A COVID-19 news coverage mood map of Europe

Frankie Robertson
University of Jyväskylä
frankie@robertson.name

Jarkko Lagus
University of Helsinki
jalagus@cs.helsinki.fi

Kaisla Kajava
University of Helsinki
kaisla.kajava@helsinki.fi

Abstract

We present a COVID-19 news dashboard which visualizes sentiment in pandemic news coverage in different languages across Europe. The dashboard shows analyses for positive/neutral/negative sentiment and moral sentiment for news articles across countries and languages. First we extract news articles from news-crawl. Then we use a pre-trained multilingual BERT model for sentiment analysis of news article headlines and a dictionary and word vectors -based method for moral sentiment analysis of news articles. The resulting dashboard gives a unified overview of news events on COVID-19 news overall sentiment, and the region and language of publication from the period starting from the beginning of January 2020 to the end of January 2021.

1 Introduction

The response to the COVID-19 pandemic and its news coverage worldwide has been marked by tension between state-level actions, and those of regional organisations such as the European Union, and the essentially borderless nature of the virus itself. This paper presents the COVID-19 news coverage mood map of Europe which visualizes sentiment in news coverage in different European languages. Within the European context, with its many small languages, this multi-lingual approach is vital for working beyond the state-level.

In order to evaluate the possibilities of automatic sentiment analysis models in this context, we first created a multilingual COVID-19 news corpus from the websites of state broadcasters. We then applied two types of automatic sentiment analysis to the articles. Finally, we created a dashboard containing a number of interactive plots based on the analysed corpus[1]. This report details work undertaken at the EMBEDDIA Hackashop.

[1] http://covidmoodmap.rahtiapp.fi/

European languages	av az ba be bg bs ca ce cs cv cy da de el en es et eu fi fo fr fy ga gl gv hr hu is it kl kv la lb lt lv mk mt nb nl nn no oc os pl pt rm ro ru sk sl sq sr sv tr tt uk yi
Langdetect	bg ca cs cy da de el en es et fi fr hr hu it lt lv mk nl no pl pt ro ru sk sl sq sv tr uk
mBERT	bg ca cs cy da de el en es et fi fr hr hu it lt lv mk nl no pl pt ro ru sk sl sq sv tr uk
MUSE	bg ca cs da de el en es et fi fr hr hu it mk nl no pl pt ro ru sk sl sv tr uk
>20 items	bg ca cs cy de el en es et fi fr it lt mk nl pl pt ro ru sv tr uk

Table 1: Pictogram with ISO 639-1 language codes summarising language coverage of multilingual techniques along with our COVID-19 news corpus. Rows are subsetted from the row above the previous rule.

2 Corpus extraction

We used the news-please extractor (Hamborg et al., 2017) on news-crawl dumps to obtain a multilingual corpus of European COVID-19 news. News-crawl is a web crawl provided by the Common Crawl organisation which is updated more frequently and contains only data from news websites[2]. In order to keep the size of the corpus manageable and the extraction time reasonable, a list of internet domain names of European state broadcasters was first obtained from Wikidata, since filtering at the domain level allows for faster processing of Common Crawl dumps. Articles without a language detected by the langdetect language detector[3] were discarded. COVID-19 keyword filtering was also applied, detailed in Section 3.1. Table 1 includes a summary of the considered European languages, their support by langdetect, and the set

[2] https://commoncrawl.org/2016/10/news-dataset-available/

[3] https://github.com/Mimino666/langdetect

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 110–115

for which 20 or more items were ultimately extracted.

The resulting corpus contains 468 thousand articles. It is thus just over a quarter of the size of the comparable AYLIEN Coronavirus News Dataset[4] corpus which has 1 673 thousand news articles. Our corpus contains news from a longer period of just over a year versus AYLIEN's, which contains just over a half a year. Most importantly, our corpus has at least 20 items in 22 languages, versus AYLIEN's corpus which is English only.

The full corpus does not include news from all European countries. Table 2 gives a coverage of the countries included in the corpus, while Table 1 gives the coverage of languages in the corpus. There are two possible reasons for the missing countries. The first is that news-crawl uses a fixed set of seeds, so one possibility is that the the websites of the state broadcaster for these countries was not on the seed list. Another possibility is that the article extraction of news-please was not able to deal with these countries. The recall of news-please is estimated at 71%. Possible future work here would be to obtain and audit the seed list from news-crawl and try other article extraction software such as Trafilatura (Barbaresi, 2019) which has an estimated recall of 88% with higher precision.

Covered	AT BE BG CZ DE EE ES FI FR GB GR IE IT LT MD NL PT RO SE SM VA
Missed	AD AL AX BA BY CH CY DK FO GG GI HR HU IM IS JE LI LU LV MC ME MK MT NO PL RS RU SI SJ SK UA

Table 2: Pictogram with ISO 3166-1 alpha-2 country codes summarising European countries covered (>20 items extracted) and missed in our corpus.

3 Analyses

All analyses were multilingual, and their coverage of different languages is compared in Table 1. During development, the full list of tools and resources given by Pollak et al. (2021) was considered.

3.1 Keyword matching

In order to detect keywords from fixed lists across languages, including those with inflectional endings which cause changes to the citation form itself, either lemmatisation or stemming must be performed. Since keyword search is also performed as a filtering step for creating the corpus, it should be fast. To achieve this goal, we applied a simple high-recall stemming-like scheme based on BPE. First we obtained the pretrained BPE (Sennrich et al., 2016) model from XLM-RoBERTa-large (Conneau et al., 2019)[5]. As a next step, all BPE tokens with length $>= 5$ are discarded from the BPE model. The full XLM tokenisation pipeline is then run as normal. The hope is that this segments the word into commonly repeating units, which are likely to include common inflectional endings. The decision to discard longer BPE tokens was made so that common longer words would still be segmented and to bound the maximum number of characters removed from the word, since removing too many is more likely to cause false positives. In cases where at least 3 BPE segments were generated, the last one is discarded. In all cases, the resulting stem or full token is at least 2 BPE characters and 5 characters, a wildcard appended to the end of the token. Matching was performed case-insensitively using the fast *pyre2*[6] library, which uses deterministic automata for matching.

While this scheme is certainly likely have lower precision than using a high quality lemmatiser, it is not language dependent beyond its central assumption: that sounds changes – if they occur – are limited towards the end of the word. The exact same procedure is applied to all languages. For comparison the Snowball stemmer[7] supports 15 languages, leaving 16 of those supported by langdetect unhandled. On the other hand, a state-of-the-art mulitlingual lemmatisers such as the Universal Lemmatiser of Kanerva et al. (2020), which supports over 50 langauges is likely to be slower. Additionally, due to Universal Lemmatiser's architecture being adapted to batch scenarios, this implies adding and extra stage to the pipeline. That said, performing keyword matching based upon Universal Lemmatiser would be a good next step for the keyword matching, and the current scheme could be kept only for the few languages not supported by Universal Lemmatiser.

Keyword lists for both COVID-19 keywords and names of European countries in all considered languages were obtained from Wikidata. For matching the topic of COVID-19, labels from the entities `Q84263196` (the COVID-19 disease), `Q82069695` (the SARS-CoV-2 virus), and

[4] https://aylien.com/blog/free-coronavirus-news-dataset

[5] https://huggingface.co/xlm-roberta-large

[6] https://github.com/andreasvc/pyre2

[7] As described at https://snowballstem.org/

Q89469904 (the hypernym of SARS-CoV-2; all corona viruses) were used. For non-English languages, the set of keywords was extended with the English keywords in case they have been used as loans, especially for example in the early stages of the epidemic. In addition, the commonly occurring trans-lingual patterns: corona*, korona*, covid* and корона* were added to all lists. The lists of country names were used as-is.

3.2 Multilingual sentiment analysis of news headlines

In recent years, deep pre-trained language models have produced state-of-the-art results in various natural language processing tasks. BERT models use self-attention layers from the transformer model (Vaswani et al., 2017), which makes them useful for detecting contextual information in text: this BERT does by looking at the neighboring words on both sides of each word in the text.

We used a multilingual BERT model (Devlin et al., 2018) fine-tuned by Pelicon et al. (2020) to classify news article headlines into the polar sentiment categories *positive*, *neutral*, and *negative*. The model was trained on a Slovenian news dataset and evaluated for zero-shot cross-lingual tasks on Croatian news. As the model was originally developed for classification of full news article texts, it is mostly trained on slices of these using the whole of BERT's maximum sequence length of 512. Here we are typically supplying much shorter sequences from the related, but distinct distribution of news headlines, which is likely to affect performance.

The results of our experiments show a near-consistent peak in news coverage of the COVID-19 pandemic in the spring of 2020 across news sources, with the exception of some languages which are less represented in our overall news dataset. Similarly, the results show a peak in news coverage in the fall of 2020, which coincides with the second wave of the pandemic. *Negative* sentiment is most prevalent in spring 2020. Overall, *negative* was the most commonly predicted label, with *neutral* as the next common, and *positive* the least common. Table 3 shows the five countries with the most news article headlines classified as *negative*. News articles from the national broadcasters of those countries also constitute a significant portion of the overall data.

Country	*Positive*	%	*Neutral*	%	*Negative*	%
Spain	22033	19	40100	36	50092	45
France	17484	17	32717	33	49836	50
U.K.	10233	17	21746	35	29261	48
Germany	9375	15	25256	40	27941	45
Belgium	5626	19	10531	35	14030	46

Table 3: Five countries with the highest absolute number of news headlines predicted as *negative*.

3.3 Multilingual moral sentiment analysis of news articles

Research shows (Kalimeri et al., 2019a; Curry et al., 2019; Mooijman et al., 2018) that accounting for moral sentiment in addition to other personality traits in natural language can give insights into many different societal phenomena. Methods based on moral sentiment have been previously used, for example, to predict street riots based on Twitter data (Mooijman et al., 2018) and analyze moral narratives in social media conversation about vaccination (Kalimeri et al., 2019b).

The basis for moral sentiment analysis is the *moral foundations theory* by Graham et al. (2013) and especially in this case the moral foundations dictionary (MFD) 2.0[8]. MFD is a word list that contains words categorized into five different sentiments: care, fairness, loyalty, authority, and sanctity.

The method we apply here follows the idea described by Kozlowski et al. (2019) in order to extract cultural dimensions from word embeddings. First, we extract antonym pairs using WordNet (Fellbaum, 1998) per sentiment by searching synsets word-by-word and fetching the list of antonyms. This list of antonyms is then filtered based on the list containing words of opposite sentiment polarity. As an example, we search antonyms for word the "peace" from the list of positive polarity words related to care sentiment. We acquire the antonym "war" based on the WordNet synset search. As this word is also found from the negative polarity list of care dimension, we add this pair to the list of antonym pairs for the care sentiment.

These pairs we use to compute vectors representing moral sentiment dimension. This is done via encoding the words as word embeddings, doing simple subtraction of these antonym pairs, and computing the mean of these vectors per sentiment. This gives us one vector per sentiment, onto which

[8] https://osf.io/ezn37/

we can then project any other word to measure the strength of that specific sentiment. As large number of words do not have antonyms that are found from the opposite polarity list, this leads to a noisy estimate of the dimension.

To compute the sentiment of a document, each word of a document is first encoded into word embedding and then projected onto each moral sentiment vector. This gives a list of scores for each word that we can then just sum to obtain the final score for the document.

This way of summing everything up creates an effect where longer documents will have stronger sentiment scores if most words are towards the same polarity per sentiment. This is a somewhat desirable effect since we do not wish to have one-sentence articles to have the same weight as a full article.

Multilinguality creates issues since the original MFD is only officially available in English. The issue is solved by using aligned word embeddings and doing approximate translation via distance-based search in the aligned word embedding space. For this, we use embeddings by Conneau et al. (2017). An alternative option would be to translate the words exactly, but since all languages might have very culture-specific semantics that do not directly translate, it might remove or even change the moral sentiment of the word. Doing approximate translation directly in word embedding space does not suffer from this, as we only search for the word embedding with the closest meaning. This should also preserve the sentiment information.

The distance-based search method does not guarantee a good translation, but should in most cases work better than exact translation. If the exact translation is good and the embedding is close to the actual meaning, the distance-based method will approximately retrieve the same embedding. If the exact translation is good, but because of cultural differences the semantic meaning has shifted, the distance-based method should retrieve an embedding that is closer to original semantics, even though the exact word might be different from the exact translation. So in both of these cases, the distance-based method should yield better results.

The results show that the strongest sentiment in the positive direction was sanctity and in the negative direction loyalty, both being clearly distinct from the other three sentiments in magnitude. Different sentiments showed fluctuation over time and countries, but overall the sentiment seemed to stay in the same polarity, suggesting no drastic changes in the way COVID-19 was covered in the news over time from the moral sentiment point of view.

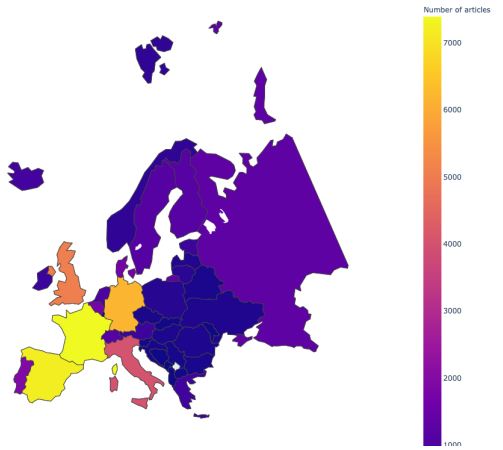

Figure 1: An example choropleth map showing the positive sentiment distribution over countries mentioned. Note that the purple color at the lower end of the spectrum does not indicate a high amount of negative news, but just the lack of positive news.

4 Visualisation

In order to visualise the results of the analyses, we created a dashboard using the Dash framework[9]. The resulting application makes heavy use of analytical queries which tend to feature range selections and grouping based on dates as well as numerical aggregates such as value summations and counts. To run these queries efficiently across the whole data set of 468 thousand articles, we used the DuckDB column database (Raasveldt and Mühleisen, 2019).

The chief dimensions visualised as independent grouping variables in space were date and either the country of production or the country mentioned in articles. For plots in which these variables could not be shown spatially, the user was given the option of filtering using them. The language used in the articles and type of sentiment were also available as filters. The main dependent variables were are either raw article counts or lumped measures. For visualising only time as a visual grouping, a bar chart was used, while for visualising only country, a choropleth map, an example of which is given in Figure 1. Finally, an animated choropleth map showing consecutive time slices corresponding to weeks taken Monday to Sunday groups by both

[9] https://plotly.com/dash/

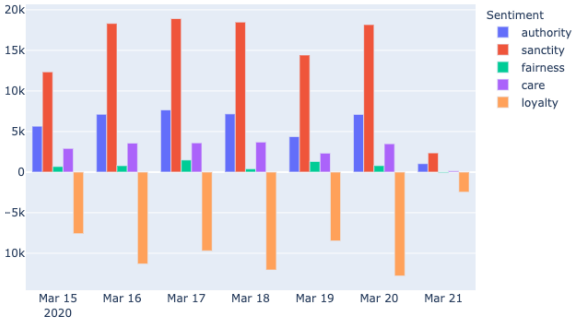

Figure 2: An example bar chart of one week of moral sentiment scores. Height indicates the strength of the sentiment component and polarity tells if it is positive or negative sentiment along that dimension.

time and country.

The lumped measure used for polar sentiment is a simple ratio, with the *neutral* class included to pull the measure towards zero:

$$\frac{\text{positive} + \frac{1}{2}\text{neutral}}{\text{positive} + \text{neutral} + \text{negative}}$$

For moral sentiments, we have already obtained a measure for each document and each moral sentiment. These are simply summed to create a single aggregate bipolar measure of sentiment strength per moral sentiment (see Figure 2 for an example). The sentiment estimate is rather noisy, so looking at the absolute values is not recommended. A better way to look at these numbers, is to look at them in relation to other countries or time spans. This tells how different countries differ in the way they represent this information and how is the overall trend progressing over time.

5 Conclusion

We have presented a COVID-19 news dashboard for exploration of reporting across time and from and about different countries during the COVID-19 epidemic. The dashboard demonstrates the potential of automated multilingual text analysis for understanding reporting on complex phenomena such as the COVID-19 crisis beyond the state-level. This type of tool could be integrated into a system used by news agencies to track news trends. Beyond COVID-19, it could be used to plan coverage of other national or global news events such as elections, international summits, or sports events.

The visualizations in the dashboard do seem to line up with the authors' preconceived ideas about sentiments during COVID-19 and their evolution. However, all analyses in the dashboard were produced automatically and have not undergone evaluation within this context. Since the analyses themselves may not be entirely accurate, the resulting plots may be misleading, and thus should not be used as a basis for decision making. Evaluation of the underlying techniques is a clear next step for this work.

Acknowledgements

The authors wish to thank the organisers of the EMBEDDIA Hackashop 2021 for organising the event and CSC – IT Center for Science, Finland, for computational resources.

References

Adrien Barbaresi. 2019. Generic web content extraction with open-source software. In *KONVENS 2019*, pages 267–268. GSCL.

Alexis Conneau, 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. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

O Curry, H Whitehouse, and D Mullins. 2019. Is it good to cooperate? testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, 60(1).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. volume 47 of *Advances in Experimental Social Psychology*, pages 55–130. Academic Press.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

Kyriaki Kalimeri, Mariano G Beiró, Matteo Delfino, Robert Raleigh, and Ciro Cattuto. 2019a. Predicting demographics, moral foundations, and human values from digital behaviours. *Computers in Human Behavior*, 92:428–445.

Kyriaki Kalimeri, Mariano G. Beiró, Alessandra Urbinati, Andrea Bonanomi, Alessandro Rosina, and Ciro Cattuto. 2019b. Human values and attitudes towards vaccination in social media. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 248–254.

Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2020. Universal lemmatizer: A sequence to sequence model for lemmatizing universal dependencies treebanks. *Natural Language Engineering*, pages 1–30.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. 2018. Moralization in social networks and the emergence of violence during protests. *Nature human behaviour*, 2(6):389–396.

Andraž Pelicon, Marko Pranji'c, Dragana Miljkovi'c, Blaž Škrlj, and Senja Pollak. 2020. Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, 10(17):5993.

Senja Pollak, Marko Robnik Šikonja, Matthew Purver, Michele Boggia, Ravi Shekhar, Marko Pranjić, Salla Salmela, Ivar Krustok, Tarmo Paju, Carl-Gustav Linden, Leo Leppänen, Elaine Zosa, Matej Ulčar, Linda Freienthal, Silver Traat, Luis Adrián Cabrera-Diego, Matej Martinc, Nada Lavrač, Blaž Škrlj, Martin Žnidaršič, Andraž Pelicon, Boshko Koloski, Vid Podpečan, Janez Kranjc, Shane Sheehan, Emanuela Boros, Jose Moreno, Antoine Doucet, and Hannu Toivonen. 2021. EMBEDDIA tools, datasets and challenges: Resources and hackathon contributions. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.

Mark Raasveldt and Hannes Mühleisen. 2019. Duckdb: an embeddable analytical database. In *Proceedings of the 2019 International Conference on Management of Data*, pages 1981–1984.

Rico Sennrich, B. Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *ArXiv*, abs/1508.07909.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.