Detecting Propaganda in Trending Twitter Topics in India—A Metric Driven Approach

K. Sree Hari, Disha Aravind, Ashish Singh, and Bhaskarjyoti Das

Abstract Analysis of conversations and communities in twitter is essential in order to derive insights on how people react to a certain trending topic. Most participants have a stance towards a topic and many of them possess a bias. Bias in the current context refers to users who endorse or oppose a particular text without bothering to understand it fully and refuse to revise their stance solely because of affiliation towards a certain group or people. Because of their activity intensity and connectivity profile, many such biased users act as propagandist. Certain propagandists are not even human i.e. they are software programs or BOTs. So, a fair analysis of stance, bias, propaganda and elimination of BOT activities around a trending topic is necessary to make any fair estimate of people's opinion about such a topic. In this paper, for a recent trending topic in India i.e. "CAA", the whole pipeline is implemented and an analysis framework is built around a set of metrics. The first metric is around tweeting and re-tweeting timing pattern whereas the second metric is around originality content of tweets of a user. In the analysis of a dataset of CAA related tweets, two stance groups (for and against) were identified, BOT users were eliminated and finally percentage of propagandist users from the two stance groups were analyzed and compared. The framework yields promising as well as insightful results and can be used as well as extended for any such analysis in future.

Keywords Sentiment · Opinion · Stance · Bias · Propaganda · BOT · Twitter · CAA

K. Sree Hari · D. Aravind (✉) · A. Singh · B. Das
PES University, Bengaluru, India
e-mail: dishaaravind1998@gmail.com

K. Sree Hari
e-mail: sreehari98@gmail.com

A. Singh
e-mail: ashish.singh.6223@gmail.com

B. Das
e-mail: bhaskarjyoti01@gmail.com

A. E. Hassanien et al. (eds.), *Emerging Technologies in Data Mining and Information Security*, Advances in Intelligent Systems and Computing 1300,
https://doi.org/10.1007/978-981-33-4367-2_62

1 Introduction

Over the recent years, we have noticed a considerable increase in the usage of social media to express opinions related to trending topics. India has a lot of users who actively express their opinion and debate about ongoing events on twitter. There are various ways in which people react to tweets i.e. likes, retweets and replies. We are interested in the retweets and replies as they account for spread of information and give an insight into the minds of the user.

By analyzing the twits, one can decide what stance a user takes towards a topic. Stance refers to what side he is on regarding the topic i.e. he could either be "for", "against", or "neutral" towards it. In most cases, the linguistic, semantic and syntactic features help us get that insight. However every user who expresses his/her view on twitter is not unbiased. Bias in this context refers to stance that is very rigid irrespective of the arguments other sides put forward. It is observed that biased users sometimes spreading his/her opinion like a propagandist. Current studies show that there are a vast number of users who affect the views of other users by either constantly retweeting, or just spamming tweets.

Recent events have shown that social media has a heavy influence on the way users think or feel about a particular situation. There are a multitude of instances where the reaction of users leads someone to believe that a certain topic is trending. These are not always accurate since there is a lot of bias in the people who tweet. This paper provides a more accurate report on the reactions towards a certain topic by classifying users into sub-categories and eliminating bias. These are highly informed biased users, blind biased users, and BOTs. In this work, we intend to come up with an approach that can classify a given user's bias and give statistical and visual analysis on the views of these users. Given a trending topic, this framework and the results will help us infer what percentage of the users are biased towards or against the given topic of discussion.

2 Related Work

2.1 Stance Detection in Twitter Conversations

Sentiment analysis, opinion mining and stance detection are closely related. Sentiment analysis focuses on finding the overall polarity of a text whereas opinion analysis is focused on doing the same for a particular aspect or feature of an entity such as product mentioned in the text. The stance detection is a special type of opinion mining that captures the disposition of the user about a topic being discussed in the text. In online forums such as twitter, often many topics are discussed and stance detection is very relevant for understanding what users think about these topics. Stance itself can be positive, negative or neutral about a topic. Hence, if framed as a supervised learning problem, the usual metrics of precision, recall and F score

are applicable. Stance detection can be multi-stance detection as well where a text has multiple topics and related stance. However, this work is limited to single stance and single topic tweets.

When structured as a multi-class classification problem, feature engineering is one approach and this uses multitude of affective features such as sentiment, emotion and argument lexicons, linguistic features (n-grams, POS tags, repetitions, cue words etc), syntactic features from dependency parsing and semantic derived features (Latent Dirichlet Allocation for example). Existing researches for stance detection have used all kinds of learning techniques:

1. Traditional supervised methods [6, 10, 12, 18, 27] when enough labeled training data is available. Typically these researchers used feature engineering routes.
2. Semi-supervised methods [24] when not enough training data is available. The key assumption here is : both labeled and unlabeled data are part of the same distribution and typically generative models are preferred methods.
3. Transfer learning [32] that takes advantage of models trained on related tasks.
4. Weakly supervised learning [3, 30, 31] that addresses incomplete, inaccurate and inexact labeling of training data. Semi-supervised learning is a strategy to address incomplete labeled data . Ensemble strategy is typically followed to address inaccurate labels. Deep learning, with its ability to learn features automatically, is a typical method to address inexact labels.
5. Deep learning [11, 31] is not really a separate learning strategy and can fit in several of the above categories.

2.2 When Stance Becomes Bias

In the context of echo chamber [7, 14], filter bubble and political homophily, bias detection is very important. Filter bubble is a term coined to signify intellectual isolation. Echo chamber is a related term used to explain the phenomena when a biased twitter user likes to toss his idea only on users who have the same bias or homophily.

Social media itself can be biased [26] in terms of what its API gives out as data, the user demographics and their sources of news stories. There is existing work [17] about detection of bias in language. Typically there are bias words in a sentence and this approach depends on finding linguistic features to detect biased language. Like in stance detection, recently neural techniques [16] have been used to detect bias in language as well.

Unlike bias in language, detection of topic oriented bias also has been attempted. Kwon et al. [21] does opinion mining about topics and detects bias based on contrasting views on topics in twitter. According to them, a biased user will be an outlier. They adopt an approach of contrasting individual preference with social preference extending concepts of opinion mining. Pinkesh Badjatiya et al. [4] say that focusing on bias words inherently biases the training phase in the learning process. The

approach followed is knowledge based generalization policies to detect and remove bias sensitive words. Nagaraju Vadranam et al. [28] adopt an interesting approach using a two stage learning model. In the first stage, each tweet is considered a vector in space determined by some derived features. Based on some manually labeled dataset, the model classifies tweet into three classes (positive, negative and neutral) with respect to bias. Having done the above, in the second stage, they do a fuzzy clustering to further refine the associated class. Abhinav Mishra et al. [23] took a network centric route to assess the bias of a node. The assumption is : bias can be considered as the propensity of the node to trust or mistrust all its neighbours in the implicit network. This framework requires an implicit network of nodes where edges are weighted by the trust score.

However, unlike stance detection, bias detection does not have a benchmark at the time of this work. Recently, a data set has been created for detecting topic bias in articles [8] but not on tweets.

2.3 When Bias Becomes Propaganda

Biased users become propagandists when they propagate their stance across the implicit network taking advantage of their activity intensity and large number of connections. Emilio Ferrara et al. in their discussion about organic and promoted campaigns [13] gives important insights into how twitter communities spread propaganda. It looks at the frequency of tweets over the duration of which the two campaigns that are compared are run. We observe a stark difference over the time it takes for the topic to become trending. The promoted campaign takes around 40–60 hours to become trending, whereas the organic one takes much longer to become trending. They explore 5 major classes which they consider in order to detect communities which perform campaigns, these are - Network and diffusion features, User account features, Timing features, Content and language features, Sentiment features. Recently, there has been lot of research [1, 2, 5, 19, 29] to detect propaganda around extremist activity, abusive adult content in twitter. Propaganda and "clickbaits" (headlines that catch attention) go hand in hand. For propaganda detection, mostly machine learning is used with feature engineering using textual, psychological and behavioural properties of twits.

Lumezanu et al. [22] have explored the characteristics of a propagandist user. It identifies twitter user patterns based on a number of parameters. These users are found to express the same opinion repeatedly. The topics in discussion are the Nevada senate elections and the debate surrounding the 2011 debt ceiling. There are some key patterns that are identified are: tweeting a high number of tweets during small periods of time, little exclusive content in the profile, retweeting quickly and associating with other unrelated users. These are some of the key features that were identified that profile a user as propagandist or not.

2.4 When BOT Takes Over the Propaganda Function

BOTs are software programs that emulate human users in social media and can perform repetitive tasks. BOTs became popular in USA presidential election [15] in 2016. This is part of "fake news" and "computational propaganda" that has become an integral part of both "activism" as well as "control". It has come to a point when "real news" can be less compared to propagandist fake news. Clearly they exhibit both stance and bias and it is important to eliminate these from any research data set to the extent possible to achieve any reasonable analysis. "Political Astroturfing" [20] takes all these to a different high by coordinated disinformation campaign and can potentially change outcomes of national events.

3 Data Set and Data Collection

In this work, data sets were used for two purposes. First, it is used for training the stance classifier and then a real world twit data set is used for a case study using the proposed framework.

The data sets that we used for training the stance classifier were the SemEval 2016 dataset [25] which contained a set of around 4500 tweets which are stance annotated. This data set deals with topics such as atheism, Hillary clinton and feminist movements. The second data set we used was the Kaggle tweet dataset on demonetisation that were collected over six months and stance annotated.

To get inferences using the proposed framework, we manually scraped over 10,000 tweets on the "CAA" topic. Since a significant number of tweets were in Hindi, we used the Google translate API to translate them to English and performed stance detection on them. We started out by scraping the latest tweets without paying any heed to the user's history. Then, on studying the nature of propaganda, bias and how it propagates through a system, we decided to change the scraping approach. We focused on the well-connected components of the graph and tried to analyse the influential users affecting their respective audience. Standard python libraries such as "twint" and "tweepy" are used in the data collection process.

3.1 User Timeline Approach

We decided to extract only the tweets on a user's timeline that were related to the topic of interest instead of trying to get all tweets from all users' timelines. This allows us to mine a large number of users. Other metrics are used to detect bias on the basis of values of propaganda being spread throughout the sub-network.

3.2 User Database Approaches

The objective of our methods was to try and add highly connected users and also to make sure that it is a very unbiased selection of users so that we can maintain that the results that are found on the sub-set of the total network is applicable to the whole network.

3.2.1 User and Follower Approach

This approach involved manually picking some important figures in the political discussion of the topic of interest and fetching their followers by the use of twint. These connections did not allow us to infer any behavior changes or opinions and was rejected.

3.2.2 Influencer Approach

This approach involves two halves. The first half is the extraction of the top tweets about said topic based on the number of times they have been retweeted. As the search terms used were only topical, the mining remains unbiased. We performed a sort of binary search to isolate the top 10–1000 tweets about the topic with maximum number of retweets. Following the trend that it is usually public figures, celebrities and active twitter users that have a large number of following and reactions on their content, we extracted all the topic-specific tweets of each of these 'Influential Users' using twint and then in the second half, used tweepy to get the users that have retweeted any tweet on the influencer's wall. Though tweepy only gives the latest 100 retweeters, we compensate by including many influencers so as to exploit varied and continuous activity on social media. This is aided by the fact that users tend to retweet tweets that are already trending. Thus, there is a high degree of completeness to the database.

3.2.3 Expansion Through Connections

After adopting the influencer approach, a boost added to this technique is to expand further using the retweets of the general users. The database is further expanded by looking at the writers of the retweets of each user. This in particular increases the connectivity of the graph to a great degree and helps in case the influencer list is too small. A continuous alternating process of expanding the retweeters of an influencer and authors of retweeted tweets is started. This should eventually give the complete graph when paired with the simple twint search. But due to the vastness of the data, we have constrained the searches till twint throws a root error.

4 Methodology

Figure 1 shows the overall methodology used. At the high level, first tweets related to "CAA" are collected and then the corresponding stances were detected using stance detection model. The tweets having neutral stance were rejected. The BOTS were then filtered and the rest of the users were then analyzed with respect to the metrics and appropriate visualization is done.

4.1 Stance Detection

The pre-processing involved replacement of symbols such as /(){}[] and numbers, lower-casing tweets and removal of the stop words. Furthermore, we resolved the hashtags to improve the results of stance. Like in case of #IndiaSupportsCAA, we resolved it to 'India Supports CAA' to make the results better. Dataset used for our model was annotated with a stance. There are three possible tags i.e. 'F' for favor, 'A' for against and 'N' for none.

We started out with a stance model that was trained on the SemEval dataset but then trained it on the Demonetisation dataset to help compensate for tweets which are in English but meant to be interpreted in Hindi. This is a strategy adopted for weakly supervised learning as the number of samples in each dataset was not very large. We attempted to create a classifier which would detect stance automatically for given input of tweets.

We used GloVe embedding for getting our vector embeddings of the tweets. After tweets are processed, we extracted and tokenized the words. Each word is

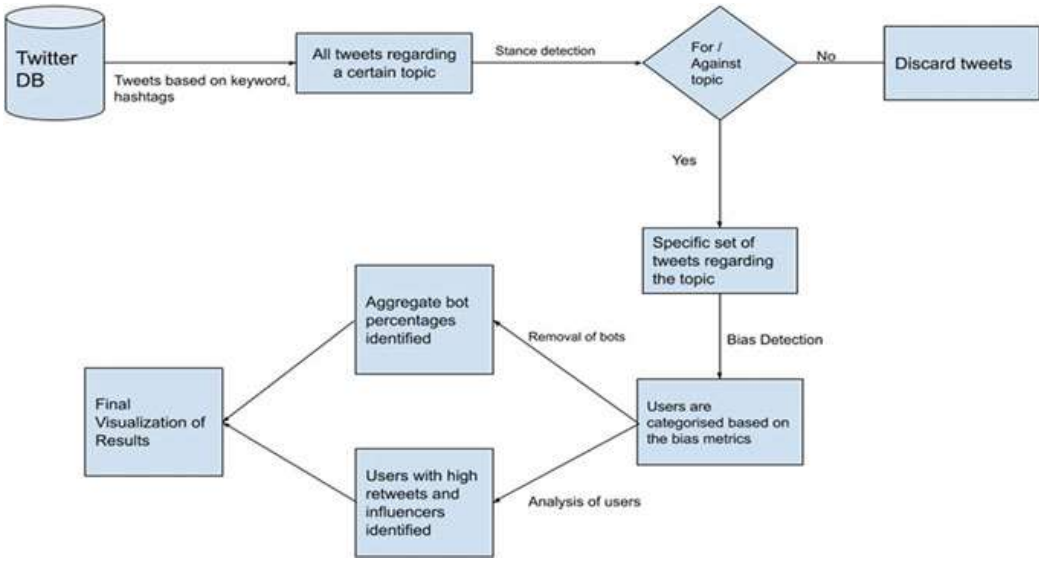


Fig. 1 Overall methodology

mapped to a word embedding. After tweets are tokenized, each word's presence was checked in the GloVe dataset. If it was present then we added its corresponding vector representation from GloVe dataset. If the word was not present then we added a vector filled with zeros of same dimensions of the vector representation in the GloVe embedding.

4.1.1 Shallow Model

We trained our model using several models like 'Gradient Boosting Classifier', 'Logistic Regression', 'Neural Network' and used the one which gives us the best accuracy. We compared the accuracy of three models. The performance was optimal when the Logistic regression classifier was used.

4.1.2 Deep Model Based on CNN and LSTM

This model implements a multi-class classifier using the combination of Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM). Each word is mapped to vector representation also known as a word embedding, such that an entire tweet can be assigned to an $m \times n$ sized matrix, where n is the dimension of the embedding space and m is the number of words in the tweet (we chose $n = 100$). We used the zero padding strategy such that all tweets have the same matrix dimension. These embeddings were fed into the convolution layer to extract features. Three different sizes of filters and concatenation of the convolution over all possible windows of words in the tweet are used. A max-pooling layer is added to each convolution to select the most important features from each convolution. CNN extracts the most significant features in the embedding space. The model now connects to a hidden neural network and add a dropout layer to reduce over-fitting. These outputs are then fed into our LSTM model. We fixed the disadvantage of the LSTM cell losing context of the sentence by using a Bidirectional LSTM to read the sentence from the beginning and from the end. After the LSTM layer, again a dense layer was added. Finally the softmax layer gives us the final probabilities for multi-class classification.

4.2 BOT Removal

For BOT removal, an open source API known as Botometer which was earlier known as BotOrNot [9] has been used. Botometer extracts over 1,000 pieces of information about each account, including measures of sentiment, time of day, account follow list, the tweet content and its Twitter network. Botometer [9] enables us to detect bot accounts by giving a Complete Automation Probability (CAP) value. This value determines if a given account is a bot or not. The results we obtained are 0.0466

for users that have stance as "Against" and 0.0598 for users that are "For". When we average this out over all the users that we tested on, we notice that 1 in 70 user accounts are bot accounts. These accounts were searched and verified to be bot accounts.

4.3 Propaganda Metrics

Based on our literature survey, we chose to identify several extreme tweeting patterns that characterize and help identify users who spread propaganda:

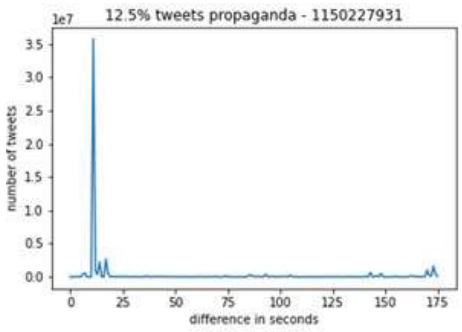
- **Sending high volumes of tweets over short periods of time**: Tweeting activities like replying and publishing own content require a minimum limit of time. Anything less than that indicates the tendency to push content that contributes to propaganda. A user that spreads propaganda has the trend in the time difference between his twitter activities as seen in Fig. 2a. Meanwhile, a genuine user tends to have the following trend in his retweet activity as seen in Fig. 2b. The user's tweets tend to be appropriately spaced out though time.
- **Quickly retweeting**: This means that the user has retweeted several tweets without truly reading or comprehending the tweet. Though we cannot find the exact time at which the users are exposed to the tweet, we can find and eliminate such users who retweeted too quickly. The reweet content per minutes in such cases follows a pattern as shown in Fig. 2c.
- **Retweeting while publishing little original content**: Users that simply retweet others tweets while not publishing a lot of original content are more likely to be uninformed propagandists. While a high amount of replies indicates that the user is more informed, it still contributes to the bias in the network. The retweet graph of a propagandist user when sorted by activity per minute looks like the Fig. 3a. Whereas the original content pushed by the same users is considerable lesser and the trend of high tweet activity per minute is maintained as seen in Fig. 3b.

5 Results and Analysis

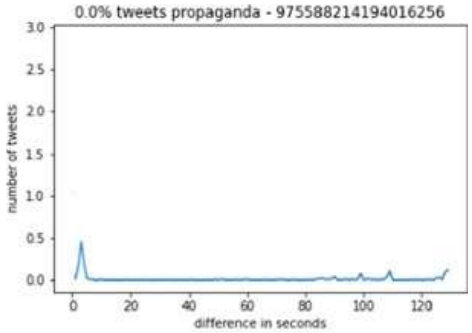
5.1 Metric Representation

Using the first two propaganda metrics, we defined a single metric to check the time difference between any two tweeting activities and found if it is a propaganda spreading tweet by introducing a lower limit on the time difference between them.

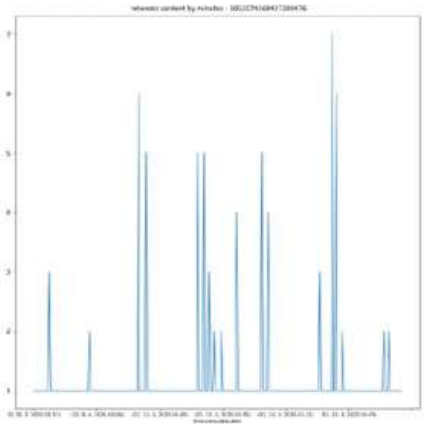
An average person reads around 200 words per minute with 60% comprehension of the content of the tweet. As we are dealing with recent and trending events, we assume that the comprehension is higher due to presence of context and information.



(a) A Biased user

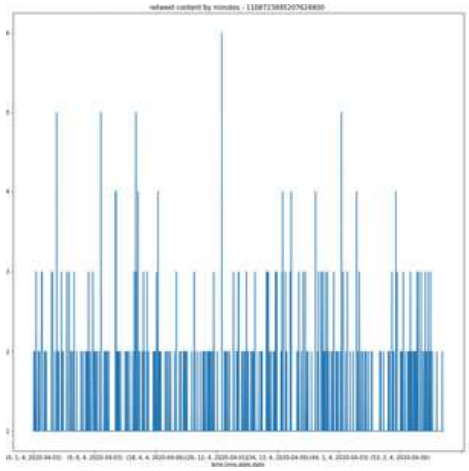


(b) An Unbiased user

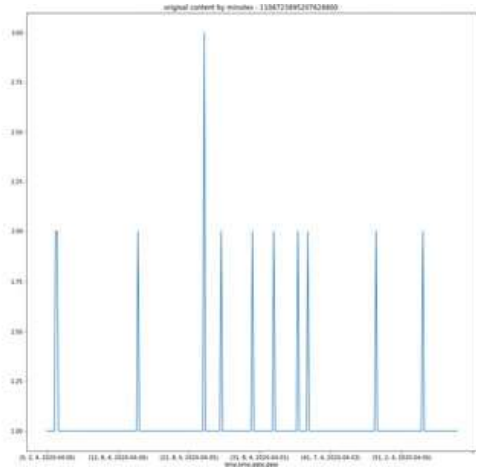


(c) Quick retweets

Fig. 2 A comparison of tweet characteristics



(a) A Biased user



(b) An Unbiased user

Fig. 3 A comparison of re-tweet characteristics

Re-reading a certain part of the tweet or regression and comprehension after reading contributes 300–500 ms to the total time. As we are dealing with political tweets with high amounts of emotion, connected pictures or articles and often an element of sarcasm or satire, we consider the higher limit of 500ms for comprehension. The average length of a political tweet or a tweet that contains Hindi words, as is the case with most of our database, is 75 characters out of the allowed 280 tweets. Taking into consideration that the average number of characters per word is approximately 5, the average number of words per tweet is 75/5 i.e. 15. Hence, the time taken to read and comprehend per tweet should be 15 * (200/60) + 0.5 s. This gives us our lower limit on the time for a well-informed and understood tweet activity as 50.5 s. According to this, the rate at which a person can read and understand tweets is 1.18 tweets per minute. If a person participates in retweeting, replying to or writing more than 2 tweets in a minute, they are participating in ill-informed propaganda.

We used the above logic to get the value of Metric 1 which indicates the level of propaganda as the percentage of propagandist tweet activity in total tweet activity per person. Metric 2 can be easily defined from the statistics of user activity that can provide insight on his originality content. These are then observed within the sub-graphs of each stance to compare the levels of propaganda.

Metric 1 = (total number of propagandist tweets) / (total number of tweets by the user.)

Metric 2 = ((% of original tweets) − (% of retweets) − (% of replies))/2.

5.2 Visualizations

We found that 1.94% of all tweets that were generated by a sample of users regarding CAA were contributing to the bias. Furthermore, 16% of all users were participating in the spread of propaganda.

Based on the Complete Automation Probability by the Botometer tool, further visualisation was done on the twitter implicit retweet-reply graph to give the following intermediate result without the stances. Figure 3 shows us the nature of the CAA related graph where many users retweet the tweets written by a small set of influencers. The four dense bunches of nodes are the sets of said influencers. In the graph shown by Fig. 5, green indicates the stance For CAA, red indicates Against CAA and blue is for users with neutral stance. This graph has edges which are either retweets or replies with retweet edges having greater weight (Fig. 4).

Table 1 lists the metrics derived after filtering users only with ten or more tweets. Note that propagandist users are those biased users who by their tweeting characteristics adopt a propagandist role. As per the metrics chosen, the biased users in the tables are the propagandist users.

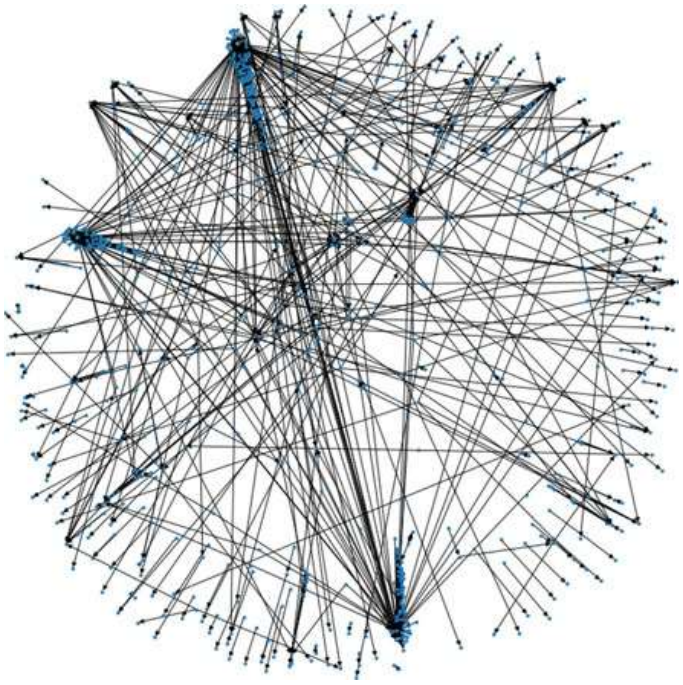


Fig. 4 Dense nodes representing influencers

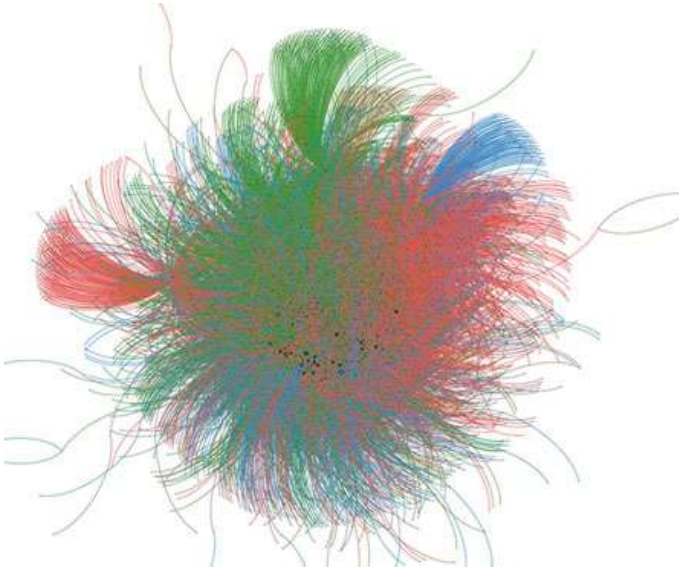


Fig. 5 Stance comparison

Table 1 Propagandist user percentages filtered over users with more than 10 tweets

For CAA	Against CAA
Total users 956	Total users 1173
Total propagandist users 487	Total propagandist users 516
Propagandist users 50.94%	Propagandist users 43.99%
Mean propaganda 3.79%	Mean propaganda 3.19%
Total unoriginal users 73	Total unoriginal users 133
Mean unoriginality 3.57%	Mean unoriginality 6.10%

The conclusions drawn from the graphs and the derived metrics are:

1. Even with a lesser number of users, the For-stanced group occupies a greater majority of the graph. This can be explained by the high value of propaganda metric and high percentage of propagandist users.
2. Another observation made by our analysis framework is that there is a high percentage of "unoriginality" in the Against-stanced community.
3. The percentage of users that are spreading a bias increases substantially when the trivial users are removed, meaning that most of the bias is contributed to the network by influencers and high content users whose presence is detected by the increased measure of originality on the removal of trivial users.
4. As expected, a few users are highlighted strongly as they are the producers of the content that is used by the rest of the network.

6 Conclusion and Future Work

Trying to establish metrics to be able to classify a user as propagandist or not is an unsolved problem and has many angles. The current study that comes close to dealing with what we aim to achieve talks about the characteristics of users who spread the same ideologies or opinions. It does not take into account the influence factor of these propagandists.

This paper establishes metric based framework and is able to provide a visual representation that can accurately display the effect of all these propagandists. The results obtained so far exemplify this intended behaviour of the metrics that were created. These metrics can further be improved to be able to profile any type of user and gain insights on what are the various types of users that are reacting to a certain topic. We are able to classify users based on the amount of propaganda that they are spreading and give a realistic estimate about how the population feels about the topic without all the bias. Although our model predicts stance, it does not take into account sarcasm. In future, this is an addition that will help improve the accuracy of the stance detection. We intend to make modifications to the stance detection model and train it with a more extensive data set to observe improvements in performance.

References

1. Abozinadah, E.A., Jones Jr, J.H.: A statistical learning approach to detect abusive twitter accounts. In: Proceedings of the International Conference on Compute and Data Analysis, pp. 6–13 (2017)
2. Ashcroft, M., Fisher, A., Kaati, L., Omer, E., Prucha, N.: Detecting jihadist messages on twitter. In: 2015 European Intelligence and Security Informatics Conference, pp. 161–164. IEEE (2015)
3. Augenstein, I., Vlachos, A., Bontcheva, K.: Usfd at semeval-2016 task 6: Any-target stance detection on twitter with autoencoders. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 389–393 (2016)
4. Badjatiya, P., Gupta, M., Varma, V.: Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In: The World Wide Web Conference, pp. 49–59 (2019)
5. Benigni, M.C., Joseph, K., Carley, K.M.: Online extremism and the communities that sustain it: Detecting the is supporting community on twitter. PloS one **12**(12), (2017)
6. Benton, A., Dredze, M.: Using author embeddings to improve tweet stance classification. In: Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, pp. 184–194 (2018)
7. Colleoni, E., Rozza, A., Arvidsson, A.: Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. J. Commun. **64**(2), 317–332 (2014)
8. Cremisini, A., Aguilar, D., Finlayson, M.A.: A challenging dataset for bias detection: the case of the crisis in the ukraine. In: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, pp. 173–183. Springer (2019)
9. Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F.: Botornot: A system to evaluate social bots. In: Proceedings of the 25th international conference companion on world wide web, pp. 273–274 (2016)
10. Dey, K., Shrivastava, R., Kaushik, S.: Twitter stance detection–a subjectivity and sentiment polarity inspired two-phase approach. In: 2017 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 365–372. IEEE (2017)
11. Dey, K., Shrivastava, R., Kaushik, S.: Topical stance detection for twitter: a two-phase lstm model using attention. In: European Conference on Information Retrieval, pp. 529–536. Springer (2018)
12. Dias, M., Becker, K.: Infufrgs-opinion-mining at semeval-2016 task 6: Automatic generation of a training corpus for unsupervised identification of stance in tweets. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 378–383 (2016)
13. Ferrara, E., Varol, O., Menczer, F., Flammini, A.: Detection of promoted social media campaigns. In: Tenth International AAAI Conference on Web and Social Media (2016)
14. Garrett, R.K.: Echo chambers online?: politically motivated selective exposure among internet news users. J. Comput.-Med. Commun. **14**(2), 265–285 (2009)
15. Howard, P.N., Bolsover, G., Kollanyi, B., Bradshaw, S., Neudert, L.M.: Junk News and Bots During the Us Election: What Were Michigan voters sharing over twitter. CompProp, OII, Data Memo (2017)
16. Hube, C., Fetahu, B.: Neural based statement classification for biased language. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 195–203 (2019)
17. Hube, C., Jäschke, R., Fetahu, B.: Towards Bias Detection in Online Text Corpora. Jo Bates Paul D. Clough Robert Jäschke p. 19
18. Igarashi, Y., Komatsu, H., Kobayashi, S., Okazaki, N., Inui, K.: Tohoku at semeval-2016 task 6: Feature-based model versus convolutional neural network for stance detection. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 401–407 (2016)

19. Jones, M.O.: The gulf information war| propaganda, fake news, and fake trends: the weaponization of twitter bots in the gulf crisis. Int. J. Commun. **13**, 27 (2019)
20. Keller, F.B., Schoch, D., Stier, S., Yang, J.: Political astroturfing on twitter: how to coordinate a disinformation campaign. Political Commun. **37**(2), 256–280 (2020)
21. Kwon, A., Lee, K.S., et al.: Opinion bias detection based on social opinions for twitter. J. Inf. Proc. Syst. **9**(4), 538–547 (2013)
22. Lumezanu, C., Feamster, N., Klein, H.: # bias: measuring the tweeting behavior of propagandists. In: Sixth International AAAI Conference on Weblogs and Social Media (2012)
23. Mishra, A., Bhattacharya, A.: Finding the bias and prestige of nodes in networks based on trust scores. In: Proceedings of the 20th international conference on World wide web, pp. 567–576 (2011)
24. Misra, A., Ecker, B., Handleman, T., Hahn, N., Walker, M.: Nlds-ucsc at semeval-2016 task 6: A semi-supervised approach to detecting stance in tweets. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 420–427 (2016)
25. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: Semeval-2016 task 6: Detecting stance in tweets. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 31–41 (2016)
26. Morstatter, F.: Detecting and mitigating bias in social media. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1347–1348. IEEE (2016)
27. Tutek, M., Sekulić, I., Gombar, P., Paljak, I., Čulinović, F., Boltužić, F., Karan, M., Alagić, D., Šnajder, J.: Takelab at semeval-2016 task 6: Stance classification in tweets using a genetic algorithm based ensemble. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 464–468 (2016)
28. Vadranam, N., George, K., Demings, S.M.: An analysis of slant in tweets: case study. In: Proceedings of the 6th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, pp. 59–62 (2019)
29. Volkova, S., Shaffer, K., Jang, J.Y., Hodas, N.: Separating facts from fiction: linguistic models to classify suspicious and trusted news posts on twitter. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 647–653 (2017)
30. Wojatzki, M., Zesch, T.: ltl. uni-due at semeval-2016 task 6: stance detection in social media using stacked classifiers. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 428–433 (2016)
31. Xu, C., Paris, C., Nepal, S., Sparks, R.: Cross-target stance classification with self-attention networks. arXiv preprint arXiv:1805.06593 (2018)
32. Zarrella, G., Marsh, A.: Mitre at semeval-2016 task 6: Transfer learning for stance detection. arXiv preprint arXiv:1606.03784 (2016)