConteCorpus test set were better than the results on the PoSTWITA-UD test set. To offer an overall view of the content of the ConteCorpus we performed topic modeling. The topics generated by the LDA model cover various aspects of the pandemic emergency, with a preponderance of political and economic issues. Unexpectedly, topics identified do not show concern regard the risk of contagion and the possibility of catching the disease.

References

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi M., & Shah, Z. (2020). Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. Journal of medical Internet research, 22(4), e19016.

Ahmed, M. E., Rabin, M. R. I., & Chowdhury, F. N. (2020). COVID-19: Social media sentiment analysis on reopening. arXiv preprint arXiv:2006.00804.

Amara, A., Taieb, M. A. H., & Aouicha, M. B. (2021). Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis. Applied Intelligence, 51(5), 3052-3073.

Andreadis, S., Antzoulatos, G., Mavropoulos, T., Giannakeris, P., Tzionis, G., Pantelidis, N., ... & Kompatsiaris, I. (2021). A social media analytics platform visualising the spread of COVID-19 in Italy via exploitation of automatically geotagged tweets. Online Social Networks and Media, 23, 100134.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003) Latent dirichlet allocation, the Journal of machine Learning research (JMach), 3, 993–1022.

Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., & Hassanien, A. E. (2020). Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. Applied Soft Computing, 97, 106754.

Crystal, D. (2001). Language and the Internet. Cambridge University Press.

Dashtian, H. and Murthy, D. (2021). Cml-covid: A large-scale covid-19 twitter dataset with latent topics, sentiment and location information. arXiv preprint arXiv:2101.12202.

De Santis, E., Martino, A., & Rizzi, A. (2020). An Infoveillance System for Detecting and Tracking Relevant Topics from Italian Tweets During the COVID-19 Event. IEEE Access, 8, 132527-132538.

Fiorentino, G. (2013). "Wild language" goes Web: new writers and old problems in the elaboration of the written code. In E. Miola (Ed.), Languages Go Web. Standard and non-standard languages on the Internet (pp. 67-90.). Alessandria, Edizioni dell'Orso.

Dozat, T. and Manning, C. D. (2017). Deep biaffine attention for neural dependency parsing. In Proceedings of the 2017 International Conference on Learning Representations (ICLR).

Duong, V., Luo, J., Pham, P., Yang, T., & Wang, Y. (2020). The ivory tower lost: How college students respond differently than the general public to the covid-19 pandemic. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 126-130).

Feng, Y. and Zhou, W. (2020). Is working from home the new norm? an observational study based on a large geo-tagged covid-19 twitter dataset. arXiv preprint arXiv:2006.08581.

Gagliardi, G., Gregori, L., & Suozzi, A. (2021). L'impatto emotivo della comunicazione istituzionale durante la pandemia di Covid-19: uno studio di Twitter Sentiment Analysis. Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy. Volume 2769 of CEUR Workshop Proceedings.

Garcia, K. and Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. Applied Soft Computing, 101, 107057.

Gozzi, N., Tizzani, M., Starnini, M., Ciulla, F., Paolotti, D., Panisson, A., & Perra, N. (2020). Collective Response to Media Coverage of the COVID-19 Pandemic on Reddit and Wikipedia: Mixed-Methods Analysis. Journal of medical Internet research, 22(10), e21597.

Gupta, V., Jain, N., Katariya, P., Kumar, A., Mohan, S., Ahmadian, A., & Ferrara, M. (2021). An emotion care model using multimodal textual analysis on COVID-19. Chaos, Solitons & Fractals, 144, 110708.

Hussain, A., Tahir, A., Hussain, Z., Sheikh, Z., Gogate, M., Dashtipour, K., et al. (2021). Artificial Intelligence–Enabled Analysis of Public Attitudes on Facebook and Twitter Toward COVID-19 Vaccines in the United Kingdom and the United States: Observational Study. Journal of medical Internet research, 23(4), e26627.

Jelodar, H., Wang, Y., Orji, R., & Huang, S. (2020). Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. IEEE Journal of Biomedical and Health Informatics, 24(10), 2733-2742.

Kruspe, A., Häberle, M., Kuhn, I., & Zhu, X. X. (2020). Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic. arXiv preprint arXiv:2008.12172.

Lamsal, R. (2020). Design and analysis of a large-scale COVID-19 tweets dataset. Applied Intelligence, 1-15.

Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. The Information Society, 30(4), 256-265.

Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020) Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. Journal of medical Internet research, 22(10), e22635.

Naseem, U., Razzak, I., Khushi, M., Eklund, P. W., & Kim, J. (2021). COVIDSenti: a large-scale benchmark Twitter data set for COVID-19 sentiment analysis. IEEE transactions on computational social systems.

Nemes, L. and Kiss, A. (2021). Social media sentiment analysis based on COVID-19. Journal of Information and Telecommunication, 5(1), 1-15.

Ordun, C., Purushotham, S., & Raff, E., (2020). Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. arXiv preprint arXiv:2005.03082.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. Association for Computational Linguistics (ACL) System Demonstrations.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. Proceedings of the eighth ACM international conference on Web search and data mining (pp. 399–408).

Sanguinetti, M., Bosco, C., Lavelli, A., Mazzei, A., Antonelli, O., & Tamburini, F. (2018, May). PoSTWITA-UD: an Italian Twitter Treebank in universal dependencies. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Sciandra, A., (2020). COVID-19 Outbreak through Tweeters' Words: Monitoring Italian Social Media Communication about COVID-19 with Text Mining and Word Embeddings. 2020 IEEE Symposium on Computers and Communications (ISCC) (pp. 1-6), IEEE.

Stella, M., Restocchi, V., & De Deyne, S., (2020). #lockdown: Network-enhanced emotional profiling in the time of Covid-19. Big Data and Cognitive Computing, 4(2), 14.

Stella, M., (2020). Cognitive network science reconstructs how experts, news outlets and social media perceived the COVID-19 pandemic. Systems, 8(4), 38.

Stella, M., Vitevitch, M. S., & Botta F., (2021) Cognitive networks identify the content of English and Italian popular posts about COVID-19 vaccines: Anticipation, logistics, conspiracy and loss of trust. arXiv preprint arXiv:2103.15909.

Su, Y., Xue, J., Liu, X., Wu, P., Chen, J., Chen, C., et al. (2020). Examining the impact of COVID-19 lockdown in Wuhan and Lombardy: a psycholinguistic analysis on Weibo and Twitter. International journal of environmental research and public health, 17(12), 4552.

Sullivan, K. J., Burden, M., Keniston, A., Banda, J. M., & Hunter, L. E. (2020). Characterization of Anonymous Physician Perspectives on COVID-19 Using Social Media Data. Pac Symp Biocomput.

Tamburini, F. (2020). EmoItaly. http://corpora.ficlit.unibo.it/EmoItaly/.

Trevisan, M., Vassio, L., & Giordano, D. (2021). Debate on online social networks at the time of COVID-19: An Italian case study. Online Social Networks and Media, 23, 100136.

Wang, J., Zhou, Y., Zhang, W., Evans, R., & Zhu, C. (2020). Concerns Expressed by Chinese Social Media Users During the COVID-19 Pandemic: Content Analysis of Sina Weibo Microblogging Data. Journal of medical Internet research, 22(11), e22152.

Vitale, P., Pelosi, S., Falco, M. (2020). #andràtuttobene: Images, Texts, Emojis and Geodata in a Sentiment Analysis Pipeline. Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy. Volume 2769 of CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2769/paper_62.pdf.