Measuring Shifts in Attitudes Towards COVID-19 Measures in Belgium Using Multilingual BERT

Kristen Scott[*]   KRISTEN.SCOTT@CS.KULEUVEN.BE
Pieter Delobelle[*]   PIETER.DELOBELLE@CS.KULEUVEN.BE
Bettina Berendt[**]   BETTINA.BERENDT@CS.KULEUVEN.BE

[*]*Department of Computer Science; Leuven.AI, KU Leuven, Belgium*
Also denotes equal contribution

[**]*TU Berlin and Weizenbaum Institute, Germany; KU Leuven, Belgium*

Abstract

We classify seven months' worth of Belgian COVID-related Tweets using multilingual BERT and relate them to their governments' COVID measures. We classify Tweets by their stated opinion on Belgian government curfew measures (too strict, ok, too loose). We examine the change in topics discussed and views expressed over time and in reference to dates of related events such as implementation of new measures or COVID-19 related announcements in the media.

1. Introduction and Related Work

Sentiment analysis or opinion mining of social media content presents the possibility of following trends in public discussion. Twitter, with an easy to use API and short, focused messages called Tweets, is often targeted for such tasks (Medhat et al. 2014, Giachanou and Crestani 2016). During the COVD-19 pandemic, quantifying which measures are supported by the general population, and which ones are not, could be useful in shaping the course of a nation's strategy. Recent work has focused on monitoring reactions to the COVID pandemic utilizing sentiment analysis (Wang et al. 2020b, Chen et al. 2020, Brandl and Lassner 2020, Kurten and Beullens 2021, Wang et al. 2020a). However, sentiment does not necessarily map to opinions on more complex opinions about measures. Wang et al. (2020a) presented initial results in workshop on classifying stances (for or against) towards Dutch government policies on masks and distancing using a neural network. Others have incorporated qualitative analysis techniques into similar work flows in an attempt to gain a more nuanced understanding of social media discussion than sentiment analysis, unsupervised machine learning or classification models alone (Jimenez-Sotomayor et al. 2020, Xue et al. 2020).

BERT (Devlin et al. 2019) is a Natural Language Processing (NLP) model that uses pre-training and fine-tuning. This makes it easier than ever to create custom, domain-adapted classifiers that capture the nuances of discussions and opinions expressed in a given domain. We characterize the discussion of Belgian COVID measures on Twitter over time using multilingual BERT models that we finetuned on manually labelled Tweets. By visualizing the change of rate of the sentiment in curfew-related tweets, we found shifts in the support for curfews over time.

2. Methodology

We used a multilingual BERT model to classify 1.3 million Tweets related to the COVID-19 pandemic, based on a manually labeled training set, as described in Subsection 2.1. The Tweets were collected through the Twitter API starting from October 13, 2020 until April 08, 2021 using a continuously running script[1]. Tweets were collected using (i) multilingual search terms related to

---

1. This script with search terms and all our code is available at `https://github.com/iPieter/bert-corona-tweets`.

Table 1: Labeling categories for each tweet.

TOPIC		MEASURE SUPPORT	GOVERNMENT SUPPORT	RELEVANCE
masks	testing	too-strict	supportive	irrelevant
curfew	closing-horeca	ok	unsupportive	
quarantine	vaccine	too-loose	not-applicable	
lockdown	other-measure	not-applicable		
schools				

COVID-19, corona and specific related topics, (ii) a language filter on Dutch, French and English, and (iii) a filter for locations in Belgium.

In Subsection 2.2, we describe how we use this dataset and the collected labels to develop multiple models. These models were developed synchronously with the labeling task. We started with model to filter irrelevant Tweets (e.g. news announcements) to save labeling time (Sieve I). We then created a second model to classify the Tweets into topics, which we use to focus on the curfew topic (Sieve II) for the more challenging labels: measure and government support. This interplay allowed us to reduce labeling cost and develop multiple useful models.

2.1 Labeling

Two manually labeled datasets were used for training. The first, consisting of 1695 Tweets was used for classifying topics. The second set of 2000 labeled Tweets was used to classify support for curfews. As described in Table 1, Tweets were labeled by topic (curfew measure), as well as by two opinion axis: opinion toward specific measures (too strict, acceptable, not strict enough and a neutral option) and measure of the overall support expressed towards the government's handling of the pandemic (supportive, unsupportive).

We developed a code book which defines the labels procedure in detail. This labeling process was tested and refined through two rounds of labeling on smaller datasets with Belgian and multilingual labelers. Each round was followed by discussion, resolving of disagreement and making minor adjustments to the code book [2].
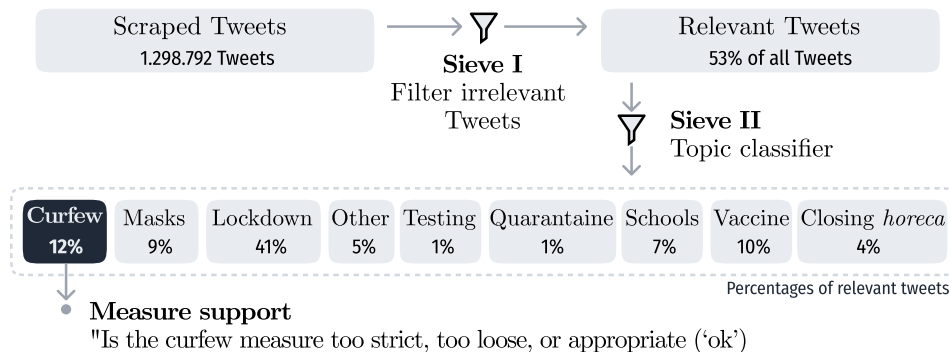
2.2 Training

We developed multiple models to classify Tweets, which correspond with the labeling rounds. Figure 1 shows how collected Tweets on the topics are filtered and that we have four models: (i) a model to filter irrelevant Tweets for Sieve I, (ii) another model to classify topics and two models to (iii) predict support for a measure and (iv) support for the government. As mentioned before, each sieve helped reduce the number of Tweets that needed to be classified each round.

Classifying relevant Tweets. For the first sieve, we focus on relevant versus irrelevant Tweets as discussed in Subsection 2.1 Only 53% of the labeled Tweets were relevant. To automatically filter these Tweets, we trained and evaluated multilingual BERT (mBERT) and XLM-RoBERTa models. Each training was run 8 times with random hyperparameters and the best-performing model—using accuracy as a selection metric—was evaluated on a held-out test set, following Dodge et al. (2019).

mBERT performed slightly better than XLM-RoBERTa, with an AUC score of 0.85 and 0.84 respectively. The mBERT model also had a higher true positive rate of 0.3 when selecting a threshold with a false positive rate of zero, as the goal of the first sieve is to remove *clearly* irrelevant Tweets. From a computational standpoint, the base mBERT model also has the benefit that it is significantly cheaper and faster to train due to a smaller model size.

---

2. Code book available at `https://github.com/iPieter/bert-corona-tweets`.

Figure 1: Schematic illustration of our contribution with domain-specific classifiers.

Classifying topics. We trained mBERT on 600 labeled Tweets to classify topics, we validated 8 models on a validation set with 64 Tweets and finally tested the best-performing model, using accuracy, on 100 Tweets. The best-performing model had an overall accuracy of 0.73 (macro-averaged accuracy of 0.95). Some classes perform very well, like `curfew` ($AUC = 0.90$), `lockdown` ($AUC = 0.85$) and `vaccine` ($AUC = 0.90$). Yet, some classes are ill-represented in the dataset and perform significantly worse, more specifically `quarantine` ($AUC = 0.50$) and `testing` ($N = 1$).

The topic model performs quite well overall and given our interest in the curfew topic specifically, this model is quite suitable. We also make the model available on the HuggingFace repository[3] for practitioners to use.

Classifying support for curfews. For the last classification model, we trained mBERT for multiclass classification on 1518[4] Tweets with support labels, of which 100 were used as held-out test set and 75 as validation split. We tested 5 hyperparameter assignments. The overall accuracy is 0.71. However, there is a significant class imbalance and despite oversampling, the performance varies from no better than random ($AUC = 0.5$ for `too-loose`) to good ($AUC = 0.74$ for `not-applicable`, $AUC = 0.69$ for `ok` and $AUC = 0.73$ for `too-strict`).

Given these results, we primarily focus on the `too-strict` label for the curfew topic in the rest of this work. We also make this model available through the HuggingFace repository[5].

3. Results and discussion

Here we focus on the topic of curfew for reporting of more detailed results. The timeline of the rate of classified Tweets with the topic of curfew, along with classified rate of support (or non-support) for curfews, is shown in Figure 2. Also included, for reference, is the rate of confirmed COVID cases in Belgium (Sciensano 2021).

November 2, 2020, Belgium entered a country-wide lockdown which included a national midnight curfew, while some regional curfews had been put in place in the days prior. We find media announcements of these upcoming curfews as well as announcements regarding the extension of these curfews (VRT NWS 2021, Johnston 2021) were accompanied by temporary increases curfew related Tweets. In October, as the rate of curfew Tweets dropped, there was no change in the opinions expressed about the curfew (with the majority remaining 'no opinion' until February). By contrast, during the 2021 increases in curfew Tweets we see a large change in opinions (particularly an increase in 'too strict'). Further research is required to determine whether the changes in rate of negative

---

3. Available at `https://huggingface.co/DTAI-KULeuven/mbert-corona-tweets-belgium-topics`.
4. This were originally 2000 Tweets of which the clearly irrelevant ones were filtered with Sieve I before labeling.
5. Available at `https://huggingface.co/DTAI-KULeuven/mbert-corona-tweets-belgium-curfew-support`.
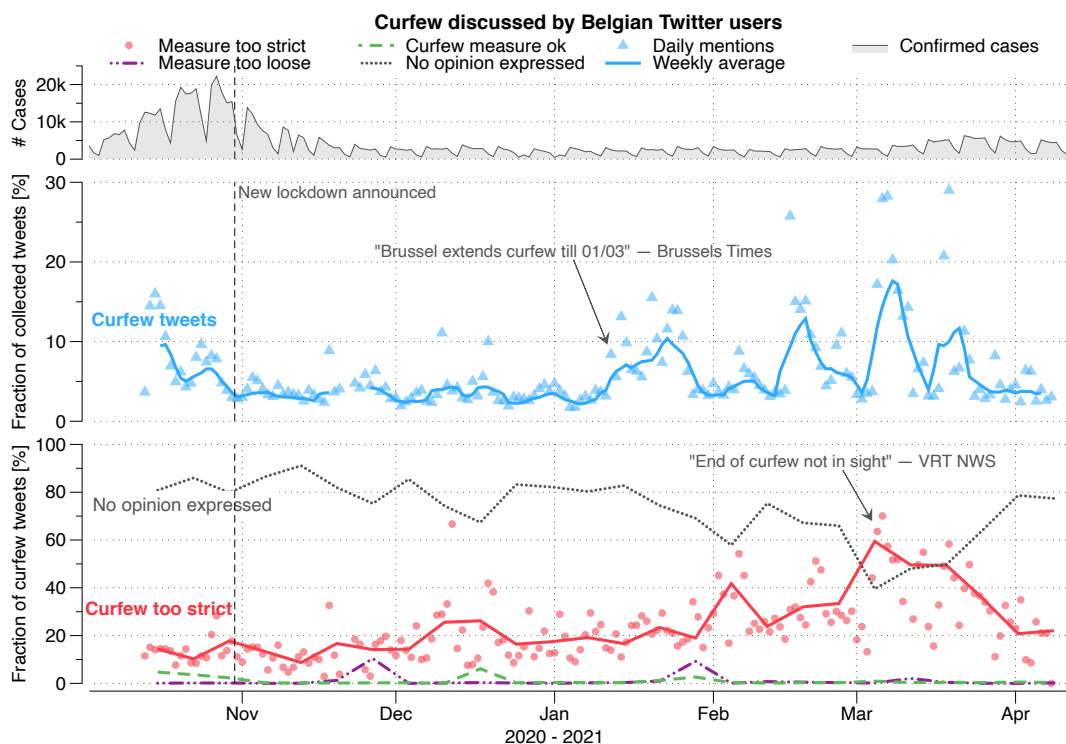
Figure 2: Timeline of the relative number of Tweets on the curfew topic (middle) and the fraction of those Tweets that find the curfew too strict, too loose, or a suitable measure (bottom), with the number of daily cases in Belgium to give context on the pandemic situation (top).

opinion observed correspond to changes in public opinion or some other effect such as increased attention to particular announcements by individuals with consistent anti-curfew opinions.

We also see that opinions on measure strictness are just one element of the discussions around COVID measures, suggesting the use of Twitter for other forms of communication, such as information sharing and humour as well as conveying complex points of views and personal stories (e.g. about the impact of the curfew). The ability of the BERT models to classify tweets based on our highly specific scale of strictness suggests that such models may be effective for categorizing based on other complex and nuanced labels when trained with carefully labeled data.

4. Conclusion and future work

We are able to observe the discussion of COVID measures on Twitter over time examine the shifting in both the topics of focus and the view points. We found that the majority of Twitter discussion of these measures is not centered on expression of specific levels of support and thus identify the need to characterize the nature of the non-opinionated Tweets as well as to understand the poorer performance of our model on opinions other than 'no opinion' and 'too strict'. We also acknowledge the limitations of treating Twitter data as representative of the viewpoints of the general population. While we do work with multiple languages, further work can be done to determine differential performance between languages, dialects and informal and slang texts.

Acknowledgment

References

Brandl, Stephanie and David Lassner (2020), Corona twitter dataset: 16 february 2020 - 03 march 2020. http://dx.doi.org/10.14279/depositonce-10012.

Chen, Emily, Kristina Lerman, and Emilio Ferrara (2020), Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set, *JMIR Public Health Surveill* **6** (2), pp. e19273. http://publichealth.jmir.org/2020/2/e19273/.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186.

Dodge, Jesse, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith (2019), Show your work: Improved reporting of experimental results, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, pp. 2185–2194. https://www.aclweb.org/anthology/D19-1224.

Giachanou, Anastasia and Fabio Crestani (2016), Like it or not: A survey of twitter sentiment analysis methods, *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2938640.

Jimenez-Sotomayor, Maria Renee, Carolina Gomez-Moreno, and Enrique Soto-Perez-de-Celis (2020), Coronavirus, Ageism, and Twitter: An Evaluation of Tweets about Older Adults and COVID-19, *Journal of the American Geriatrics Society* **68** (8), pp. 1661–1665. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jgs.16508. http://agsjournals.onlinelibrary.wiley.com/doi/abs/10.1111/jgs.16508.

Johnston, Jules (2021), Brussels extends curfew to 1 March, *The Brussels Times*. https://www.brusselstimes.com/brussels/149404/brussels-extends-curfew-to-1-march-rudi-vervoort-epidemiological-situation/.

Kurten, Sebastian and Kathleen Beullens (2021), #coronavirus: Monitoring the belgian twitter discourse on the severe acute respiratory syndrome coronavirus 2 pandemic, *Cyberpsychology, Behavior, and Social Networking* **24** (2), pp. 117–122. PMID: 32857607. https://doi.org/10.1089/cyber.2020.0341.

Medhat, Walaa, Ahmed Hassan, and Hoda Korashy (2014), Sentiment analysis algorithms and applications: A survey, *Ain Shams Engineering Journal* **5** (4), pp. 1093–1113. https://www.sciencedirect.com/science/article/pii/S2090447914000550.

Sciensano (2021), Covid-19 reports, *Epistat*. https://epistat.wiv-isp.be/covid/.

VRT NWS (2021), Liveblog - Einde van avondklok nog niet in zicht: "Ook na heropening horeca", *vrtnws.be*. https://www.vrt.be/vrtnws/nl/2021/03/05/liveblog-corona-5-maart-2021/.

Wang, Shihan, Marijn Schraagen, Erik Tjong Kim Sang, and Mehdi Dastani (2020a), Public Sentiment on Governmental COVID-19 Measures in Dutch Social Media, *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Association for Computational Linguistics, Online. https://www.aclweb.org/anthology/2020.nlpcovid19-2.17.

Wang, Tianyi, Ke Lu, Kam Pui Chow, and Qing Zhu (2020b), COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model, *IEEE Access* **8**, pp. 138162–138169.

Xue, Jia, Junxiang Chen, Chen Chen, Chengda Zheng, Sijia Li, and Tingshao Zhu (2020), Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter., *PloS one* **15** (9), pp. e0239441. http://search.proquest.com/docview/2446668077?pq-origsite=primo.