Hate Speech and Topic Shift in the Covid-19 Public Discourse on Social Media in Italy

Komal Florio, Valerio Basile, Viviana Patti
Department of Computer Science, University of Turin, Italy
{komal.florio,valerio.basile,viviana.patti}@unito.it

Abstract

The availability of large annotated corpora from social media and the development of powerful classification approaches have contributed in an unprecedented way to tackle the challenge of monitoring users' opinions and sentiments in online social platforms across time but also arose the challenge of temporal robustness of such detection and monitoring systems. We used as case study a dataset of tweets in Italian related to the COVID-19 induced lockdown in Italy to measure how quickly the most debated topic online shifted in time. We concluded that it is a promising approach but dedicated corpora and fine tuning of algorithms are crucial for more insightful results.

1 Introduction

The task of abusive message detection is a very challenging one and from multiple perspective. From the computational point of view, despite the increasing interest and effort of the community on developing automatic systems abusive language detection and related tasks for different languages (Poletto et al., 2021; Vidgen and Derczynski, 2021), the robustness of detection and monitoring systems emerges as a crucial factor to be addressed, where one of the main limitations observed is to consider the Natural Language Processing (NLP) task of detecting abusive language in isolation, without taking into account the intersection with the contextual or social dimensions, that could contribute to a more holistic comprehension of the abusive phenomena in language. In fact, it is becoming increasingly evident that the goodness of hate speech prediction systems, and of NLP algorithms in general, is rooted in how well they capture and model all the relevant characteristics of language in the context of a specific phenomenon and its evolution over time (Jurafsky and Martin, 2000; Nadkarni et al., 2011; Feldman, 2013; Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). This brought us to intersect our NLP research with the field of Computational Social Science.

The recent availability of long-term and large-scale digital corpora and the effectiveness of methods for representing words over time can play a crucial role in the recent advances in this field. In particular, social media have recently become one of the predominant sources of linguistic data, being the venue for noticeable phenomena in the domain of NLP tasks. They represent the ideal communication context to address the challenges we have outlined.

This paper aims to characterize how the online conversation on the Italian Twitter around the first Covid-19 lockdown, imposed in Italy in 2020, shifted very quickly from one heated debate to another one, following the quick succession of news reports on both news cases and institutional advice and rules on how to navigate everyday life as the crisis was unfolding in the entirety of the world. At first we tried to identify the most polarizing conversation by analyzing the presence of hate speech using AlBERTo (Polignano et al., 2019) but we found that this BERT (Devlin et al., 2019) based algorithm, trained on Italian Social Media language, seemed to under-perform, in comparison with similar case studies (Capozzi et al., 2019). We hence performed the same task using an abusive language computational lexicon, Hurtlex (Bassignana et al., 2018). We discovered the most recurrent types of abusive language, their distribution over time and correlation with real life events regarding the ongoing pandemic.

To identify the most debated topics we resorted to topic modeling and in particular the Dynamic Topic Modeling allowed us to describe how the most frequent topics evolved over time and shed lights on the interplay with the governmental measures that sparked the most debated conversations.

2 Abusive Speech Prediction

In this work we use as case study a dataset of tweets related to the COVID-19 induced lockdown in Italy, as this was the perfect example of government measures that deeply affected everyday life of citizens and hence had the potential to spark very heated debates online. We rely on a recently developed resource, named 40wita [1] (Basile and Caselli, 2020), created by means of filtering with a set of dedicated keywords the publicly available TWITA dataset (Basile et al., 2018), a long term collection of tweets in Italian. The filtering was run from 1^{st} February 2020 to 30^{th} April 2020 and resulted in the collection of 3309704 tweets.

AlBERTo Our first experiment to detect the most debated conversation consisted in a hate speech prediction with AlBERTo, using the same set of hyper-parameters as in (Florio et al., 2020). The findings show a peak of 6% of daily abusive messages around mid February 2020 and at the end of April 2020, while for the rest of the timestamps the rates were much lower (in some cases almost close to zero) than those found in other Twitter-based datasets (see for example (Capozzi et al., 2020)).

Even allowing for the influence of a different context, this finding induced us to conclude that an unknown but not negligible percentage of hateful messages were left undetected. We believe that increasing the training dataset size and quality could lead to better results. For this experiment the data were annotated using guidelines developed for an hate speech detection task, while a set of new guidelines developed specifically for this context could be a significant improvement in the quality of the labelled data. Another possible adjustment relies on the number of annotators and the exploration of the best metric to compute their disagreement, following the latest published work on annotating subjective tasks (Basile et al., 2021).

Hurtlex In order to get a broader insight of the hateful messages in this dataset that were potentially left out by AlBERTo, we performed the same task by means of Hurtlex (Bassignana et al., 2018)[2], a multilingual computational lexicon that contains 17 different categories of abusive language, each of them consisting of a list of characterising words.

The predominant categories of hate speech are represented by tweets containing derogatory words, abusive terms related to moral and behavioural defects, and words indicating cognitive disabilities and diversity. To gain a deeper insight on how this classification has unfolded we analysed which were the most common words that classified a tweet into a specific category. Quite often the words that determine whether a tweet falls or not into a category, and independently on the category, are very generic (e.g., "problema"="problem", "storia"="history") or can assume very different meaning depending on the context (e.g.: "cane"="dog" can be used as a derogatory term or with a neutral meaning), and this contributes in creating a noisy tweets classification. This insight is meaningful in showing why HurtLex presents some struggles in the accuracy of this task. For this reason, the division into pre-defined categories turned out to be not as informative as we were hoping at the beginning. An improvement on this would encompass a manual revision of the list of words for each category, in order to exclude the most generic ones and retain only those which can potentially improve the accuracy of the result. We also conducted a manual revision of all the tweets belonging to the categories with less than 30 tweets, while for the other categories we choose a random sample of 30 tweets, for consistency with the previous case. One of the most interesting findings was that in the category "rci - locations and demonyms", in contrast to the global dimension of the pandemic, our data counter-intuitively showed that the debate was centered strictly around the measures taken in Italy and the differences between national and local rules.

This lexicon-based approach, even though it did not lead to the desired outcome, was nevertheless important to gain more information on our corpus and experience for future directions. In the next sections we will focus on the most powerful classification tool that we employed on this dataset: two

[1] https://osf.io/n39ks/

[2] http://hatespeech.di.unito.it/resources.html

different algorithms for unsupervised topic modeling.

3 Topic Modeling

We implemented two different classification algorithms. At first we run an exploratory topics analysis with a Latent Dirichlet Allocation (or LDA) and then a Dynamic Topic Modeling (or DTM) to better capture the temporal evolution of topics in the discourse.

Latent Dirichlet Allocation The first topic model algorithm that we applied to our dataset is the Latent Dirichlet Allocation, which was first introduced by Blei (Blei et al., 2003). The popularity and versatility of such algorithm relies on the human-interpretable form of the extracted topics and on being, by construction, very robust when deployed on unseen documents.

This model was able to correctly and precisely identify the conversations around the first relevant news around the incoming pandemic. Examples of this include the first restrictions on movements following the first Covid-19 outbreak in Lombardy and Veneto, the national lockdown issued in March and the consequent gradual shift of the conversation towards the difficulties of normal life in such a new context.

As powerful as this model is, it showed a fundamental limit for our perspective and purpose. The relevant topics were punctual but, as expected, not consistent over time because the model was completely re-trained on data from every single week, hence the results for each single time slice were agnostic of the result for every other time slices, and therefore not time-consistent, or comparable, by design. To overcome this issue we implemented a Dynamic Topic Modeling.

Dynamic Topic Modeling The Dynamic Topic Modeling (Blei and Lafferty, 2006) allows to split the datasets into custom time slices and extracts the same exact topics over all of them, thus enabling an analysis on how topics evolve over time.

At first we fine tuned the model by optimizing the perplexity and the coherence score. The first score captures the behaviour of the model towards data which were previously unknown by means of a normalised log-likelihood of a held-out test set. However there are relevant studies (for example (Chang et al., 2009)) proving that perplexity and human judgement not only often do not correlate, but sometimes they even anti-correlate. For this reason a second metric was elaborated: the coherence score, to better model human judgement. This measure captures the degree of semantic similarity between the words related to each single topic (i.e., a measure of the likeness of their meaning). We did not have an annotated corpus that can serve as a training set, hence we only explored the trend of the coherence score with reference to changes in the number of topics, chunksize of data, number of passes and evaluation score. We then concluded for 5 topics and 20 words per topics, as listed in the following Table 1. We chose to leave one topic undetermined ("Topic 1 - Other") to label all the messages that the algorithm struggled to correctly assign to a specific topic.

Topic No.	Italian	English
Topic 0	quarantena	quarantine
Topic 1	altro	other
Topic 2	lavoro	work
Topic 3	governo	government
Topic 4	sanità	healthcare

Table 1: Topics Extracted using the Dynamic Topic Modeling.

The DTM outputs each unlabelled topic as a list of words with a relevance value. This value, between 0 and 1, represents the probability of a single word to be affiliated with a specific topic. The rationale behind the decision of choosing only 5 topics is that a higher number did not improve the understanding of the corpus as it led to a noisier classification. Each additional topic consisted of a list of words that were either very general in their meaning, or not very close semantically, or both, which made it very difficult to find a topic label that properly represented all the listed tokens.

The most powerful feature of the DTM is that, for each topic, it is possible to rank the most relevant words based on their attached probability value (of referring to the specific topic) and see how they evolve over time. In the following Figure 1, the change in ranking for all the 20 words involved is presented as a coloured heatmap, where the blue values represents words with higher ranking while the red ones are at the lower end of ranking.

There are two main insights we can gain from this visualization. The first one is that topics

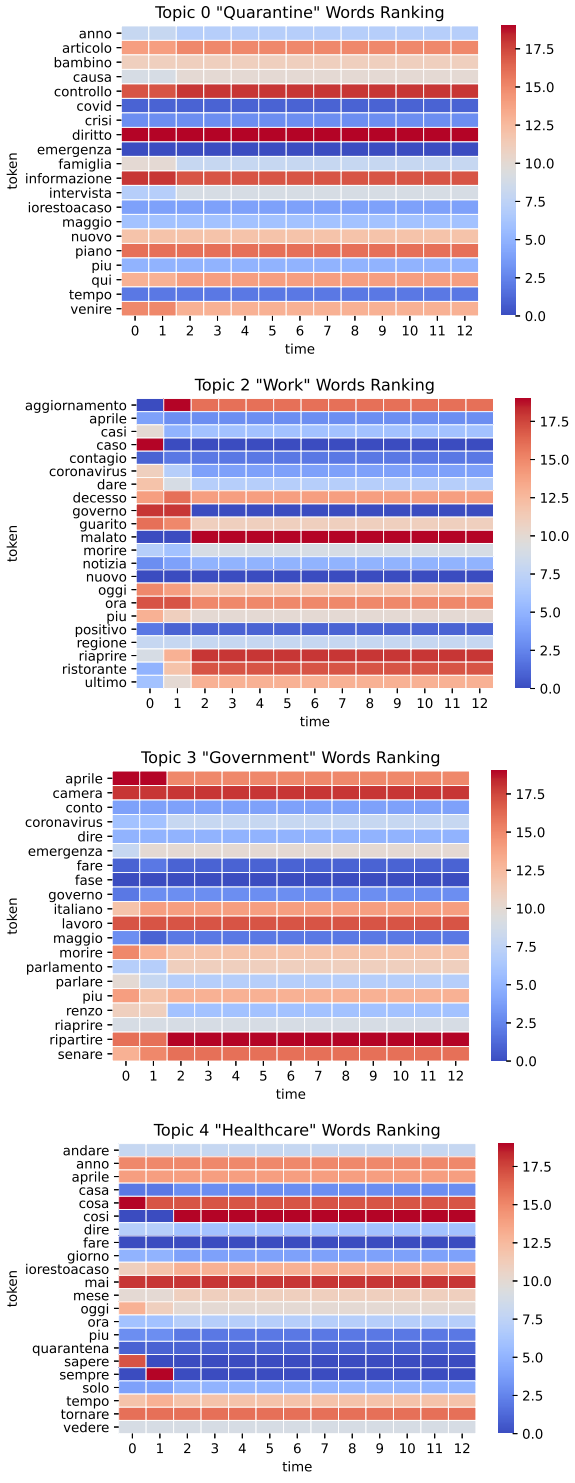

Figure 1: Time evolution of words relevance ranking for all the topics.

are lists of pretty common words, which proves how hard of a task topic detection is, because of the complexity and versatility of human language, where general words can be used in different contexts with different meanings. The second insight is that the biggest changes in the word ranking happen within the first time slices. A possible explanation may be traced back to how this dataset was created. The list of hashtags and trends used to filter the tweets was compiled in February and was fixed in time. This means that potentially interesting tweets were left out because they contained hashtags that emerged as relevant later in time but hence were not captured by the keywords used for selecting relevant tweets.

In order to measure the temporal trend of predominance for each topics, we computed, for each of the 13 time slices, the ratio of documents labeled as predominantly referring to each of the topic.

We plotted in Figure 2 the normalized share of documents classified as containing each of the topics in each time slices, to highlight the relative trends over time.

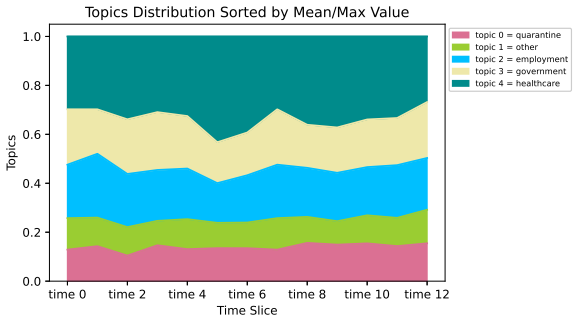

Figure 2: Evolution over time of mean and maximum values of the share of documents related to each of the topics.
We listed both metrics in the same chart as the difference in values was below our error threshold.

It is aligned with our intuition that the largest share of documents across time refer to topic ”healthcare”. But more in detail it is interesting to analyse the relation between the timestamp of the spikes and relevant Covid-19 events in Italy, as presented in Table 2.

The spikes in shares of documents related to the most predominant topic ”quarantine” do follow temporally major events about public health announcement and measures, as shown in Table 2. This proves the point of this research, which

Topic	Time- Slice	Start Day	End Day	Relevant event
4-healthcare	2	16/2/20	22/2/20	public discussion around the first red zones in Veneto
	5	8/3/20	14/3/20	Announcement of the arrival of a medical task force from Cuba in Lombardy (14/3/20). Appointment of a special consultant for the emergency in Lombardy (16/3/20).
2-work	1	9/2/20	15/2/20	public discourse around the Chinese community in Italy
	5	8/3/20	14/3/20	Announcement of the arrival of a medical task force from Cuba in Lombardy (14/3/20). Appointment of a special consultant for the emergency in Lombardy (16/3/20).
	11	19/4/20	25/4/20	First positive news about the Oxford vaccine AstraZeneca
3-government	2	16/2/20	22/2/20	public discussion around the first red zones in Veneto
0-quarantine	3	23/2/20	29/2/20	first red zones issued in Lombardy and Veneto
	9	5/4/20	11/4/20	Economical measure announced. Public discourse around lifting the strict lockdown measures.

Table 2: Relevant Covid-19 events occurred around spikes in the chart.

is that the discourse on Twitter does not only follow closely the most recent and relevant news but it quickly shifts from one topic to the other. In fact, all major peaks in Fig. 2 are followed by a sharply decreasing trend, indicating an immediate loss of predominance and hence an alternation of the dominant arguments of debates.

We explored in a similar way also the temporal evolution of the share of tweets labelled with the Hurtlex categories.

For each of the time slices we computed the relative frequency of tweets labeled with every categories and then created a stacked plot of their maximum values (shown in Figure 3) and the normalized mean values (shown in Figure 4) of their frequencies, to identify both peaks and categories that were consistently predominant through the time.

The relevance of the Hurtlex category related to derogatory words detected over the whole dataset, as described in Section 2, confirms its validity also at a weekly time granularity, as shown by Figure 3. Looking at the chart as a whole it is important to notice that, as we have already highlighted before, the peaks occur in time slices 3 and 5, which respectively correspond the the issue of the first red zones in Italy and two major public health news regarding Lombardy, the hardest hit region of Italy in the first phases of the pandemic (see Table 2 for details).

It is relevant to notice that these peaks occur exactly in the same time slices as the peaks in Figure 2 for the topics ”quarantine” and ”healthcare”, showing that the most heated debates happened around public measures that affected directly and immediately on both the collectivity (”healthcare”) and personal life (”quarantine”). Analysing the mean value of the frequencies, in Figure 4, we can see that categories rank differently from Figure 3. More specifically we see that

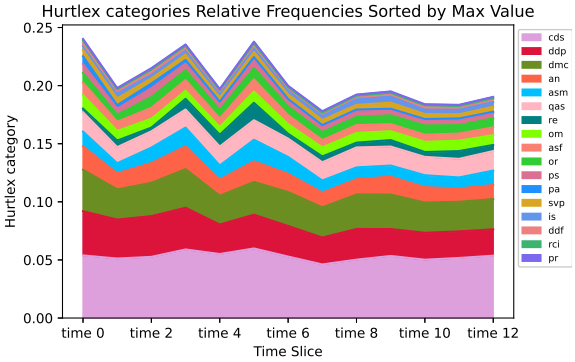

Figure 3: Hurtlex categories maximum frequencies values over time.

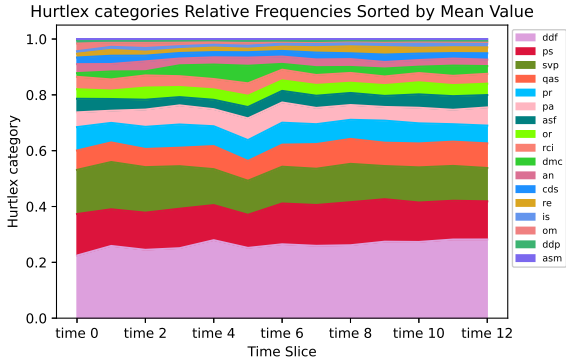

Figure 4: Hurtlex categories mean frequencies values over time.

for example ”ddf - physical disabilities and diversity” is by far the most consistent over time but it represents somehow a generic type of offensive language, not correlated with the pandemic, and to some extent this is as a noisy classification of tweets and it would be interesting to investigate further how to improve on this result.

4 Conclusion and Final Remarks

In this work we tried to tackle the challenge of measuring and quantifying the topic shift in the

public discourse on Social Media, using as a case study the online debate on Twitter following the Covid-19 related lockdown in Italy in 2020, by means of a dedicated dataset. By combining multiple classification methods we gathered insights into which governmental measures generated the most debated online conversation but we also concluded for the need of deeper investigation on how to build ad hoc corpora and methods to investigate specific linguistic phenomena as online conversation with rapid topic shift following the flow of news coming from both online and traditional media outlets. We also tried to inform AlBERTo with information extracted from topic modeling but the results were far from satisfying. This is a promising way to enhance the accuracy of hate speech prediction, but we concluded that a further investigation on size and characteristics of datasets is essential to gain better results.

References

Valerio Basile and Tommaso Caselli. 2020. 40twita 1.0: A collection of Italian Tweets during the COVID-19 Pandemic.

Valerio Basile, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term Social Media Data Collection at the University of Turin. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2253 of *CEUR Workshop Proceedings*, pages 1–6, Torino,Italy. CEUR-WS.org.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021. We need to consider disagreement in evaluation. In *1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21. Association for Computational Linguistics.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*, pages 1–6. CEUR-WS.org.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March.

Arthur TE Capozzi, Mirko Lai, Valerio Basile, Fabio Poletto, Manuela Sanguinetti, Cristina Bosco, Viviana Patti, Giancarlo Ruffo, Cataldo Musto, Marco Polignano, et al. 2019. Computational linguistics against hate: Hate speech detection and visualization on social media in the" contro l'odio" project. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR-WS.

Arthur TE Capozzi, Mirko Lai, Valerio Basile, Fabio Poletto, Manuela Sanguinetti, Cristina Bosco, Viviana Patti, Giancarlo Ruffo, Cataldo Musto, Marco Polignano, et al. 2020. "contro l'odio": A platform for detecting, monitoring and visualizing hate speech against immigrants in italian social media. *IJCoL. Italian Journal of Computational Linguistics*, 6(6-1):77–97.

Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09, page 288–296, Red Hook, NY, USA. Curran Associates Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, USA. Association for Computational Linguistics.

Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12):4180.

Paula Fortuna and Sergio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, USA, 1st edition.

Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. 2011. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55(2):477–523.

Marco Polignano, Valerio Basile, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. AlBERTo: Modeling Italian Social Media Language with BERT. *Italian Journal of Computational Linguistics - IJCOL*, -2, n.2.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April. Association for Computational Linguistics.

Bertie Vidgen and Leon Derczynski. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32, 12.