Computational Linguistics in the Netherlands Journal 11 (2021) 161-171 Submitted 10/2021; Published 12/2021

Measuring Shifts in Attitudes Towards COVID-19 Measures in Belgium

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; Leuven.AI and KU Leuven, Belgium

Abstract

With the COVID-19 pandemic and subsequent measures in full swing, people voiced their opinions of these measures on social media. Although it remains an open problem to correctly interpret these voices and translate this to public policy, we work towards this by tracking support for corona-related measures in Belgium, a densely-populated trilingual country in Western Europe. To this end, we classify seven months' worth of Belgian COVID-related tweets using multilingual BERT and a manually labeled training set. The tweets are classified by which measure they refer to as well as by their stated opinion towards the curfew measure, for which we introduce a custom classification scheme (too strict, ok, too loose). Using this classification, we examine the change in topics discussed and views expressed over time and in reference to dates of related events such as the implementation of new measures or COVID-19 related announcements in the media. With these promising results, our contributions include (i) multiple multilingual BERT models trained on manually labeled data accompanied by (ii) historical analysis of the support for the curfew measure on Twitter and (iii) a thorough analysis of limitations and risks, together with best practices and a reference code book.

1. Introduction

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 COVID-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-19 pandemic utilizing sentiment analysis (Wang et al. 2020b, Chen et al. 2020, Brandl and Lassner 2020, 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 a 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 workflows 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).

Inspired by these initiatives, we set out to create a method focused on the Belgian situation that can also serve as case-study on the risks and limitations inherent to this approach. Concretely, we present the following three contributions:

Firstly, we develop a set of BERT models on a real-world use-case in the context of the COVID-19 pandemic, allowing for testing the viability of its use in providing insight into opinions being expressed on Twitter in Belgium, a multi-lingual country. We find evidence of promise for identifying specific opinions on specific topics within a large corpus of tweet data and publish the models.

Secondly, we visualize the trends in tweets on the topic of curfews over time, along with the corresponding shift in views towards Belgian curfew measures. This allows us to compare these tweets with dates of curfew policy shifts and their accompanying announcements. We reckon this could be useful for policy makers in the future.

Thirdly, we explicitly discuss the limitations and risks in the practice of utilizing social media data as a measure of public opinion or as an information source for policy makers. We provide an analysis of model results with an eye towards these known limitations, including an assessment of some specific potential biases of the BERT model itself. Communicating bias may improve the value of the research by adding specificity to a claim (Elish and boyd 2018). We need to be able to explicate potential pitfalls and develop best practices when working on such sensitive real-world use cases. We cannot separate the question of what we learned about COVID-19 restrictions opinions from questions of BERT model performance, data characteristics and labeling choices. Given the current confluence of advances in NLP language models, availability of social media content, and presence of globally discussed topics, our focus on limitations and risks is timely and needed, as can be seen by the explosion of papers applying NLP tools in a similar workflow.

2. Related Work

The recent COVID-19 pandemic has motivated a significant amount of research using Natural Language Processing (NLP) methods to analyze social media data for information about the public's response to the unprecedented pandemic. A large portion of this research utilizes sentiment analysis tools as a method for monitoring changing reactions of the public over time to the COVID pandemic, for example, (Wang et al. 2020b, Chen et al. 2020, Brandl and Lassner 2020, Kurten and Beullens 2021). However, sentiment analysis does not necessarily provide insight into specific opinions about a given government measure—a point that we will address in Subsection 3.1.

Some of these analysis tools utilize BERT (Devlin et al. 2019), which is a state-of-the-art 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. BERT has been used for classification tasks, including sentiment, gender of writer and stance detection. Müller et al. (2020) have created a T-BERT, a model pre-trained on a large corpus of unclassified English language Twitter messages on the topic of COVID-19, which is not suitable for use in the context of countries such as Belgium, with multiple non-English official languages. Additionally, no previous work known to us has created a BERT-based model pre-trained on manually labeled multilingual tweets on the topic of COVID-19.

It is important to recognize that even a highly accurate model for classifying opinions expressed in tweets may not actually be sufficient for ensuring that such a model can act as a valid source of information about the public's opinions or accurately inform policy choices. The risks of using social media data to draw conclusions about the general population have been described in detail by Olteanu et al. (2016). Some points made in that work particularly salient to the current work include the demonstrated fact that Twitter users do not represent the population as a whole, which can create a biased representation of public opinion. The nature of social networks such as Twitter also provokes a tendency towards strong opinions and language that can polarize groups (Garimella and Weber 2017). People are motivated by a variety of goals when using social media, including gaining attention, or followers, causing disruptions ('trolling') or advertising a product. For example, Kıcıman (2012) found that network attributes, such as network size, correlate with different user behaviors on Twitter. There are also significant ethical considerations to be made regarding collecting and analyzing tweets in terms of privacy and surveillance concerns (franzke et al. 2020). Despite the extensive discussion, the perception that continued research of social data has a valuable role persists.

Olteanu et al. (2016) give specific recommendations for a path forward for conducting valuable research with social data, which include: documenting in detail the process by which datasets and

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				

† *horeca* = *ho*tels, *re*staurants and bars ("*ca*fé" in Dutch)

models are created, including identifying sources of bias, broadening studies to varied contexts and extending the research on guidelines, standards, methodologies, and protocols, as well as to encouraging their adoption. To this end, in addition to providing open access to the search queries and models used for the current research, as is the standard expectation, we give particular attention to the process of documenting our work (including our code book and labeling methodologies). We also address these recommendations further in Section 5.

3. 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 3.1. The tweets were collected through the Twitter API from October 13, 2020 until April 08, 2021 using a continuously running script.[1] Tweets were collected using (i) multilingual search terms related to COVID-19, corona and specific related topics[2], (ii) a language filter on Dutch, French and English, and (iii) a filter for locations in Belgium.

Given that the location field in Twitter is a free-form input and that some users do not use this feature, we relied on the occurrences of 'Belgium', translated versions and Unicode emoji flag in the location and free-form description field and the occurrence of a city or region in the location field[3]. We also considered filtering using GPS coordinates that are available on some tweets, but during initial testing we saw no tweets that contained this metadata originating from Belgium.

In Subsection 3.2, 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 a model to filter out 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 models.

3.1 Labeling

Two manually labeled datasets were used for training. The first, consisting of 1695 tweets was used for training the topics classifier. 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 axes: 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 multilin-

1. This script with search terms and all our code is available at `https://github.com/iPieter/bert-corona-tweets`.
2. These search terms are also available on our repository.
3. We queried DBPedia for Dutch, French and English names of cities (e.g. 'Leuven') and regions (e.g. 'Flanders').

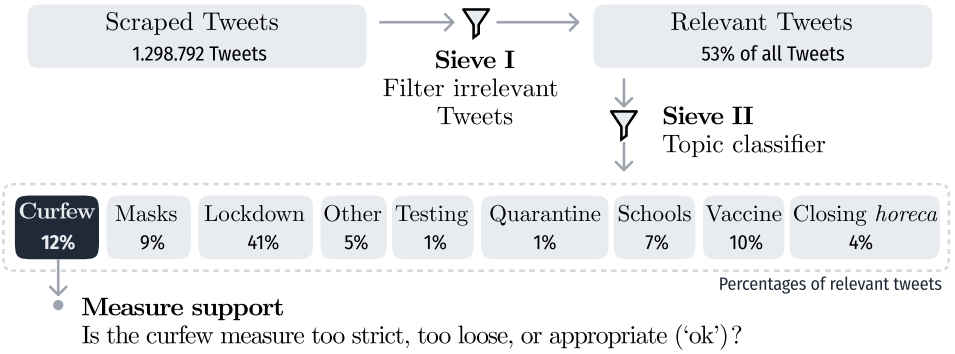


Figure 1: Schematic illustration of our contribution with domain-specific classifiers.

gual labelers. Each round was followed by discussion, resolving of disagreement and making minor adjustments to the code book[4].

The specific measures labeled for were selected based on what was prevalent in public discourse, recent and upcoming regulations, and what we saw in the Twitter data. Additional topics were added as collecting continued, as various topics shifted in prevalence or emerged over time.

Two opinion axes were used to label tweets. The first captures the opinion toward specific measures (too strict, acceptable, not strict enough); the second axis is a measure of the overall support expressed towards the government's handling of the pandemic overall (supportive, unsupportive). For both of these axes, a *non-applicable* option was available to the labelers, for use if a tweet did not express a clear opinion towards a specific measure or towards the government response, see Table 1.

To build an evaluation and test set, 200 tweets were labeled by three labelers; Krippendorff's alpha, a measure for inter-annotator agreement based on the fraction of equal and total labels, was unsatisfactorily low, at 0.40. This was followed by a round of discussion of disagreement, which led to clarification of definitions between coders and minor changes to the code book. A second round of labeling followed, during which which 400 tweets were labeled by two of the labelers and Krippendorff's alpha increased to 0.62. A second round of discussion and resolving of disagreement occurred. All subsequent labeling was done by one of these labelers.

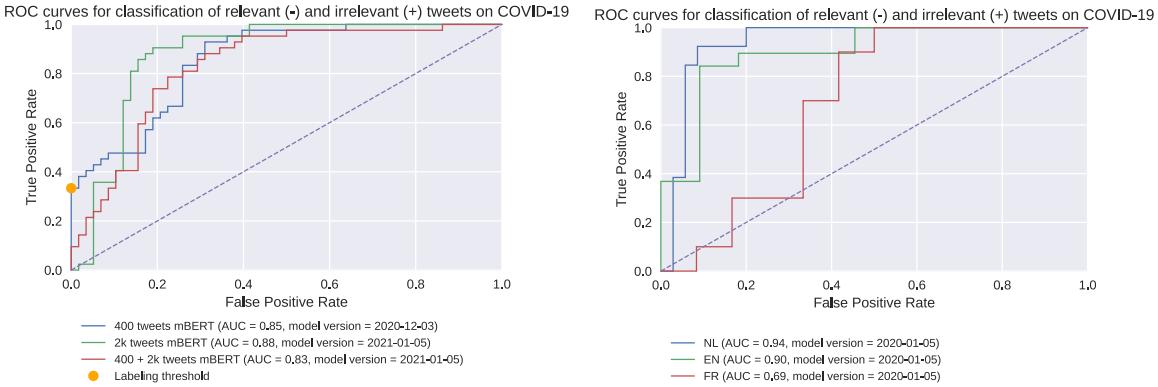
Mapping support levels to sentiment. Our labeling axes differ from the classic sentiment analysis setting, as we have two 'negative' sentiments (too strict and too loose), a positive sentiment (measure ok) and additionally an ambivalent sentiment (no opinion expressed). During our test rounds, we found that a binary classification into positive and negative sentiment does not fully capture the stances that Twitter users took. There is quite a difference between stance that a measure is too strict and too loose, with the former being much more popular in our data. For this reason, we decided to not lump both together in a negative sentiment category.

Ambiguous tweets. The topic or context of a tweet is often ambiguous; it may be part of a conversation that is not included as a thread, or in reference to some topic that is no longer clear when reading it in a different time or place, or they may contain slang, hashtags or references that the coder may not be familiar with. Our coding policy was to conduct a minimal level of research when required: this includes clicking on any included links, and looking up unfamiliar terms.

3.2 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

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



(a) ROC curves for different model versions, including the threshold set on the first (400 tweets) model used as Sieve 1.

(b) ROC curves conditioned on language (English, Dutch and French) for the best-performing model: mBERT trained on 2k tweets.

Figure 2: ROC curves of different models trained (left) and the performance per language for the best-performing model (right).

filter irrelevant tweets for Sieve I, (ii) another model to classify topics and two models to (iii) predict support for a measure, curfew, 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 3.1 Only 53% of the labeled tweets were relevant. To automatically filter these tweets, we trained and evaluated multilingual BERT (mBERT) (Devlin et al. 2019) and XLM-RoBERTa (Conneau et al. 2019) models.

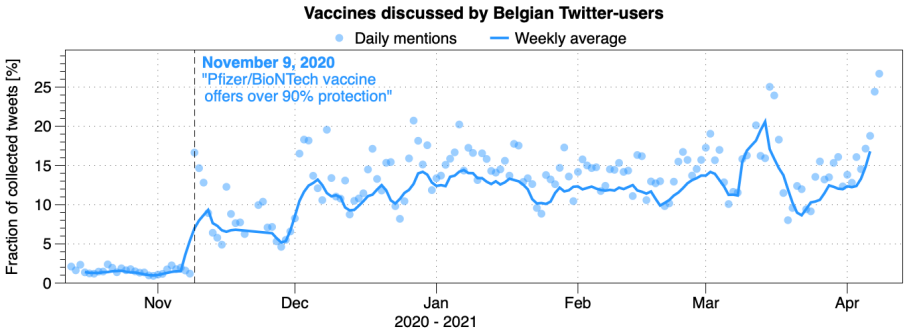
Each training was run 8 times with random hyperparameters and the best-performing model—using accuracy as a selection metric—was evaluated on a test set, following Dodge et al. (2019).

The mBERT model 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 0.0. This is acceptable, since the goal of this model is to filter out irrelevant, mostly automated, tweets in a first iteration. 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.

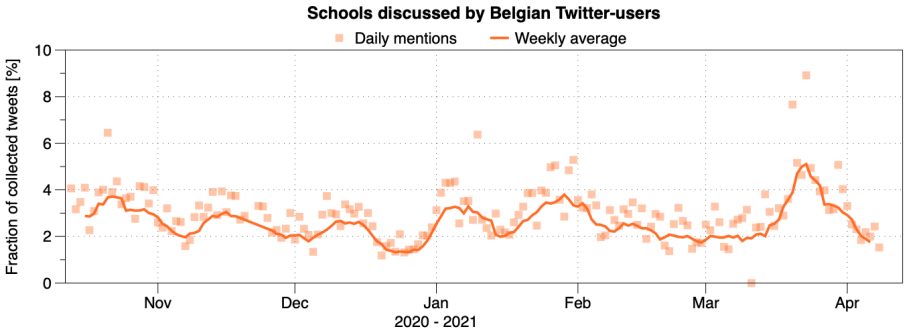
We also developed this model in two iterations: (i) first by focusing on developing a model to remove irrelevant tweets that are usually automated and follow a fixed template, e.g. emergency service calls. We did this by training mBERT and XLM-RoBERTa on 400 labeled tweets. (ii) Using this model as a filter a threshold set for almost no false positives (see the threshold in Figure 2a), we labeled an additional 1695 tweets (this labeling round included all labels, not just relevance) and retrained the irrelevance model on the new dataset.

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. 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 as a filter (Sieve II in Figure 1). Figure 3 illustrates the result of the



(a) Fraction of collected tweets talking about vaccines.



(b) Fraction of collected tweets talking about schools in relation to the COVID-19 pandemic.

Figure 3: Two topics, vaccines and schools, that are discussed by Twitter users plotted over time.

topic classifier on our scraped dataset, where we visualized two topics (vaccines and schools) over time. We also make the model available on the HuggingFace repository[5] for practitioners to use.

Classifying support for curfews. For the last classification model, we trained mBERT for multiclass classification on 1518[6] 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[7].

Classifying support for the government. As mentioned in Subsection 3.1, we also introduced a labeling axis on support for the government—which is different from support for a specific measure. After labeling, our data analysis revealed two trends: (i) the majority of tweets (82%) does not express any opinion on the government, which is very different from opinions on specific measures. (ii) Only 1.8% of the tweets is supportive of the government, whereas 12% is not. With these imbalances, most classification systems would struggle at best, which makes us question the usefulness of this classifier. Therefore, we opted to drop this labeling axis.

5. Available at `https://huggingface.co/DTAI-KULeuven/mbert-corona-tweets-belgium-topics`.
6. This were originally 2000 tweets of which the clearly irrelevant ones were filtered with Sieve I before labeling.
7. Available at `https://huggingface.co/DTAI-KULeuven/mbert-corona-tweets-belgium-curfew-support`.

4. Results

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 4. Also included, for reference, is the rate of confirmed COVID-19 cases in Belgium (Sciensano 2021).

On 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 in 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 negative 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. Interestingly, we also see that peaks in one polarity of the opinion about the measure, whether too strict or too loose, are not necessarily accompanied by a peak in the opposing opinion, as we might expect if increased discussion of the topic is due to an increase in contentious disagreement.

We also observed 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 humor as well as conveying complex points of views and personal stories (e.g. about the impact of the curfew). The ability of 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.

5. Discussion

We were able to characterize the discussion on Twitter about different Belgian COVID-19 measures over approximately 7 months using multilingual BERT models. Interestingly, we found that the fraction of tweets expressing opinion that the curfew measure was 'too strict' remained stable for the first months that the curfew was in place. We saw an increase in the fraction of tweets expressing this opinion in early 2021, corresponding with announcements that the curfew would be continued. However, the fraction of 'curfew too strict' tweets then dropped, even though the curfew remained in place. This provides evidence for the common perception that the opinions expressed on Twitter are highly impacted by salient current event discussions.

Following the recommendations of (Olteanu et al. 2016), in this section, we will address some of the lessons and limitations of our work. More specifically, we discuss (i) challenges we encountered during labeling in Subsection 5.1, (ii) limitations of the models in this paper in Subsection 5.2 and (iii) our thoughts on monolingual versus multilingual models for this setting in Subsection 5.3.

5.1 Challenges with labeling

As discussed in Subsection 3.2, we iteratively improved our code book during multiple feedback sessions between labelers. This methodology highlighted some issues with manually labeling data, such as missing or limited available context or subjectivity in labeling. In particular, tweets with hyperlinks or images or that are part of a long thread, pose additional challenges to correctly interpret within this context. Unraveling long threads is time consuming and following hyperlinks, which are obfuscated by Twitter's URL shortener, has some risks. We acknowledge that the point of view

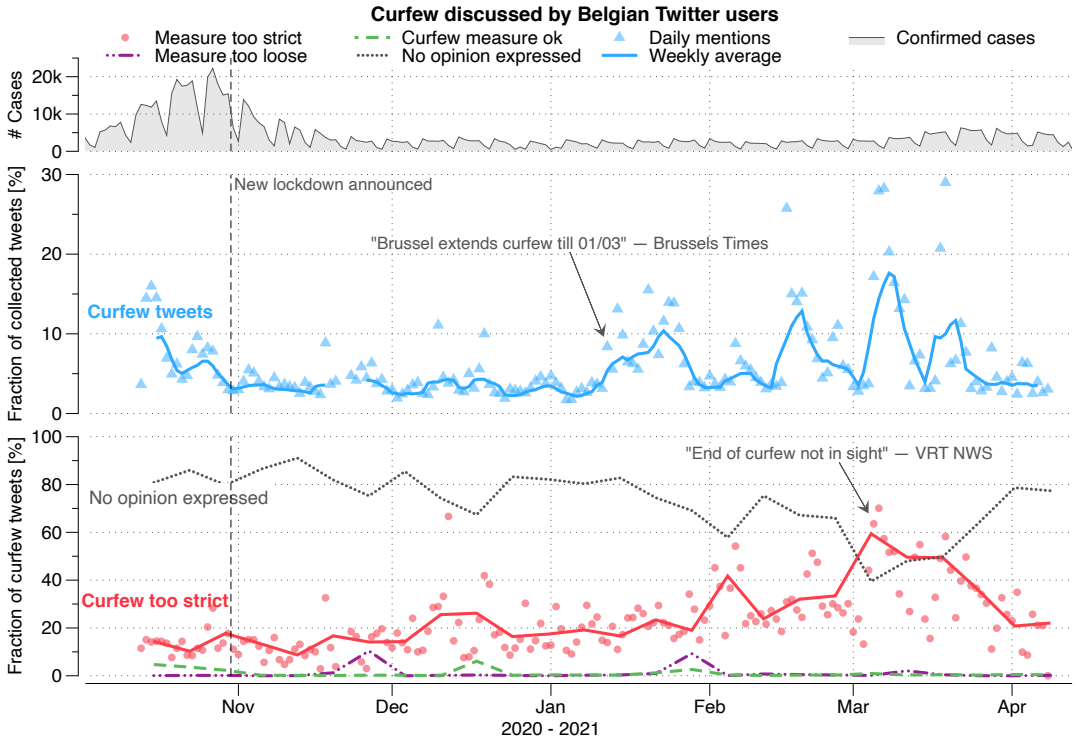


Figure 4: Timeline of the relative number of tweets on the curfew topic (middle) and the fraction of those tweets that find the curfew too strict, too loose, or a suitable measure (bottom), with the number of daily cases in Belgium to give context on the pandemic situation (top).

of our labelers—and ours—is not the universal one. The code book is available to indicate what decisions we made as to how we have defined the labels.

5.2 Limitations

Again, we reiterate that the Twitter activity measured and analyzed here are unlikely to be representative of the population of Belgium as a whole. Additionally, as we have demonstrated, the error rates of these models are unequal across categories, including across topics between languages. The issue of unequal performance among specific groups is likely to persist in any model that does not have 100% accuracy. In the context of using tweets to impact public policy, these differences can have real-world impact on which groups' opinions are able to have impact. Finally, if the models presented in this work were to be utilized to inform public policy, they would be highly vulnerable to gaming in their current state, through methods such as bot attacks or coordinated influence campaigns. This work does not claim to address those concerns.

5.3 On the trade-offs between monolingual and multilingual models

In this work, we created multilingual models focusing on Dutch, French and English using mBERT and XLM-RoBERTa. Despite an overall acceptable performance, there are differences in performance between languages. Noticeably, French tweets are significantly more misclassified, as highlighted in Figure 2b. One way of addressing this discrepancy, would be to create and use multiple monolingual

models for each language, such as RobBERT (Delobelle et al. 2020) for Dutch, CamemBERT (Martin et al. 2020) for French and BERT (Devlin et al. 2019) or RoBERTa (Liu et al. 2019) for English.

However, this approach would also reduce the number of training examples available for each model. Our labeled dataset has no extreme skew towards one language, but French (24%) is slightly underrepresented in comparison to Dutch (41%) and English (35%). This could lead to a further reduction of performance for the French model and all others, since only the monolingual training data is used. For this reason, we opted for multilingual models where the same representations can be used and we can benefit from easier, non-separated model development. Yet, further research in this trade-off might be warranted.

6. Conclusion and future work

We were able to observe the discussion on Twitter of Belgian COVID measures in three languages, over a 7-month time period. Using the models we developed, we observed some interesting patterns, like an increasing sentiment that the Belgian curfew was too strict, which also quickly faded away when cases rose again. However, we also found that the majority of the collected tweets did not express specific levels of support and thus we identified the need to characterize the nature of these non-opinionated tweets.

An important direction for future work is to understand the poorer performance of our model on opinions other than 'no opinion' and 'too strict'. While we do work with multiple languages, further work can be done to determine differential performance between languages, dialects and informal and slang texts. Given the small minority of German speakers in Belgium, introducing German tweets would have had detrimental effects on our data quality and model performance. However, future work should also aim to include German tweets when tracking support for corona-related measures in Belgium. Related to this, language-specific classifiers can potentially increase performance, although more general multilingual models have the benefit of using all English, Dutch and French data. A comparative study of these trade-offs would be an interesting research direction.

Acknowledgment

We thank Robin Cuypers for his assistance with data labeling. Kristen Scott was supported by the NoBIAS — H2020-MSCA-ITN-2019 project GA No. 860630. Pieter Delobelle was supported by the Research Foundation - Flanders (FWO) under EOS No. 30992574 (VeriLearn). Pieter Delobelle also received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. Bettina Berendt received funding from the German Federal Ministry of Education and Research (BMBF) – Nr. 16DII113f.

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/.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2019), Unsupervised cross-lingual representation learning at scale, *CoRR*. http://arxiv.org/abs/1911.02116.

Delobelle, Pieter, Thomas Winters, and Bettina Berendt (2020), RobBERT: a Dutch RoBERTa-based Language Model, *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, Online, pp. 3255–3265. https://www.aclweb.org/anthology/2020.findings-emnlp.292.

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.

Elish, M. C. and danah boyd (2018), Situating methods in the magic of Big Data and AI, *Communication Monographs* **85** (1), pp. 57–80. https://www.tandfonline.com/doi/full/10.1080/03637751.2017.1375130.

franzke, aline shakti, Anja Bechmann, Michael Zimmer, Charles Ess, and the Association of Internet Researchers (2020), Internet Research: Ethical Guidelines 3.0. https://aoir.org/reports/ethics3.pdf.

Garimella, Venkata Rama Kiran and Ingmar Weber (2017), A long-term analysis of polarization on twitter, *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.

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/.

Kıcıman, Emre (2012), Omg, i have to tweet that! a study of factors that influence tweet rates, *Sixth International AAAI Conference on Weblogs and Social Media*.

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.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019), Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692*.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot (2020), CamemBERT: a tasty

French language model, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, pp. 7203–7219. https://www.aclweb.org/anthology/2020.acl-main.645.

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.

Müller, Martin, Marcel Salathé, and Per E. Kummervold (2020), COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter, *arXiv:2005.07503 [cs]*. arXiv: 2005.07503. http://arxiv.org/abs/2005.07503.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kiciman (2016), Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. https://papers.ssrn.com/abstract=2886526.

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.