Detecting Social Media Manipulation in Low-Resource Languages

Samar Haider
Information Sciences Institute,
University of Southern California
USA
samarhai@usc.edu

Luca Luceri
Information Sciences Institute,
University of Southern California
USA
lluceri@isi.edu

Ashok Deb
Information Sciences Institute,
University of Southern California
USA
adeb@usc.edu

Adam Badawy
Information Sciences Institute,
University of Southern California
USA
abadawy@usc.eadu

Nanyun Peng
Information Sciences Institute,
University of Southern California
USA
nanyunpe@usc.edu

Emilio Ferrara
Information Sciences Institute,
University of Southern California
USA
emiliofe@usc.edu

ABSTRACT

Social media have been deliberately used for malicious purposes, including political manipulation and disinformation. Most research focuses on high-resource languages. However, malicious actors share content across countries and languages, including low-resource ones. Here, we investigate whether and to what extent malicious actors can be detected in low-resource language settings. We discovered that a high number of accounts posting in Tagalog were suspended as part of Twitter's crackdown on interference operations after the 2016 US Presidential election. By combining text embedding and transfer learning, our framework can detect, with promising accuracy, malicious users posting in Tagalog without any prior knowledge or training on malicious content in that language. We first learn an embedding model for each language, namely a high-resource language (English) and a low-resource one (Tagalog), independently. Then, we learn a mapping between the two latent spaces to transfer the detection model. We demonstrate that the proposed approach significantly outperforms state-of-the-art models and yields marked advantages in settings with very limited training data—the norm when dealing with detecting malicious activity in online platforms.

CCS CONCEPTS

• **Information systems** → **Social networks**.

KEYWORDS

social media, disinformation, language processing

ACM Reference Format:
Samar Haider, Luca Luceri, Ashok Deb, Adam Badawy, Nanyun Peng, and Emilio Ferrara. 2023. Detecting Social Media Manipulation in Low-Resource Languages. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3543873.3587615

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04.
https://doi.org/10.1145/3543873.3587615

1 INTRODUCTION

Disinformation and political manipulation have a long history: for example, in 1984, long before the social media era, a story claiming that the HIV virus was created by the US government as a biological weapon became viral worldwide. Nowadays, social media amplify and accelerate information spread to an unprecedented pace. Online Social Networks (OSNs) like Twitter and Facebook have been facing a copious growth of malicious content, which undermine the truthfulness and authenticity of online discourse [1, 21, 24, 32, 74, 77].

Various studies showed that OSNs have been used for malicious purposes harming several constituents of our society [42, 75], ranging from geo-political events [22, 27, 44, 58, 63] to public health [14, 25, 28, 52, 78]. Bots and trolls act as main actors in social media manipulation and disinformation campaigns [3, 11, 26, 45, 62], often in a coordinated fashion [29, 51, 53, 64, 67, 76].

Particular attention has been devoted to the risk of mass manipulation of public opinion in the context of politics, whose prime example is the online interference in the 2016 US Presidential discussion election [4, 6]. Since then, OSNs have been trying to fight abuse and maintain a trustful and healthy conversation on their platforms. Despite the effort, the activity of trolls and bots appears to persist [36, 43, 72]. For instance, Twitter identified and suspended malicious accounts originating from diverse countries, including Russia, Iran, Bangladesh, and Venezuela [71], suggesting the presence of coordinated efforts to manipulate online discourse across countries and languages. Recently, Pierri et al. [58] documented evidence of platform abuse and subsequent Twitter interventions [57] in the context of the ongoing conflict between Ukraine and Russia. While others have explored the various strategies of malicious users in high-resource languages [41, 43, 69, 70] to enable their detection [12, 13, 33, 49], here we present a novel approach using transfer learning to empower the automated identification of misbehaving accounts in low-resource languages.

Contributions of this work

Our aim is to investigate whether and to what extent textual content can be used as a proxy to detect malicious activity on social media, with a particular focus on accounts sharing messages in low-resource languages. Overall, we aim at answering two main research questions:

RQ1: *Can we classify an account as malicious based only on their shared content?* We explore the effectiveness of learning word representations from tweets to identify suspended accounts.

RQ2: *Can we learn a model from a high-resource language (English) and transfer knowledge to a low-resource one (Tagalog) for detecting suspended accounts?* We investigate whether learning a mapping between two independently-trained word embeddings can be beneficial to identify misbehaving accounts.

2 RELATED WORK

2.1 Political Manipulation

Social media has provided an online venue for users to connect and share ideas. However, social media networks like Twitter and Facebook can be used for malicious purposes [21]. Social media mass manipulation of public opinion is based on disinformation campaigns carried out by malicious actors, including social media bots and state-backed trolls [6, 7, 9, 20, 34, 43, 45, 50, 61, 65, 77].

Despite the mitigation efforts, recent election-related research shows that the number of bots has not significantly reduced [16, 43] and indeed bots are becoming more sophisticated [44]. Thus, potentially malicious activity on social media can become an even more pervasive problem for political discourse. Especially for the spread of fake news, various studies showed how political leaning [1, 43], polarization [2, 5], age [32], and education [61] can greatly affect fake news spread, alongside with other mechanisms that leverage emotions [30, 31], cognitive limits [54, 55], and social network vulnerabilities [10, 60, 68]. Other work established that social media platforms were used as well to distort other elections [17, 22, 35, 47, 59, 66] and other real-world events [23, 52, 58].

2.2 Bots & Trolls

The article *The Rise of Social Bots* [29] initially highlighted the issue of bots, or algorithmic automated accounts, on social media platforms. [6] focused on bot detection during the 2016 US Presidential election, finding that an estimated 400K accounts were likely automated and produced nearly 1 in 5 tweets in that political conversation. Since the 2016 US Presidential election, the US election system has been under scrutiny. Since then, social media networks have been trying to fight malicious actors to maintain an healthy conversation on their platform. Foreign actors were shown in [19] to influence unsuspecting users on social media for the expressed purpose of sowing discord. Badawy *et al.* [3] analyzed the Russian troll accounts on Twitter to understand their information warfare campaign, while Im *et al.* [36] showed that the accounts are remaining. Indeed, Zannettou et al. [79] described how the automated identification of human operators, as state-backed troll accounts, is a challenging, yet-unsolved task.

These and other studies that corroborated the issue of automated accounts and trolls being used for malicious purposes have been used to inform policy and new regulations. For example, the *Bot Disclosure and Accountability Act* of 2018[1] first directs the Federal Trade Commission to implement controls on the social media companies and secondly, it amends the Federal Election Campaign Act of 1971 to prohibit a campaign from impersonating human activity online.

Hashtag	# of Tweets
#election2016	422,952
#VPdebate	1,031,972
#hillary	1,516,318
#trump	3,290,636
#neverhillary	1,063,545
#nevertrump	746,430
#garyjohnson	58,832
#jillstein	51,831

Table 1: A representative subset of hashtags used in the data collection, along with the number of tweets emdedding the hashtag

3 DATA: US 2016 PRESIDENTIAL ELECTION

In this study, we use Twitter as a test-bed to detect the activity of malicious accounts focusing on the 2016 US presidential election. The dataset, about 42 million tweets posted by almost 6 million distinct users, was first published by [6]. Tweets were collected through the Twitter Streaming API using 23 election keywords (5 for Donald Trump, 4 for Hillary Clinton, 3 for third-party candidates, and 11 for the general election terms). The collection was carried out between September 16, 2016 and October 21, 2016. From the set of collected tweets, duplicates were removed, which may have been captured by accidental redundant queries to the Twitter API. A list of the most popular keywords and associated number of tweets is given in Table 1. Although all the keywords are in English, tweets in other languages were collected.

We identified over 60 different languages, with the highest number of tweets written in European languages. In particular, there were over 37.6 million English tweets posted by nearly 5 million users. We found a noticeable number of tweets in Tagalog, an Austronesian language which is the first language of a quarter of the population of the Philippines, and the second language of more than half of the rest. As the fourth-most common language by number of speakers in the United States [73], behind only English, Spanish, and Chinese, Tagalog represents the top low-resource language in our data by number of tweets. The US is also home to one of the largest population of Filipino emigrants living outside of the Philippines. Additionally, Tagalog's low-resource status is further confirmed by an analysis of the size of its Wikipedia—a common proxy for estimating the amount of digital resources in a language. Tagalog's Wikipedia is currently ranked 101st by number of articles,[2] in sharp contrast to its prevalence in our dataset. For this reason, we focus our attention on Tagalog as the target language in this work.

4 METHODOLOGY

4.1 Word Representations

To learn word embeddings and train classification models, we use the FastText[3] framework. Instead of treating words as atomic units

[1] https://www.congress.gov/bill/115th-congress/senate-bill/3127/text

[2] https://meta.wikimedia.org/wiki/List_of_Wikipedias

[3] https://github.com/facebookresearch/fastText

of text, FastText represents words as a bag of *character n-grams* [8], wherein each n-gram has its own vector representation and a word is represented as the sum of its constituent character n-grams. This allows the model to adapt to morphologically rich languages with large vocabularies as well as generalize better from smaller training corpora.

Although neural network-based models have achieved considerable success at text classification tasks, they remain quite expensive to train and deploy. FastText utilizes a hierarchical softmax to serve as a fast approximation of the softmax classifier to compute the probability distribution over the given classes [38]. Using feature pruning, quantization, hashing, and retraining to substantially reduce model size without sacrificing accuracy or speed, this approach allows the training of models on large corpora of text much faster than neural network-based methods [37].

4.2 Transfer Learning

Traditional machine learning approaches for natural language processing focus on training specialized models for specific tasks. However, this requires significant amounts of data which is hard to acquire for low-resource languages. This has historically elicited more research on high-resource languages (primarily European), which leads to more resources created for these languages, thus feeding the cycle. Transfer learning has recently arisen as a way to leverage knowledge learned from a source language (or source task) and utilize it to improve performance on a target language (or target task).

To address the scarcity of data in the target language under analysis in this work, we use MUSE[4], a framework for aligning monolingual word embeddings from different languages in the same space and allowing transfer of knowledge between them. MUSE learns a mapping from the source to target space using Procrustes alignment to minimize the distance between similar words in the two languages [40]. It accepts as input two sets of pretrained monolingual word embeddings (such as those learned by FastText), one for each language, and can learn a mapping between them in either a supervised or unsupervised fashion. The supervised method requires the use of a bilingual dictionary to assist in aligning the two embeddings together by identifying similar word pairs that should be close together in the shared space. In the absence of such a dictionary, the unsupervised alternative utilizes adverserial training to initialize a linear mapping between a source and a target space and to produce a synthetic parallel dictionary. [15] showed that this approach can be used to perform unsupervised word translation without the use of any parallel data, with results that in some cases outperform even prior supervised methods.

4.3 Learning Tasks

Monolingual text classification. In the first approach, we train independent text classification models for each language from scratch using their respective datasets. For classification purposes, we use the FastText framework, which represents text as a bag of words (BoW) and averages their individual representations into a combined text representation. This text representation is then used as input to a linear classifier with a softmax function that computes the probability distribution over the label classes in order to make predictions.

Transfer learning. In the second approach, we use transfer learning from the high-resource language with more data (English) to improve text classification accuracy on the low-resource language with fewer data (Tagalog). We first train unsupervised monolingual word embeddings for each language using FastText's skipgram model [48]. We then obtain multilingual word embeddings by mapping the embeddings for both English and Tagalog in the same space with MUSE by using a bilingual English-Tagalog dictionary to establish correspondence between words in the two languages, and eventually using these words as anchors to align the embeddings of both languages in the same latent space. This allows to maximize information learned from one language to another. The multilingual embeddings are then used as pretrained vectors to initialize a FastText model, trained on the target language using its dataset to make predictions over users' account status.

4.4 Baseline Models

We compare our work with a number of different baselines, both traditional and deep learning-based approaches, which we detail as follows.

Bag-of-Words and their TFIDF. We create a bag-of-words model by extracting a vocabulary of words in the corpus. We then calculate the counts of the words in the examples as features for our model. For the TFIDF (term frequency—inverse document frequency) variant, we normalize the aforementioned counts by dividing them by the total number of words in a document, which gives us the term-frequency. The inverse document frequency measures how common or rare a word is, and is equivalent to the logarithm of the total number of documents divided by the number of documents containing that word. The TFIDF is then given by the product of the term frequency and the inverse document frequency.

Bag-of-n-grams and their TFIDF. Often, sequences of words (*n*-grams) carry more information than words taken individually, specifically because n-grams carry contextual information that is lost when single words are considered. We construct a bag-of-ngrams model by extracting n-grams, ranging from unigrams to 5-grams, ($n = 1 \ldots 5$) from the corpus, and use those as features. For the TFIDF variant, we apply the same normalizing scheme as described above.

BERT contextual embeddings. Traditional word embeddings, while very efficient to compute and use, lack of context. As words exhibit polysemy, they can have different meanings based on the context in which they are used. An example is the word 'bank', which can either mean a financial institution or the land alongside a river or lake. Recently, there has been a significant interest in the use of contextual word embedding models such as ELMo (Embeddings from Language Models) [56] and BERT (Bidirectional Encoder Representations from Transformers) [18], which are trained by a class of deep neural network models called *transformers*. The final layers of these models have been shown to effectively capture a high degree of semantic knowledge from the input text, which can subsequently be used for downstream or auxiliary tasks. For our work,

[4]https://github.com/facebookresearch/MUSE

Description	English	Tagalog
# of accounts	4,872,565	23,979
% of suspended	31%	31%
# of tweets	37,623,535	29,887
% from suspended	29%	31%

Table 2: Statistics of the monolingual datasets. The amount of data for Tagalog is an order of magnitude less than that for English, but both languages contain a significant fraction of accounts that got suspended in the wake of the campaign. Furthermore, Tagalog data is also very sparse as we only have around 1 tweet per account, which also contributes to the difficulty of our classification task.

we use the contextualized word representations produced by the multilingual variant of the BERT model, which has been pre-trained on 104 languages using their respective *Wikipedia*s as corpora. We extract the features generated by the final layer of BERT and train a softmax layer on top for our binary classification task.

5 EXPERIMENTAL SETUP

5.1 Design

For the purpose of this study, we define a malicious account as a user account that has been suspended from the Twitter platform. Although this is an approximation, Twitter systematically reviews suspicious accounts, either in an algorithmic fashion or based on reports of inappropriate behaviors, such as spam, abuse, etc. Given the negative cost associated with suspending a possibly legitimate account, Twitter's suspensions are typically associated with serious, repeated instances of misbehavior. In the context of political discussions, for example, reasons for suspension include the use of automated accounts, or orchestrated attempts to bolster the visibility or support of a political candidate.

We assign a label to each user as either "Suspended" or "Not suspended" based on querying the Twitter Search API, whose response can be: (i) *account_suspended*, if Twitter has removed the account; (ii) *not_found*, if the account has been deleted by the owner; (iii) *protected*, if the owner has been made it private (i.e., invisible to the public); or, otherwise (iv) *active* (i.e., not suspended). We hence use class (i) *account_suspended* as the positive label (i.e., "Suspended") and the other classes *active*, *not_found* and *protected* as the negative label (i.e., "Not suspended").

5.2 Tweet Aggregation

We then aggregate all the tweets written in either English or Tagalog and build two monolingual datasets accordingly. In Table 2, we list some statistics about users and tweets in the two sets of data. For each dataset, we then aggregate tweets by user account by concatenating all text that a particular account has tweeted. This results in a tweet "document" for each user account. Such a set of tweets is then collectively used to make the prediction of whether a user was suspended or not. We minimize pre-processing the text in order to preserve characteristics of the tweet, such as punctuation, URLs, and hashtags, which can help flag potentially malicious users.

5.3 Hyperparameters

For both languages, we randomly sample 80% of the data for training and retain the remainder for testing.

For the bag-of-words and n-gram-based models, we use approximately 35,000 features that are obtained from the text and train a logistic regression classifier with L2 regularization for 100 epochs on top of them.

For the BERT-based model, we use the the 768-dimensional contextual word embeddings obtained from the final layer of the model. We average the embedding for each token to produce a representation of the entire sequence. We feed this representation to a softmax layer with binary outputs, which we train for 100 epochs using the Adam optimizer [39] with a learning rate of 0.001.

For the monolingual learning task, we train separate supervised FastText classification models for each language. For the transfer learning task, we train separate unsupervised FastText word embeddings of dimensionality 100 using the skipgram model. We then align the two sets of word embeddings into the same space with MUSE and perform 5 iterations of refinement for alignment. The procedure uses the provided English-Tagalog dictionary which contains 5000 word pairs for training and 1500 for evaluation. We then use the aligned embeddings as pretrained vectors to initialize another FastText model for the target language (Tagalog) and compare it with a classic monolingual model. To evaluate the effectiveness of transfer learning in a low-resource setting, we train both models on only 10% of the original Tagalog training set.

5.4 Metrics

We use F1, Precision, and Recall with binary averaging to evaluate the performance of our models. While *macro* and *micro averaging* give high scores, they are calculated globally and disregard class imbalance (the majority of unsuspended accounts lead to inflated results). We are, however, focused on the task of accurately identifying malicious users (the positive label in our setup), and thus we use binary averaging which reports results for the positive class of suspended accounts.

6 RESULTS

In this section, we present the results related to both the monolingual and cross-lingual approaches.

Let us first consider the monolingual models for both English and Tagalog languages. Table 3 shows the performance in terms of precision, recall, and F1 score. Note that these evaluation metrics have been computed by considering a binary classification scenario, where the positive label is represented by the class "Suspended". What stands out is that precision is higher than recall in both languages, suggesting that the models learn a conservative classification schema that minimizes the costly false positives. This yields models missing a large number of suspended accounts in both languages, which is most apparent in the case of English, where the precision is high (>70%), but the recall is quite low (<20%). For Tagalog, the figures are somewhat closer to each other, resulting in a slightly higher F1 score w.r.t. English. Despite the low recall scores may appear as indicative of large margins of improvement, which is addressed by our transfer learning model, it is also worth noting that in a practical setting, a conservative model shall be preferred over a more aggressive detection system that may yield a high false positive rate.

Language	Precision	Recall	F1
English	0.708	0.184	0.292
Tagalog	0.448	0.218	0.293

Table 3: Performance of Monolingual Models. While F1 scores for both models are similar, the Tagalog model suffers in terms of precision in identifying suspended accounts.

In Figure 1, we show the performance of transfer learning compared to the monolingual model at a varying training set size.

Two facts are worth noting. First, as we expected, as the percentage of training data increases, the F1 score of both models improves. Second, in cases where up to 30% of the training set (5,754 instances) is used to train the models, transfer learning performs comparable to the monolingual model and achieves a higher F1.

To further investigate this finding, in Table 4, we show precision and recall of both the models in the low-resource setting where we use only 10% of the Tagalog data as the training set. The monolingual model achieves better precision than transfer learning, but it performs poorly in terms of recall. However, the transfer learning model offers a more balanced trade-off between precision and recall, and appears to be more suitable in a low-resource setting.

We also evaluate the performance of our approach by comparing it with baseline models, as displayed in Table 5. To resemble the training setup for the baselines, here we train our proposed models for 100 epochs with a learning rate of 1.0. We see that the transfer learning approach outperforms all other methods with a higher F1 score and recall. While the baselines have good precision, they are unable to accurately detect malicious users, which is evident from their low recall. Our proposed approach shows promising performance especially when compared to a sophisticated model that uses contextualized word representations (BERT), probably because of the different nature of social media discourse with respect to the corpora BERT has been pre-trained on. This shows how the analysis of social media content, in general, and the detection of discourse manipulation, in particular, can not be easily conducted by using models trained on other, more formal bodies of text, further highlighting the downstream challenges of designing language-agnostic tools that can generalize to low-resource languages.

Finally, we perform dimensionality reduction of text embeddings using *Uniform Manifold Approximation and Projection* (UMAP) [46]. After aggregating the tweets for each user, we use FastText to generate vector embeddings from their text. We then use UMAP to map these embeddings into a lower-dimensional space, retaining only the components that explain most of the variation in the data. Figure 2 shows the results of this projection onto a two-dimensional space. A separation is seen between the embeddings of the suspended and non-suspended tweets. This suggests that the model can, to some degree, capture the distinction between the two classes of accounts in the target low-resource language.

7 CONCLUSIONS

Since billions of users populate social media platforms around the world, these represents ripe targets for malicious actors and deceptive behaviors. We can leverage NLP to assist in an automated way the detection of manipulation efforts. Importantly, the research community has predominantly been focusing on the study of online

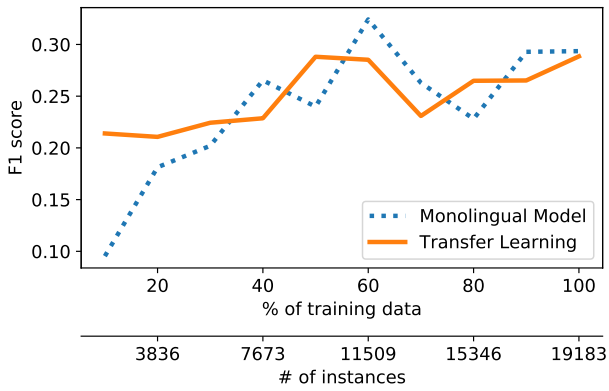

Figure 1: Monolingual vs. transfer learning at varying training set size. While enough data allows both to converge to similar performance, in low-resource cases, transfer learning vastly outperforms the monolingual approach.

Model	Precision	Recall	F1
FastText	0.847	0.051	0.095
FastText + MUSE	0.390	0.147	0.214

Table 4: Models comparison in the low-resource setting. Evaluated on 10% of the Tagalog training set, the transfer learning approach achieves more than twice the F1 score and almost three times the recall of the more conservative monolingual approach.

Model	Precision	Recall	F1
Bag-of-Words	0.502	0.172	0.256
Bag-of-Words + TFIDF	0.580	0.125	0.206
N-grams	0.546	0.166	0.247
N-grams + TFIDF	**0.635**	0.104	0.179
BERT embeddings	0.513	0.136	0.215
FastText	0.424	0.237	0.337
FastText + MUSE	0.416	**0.280**	**0.347**

Table 5: Model comparison on Tagalog. Among our proposed methods (in bold), transfer learning outperforms all other models in terms of F1 and recall.

platforms in high-resource languages, for which many NLP tools exist, e.g., sentiment analysis, semantic parsers, etc. However, a need for language-agnostic frameworks exists to allow the study of discourse in low-resource languages and enable the automated identification of malicious activity. In this paper, we posed the problem of detecting social media abuse in low-resource languages.

Using a backdrop of the 2016 US Presidential election Twitter discussion, and by drawing a parallel between abuse in English and Tagalog, we proposed a framework to detect suspended, misbehaving accounts leveraging only their shared content in Tagalog. Although the task was proven to be challenging even in a high-resource language setting, we showed that our proposed framework

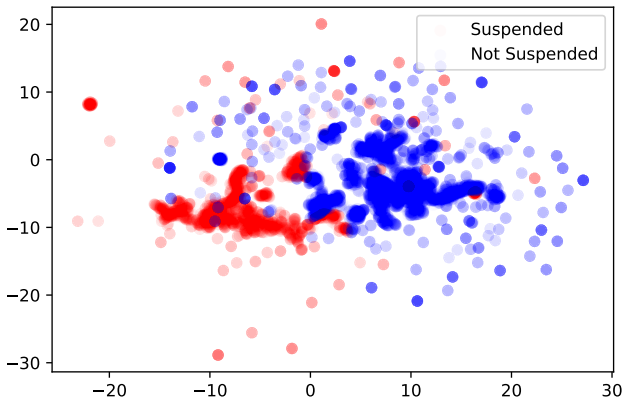

Figure 2: Visualization of Tweets. UMAP shows a clear separation between the text embeddings of suspended (red) and not suspended (blue) accounts.

can build a conservative model to detect malicious actors manipulating the discourse in a low-resource language. Much more work is needed to guarantee healthy conversations in the online landscape of low-resource languages. New language-agnostic tools can spur from our work. We seek to initiate an agenda and stimulate research on social media in low-resource languages, including for the study of misbehavior, manipulation, and abuse.

ACKNOWLEDGMENTS

Work supported in part by AFOSR (grant #FA9550-20-1-0224).

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36.
[2] Marina Azzimonti and Marcos Fernandes. 2018. *Social media networks, fake news, and polarization.* Technical Report. National Bureau of Economic Research.
[3] Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. In *Int. Conference on Advances in Social Networks Analysis and Mining*. 258–265.
[4] Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Who Falls for Online Political Manipulation?. In *Companion Proceedings of the 2019 World Wide Web Conference.*
[5] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *PNAS* 115, 37 (2018), 9216–9221.
[6] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 21, 11 (2016).
[7] Olga Boichak, Sam Jackson, Jeff Hemsley, and Sikana Tanupabrungsun. 2018. Automated Diffusion? Bots and Their Influence During the 2016 US Presidential Election. In *International Conference on Information*. Springer, 17–26.
[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.
[9] Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 10, 1 (2019), 7.
[10] Gizem Ceylan, Ian A Anderson, and Wendy Wood. 2023. Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences* 120, 4 (2023), e2216614120.
[11] Ho-Chun Herbert Chang and Emilio Ferrara. 2022. Comparative analysis of social bots and humans during the COVID-19 pandemic. *Journal of Computational Social Science* (2022), 1409–1425.
[12] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. 2016. DeBot: Twitter Bot Detection via Warped Correlation.. In *ICDM*. 817–822.
[13] Nikan Chavoshi and Abdullah Mueen. 2018. Model Bots, not Humans on Social Media. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 178–185.
[14] Emily Chen, Julie Jiang, Ho-Chun Herbert Chang, Goran Muric, and Emilio Ferrara. 2022. Charting the information and misinformation landscape to characterize misinfodemics on social media: COVID-19 infodemiology study at a planetary scale. *Jmir Infodemiology* 2, 1 (2022), e32378.
[15] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word Translation Without Parallel Data. *arXiv preprint arXiv:1710.04087* (2017).
[16] Ashok Deb, Luca Luceri, Adam Badawy, and Emilio Ferrara. 2019. Perils and Challenges of Social Media and Election Manipulation Analysis: The 2018 US Midterms. In *Companion Proceedings of the 2019 World Wide Web Conference.*
[17] Michela Del Vicario, Fabiana Zollo, Guido Caldarelli, Antonio Scala, and Walter Quattrociocchi. 2017. Mapping social dynamics on Facebook: The Brexit debate. *Social Networks* 50 (2017), 6–16.
[18] 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* (2018).
[19] Ritam Dutt, Ashok Deb, and Emilio Ferrara. 2018. "Senator, We Sell Ads": Analysis of the 2016 Russian Facebook Ads Campaign. In *International Conference on Intelligent Information Technologies*. Springer, 151–168.
[20] Fatima Ezzeddine, Luca Luceri, Omran Ayoub, Ihab Sbeity, Gianluca Nogora, Emilio Ferrara, and Silvia Giordano. 2022. Characterizing and Detecting State-Sponsored Troll Activity on Social Media. (2022).
[21] Emilio Ferrara. 2015. Manipulation and abuse on social media. *ACM SIGWEB Newsletter* Spring (2015), 4.
[22] Emilio Ferrara. 2017. Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election. *First Monday* 22, 8 (2017).
[23] Emilio Ferrara. 2018. Measuring social spam and the effect of bots on information diffusion in social media. In *Complex Spreading Phenomena in Social Systems*. Springer, 229–255.
[24] Emilio Ferrara. 2019. The history of digital spam. *Commun. ACM* 62, 8 (2019), 82–91.
[25] Emilio Ferrara. 2020. What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday* (2020).
[26] Emilio Ferrara. 2022. Twitter spam and false accounts prevalence, detection, and characterization: A survey. *First Monday* 27, 12 (2022).
[27] Emilio Ferrara, Herbert Chang, Emily Chen, Goran Muric, and Jaimin Patel. 2020. Characterizing social media manipulation in the 2020 US presidential election. *First Monday* (2020).
[28] Emilio Ferrara, Stefano Cresci, and Luca Luceri. 2020. Misinformation, manipulation, and abuse on social media in the era of COVID-19. *Journal of Computational Social Science* 3 (2020), 271–277.
[29] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.
[30] Emilio Ferrara and Zeyao Yang. 2015. Measuring emotional contagion in social media. *PloS one* 10, 11 (2015), e0142390.
[31] Emilio Ferrara and Zeyao Yang. 2015. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science* 1 (2015), e26.
[32] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 US presidential election. *Science* 363, 6425 (2019), 374–378.
[33] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*. Springer, 228–243.
[34] Philip N Howard, Gillian Bolsover, Bence Kollanyi, Samantha Bradshaw, and Lisa-Maria Neudert. 2017. Junk news and bots during the US election: What were Michigan voters sharing over Twitter. *CompProp, OII, Data Memo* (2017).
[35] Philip N Howard, Bence Kollanyi, and Samuel Woolley. 2016. Bots and Automation over Twitter during the US Election. *Computational Propaganda Project: Working Paper Series* (2016).
[36] Jane Im, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens, and Eric Gilbert. 2019. Still out there: Modeling and Identifying Russian Troll Accounts on Twitter. *arXiv preprint arXiv:1901.11162* (2019).
[37] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).
[38] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* (2016).
[39] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[40] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised Machine Translation Using Monolingual Corpora Only. *arXiv preprint arXiv:1711.00043* (2017).
[41] Luca Luceri, Felipe Cardoso, and Silvia Giordano. 2021. Down the bot hole: Actionable insights from a one-year analysis of bot activity on Twitter. *First Monday* (2021).
[42] Luca Luceri, Stefano Cresci, and Silvia Giordano. 2021. Social media against society. *The Internet and the 2020 Campaign* (2021), 1.

[43] Luca Luceri, Ashok Deb, Adam Badawy, and Emilio Ferrara. 2019. Red Bots Do It Better: Comparative Analysis of Social Bot Partisan Behavior. In *Companion Proceedings of the 2019 World Wide Web Conference*.

[44] Luca Luceri, Ashok Deb, Silvia Giordano, and Emilio Ferrara. 2019. Evolution of bot and human behavior during elections. *First Monday* 24, 9 (2019).

[45] Luca Luceri, Silvia Giordano, and Emilio Ferrara. 2020. Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 417–427.

[46] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[47] Panagiotis T Metaxas and Eni Mustafaraj. 2012. Social media and the elections. *Science* 338, 6106 (2012), 472–473.

[48] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[49] Amanda Minnich, Nikan Chavoshi, Danai Koutra, and Abdullah Mueen. 2017. BotWalk: Efficient adaptive exploration of Twitter bot networks. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 467–474.

[50] Bjarke Mønsted, Piotr Sapieżyński, Emilio Ferrara, and Sune Lehmann. 2017. Evidence of Complex Contagion of Information in Social Media: An Experiment Using Twitter Bots. *Plos One* 12, 9 (2017), e0184148.

[51] Leonardo Nizzoli, Serena Tardelli, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. 2021. Coordinated behavior on social media in 2019 UK General Election. In *AAAI ICWSM*.

[52] Gianluca Nogara, Padinjaredath Suresh Vishnuprasad, Felipe Cardoso, Omran Ayoub, Silvia Giordano, and Luca Luceri. 2022. The disinformation dozen: An exploratory analysis of covid-19 disinformation proliferation on twitter. In *14th ACM Web Science Conference 2022*. 348–358.

[53] Diogo Pacheco, Alessandro Flammini, and Filippo Menczer. 2020. Unveiling Coordinated Groups Behind White Helmets Disinformation. *ACM WWW Companion* (2020).

[54] Gordon Pennycook and David G. Rand. 2018. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* (2018).

[55] Gordon Pennycook and David G Rand. 2019. Cognitive reflection and the 2016 US presidential election. *Personality and Social Psychology Bulletin* 45, 2 (2019).

[56] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

[57] Francesco Pierri, Luca Luceri, and Emilio Ferrara. 2022. How does Twitter account moderation work? Dynamics of account creation and suspension during major geopolitical events. *arXiv preprint arXiv:2209.07614* (2022).

[58] Francesco Pierri, Luca Luceri, Nikhil Jindal, and Emilio Ferrara. 2022. Propaganda and Misinformation on Facebook and Twitter during the Russian Invasion of Ukraine. *arXiv preprint arXiv:2212.00419* (2022).

[59] Jacob Ratkiewicz, Michael Conover, Mark R Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. Detecting and tracking political abuse in social media. *ICWSM* 11 (2011), 297–304.

[60] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 131–141.

[61] Dietram A Scheufele and Nicole M Krause. 2019. Science audiences, misinformation, and fake news. *PNAS* (2019), 201805871.

[62] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications* 9, 1 (2018), 4787.

[63] Karishma Sharma, Emilio Ferrara, and Yan Liu. 2022. Characterizing Online Engagement with Disinformation and Conspiracies in the 2020 US Presidential Election. In *16th International AAAI Conference on Web and Social Media*.

[64] Karishma Sharma, Yizhou Zhang, Emilio Ferrara, and Yan Liu. 2021. Identifying Coordinated Accounts on Social Media through Hidden Influence and Group Behaviours. In *KDD'21*.

[65] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[66] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences* 115, 49 (2018), 12435–12440.

[67] Vishnuprasad Padinjaredath Suresh, Gianluca Nogara, Felipe Cardoso, Stefano Cresci, Silvia Giordano, and Luca Luceri. 2023. Tracking Fringe and Coordinated Activity on Twitter Leading Up To the US Capitol Attack. *arXiv preprint arXiv:2302.04450* (2023).

[68] Antonela Tommasel and Filippo Menczer. 2022. Do Recommender Systems Make Social Media More Susceptible to Misinformation Spreaders?. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 550–555.

[69] Christopher Torres-Lugo, Manita Pote, Alexander C Nwala, and Filippo Menczer. 2022. Manipulating Twitter Through Deletions. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1029–1039.

[70] Christopher Torres-Lugo, Kai-Cheng Yang, and Filippo Menczer. 2022. The Manufacture of Partisan Echo Chambers by Follow Train Abuse on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1017–1028.

[71] Twitter. 2018. Enabling further research of information operations on Twitter. https://t.co/PvDKC3mPuu.

[72] Twitter. 2018. How Twitter is fighting spam and malicious automation. https://t.co/GsZawFVWpd.

[73] U.S. Census Bureau. 2015. 2009-2013 American Community Survey. https://www.census.gov/data/tables/2013/demo/2009-2013-lang-tables.html. (2015).

[74] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[75] Emily L Wang, Luca Luceri, Francesco Pierri, and Emilio Ferrara. 2022. Identifying and characterizing behavioral classes of radicalization within the QAnon conspiracy on Twitter. *arXiv preprint arXiv:2209.09339* (2022).

[76] Derek Weber and Frank Neumann. 2021. Amplifying influence through coordinated behaviour in social networks. *Social Network Analysis and Mining* 11, 1 (2021), 1–42.

[77] Samuel C Woolley and Douglas R Guilbeault. 2017. Computational propaganda in the United States of America: Manufacturing consensus online. *Computational Propaganda Research Project* (2017), 22.

[78] Kai-Cheng Yang, Francesco Pierri, Pik-Mai Hui, David Axelrod, Christopher Torres-Lugo, John Bryden, and Filippo Menczer. 2021. The covid-19 infodemic: Twitter versus facebook. *Big Data & Society* 8, 1 (2021), 20539517211013861.

[79] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Who let the trolls out? towards understanding state-sponsored trolls. In *Proceedings of the 10th acm conference on web science*. 353–362.