# Press Freedom Monitor: Detection of Reported Press and Media Freedom Violations in Twitter and News Articles

**Tariq Yousef** [1] **Antje Schlaf** [2] **Janos Borst** [3]

**Andreas Niekler** [3] **Gerhard Heyer** [1]

[1]Institute for Applied Informatics (InfAI)
[2]European Centre for Press and Media Freedom (ECPMF)
[3]Leipzig University

{yousef,heyer@infai.org} {antje.schlaf@ecpmf.eu} {borst,aniekler@informatik.uni-leipzig.de}

## Abstract

Freedom of the press and media is of vital importance for democratically organised states and open societies. We introduce the *Press Freedom Monitor*, a tool that aims to detect reported press and media freedom violations in news articles and tweets. It is used by press and media freedom organisations to support their daily monitoring and to trigger rapid response actions. The *Press Freedom Monitor* enables the monitoring experts to get a swift overview of recently reported incidents and it has performed impressively in this regard. This paper presents our work on the tool, starting with the training phase, which comprises defining the topic-related keywords to be used for querying APIs for news and Twitter content and evaluating different machine learning models based on a training dataset specifically created for our use case. Then, we describe the components of the production pipeline, including data gathering, duplicates removal, country mapping, case mapping and the user interface. We also conducted a usability study to evaluate the effectiveness of the user interface, and describe improvement plans for future work.

## 1 Introduction

Press freedom is under constant and increasing attack, even in Europe. Therefore, now more than ever, it is important to monitor developments and advocate for measures to protect press and media freedom. Mapping Media Freedom[1] (MMF) is a project and platform which identifies and documents threats, violations and restrictions faced by media workers across Europe and beyond. The documented incidents include physical attacks, threats of violence made online and offline, legal actions aimed at silencing critical coverage and moves to block access to information or reporting on incidents or denying access to independent and government-critical media platforms. These incidents are published as alerts on MMF and combined with analysis reports they provide an overview of the current state and development of press and media freedom in Europe. This project is run by the Media Freedom Rapid Response (MFRR[2]), a rapid response mechanism against press and media freedom violations in the European Union member states and candidate countries[3]. It provides legal support, shelter, public advocacy and information to protect journalists and media workers. The alliance is led by the European Centre for Press and Media Freedom (ECPMF) in conjunction with ARTICLE 19, the European Federation of Journalists (EFJ), Free Press Unlimited (FPU), the Institute for Applied Informatics at the University of Leipzig (InfAI), International Press Institute (IPI) and CCI/Osservatorio Balcani e Caucaso Transeuropa (OBCT). The project commenced in 2020. It is funded by the European Commission. The MMF alerts guide MFRR to directly engage with and help at-risk journalists and media workers The alerts are submitted mainly by the MFRR monitoring experts, as well as an international network of local partners. However, MMF is also a crowdsourced platform that enables anyone to upload an alert, which is then verified by the expert network before publication to guarantee the reliability of the cases and the comprehensiveness of the published details. In order to support the labour-intensive manual monitoring of incidents, we developed the *Press Freedom Monitor* which regularly monitors and automatically detects reports about press and media freedom violations in vast amounts of online published text sources, namely news articles and tweets. The automatic detection is based on a trained deep learning model. These detected incident reports are then verified by the monitoring experts who create and publish an MMF alert in-

[1]https://www.mappingmediafreedom.org

[2]https://www.mfrr.eu

[3]Further referred here as "MFRR" region



*Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 153–159
November 7–11, 2021. 

cluding comprehensive details and trigger further response actions.

The advantage of integrating automatic extraction processes is that they can use a wider range of sources and provide a faster alert mechanism. This in turn means that more violations can be found and more sources can be provided for each case, allowing a more realistic and reliable assessment of the press freedom situation. Table 1 shows some examples of tweets or news headlines which are considered to be reports of attacks and violations and thus of interest for the experts. We divide our work into a training phase where we describe the data collection and the evaluation of the training data created specifically for our use case, and a production phase where we describe details of the architecture.

| |
|---|
| This is a #BBC reporter being harassed and chased by a mob. Scary stuff. |
| Orban-friendly owner gets Hungary independent radio frequency |
| #Serbia's #Govt #Group #condemns death #threat to #writer and #journalist |
| A new law in #Germany makes journalists vulnerable to hacking and surveillance. |
| #Italy #Lazio Incredible, the regional administrative court (TAR) order journalists to reveal their sources! |
| NEWS \| NI journalist Patricia Devlin has - for the second time - received a threat via social media to rape her young son |
| Demonstrators attack, obstruct #journalists covering #protests against #COVID19 lockdown in #Germany |

Table 1: Examples of Tweets and News Headings Considered as reported Attacks.

## 2 Related Work

Monitoring content published in social networks to detect abuse, harassment, or freedom violations has been the subject of several research projects. (Hewitt et al., 2016), (Anzovino et al., 2018), (Şahi et al., 2018), and (Rodríguez-Sánchez et al., 2020) worked on the detection of misogynistic language and hate speech towards women on Twitter. (O'Dea et al., 2015) developed an approach to automatically detect suicide-related tweets. (Bourgonje et al., 2018) and (Rodríguez-Sánchez et al., 2020) aimed to detect and analyse tweets that contain racism, sexism and abusive language. The starting point was always to define a list of keywords, hashtags, and accounts related to the topic and to collect the relevant tweets using Twitter API. However, the absence of domain-specific labelled corpora drove many projects to manually create their own training dataset to meet their needs.

All the works listed above consider historic or archived data, whereas our crawling process runs continuously and collects almost live tweets and news multiple times a day. Furthermore, our project is distinguished from others by not aiming to detect direct abuses, harassment or freedom violations but reported attacks and violations directed to the specific group of media actors, which include journalists, media workers and media companies.

## 3 Training Phase

### 3.1 Data Collection

For the data gathering we use the free Twitter API[4] and the free version of NewsAPI[5]. This enables us to detect violations reported by various kinds of stakeholders, including official accounts with high publicity as well as private accounts, which can be especially helpful in countries where press freedom is restricted, and violations do not get reported in publicly available news media. The filters to query the Twitter API and the NewsAPI were defined in collaboration with the monitoring team at ECPMF, EFJ and IPI. We defined 147 keywords and hashtags based on three groups: (A) hashtags that are directly related to violations of press and media freedom[6], (B) keywords and hashtags related to media actors[7] and press and media freedom[8] and (C) keywords and hashtags related to attacks or violations[9]. The APIs were queried to require a match from a group A element or a combined match of group B and C elements. This combination was defined in order to exclude general content on media actors and press freedom as well as attacks that were not related to press and media freedom. Based on the experience of the monitoring experts, we selected an additional 66 Twitter accounts which frequently report violations

[4] developer.twitter.com

[5] https://newsapi.org

[6] Such as #journalismisnotacrime or #JournoSafe

[7] Such as editor, journalist, reporter, photographer, camera team, blogger, whistleblower, journalism, media company

[8] Such as #mediafreedom, #pressfreedom

[9] Such as arrested, attacked, censored, blocked access, defamation, harassed, insulted surveillance, threatened

| | Training | | | Validation | | |
|---|---|---|---|---|---|---|
| Model | Precision | Recall | F1 | Precision | Recall | F1 |
| Logistic Regression | 0.5802 | 0.7333 | 0.6479 | 0.9482 | 0.6868 | 0.7966 |
| Decision Tree | 0.5486 | 0.6494 | 0.5947 | 0.8809 | 0.5024 | 0.6398 |
| k-Nearest Neighbors | 0.6372 | 0.6000 | 0.6180 | 0.8799 | 0.4421 | 0.5885 |
| Linear SVM | 0.6640 | 0.6861 | 0.6749 | 0.9627 | 0.5889 | 0.7308 |
| Random Forest | 0.8426 | 0.4106 | 0.5521 | 0.9932 | 0.2286 | 0.3716 |
| Transformer | 0.7768 | 0.8274 | 0.8010 | 0.9736 | 0.7659 | 0.8571 |
| Transformer + CNN | 0.7769 | 0.8324 | **0.8036** | 0.9559 | 0.8059 | **0.8785** |

Table 2: Training and Validation Results for the Proposed Methods.

of press and media freedom.

The data collection process runs multiple times a day and collects approximately 1,000 news articles and 4,000 tweets per day.

### 3.2 Training

We aimed to train a binary classifier to be able to tell if a tweet or news article is reporting about press and media freedom violations or not. Due to the absence of any publicly available training dataset related to our purpose, we created our own training data with the help of an annotation tool developed particularly for this purpose.

ECPMF manually classified 6,005 news articles and 8,192 tweets. Around 26% of them are classified as relevant. The inter-annotator agreement was assessed for 997 news articles and 996 tweets classified by two different human annotators. We achieved a relative agreement of 84.55% for news and 86.35% for tweets, and a substantial agreement regarding Cohen's kappa (Cohen, 1960) with 0.62 for news and 0.63 for tweets. We excluded from the dataset the examples where the human annotators disagreed.

In the training process, the models were trained on news articles and tweets together. The size of the training dataset was 13,809 texts (5,822 news articles, 7,987 tweets) in total. The models were trained on 90% of the training dataset and tested on the remaining 10%. We trained each model on four different training/testing splits and guaranteed approximately the same percentage of examples from each category. The models were validated on 1,007 feedback items of unseen data created manually by the monitoring experts.

We evaluated several classic machine learning models such as Logistic Regression, Decision Tree, k-Nearest Neighbors, Linear SVM, and Random Forest using TF-IDF representation. However, The results were not satisfactory. Additionally, we experimented with convolutional neural networks (CNN) trained on top of a distilled (Sanh et al., 2019) version of the RoBERTa (Liu et al., 2019) model in addition to the vanilla finetuning of the transformer model. The structure of the CNN follows the architecture in (Kim, 2014). Table 2 shows the results. As we can see, the deep learning models clearly outperform the classic machine learning approaches. The CNN model achieved the highest average *f1-score* and *recall* during the training and the validation on the new unseen texts and was selected to be deployed and integrated into our pipeline.

## 4 Production Phase

The aim of the *Press Freedom Monitor* as an application in production is threefold: First, and most importantly, it should constantly monitor news and tweets and present automatically-detected reports about press freedom violations to the monitoring experts in a convenient format. Second, the tool should provide support for the monitoring experts when searching for further items that report about the same incident during the verification process. Third, it should help to improve the model by collecting manual feedback about the classified items, which can be used as additional training data.

Figure 1 shows the process pipeline of the *Press Freedom Monitor* in production. The data collection is performed continuously as described above. The trained model is used to detect reported violations within the gathered data. In parallel, duplicate removal, case mapping, and a country mapping based on geocoding is performed to increase the usability of the tool.

### 4.1 Geocoding and Country Mapping

MFRR is mainly interested in incidents happening in European Member States and Candidate coun-

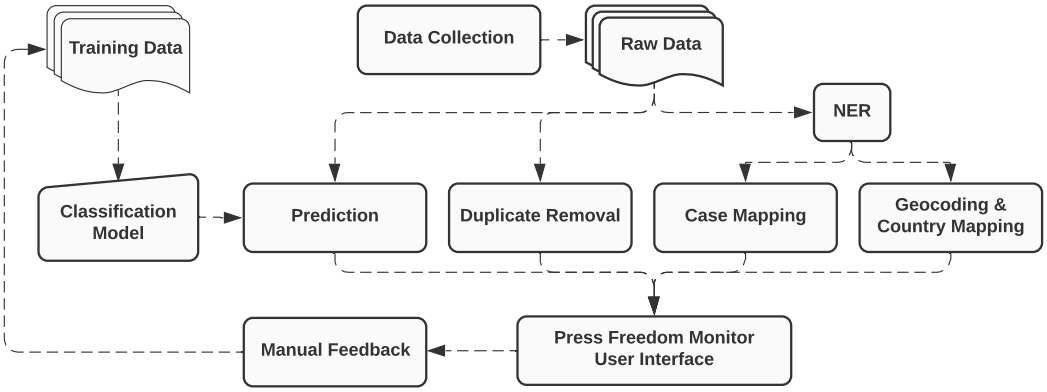


Figure 1: System Architecture.

tries. As the location of the author might be completely unrelated to the location of the reported incident, we aimed to identify locations mentioned within the text. We use SpaCy[10] to detect locations via named entities recognition and translate them to latitude-longitude coordinates using the OpenStreetMap API[11]. The coordinates can then be mapped to countries and to the region of interest. This country mapping was further extended by recognising country adjectives (such as *Italian* or *French*) as well as country names mentioned in hashtags.

### 4.2 Duplicate Detection

Text reuse is very common among news agencies (Clough et al., 2002). Similarly, twitter users tend to reuse texts posted by other users or republish their own tweets (Castillo et al., 2011). Since these duplicates do not provide an added value to our purpose, but require more resources for analysis and review, they had to be removed from our production pipeline. For this purpose, we employed Locality Sensitive Hashing (LSH) (Rajaraman and Ullman, 2011) based on min-hash signatures and the Jaccard similarity.

As a similarity threshold, we chose 95% to tolerate minor differences. This guided us to detect 1.38% of the crawled news articles as duplicates. Moreover, the fraction of duplicates in the collected tweets was more significant, around 7.9%. The duplicate removal process runs parallel to the data collection process multiple times a day, and it considers all the texts that have been crawled in the last ten days.

[10]https://spacy.io

[11]https://nominatim.org

### 4.3 Case Mapping

It is typical to find multiple news articles published by several news agencies reporting on the same incident or publishing updates on previous incidents. Similarly, numerous people post tweets on the same incident. Thus, when verifying a certain reported incident, it would be helpful to see related items that are reporting about the same incident. We call this process "Case Mapping". For this purpose, we employed semantic similarity to capture meanings relatedness between each pair of texts even if there are no exact matches among the tokens. Moreover, we used SpaCy, which creates embeddings for each text by averaging the embeddings of all tokens. Then it employs cosine similarity between the two vectors to compute the similarity score. We set 0.97 as the threshold score to decide if two texts are related or not. We chose this value experimentally after analysing the data we have. For now, related news articles/tweets are listed for every single item in the front end and can be accessed by clicking on the button *Show Related Articles/Tweets*.

### 4.4 User Interface

The user interface as shown in figure 2 is implemented as a web application with restricted access via login. The frontend is designed with regard to the threefold aim of the final tool described above. First, it presents the latest news articles and tweets that have been classified as relevant, showing the most recent at the top. A click on the item shows the full article or tweet in its original context. Furthermore, it provides several convenience filters:

A date filter allows the user to restrict the items to certain time spans. The default threshold for

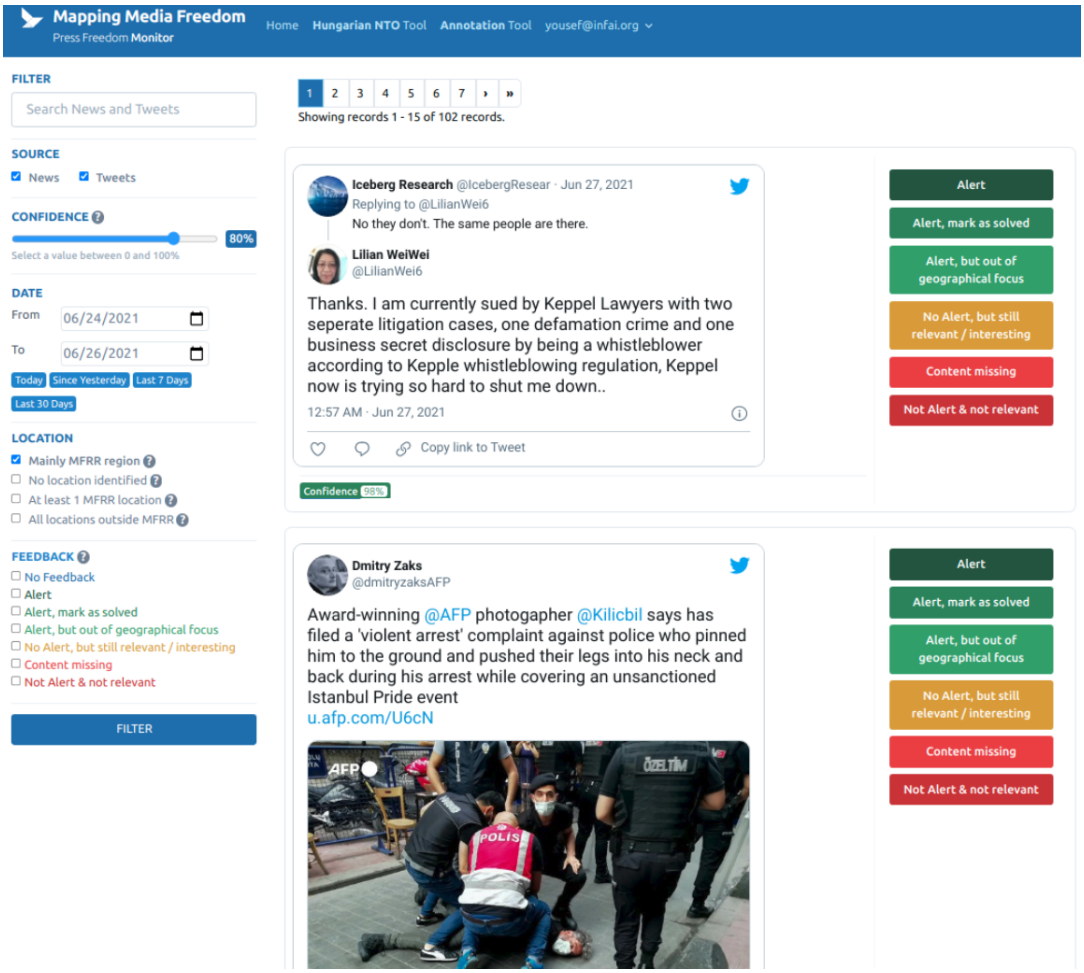


Figure 2: Press Freedom Monitor User Interface.

the prediction confidence is set to 80% and can be adjusted via a slider based on personal preferences connected to the time available to invest in order to find relevant incidents. We developed a location filter to support the filtering for MFRR's region of interest based on the country mapping described above. As multiple countries or also no countries might be mapped to a text, the location filter groups them according to the number and proportion of locations mentioned that fall within the MFRR region or outside the MFRR region. Though all groups might contain incidents within the region of interest, they differ highly in their contained portion. Multiple choice checkboxes allow the expert to adjust the items that are presented, based on their personal preferences and time available.

The second aim of the tool is to provide support for the monitoring experts when they verify an incident and need multiple sources reporting about this incident. If the case mapping analysis described above identifies items that are similar, a button named *Show related tweets/articles* is shown and presents the related items via a pop-up when clicked. The search field can further help to find items for a specific incident when e.g. searching for names of persons involved in the incident.

The third aim of the tool is to collect manual feedback to use it as further training data in order to extend the training dataset and to prevent topic drift. Therefore we implemented feedback buttons which are shown directly beside the item, and which allow the experts to give feedback in a convenient way during their daily work. Beside the possibility to rate an item as being relevant with regard to reporting about a press freedom violation (green coloured feedback buttons) or not (red), there is a feedback button for *No alert but still relevant/interesting* (orange). Items rated as the latter, mainly contain news, events or statements about press freedom violations in general, which can neither be rated as reporting about an explicit incident nor as completely irrelevant. These contents were excluded from the evaluation and will be the subject of discussions with the monitoring experts on whether

or how to include such content in the future. As the items can also be filtered based on their feedback, the feedback can also support the workflow of multiple users. One expert can label the latest detected items, whereas the other expert can filter just for the items already rated as relevant and can concentrate on verifying only these incidents. All feedback manually labeled via the green feedback buttons (reporting about a press freedom violation), can be used as positive examples for future training. However, this positive feedback can be further distinguished by the experts to differentiate between relevant and irrelevant items regarding other aspects of their workflow.

### 4.5 System Evaluation

After the tool was implemented, the front end was used to evaluate our real case scenario of content that held the most interest for the monitoring partners, namely content classified as relevant by the model and reporting about press and media freedom violations within the MFRR region. However, the evaluation went beyond this by manually applying different filter settings to the filters described before. This was performed in order to also include some evaluation of items not classified as relevant by the model and potentially missed, as well as items without detected locations or located outside MFRR countries.

Altogether, ECPMF evaluated 2,572 items via the implemented tool. This evaluated data contains 62% of items classified as relevant by the model and 32% classified as not relevant based on the default confidence threshold of 0.8. *Recall* is important to detect as many reported incidents as possible. A low *precision* would lead to too much evaluation effort for the monitoring partners. Using the keyword filter only, we achieved a baseline of 0.64 for precision and 0.31 for recall. When analysing the evaluated data for a confidence threshold of 0.8, the trained model achieved a *recall* of 0.87 and a *precision* of 0.96 regarding the data that are relevant to our interest. The evaluation showed that the lowering of the threshold from 0.8 to 0.5 would lead to a higher *recall* of 0.91 whilst retaining a *precision* of 0.94, which is still excellent.

## 5 Usability Study

To evaluate the usability of the *Press Freedom Monitor*, we conducted a usability study with 7 participants. For this purpose, we used the Computer System Usability Questionnaire (CSUQ) version3 (Lewis, 1995). The study showed a 79% overall satisfaction among the monitoring experts; similarly, the system usefulness achieved 79%, and information quality 76%. Moreover, the user interface quality achieved the highest score with 84%.

## 6 Language extension: Hungarian

By monitoring English language content, we already have broad geographical coverage. Extending the monitoring with additional languages can detect incidents not detected via the monitoring in the English language. We selected Hungarian as the test language, based on the ongoing deterioration in the field of press and media freedom in Hungary. The creation of specific training data for additional languages and the adaptation to language-specific processes (such as NER or geocoding) was not feasible in our project. Instead, we use Google Translate[12] to translate the texts to English and classify them with the model trained for English. To cope with the challenge that the monitoring experts might not speak Hungarian and therefore can not understand the news articles or tweets that are presented, the frontend also shows the translated text for each item first. This allows the experts to assess the content even if they do not speak the language. Again, each item can be clicked to show it in its original context and language for further inspection. Though the Hungarian version detects less content than the English one based on the smaller amount of news and tweets in Hungarian, it has already proved to be a valuable source of information during initial tests and usage. It detected an incident in Hungary and an incident in Germany that were not already known to the monitoring team and also not detected by the English monitoring. The first evaluation of 237 items resulted in a *recall* of 0.8 and *precision* of 0.95 when setting the confidence threshold to 0.5.

## 7 Future Work

Future work includes the extension of the manual feedback as well as the retraining of the models including this feedback data. A common improvement request by the experts was to further invest in displaying texts about the same incident together. Thus, future work will involve clustering based on incidents to enhance the case mapping performance. However, we need to evaluate how

[12] https://cloud.google.com/translate

well clustering can perform in this already narrow use case with high similarities between different incidents. As the first step, we plan to use the country mapping to present the incidents based on mapped countries. This is already expected to provide a better distinction between different incidents and might result in a first improvement with a good cost-benefit ratio. Furthermore, we want to extend the *Press Freedom Monitor* with additional languages.

**Acknowledgment**

This work is funded by the European Commission within the Media Freedom Rapid Response project and co-financed through public funding by the regional parliament of Saxony, Germany.

## References

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems*, pages 57–64, Cham. Springer International Publishing.

Peter Bourgonje, Julian Moreno-Schneider, Ankit Srivastava, and Georg Rehm. 2018. Automatic classification of abusive language and personal attacks in various forms of online communication. In *Language Technologies for the Challenges of the Digital Age*, pages 180–191, Cham. Springer International Publishing.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. WWW '11, page 675–684, New York, NY, USA. Association for Computing Machinery.

Paul Clough, Robert Gaizauskas, Scott S.L. Piao, and Yorick Wilks. 2002. Measuring text reuse. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 152–159, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Havvanur Şahi, Yasemin Kılıç, and Rahime Belen Sağlam. 2018. Automated detection of hate speech towards woman on twitter. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, pages 533–536.

Sarah Hewitt, T. Tiropanis, and C. Bokhove. 2016. The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, page 333–335, New York, NY, USA. Association for Computing Machinery.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

James R. Lewis. 1995. Ibm computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *Int. J. Hum.-Comput. Interact.*, 7(1):57–78.

Yinhan Liu, 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. *CoRR*, abs/1907.11692.

Bridianne O'Dea, Stephen Wan, Philip J. Batterham, Alison L. Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188.

Anand Rajaraman and Jeffrey David Ullman. 2011. *Mining of massive datasets*. Cambridge University Press.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.