# COVID-19 and Misinformation: A Large-Scale Lexical Analysis on Twitter

**Dimosthenis Antypas, David Rogers, Alun Preece, Jose Camacho-Collados**
School of Computer Science and Informatics & Crime and Security Research Institute
Cardiff University, United Kingdom
{antypasd,rogersdm1,preecead,camachocolladosj}@cardiff.ac.uk

## Abstract

Social media is often used by individuals and organisations as a platform to spread misinformation. With the recent coronavirus pandemic we have seen a surge of misinformation on Twitter, posing a danger to public health. In this paper, we compile a large COVID-19 misinformation-related Twitter corpus and perform an analysis to discover patterns with respect to vocabulary usage. Among others, our analysis reveals that the variety of topics and vocabulary usage are considerably more limited and negative in tweets related to misinformation than in randomly extracted tweets. In addition to our qualitative analysis, our experimental results show that a simple linear model based only on lexical features is effective in identifying misinformation-related tweets (with accuracy over 80%), providing evidence to the fact that the vocabulary used in misinformation largely differs from generic tweets.

## 1 Introduction

Social media has created a landscape where vast amounts of information on various topics is shared daily between users all around the world. Unfortunately, not all information shared is legitimate. As seen in recent events such as the Brexit referendum in the UK (Bastos and Mercea, 2019) and the 2016 US Presidential Election (Bovet and Makse, 2019), there are many cases where people, either unintentionally or deliberately (Fetzer, 2004), share unreliable information which causes confusion and suspicion in the general population. For instance, individuals and organisations share 'facts' on how the earth is flat, that vaccines cause autism, or that chlorine is treatment against COVID-19.

The spread of misinformation through social networks is made easier by the structure of these platforms. By personalising their users' news feeds and creating echo chambers, where users share believes and biases, social media provide the perfect field for spreading misinformation. Moreover, the fact that most social media platforms either do not filter misinformation or filter it inefficiently (Wardle and Singerman, 2021) means that there is no essential check on what people share online. Examples of misinformation include fabricated content, where the information is completely false; manipulated content, where there has been some distortion of genuine information; and imposter content, where someone is impersonating genuine sources (publications.parliament.uk, 2018).

Even though misinformation spread is not only related to scientific facts, health related misinformation holds an immediate danger to the public (Chou et al., 2018). Specifically, public health misinformation can be defined as a health-related claim that is currently unsupported by scientific evidence, with detrimental effects on public health (Memon, 2020). Along with the recent emergence of the COVID-19 pandemic, a number of conspiracy theories have arisen in social media; from fake and dangerous treatments to schemes that the virus is a part of a plan of the global elite to take over the world (Shahsavari et al., 2020).

The main aim of this paper is to explore whether there is a recognisable difference in the vocabulary usage between tweets conveying misinformation and random tweets present within COVID-19 discourse. To this end, we collected two corpora, one corpus consisting of misinformation-related tweets and a balancing corpus consisting of 'generic' (i.e., randomly-selected tweets) where we ran a comparative analysis. This analysis is complemented with a machine learning experiment in which we analyse to what extent misinformation-related tweets can be retrieved by using lexical features only.



*Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 119–126

## 2 Lexical Analysis of COVID-19 Misinformation Tweets

In this section, we describe our corpus collection efforts (Section 2.1) and provide a qualitative analysis on the same collected corpus (Section 2.2).

### 2.1 Corpus collection

We collected a continuous collection of tweets identified as related to the coronavirus pandemic from January to April 2020. The corpus was derived from two sources of Twitter data for the English language with a misinformation-related corpus collected via the Social Media analysis platform Sentinel (Preece et al., 2017) and a corpus of random tweets ('generic') for the same period. The tweets were tracked and selected using a list of keywords related to the pandemic[1]. Both sets ('misinformation-related' and 'generic') are balanced following the same distribution: 8,911 tweets from January, 596 from February, 411,412 from March and 20,434 from April.

Gathering a corpus of truly misinformation content is a challenging and time-consuming endeavour (Helmstetter and Paulheim, 2018) and the assumption here is that the 'generic' set contains a more diverse set of information related to COVID-19.

#### 2.1.1 Misinformation-Related corpus

The misinformation-related corpus was extracted from an existing collection of tweets gathered as part of a longitudinal study of misinformation-related call-outs in multiple languages. The tweets were collected using a set of search terms focused on misinformation in multiple languages such as 'fake news', 'disinformation', and 'misinformation'. The objective of this collection is to focus on the calling out of misinformation by Twitter users, with the assumption that users will be tagging and replying to content with the statement that something is fake news, disinformation, or consists of lies. In this way the user base acts as social sensors (Sakaki et al., 2010) to misinformation, allowing for a proactive rather than reactive collection of tweets relating to misinformation, as terms relating to particular pieces of misinformation narrative will not be known at the time of collection.

Our data was extracted, using the COVID-19 related terms, from the larger longitudinal collection which covered English language tweets from the first four months of 2020 (January to April). Finally, as the Sentinel data included tweets relating to a variety of different subjects the same list of keywords used to identify the 'generic' set were utilised to filter down the collected tweets to those relevant to coronavirus. From a total of 9.5 million tweets in the longitudinal collection as of April 2020, 441,353 tweets were used.

#### 2.1.2 Generic corpus

In order to get related data points that do not necessarily contain misinformation, we used Tweepy (Roesslein, 2009) to obtain COVID-19 related tweets from a collection of tweet IDs provided in Chen et al. (2020), retrieving the tweets directly from Twitter's API services.

An equal amount of random COVID-19 tweets (441,353), that did not contain any of the same specific set of terms employed in Sentinel for the collection of the misinformation corpus, were gathered.[2] Clearly, however, there would be a small but non-trivial number of tweets that could also contain misinformation.

### 2.2 Data exploration

#### 2.2.1 Lexical features & statistics

As an initial analysis of the dataset, we extracted relevant features for each subset. Table 1 displays some statistics about features gathered across the two different tweet classes, i.e., misinformation and generic. In particular, we include the average relative frequency of tokens, emoji, hashtags, user mentions (@), uppercase letters, punctuation and exclamation marks.

In general, the misinformation-related tweets tend to be a bit longer with average 2.28 words more than the *generic* tweets. One of the most defining differences between both classes is the amount of user mentions (represented as @), which are on average more than double in the misinformation set 1.32 to 0.59. Another interesting observation is that even though both classes use generally the same amount of punctuation, the average use of exclamation marks in the misinformation-related tweets is on average 62% higher than those of the *generic* set, 0.27 to 0.17.

[1] https://github.com/echen102/COVID-19-TweetIDs/blob/master/keywords.txt

[2] In both subsets, retweets were only considered when the original tweet was not already available. This was done on the assumption that most of the times when users retweet content they do not add additional information.

| | Tokens | Emoji | Hashtags | @ | Uppercase | Punctuation | Exclamation |
|---|---|---|---|---|---|---|---|
| **Generic** | $14.76 \pm 0.2\%$ | $0.31 \pm 1.7\%$ | $0.87 \pm 0.6\%$ | $0.59 \pm 0.9\%$ | $13.4 \pm 0.3\%$ | $9.41 \pm 0.2\%$ | $0.17 \pm 1.2\%$ |
| **Misinformation** | $17.04 \pm 0.1\%$ | $0.21 \pm 1.7\%$ | $0.76 \pm 0.7\%$ | $1.32 \pm 0.8\%$ | $15.26 \pm 0.4\%$ | $9.23 \pm 0.2\%$ | $0.27 \pm 1\%$ |

Table 1: Set of features from the COVID-19 Twitter Misinformation dataset: quantities represent the average numbers (95% confidence intervals) of instances per tweet.

We also attempted to measure the vocabulary richness and perform a comparison between the misinformation and generic sets as text containing misinformation has often less complex vocabulary and tends to be repetitive (Horne and Adali, 2017). To accomplish this two different statistics were utilised, the Type-Token Ratio (TTR) which is the ratio of unique terms against all terms, and the Measure of Textual Lexical Diversity (MTLD), a more complex metric that is not very sensitive to text length (McCarthy, 2005). MTLD is calculated as the mean length of sequential word strings in a text that maintain a given TTR value. In general, a higher MTLD score indicates a more diverse corpus. For example, the MTLD score for an equal size, random set of tweets is 913.62 whereas the score for our corpus (misinformation-related and generic tweets) is 362.10. Additionally, three subtopics were identified (using relevant keywords [3]) and deemed interesting to investigate further. The subtopics include 1) 'Covid/Weapon' with tweets mentioning COVID-19 along the lines of "bioweapon" and "human created weapon" 2) '5G' with tweets talking about the conspiracy theory of how the 5G network is responsible for the pandemic and 3) 'Politics' where the content of the tweets is revolving around US politics. The keywords used

Table 2 displays the lexical diversity statistics for the whole corpus as well as for three different subsets (covid as a weapon, 5G and Politics)[4]. The results indicate that the misinformation subset has indeed a less diverse vocabulary, with an MTLD score of 268.83 opposite to 593.74 of the generic subset. The same pattern continues when looking at the 'Covid/Weapon' and '5G' subtopics where the generic tweets have an MTLD score that is more than double of that of the misinformation tweets. In the case of the 'Politics' subtopic the lexical diversity difference is small to nonexistent with the generic and misinformation tweets achieving the same TTR score and the generic tweets having a slightly better MTLD score.

[3] **5G:** 5G **Politics:** trump, democrat, republican, obama, ted cruz, tedcruz, joebiden, joe biden, leftwing, rightwing, left wing, right wing, left wing, right wing **Covid/Weapon:** weapon, bioweapon, weaponizing, biological weapon

[4] The comparison was made between equal size subsets.

### 2.2.2 Lexical Specificity

Even though the tweets are not equally distributed through time, an attempt was made to identify trends between each month (reminder that we randomly extracted a subset of equal number of tweets per month for each of the two classes). This was achieved by computing the lexical specificity value of each word. Lexical specificity is a statistical measure which calculates the set of most representative words for a given text based on the hypergeometric distribution (Lafon, 1980; Camacho-Collados et al., 2016). In contrast to similar scores used to calculate importance of terms, such as TF-IDF, lexical specificity is not especially sensitive to different text lengths.

Table 3 displays, for each month, the top five relevant terms according to lexical specificity with respect to the whole corpus when considering the misinformation and generic subsets separately. To gain a better understanding of tweets' content, Table 3 does not include words that were present in the top 100 most relevant terms according to lexical specificity for each class. For both groups the tweets from January are focused on China (terms not displayed), which was the initial centre of the epidemic, and the following months become more diverse. Then, as can be observed in the table misinformation-related tweets tend to be more focused around conspiracies and rumours with terms such as 'uncover', 'theory' or 'lie', while generic tweets appear to be more neutral, also including government advice such as 'stay at home'.

| | Generic | | Misinformation | |
|---|---|---|---|---|
| | **TTR** | **MTLD** | **TTR** | **MTLD** |
| **Whole Corpus** | 0.03 | 593.74 | 0.02 | 268.83 |
| **Covid/Weapon** | 0.23 | 294.81 | 0.19 | 185.12 |
| **5G** | 0.25 | 648.48 | 0.15 | 151.74 |
| **Politics** | 0.04 | 393.67 | 0.04 | 337.53 |

Table 2: Lexical diversity of generic and misinformation tweets Metrics used: Type Token Ratio (TTR) and Measure of Textual Lexical Diversity (MTLD).

We further explored the three subtopics (i.e., Covid/Weapon, 5G, Politics) identified and extracted the most relevant terms based on lexical specificity. For each subtopic we compare the generic and misinformation subsets against their combined subsets in the particular subtopic. Table 4 displays the five most relevant terms for each class (misinformation/generic) in each subtopic. Similar with the terms extracted when considering the whole corpus (Table 3) there is a trend that in misinformation tweets appear more negative/intimidating terms (e.g., 'policestate', 'chemtrail', 'deep') and also terms related to mainstream news media which are often the 'enemy' of conspiracy theorists and hyperpartisan groups.

## 3 Identifying COVID-19 related misinformation tweets

Upon collecting our dataset we aimed to explore whether the lexical features of tweets can provide a strong signal for identifying misinformation. To test our hypothesis, we built multiple models using different classification approaches based on lexical features to distinguish the misinformation-related and generic sets of tweets.

### 3.1 Experimental setting

**Data pre-processing.** Non-linguistic content, such as references to web sites and special characters referring to other users were removed from the dataset. Similarly, stopwords were removed from the vocabulary. Finally, all words involved in the construction of each of the subsets (see Section 2.1) were not considered for this experiment.

**Features.** As our main goal is to test whether models can retrieve misinformation-related content using lexical features only, we use three different types of lexical features: (1) Frequency features based on TF-IDF (TF)[5]; (2) semantic based on the average of word embeddings[6] within the tweet (WE); and (3) the extra-linguistic features listed in Table 1 (EL).

**Models.** As linear machine learning models exploiting the features, we used both Naive Bayes (as a baseline model) and SVM (as a non Deep Neural Network option) classifiers following their default implementations in scikit-learn. Moreover, a Convolutional Neural Network (CNN) was implemented. Even though CNNs have been traditionally used in computer vision, they have proved to be effective for various NLP tasks, including text classification (Kim, 2014). In the present work, we trained a CNN with three layers of convolution using the same Twitter pre-trained word embeddings as initialisation. All models were evaluated using 10-fold cross validation. Finally, as current state-of-the-art NLP system we trained the base uncased version of BERT (Devlin et al., 2018) on our dataset using the implementation provided in Simple Transformers (Rajapakse, 2019).

### 3.2 Results

Table 5 shows the results of the classification models in our collected dataset. As expected, the CNN and BERT models perform better with BERT attaining the best results, with an overall accuracy of 0.91. Nonetheless, a simple SVM using lexical and semantic features attains 0.82, which shows the marked differences of the two datasets in terms of vocabulary and topics. This is surprising given the specificity of the topic and the fact that the linear models neglect linguistic properties such as word order or syntax (which are captured by the context vectors of BERT and up to some degree from the CNN), as they only rely on tokens represented as a bag of words. In a way it also confirms some of the statistics analysed in Section 2.2 and previous general findings related to misinformation in Twitter (Castillo et al., 2011) in this particular COVID-19 domain.

### 3.3 Analysis

In addition to the main results from the previous subsection, we perform two types of analysis: error and out-of-distribution analysis.

#### 3.3.1 Error analysis: Examples

In this section, we provide some examples of the errors made by the classifiers, which we attempt to digest. First, we should note that not all errors are due to the automatic model per se, and rather to the way the corpora were collected (see Section 2.1) – there is no certainty that generic tweets do not convey a message related to misinformation. For example, both the SVM and BERT models 'misclassify' the tweet *'Take care of your health...not a good time to be run down...and stay away from Corona beer, I hear from mainstream media that it causes a virus or something.'* as generic. Exclud-

[5] We considered the 500 most frequent words for the evaluation.

[6] As pre-trained words embeddings, we used the 100-dimensional fasttext embeddings (Bojanowski et al., 2017) trained on Twitter from Camacho-Collados et al. (2020).

| January | February | March | April |
|---|---|---|---|
| **— GENERIC —** | | | |
| confirm - 375.17 | suga - 10.75 | case - 4457.29 | home - 215.29 |
| flight - 255.54 | pence - 9.07 | home - 2732.56 | stay - 195.40 |
| case - 253.73 | confirm - 6.14 | test - 2139.79 | distancing - 104.21 |
| novel - 206.65 | disease - 5.51 | positive - 1776.12 | day - 62.12 |
| health - 157.46 | border - 5.35 | stay - 1748.64 | worker - 61.52 |
| **— MISINFORMATION —** | | | |
| uncover - 846.75 | deep - 16.09 | medium - 5549.91 | lie - 245.83 |
| russia - 495.33 | rosenstein - 13.84 | lie - 5491.33 | fox - 168.14 |
| awash - 347.44 | theory - 12.50 | trump - 4682.74 3 | medium - 149.72 |
| iran - 248.20 | rod - 11.05 | spread - 4078.4 | cnn - 132.21 |
| election - 236.32 | heil - 9.31 | deep - 4053.62 | fool - 110.00 |

Table 3: Top words per class based on lexical specificity not present in the top 100 of the other class.

| Covid/Weapon | | 5G | | Politics | |
|---|---|---|---|---|---|
| **Generic** | **Misinformation** | **Generic** | **Misinformation** | **Generic** | **Misinformation** |
| denver - 48.47 | news - 52.26 | case - 32.76 | news - 101.51 | test - 334.27 | news - 3115.22 |
| attend - 44.44 | deep - 42.16 | test - 26.02 | medium - 95.72 | response - 216.37 | deep - 1238.16 |
| supporter - 38.35 | chemtrail - 38.38 | confirm - 20.77 | vaccination - 86.30 | bill - 213.14 | lie - 691.22 |
| rally - 37.59 | establishment - 38.38 | home - 19.17 | policestate - 83.63 | president - 173.50 | medium - 683.58 |
| deadly - 36.08 | vaccination - 36.67 | patient - 18.90 | drill - 83.49 | vaccine - 169.08 | state - 506.51 |

Table 4: Top words per class based on lexical specificity for subtopics identified.

ing this type of example that makes a small portion of the dataset, other mistakes of the SVM model using lexical features include *'Nonsense. I done believe this disinformation campaign - the secret services are born to capitalise on crisis. They are not army or Police.The truth is #Covid19 outbreak is the rarest golden opportunity for them to test - 1. Expand Infrastructure. 2. New Tools. 3. Scalable ops.'*.

These examples show that lexical features are not enough for this task, and other type of model capturing other features (e.g., word order or syntax) such as the BERT model (or even a simpler CNN model) can provide a performance boost, as we showed in Table 5. While both the SVM and CNN struggle with linguistic phenomena such as sarcasm, as exemplified by this error made by the CNN model: *'CHINA: *covers up all evidence of biblical plague unleashed by underground farmer's market* HA let's see you top that. USA: *multiple senators dump stocks day after learning of looming biblical plague and tell everyone things are awesome while they do nothing* CHINA: touché'*, the BERT model does seem to perform better with such entries. Finally, all models struggle with tweets where the user is calling out other users actions or behaviours, for example: *'ppl out here like when is the coronavirus cure!! but wont even vaccinate their kids. i wish ppl freaked out about the flu or measles like they are the coronavirus maybe they wouldnt be such big issues otherwise'* which is misclassified as misinformation by all the models.

### 3.3.2 Out-of-distribution analysis

To test the robustness of our SVM and BERT models, an additional set of tweets from a different time period (May, June, July 2020) was collected. The new dataset is balanced, each month containing 63,468 tweets. In total, it contains 190,404 tweets using the same methodology as described in Section 2.1.

Table 7 displays the results for BERT and the best performing SVM classifier when tested on the new dataset (see Table 6 for detailed results). The SVM classifier which used TF+WE was selected as it achieved the best F1 score on the original data. It is observable that there is no substantial difference on the average performance of the models. Therefore, this may suggest that the methods (including a simple one based on lexical features and a SVM) are still robust to detect misinformation in real time. However, these results may not be generalisable as we should also reiterate the limitations of our

| Classifier | Features | Misinfo class | | | Generic class | | | Overall | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Prec | Rec | F1 | Prec | Rec | F1 | Prec | Rec | F1 | Acc |
| **Naive Bayes** | **TF** | 0.76 | 0.75 | 0.76 | 0.75 | 0.77 | 0.76 | 0.76 | 0.76 | 0.76 | 0.76 |
| | **WE** | 0.69 | 0.77 | 0.73 | 0.74 | 0.66 | 0.70 | 0.72 | 0.72 | 0.71 | 0.72 |
| | **TF+WE** | 0.77 | 0.75 | 0.76 | 0.76 | 0.78 | 0.77 | 0.76 | 0.76 | 0.76 | 0.76 |
| | **TF+WE+EL** | 0.78 | 0.76 | 0.77 | 0.76 | 0.77 | 0.76 | 0.77 | 0.77 | 0.77 | 0.77 |
| **SVM** | **TF** | 0.86 | 0.74 | 0.80 | 0.77 | 0.88 | 0.82 | 0.82 | 0.81 | 0.81 | 0.81 |
| | **WE** | 0.77 | 0.76 | 0.76 | 0.76 | 0.77 | 0.76 | 0.76 | 0.76 | 0.76 | 0.76 |
| | **TF+WE** | 0.87 | 0.80 | 0.83 | 0.78 | 0.85 | 0.82 | 0.82 | 0.82 | 0.83 | 0.82 |
| | **TF+WE+EL** | 0.89 | 0.74 | 0.80 | 0.67 | 0.89 | 0.75 | 0.78 | 0.81 | 0.78 | 0.78 |
| **CNN** | - | 0.88 | 0.86 | 0.87 | 0.87 | 0.89 | 0.88 | 0.88 | 0.87 | 0.87 | 0.87 |
| **BERT** | - | **0.90** | **0.92** | **0.91** | **0.91** | **0.90** | **0.91** | **0.91** | **0.91** | **0.91** | **0.91** |
| ***Naive baseline*** | | 0.5 | 1.0 | 0.67 | 0.0 | 0.0 | 0.0 | 0.25 | 0.5 | 0.33 | 0.5 |

Table 5: Classification results in our COVID-19 Twitter Misinformation Dataset. Evaluation metrics: accuracy and macro-averaged precision, recall and F1. *Naive baseline* refers to a system that detects misinformation for every tweet.

| | SVM | | | | | | BERT | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | misinformation | | | generic | | | misinformation | | | generic | | |
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| **May** | 0.87 | 0.80 | 0.83 | 0.82 | 0.88 | 0.85 | 0.91 | 0.88 | 0.89 | 0.88 | 0.91 | 0.89 |
| **June** | 0.87 | 0.81 | 0.84 | 0.82 | 0.87 | 0.85 | 0.90 | 0.89 | 0.90 | 0.90 | 0.90 | 0.90 |
| **July** | 0.86 | 0.78 | 0.82 | 0.80 | 0.87 | 0.83 | 0.89 | 0.88 | 0.89 | 0.88 | 0.90 | 0.89 |
| **Total** | 0.87 | 0.80 | 0.83 | 0.81 | 0.88 | 0.84 | 0.90 | 0.88 | 0.89 | 0.89 | 0.90 | 0.89 |

Table 6: Classification results of the SVM (TF+WE) and BERT models for May - July period.

analysis that was performed on a limited set of data from a single year.

| | Precision | Recall | Accuracy | F1 |
|---|---|---|---|---|
| **SVM** | 0.84 | 0.84 | 0.84 | 0.84 |
| **BERT** | 0.89 | 0.89 | 0.89 | 0.89 |

Table 7: Overall classification results for May - July period. Evaluation metrics: accuracy and macro-averaged precision, recall and F1. SVM model used: TF+WE.

In order to better understand the behaviour of the classifiers, we further investigated how the models perform in each individual month. Figure 1 displays the precision and recall results for the misinformation class. In each month BERT outperforms the SVM model. While the performance of both is mostly consistent, there is a drop in Recall for the SVM model in July (May:0.8, June:0.81, July:0.78). This may be indicative of a change in the misinformation corpus vocabulary for July that the SVM model fails to recognise. Despite this, the results remain a strong indication that there is indeed a recognisable difference between the vocabulary used in the misinformation and generic tweets.

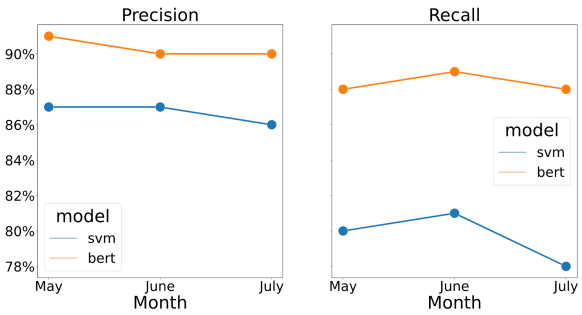


Figure 1: Monthly precision and recall results for the misinformation class.

## 4 Conclusion

In this paper, we have presented an analysis on the lexical features present in misinformation about COVID-19 in social media, and compare it with

those present in *generic* or random tweets. To this end, we compiled two different Twitter corpora from early 2020 when the pandemic emerged. Our analysis shows that there is a clear distinction in the general vocabulary used in each type of corpus and that a simple linear classifier based on lexical features can retrieve misinformation-related tweets to a high degree of accuracy. While this paper represents an initial reference point in this aspect, further analysis would be required to investigate the main features present in misinformation. On this respect, our work can also be added to the increasing evidence that shows that misinformation focuses on a specific vocabulary that does not reflect on the overall distribution of what can be found in general social media content for a certain topic (Castillo et al., 2011). Finally, it would be interesting to evaluate and compare the models' performance on other datasets that are manually labelled and are not collected based on the "call out" principle (Alam et al., 2020).

## References


Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, and Preslav Nakov. 2020. Fighting the covid-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society.

Marco T Bastos and Dan Mercea. 2019. The Brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, 37(1):38–54.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 us presidential election. *Nature communications*, 10(1):1–14.

Jose Camacho-Collados, Yerai Doval, Eugenio Martínez-Cámara, Luis Espinosa-Anke, Francesco Barbieri, and Steven Schockaert. 2020. Learning Cross-lingual Embeddings from Twitter via Distant Supervision. In *Proceedings of ICWSM*.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information Credibility on Twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus Twitter data set. *Journal of Medical Internet Research Public Health and Surveillance*, 6(2):e19273.

Wen-Ying Sylvia Chou, April Oh, and William MP Klein. 2018. Addressing health-related misinformation on social media. *Journal of the American Medical Association*, 320(23):2417–2418.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

James H Fetzer. 2004. Disinformation: The use of false information. *Minds and Machines*, 14(2):231–240.

Stefan Helmstetter and Heiko Paulheim. 2018. Weakly supervised learning for fake news detection on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 274–277. IEEE.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

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.

Pierre Lafon. 1980. Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1(1):127–165.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Shahan Ali Memon. 2020. *Characterizing Misinformed Online Health Communities*. Ph.D. thesis, Carnegie Mellon University.

Alun Preece, Irena Spasić, Kieran Evans, David Rogers, William Webberley, Colin Roberts, and Martin Innes. 2017. Sentinel: A codesigned platform for semantic enrichment of social media streams. *IEEE Transactions on Computational Social Systems*, 5(1):118–131.

publications.parliament.uk. 2018. Disinformation and 'fake news': Interim report. https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/36304.htm#_idTextAnchor002.

Thilina Rajapakse. 2019. Simple transformers. https://github.com/ThilinaRajapakse/simpletransformers/.

Joshua Roesslein. 2009. Tweepy documentation. *Online] http://tweepy. readthedocs. io/en/v3*, 5.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860.

Shadi Shahsavari, Pavan Holur, Timothy R Tangherlini, and Vwani Roychowdhury. 2020. Conspiracy in the time of corona: Automatic detection of covid-19 conspiracy theories in social media and the news. *arXiv preprint arXiv:2004.13783*.

Claire Wardle and Eric Singerman. 2021. Too little, too late: social media companies' failure to tackle vaccine misinformation poses a real threat. *British Medical Journal*, 372.