Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 9, August 2021: 7933 - 7941

Survey On Detection And Mitigation Of Hate Spreads In Social Networks.

Ms.Amruta Vijay Surana
Research Scholar
School of Computer Science & Engg.,
Poornima University Jaipur, INDIA

Dr.Geeta Chhabra Gandhi
Professor
School of Computer Science & Engg.,
Poornima UniversityJaipur, INDIA

Abstract

Spread of hate contents through any medium creates serious consequences in terms social unrest and affects social harmony. Nowadays social media is being used as tool to launch hate spreads. Due to limited control of content creation and faster propagation, social media has become a easy target for hate spreads. This work surveys the existing methods to detect spread of hate contents and mitigate the effect of hate spreads in social network. The aim of this work is to identify the gaps in the existing works and discuss the prospective solutions to address these gaps. .

I. INTRODUCTION

Social networks have connected billions of people allowing them to express their ideas and share it to wide group of audiences. Any information can spread faster in social media due to its global reach and connectivity. Though these information sharing creates many benefits in terms of advertisements, education, entertainment, public awareness etc, it is also being used to launch hate spreads. These hate spreads can create social unrest and disrupt social harmony. Much of the world now communicates on social media, with nearly a third of the world's population active on Facebook alone. As more and more people have moved online, experts say, individuals inclined toward racism, misogyny, or homophobia have found niches that can reinforce their views and goad them to violence. Social media platforms also offer violent actors the opportunity to publicize their acts. Now a day's there is increased instances of hate spreads promoting violence against other people on basis of race, national origin, language, religion etc all over the world. The role of hate spreads in genocide of Rohingya community in Myanmar [1] and anti Muslim mob violence in Srilanka [2] are well documented. Though Government and social media sites are making various efforts to curb hate speech, it is still plaguing the society and every year there is a increase in number of hate spreads. Social media platforms rely on a combination of artificial intelligence, user reporting, and staff known as content moderators to enforce their rules regarding appropriate content. Moderators, however, are burdened by the sheer volume of content and the trauma that comes from sifting through disturbing posts, and social media companies don't evenly devote resources across the many markets they serve.

In this situation, understanding the spread of hate and various mechanisms to mitigate the spread of hate speeches has become a research interest. Many solutions have been proposed for autonomous detection of hate spreads and mitigation mechanisms. This survey analyzes the existing works on detection and mitigation of hate spreads across social media. The aim is to identify the gaps in the existing works and discuss prospective solutions.

II. SURVEY

The survey is conducted in two categories of content based and behavior based hate spread detection.

A. Content based analysis

Zheng et al (2021) proposed a rumor detection method in social media based on improved transformer. Positional encoding features extracted from contents are used to detect rumor. The

method is dependent of position of words and needs large volume of training dataset for higher accuracy.

Wang et al (2021) proposed a reinforcement learning based rumor detection model. Deep learning features are extracted from tweets and replies and a Deep recurrent Q learning network is trained to classify rumor. The training and reward function is not optimized , due to which there is higher false positives in this method. Guo et al (2021) proposed a fuzzy detection system for rumors using explainable adaptive learning. It is unsupervised model and can work best for zero day problem. However the false positives are higher in this method. Kumar et al (2021) proposed a hybrid model for rumour classification using deep learning (Convolution neural network) and a filter-wrapper (Information gain—Ant colony) optimized Naive Bayes classifier. The textual features learnt using CNN are combined with optimized feature vector generated using filter wrapper technique to classify rumor. The accuracy is low in this method and this can be improved by combining with meta features like re-tweet and user based features. Salminen et al. (2020) examined several classification techniques (Logistic Regression, Naïve Bayes, Support Vector Machines, XGBoost, and Neural Networks) representation methods (Bag-of-Words, TF-IDF, Word2Vec, BERT, and their combinations) for analyzing hate mail. Though all the models are significantly more effective than the keyword-based baseline classifier, XGBoost consistently performs best (F1 = 0.92). In the analysis of feature importance, predictability is most strongly influenced by BERT features. The best model can be generalized because Twitter and Wikipedia's platform-specific results are comparable to the sources of those results. But this model cannot learn hate-related content dynamically because it needs to be trained with hate-related content each time. Castelle et al. (2018) integrate sociolinguistic and linguistic anthropological theories to examine the concept of “language ideologies” — beliefs about language and ways to speak about language, which are used in current machine learning-based NLP practices that abuse linguistic classification. A conceptual and empirical argument is made, by reviewing abusive language approaches from a variety of fields, and use two neural network techniques to investigate three datasets developed for abusive language classification tasks (derived from Wikipedia, Facebook, and StackOverflow). Evaluation and comparison of these results suggest integrating pragmatics & metapragmatism both in classification task design and machine learning architectures. This approach has the problem of making neural networks classifiers complex and labor-intensive to process because a vast volume of abusive text knowledge is involved. Salminen et al. (2018) examined 137,098 comments from online news outlets and manually categorized 5,143 hateful expressions in YouTube and Facebook videos. Researchers then compiled a granular classification of the different types and targets of online hate and then employed machine learning techniques to detect and categorize the hateful remarks across the entire dataset. Their contribution consists of two parts: 1) defining granularly the types and targets of hateful online communications, and 2) building a machine learning model that can automatically detect and categorize hateful comments in the context of online news media using Logistic Regression, Decision Tree, Random Forest, Adaboost, and Linear SVM. In the authors' study, Linear SVM was found to have the best performance, with an average F1 score of 0.79 based on TF-IDF features. However, the method produces lots of false positives, and it is not compatible with quickly spreading hate. Watanabe et al. (2018) presented an algorithm for detecting hate speech on Twitter. The approach relies on the automatic collection of unigrams and patterns on the training data. Unigrams and patterns are then used in machine learning algorithms as features. The results of experiments with 2010 tweets show that this method performs to 87.4% accuracy in classifying tweets as offensive or not (binary classification), with an accuracy of 78.4% in predicting whether the tweets are rude, aggressive, or clean (ternary classification). However, it may not be able to classify zero-day patterns. Lee et al. (2018) developed a decision support system that is capable of detecting (obfuscated) abusive text by detecting its skip-gram and cosine similarity and detecting it through unsupervised learning. As part of the company's abusive text filtering system, numerous efficient tools are used, including blacklists,edit-distance metrics, n-grams, abbreviations, mixed languages, punctuation, and words with specialized characters allow for the intentional obscuration of abusive content. Our system enhances abusive and non-abusive word lists by combining unsupervised learning and efficient gadgets. Despite having unsupervised learning ability, the system only learns from word similarities, omitting contextual similarity and online social network message dynamics. Pitsilis et al. (2018) addressed the issue of detecting racist or sexist statements by using a classifier ensemble of recurrent

neural networks (RNNs), with various integrated features associated with user-related information, including the tendency of users to be racist or sexist. Together with the vectors of word frequency derived from the textual content, these data are then fed into the above classifiers. Through the history of tweets, users are able to determine their tendency toward neutrality, racism, and sexism. Vectorization of the word-based frequency of input tweets is used in the modelling process.

That is, the index values of each word in a tweet are based on how often each word appears in a tweet based on its index value. The index value is derived from the vector of elements that describe the tweet in question. This work has the disadvantage that tendency correlation must be done manually and supervised. Mathew et al. (2019) outlined a system to counter hate speech without affecting individuals' freedom of speech. An author has created a dataset of YouTube comments as counterspeech for the first time. A total of 13,924 annotations are included in the data, which include labels describing whether an annotation is a counter-speech. It is the first time that the linguistic structure of counterspeech can be characterized with such rigor from this data. The analysis reveals several interesting insights, including that the counterspeech comments receive a greater number of likes as compared with the comments from non-counterspeeches, non-counterspeeches tend to be more hateful in some communities, counter-speech tactics vary ineffectiveness and the language used by users posting counterspeeches differs significantly from those posting non-counterspeeches, psycholinguistic research found. To conclude the study, a set of machine learning models is developed in order to detect counterspeech automatically. Using counterspeech effectively can reduce hate speech, but this research ignored ways to make counterspeech more reliable. Qian et al. (2018) presented a novel approach to automated hate speech detection on Twitter using two sources of representation learning: intra-user and inter-user. To model intra user representations of tweets, authors analyze and collect user history in addition to the target tweet. In order to reduce the noise in a single Twitter post, the authors also modeled all of the similar Tweets from other users using what is called 'reinforced inter-user representation learning'. Besides the detection of potential hate speech, this method is also helpful in identifying suspicious social media accounts. As online hate speech has been correlated to real-life hate actions, this solution can provide insight into real-life hate groups and extremists. There are zero-day problems with this method, so its accuracy is limited until the user reaches a certain threshold.

Salminen et al. (2018) investigated hate speech's subjectivity as well as how it varies by country and individual. The authors enlist crowd workers from 50 countries, ask them to score the comments on social media for toxicity, after which they compare the scores, which results in 18,125 points. In their study, they found significant differences among countries in interpretation scores. Nevertheless, the interpretations of hate vary more between individual raters than between countries. In light of these findings, users' characteristics should be considered when scoring and processing hate speech online. Study findings show that hate detection systems need to incorporate user-level features. As crowd workers are anonymous, this approach has a limitation in that author's often miss crucial background information, like age, occupation, gender, education level, ideology, and other variables that may affect the interpretation of hate. We were not able to obtain these variables for the study because of the anonymity of the crowd.

Kshirsagar et al. (2018) describe a neural network method for identifying online hate speech overall, including racist and sexist content. Using word embeddings that have been previously trained and combining the maximum/average of the fully connected transformations of the embeddings, authors are able to predict hate speech occurrence as well as the frequency thereof. Although the study found that word embeddings were robust, the results were based on a very small dataset.

Kohatsu et al. (2019) present HaterNet, an intelligent system developed by the Spanish National Office Against Hate Crimes of the Spanish State Secretariat for Security with the purpose of identifying and monitoring hate speech on Twitter. Several contributions are made by this study: (1.) It presents a first system that monitors the use of hate speech in social media, using analysis of social networks. (2) It introduces 6000 expert-labeled tweets of hate speech in Spanish, which are publicly available for the first time. (3) This study compares different methods of document classification and document representation based on text classification. (4) A neural network that uses LTSM and MLP is the best approach. By analyzing the tweets' words, emojis, and expression tokens embedded in the TF-IDF, the network can analyze the content. The system uses content analysis to detect variations in content but is unable to detect semantic variations. Watanabe et al. (2018) developed a method for

detecting hate expressions on Twitter. Patterns and unigrams can be automatically collected as part of this process. Machine learning algorithms are later trained on these patterns and unigrams, amongst others. Based on the number of unigrams in the hate speech vocabulary, this system's accuracy is determined.

Zhou et al. (2020) federated the results of different classifiers to improve hate speech detection accuracy. These authors analyze several well-known machine learning methods for text classification, including Embeddings from Language Models (Elmo), Bidirectional Encoder Representation from Transformers (BERT), and Convolutional Neural Network (CNN), based on the data included in SemEval 2019 Task 5.

To improve the overall classification performance, the authors then employ fusion strategies to combine the classifiers. Fusion after two separate classifications was the focus of this work; however, it was not sufficiently integrated. Mozafari et al. (2018) conducted a study using a pre-trained language model based on the use of a novel transfer learning approach called BERT (Bidirectional Encoder Representations from Transformers). In their research study, the authors examined the capability of BERT in capturing hateful context embedded within social media content through fine-tuning to account for transfer learning. Biases of different types cannot be detected with this method.

Mondal et al. (2017) present the first systematic study of hate speech in online social media, measuring and analyzing the amount and type of hate speech. In this study, researchers examined the prevalence of hate speech in social media, the most common forms of hate speech, the effect of anonymity on hate speech, and the most hated groups throughout the world. Specifically, two social media platforms are used by the authors: Whisper and Twitter. Both systems are then tested to determine whether hate speech is detected. Findings suggest that hate speech comes in many different forms, and reveals important patterns, not only offering directions for detection and prevention approaches but also revealing important trends in hate speech. Although no detection system was implemented, hate speech forms were mined and identified. Alshalan et al. (2020) built a network in which we searched for hate speech related to the COVID-19 pandemic posted on Twitter by Arab users in order to analyze the content of the hate speech posted. An analysis of tweets was performed to identify hate speech using Convolutional Neural Networks (CNN) algorithms; tweets were scoreable from 0 to 1, with 1 being the most hateful. To discover how hate tweets are discussed, the authors also used nonnegative matrix factorization. In this study, content analysis was used exclusively and the dynamics of message spread were not taken into account. As well, hate speech can't be quantified. Roy et al. (2020) designed an automated system to use Deep Convolutional Neural Networks (DCNN) to detect hate speech. We propose a DCNN model that uses tweet text combined with GloVe embeddings to extract tweet semantics through convolutional neural networks. As a result of using a limited training set, accuracy is lower at 64%.

Unsvag et al. (2018) identified user features that could be influential in hate speech classification.

In order to better understand Twitter users, Twitter data was analyzed quantitatively, but an insignificant correlation was found between hate speech and the characteristics of the people posting it. Combining certain user factors with textual attributes could however result in improved performance, which was demonstrated in experiments with a hate speech classifier based on datasets from three different languages. The impact of incorporating user features on performance varied according to the datasets, but features relating to users' networks tended to have the greatest impact. The proposed work focuses on four aspects of the user, but can also be extended to many features and relationships among those features. Berglind et al. (2019) propose a method for automatically determining hate levels, by using transfer learning on annotated language models combined with pre-trained language models. Currently, the solution does not include a more detailed evaluation of the nature and level of hatred in a given environment. ElSherief et al. (2018) explored an aspect of hate speech that has been neglected – its target: generalized expressions directed toward groups that share a common characteristic. These two forms of hate speech are analyzed for the first time linguistically and psycholinguistically in order to be able to identify interesting differences between them. The analysis reveals that redirected hate speech is more personal and directed, so on top of being more personal and directed, it is more informal, angrier, and works to attack the target (via name-calling), relying less on analytical and more on authority-related words. Hate speech dominated by religious hatred, on the other, consists mainly of lethal language such as kill, exterminate, murder, and words with large quantities such as million and many. This study, combined with the work of others, allows

for a comprehensive analysis of online hate speech providing a deeper understanding of hate speech and its social repercussions as well as its detection. Due to keyword-based matching, this approach misses thousands of instances of hateful speech.

.

B. Behavior based analysis

Weiss et al (2021) proposed a rumor detection system based on time series characteristics like volume of tweets, number of followers, followees of users involved in the story etc. A Siamese network is trained with time series of training samples. Though the method performs better than lexical and structural characteristics, it needs large volume of training data set and it does not address the zero day problems. Han et al. (2019) investigated the spread of the model rumor. Several security threats already exist in social networks related to rumors. Unidirectional networks are even more prone to malicious intrusion and propagation because the unidirectional connections, created by partnering with social networks and E-commerce websites, provide an extra opportunity for malicious penetration. There are some unscrupulous car dealers who claim that fuel cars will become more expensive due to the upcoming environmental protection policy. Rumors about fuel cars made some consumers purchase them on impulse. The speculators were delighted and devoted themselves to spreading the rumor among business circles. The rumor was soon believed by more people who were trapped. Last but not least, the rumor caused a great deal of controversy both on social networks and in the business world. The market order and stability would be severely harmed. According to the S2I2R model, rumor spreading evolves over time. There follows a mean-field formula for the model. The approach, however, has not been implemented in any concrete way. Yan et al. (2019) investigated the minimization of rumor influence (MIR) problem, which is finding a blocker set B with k nodes in which the activation probability of users in the user set S reaches a minimum. A classical independent cascade (IC) model is employed by the authors as a model for the diffusion of information. Using the IC model, we demonstrate that the objective function is monotone decreasing and non-submodular. In order to solve the MIR problem, the authors suggest a two-stage approach based on generating candidate sets for general networks and selecting their blockers. In addition, the authors examine the MIR problem on the tree network and present a dynamic programming approach that guarantees an optimum result. The authors then employ simulations to evaluate proposed algorithms using simulated synthesized social networks based on real-life examples. The rumored model does not take into account severity or rumor spread, and it is not content-specific. Tong et al. (2020) investigated the problem of rumor blocking in online social networks. Optimal rumor blocking can be achieved with a randomized sampling method based on R-tuples. Compared with existing rumor blocking algorithms, the proposed RBR algorithm is theoretically superior, as the experiments have shown, it is greatly efficient without sacrificing blocking effectiveness. The model, however, is generic and does not specifically relate to pandemic-related rumor spread. Ye et al. (2020) simulate real-world social network behavior by combining information spreading mechanisms. This allows the author to estimate the risk level at which each node will be exposed during the hazard period and analyze the threat level a node's infection might pose to other nodes in the network. To analyze the rumor-spreading path, authors use Rumor Path Trees (RPTs). If the propagation of rumor and truth is compared, then it is possible to estimate the steps the rumor node has taken to propagate. The authors calculate the effective influence node using a fractional function then select it from the pool of generated RPTs with the highest score. Authors are able to prevent rumors from spreading using the truth node. The model does not focus on modeling rumors spread in pandemics. Huang et al. (2020) developed a cost-effective approach to insulating online social networks from rumor spreading by developing an optimal control strategy. The first step is to propose a new model of rumor spreading that takes individuals into account, taking the external environment into consideration for the first time in rumor spreading. The cost-effectiveness of rumor-containing schemes is determined by balancing losses from rumors with the cost of a rumor-containing scheme. In this respect, the authors reduced the original problem to an optimal control model. This model is then proven to be solvable and provides the optimality system for the model. Wang et al. (2019) examine the dynamics of positive and negative information diffusion and analyze the ways to curb negative information diffusion. The authors determine the critical condition for the diffusion of negative information when

both positive and negative information is simultaneously present, and calculate the critical condition for the diffusion of negative information, as well as certify the accuracy of the diffusion model. A second collaborative control strategy is proposed to help persuade users that positive information should be spread simultaneously. A control problem emerges from the issue of minimizing the total cost of a system. Last but not least, the authors demonstrate that optimal control strategies exist and are unique, as well as determining their distribution overtime over different time periods to minimize system costs. The collaborative nature of the process means, if trust among collaborators is not taken into account, the spread of negative information can be difficult to control. Yin et al. (2020) introduced the multiple-information-vulnerability-discussing-immune model (M-SDI) aimed at understanding how key information propagates in social networks. The M-SDI model is developed by looking at the volume of public discussion, as well as considering users' behavior to re-enter related topics or post to Weibo after a discussion. By fitting the model with COVID-19 opinion data obtained from public Chinese blogs can be parameterized accurately to predict public opinion until the next major news event takes place. The model takes into account the fact that users are likely to re-enter the susceptible (to a news item) state of information in relation to that item after discussing an item in detail. Models such as this cannot predict how public opinions will change. Zan et al. (2018) examined the spread of double rumors with different launch times and proposed two models: double-susceptible-infected-recovered (DSIR) model and comprehensive-DSIR (C-DSIR) model, taking into account the interactions between old and new rumors, as well as two rumors published consecutively. Studying the expressions and attraction of various state-vector rumors, they came up with the double-rumors dissemination mechanism. Models like this do not undergo testing for real time spreading scenarios

III. SURVEY SUMMARY

Following are the findings from the survey.

- Content-based analysis approaches suffers from zero day problems for new attacks, as they can reason only based on training. For reliable content based analysis, a sufficient training volume must be available. Given the diversity of traffic generated every data in social network, it becomes difficult to construct the training volume. Even if constructed, they have lot of gaps, due to which many cases failed to be detected.
- Most behaviour based analysis model are short term and are not suited for pandemic spreads like COVID-19 which lasts for many months. Most existing works on trend analysis are short term in day or week. But pandemic like tends last for years. So fine tuning of behaviour based models for long term trends is needed.
- Counter hate messages spreading are not automated and are not targeted for maximum hate removal efficiency. There is also no metric to quantify the effectiveness of counter hate messages. Counter hate messages is the solution adopted to compensate for hate. But currently there is no standard way to decide the effectiveness of counter hate message and decide the propagation spread. This is needed for constructing a effective defense mechanism against hate spread.
- Most approaches does not consider various population attributes like demographics, user level features etc. Hate is dependent on many factors like user profile, demographics, context etc. The methods to detect hate must consider all these factors. The current mechanism for hate detection only analyses the content features without consideration for external factors like demographics, user level features etc.
- Automated generation of counter messages based on content analysis is still lacking. There is no way to generate the counter messages automatically based on analysis of contents. Machine learning based techniques can be designed to generate counter messages automatically. There are many challenges in it like emotions expression, multi lingual, negation etc.
- Preventing rumour spread at source level based on content and context is still lacking. Rumors or hate speeches must be detected earlier to prevent the ill effects of it. To detect rumor at source level, content, context and its effect must be analyzed. Machine learning techniques can be used for detection of rumor spreads at source level.

IV. RESEARCH GAP

Absence of long term behavior analysis

There is no long term behavior analysis method. This is needed for COVID like pandemic situations which last over many years. Most of the models for predicting rumor spread are short time time-series forecast on volume of tweets. But they don't consider chain of rumors created from same stem rumor. Also the existing methods cannot identify behaviors which are dormant for some intermittent times and becomes active in same form or in cloned form. Hate messages can also propagate cross platform.

Lack of measurement for effectiveness of counter hate spread

There is no way to measure the effect of counter hate spreads and effective strategies to improve the effect of counter attack. Currently, social media companies adopt two approaches to fight misinformation. The first one is to block such content outright. For example, Pinterest bans anti-vaccination content and Facebook bans white supremacist content. The other is to provide alternative information alongside the content with fake information so that the users are exposed to the truth and correct information. This approach, which is implemented by YouTube, encourages users to click on the links with verified and vetted information that would debunk the misguided claims made in fake or hateful content.

But there is no mechanism to find the effectiveness of counter hate spreads and selection of suitable strategy for countering the hate propaganda. The current mechanism for blocking the account based on report is not effective and any malicious group of user can block the legitimate account. This necessitate exploring different mechanisms for counter hate spread and automatic strategies for selection of suitable counter spread strategy based on the effectiveness measured lively from the social network.

Prevent hate spread at source level

There is no effective way to prevent hate spread at source level. If this can be done, the effect of hate spread can be minimized. The latency time between hate origination and time it detected is very high, by that time hate spread is compounded and gets cloned cross platform. The latency must be minimized to avoid this compounding and hate spread must be prevented at source level or at some intermediate levels from source. There are currently no mechanisms to provide a guarantee of reducing the latency in detection and preventing at source level.

V. CONCLUSION

The survey analyzed the existing solutions on detection and mitigation of hate spreads across social network. The survey analyzed both content based and behavior based methods. The research gaps in existing works were found. The study also presented prospective research directions to address the research gaps.

.

REFERENCES

1. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate
2. https://goo.gl/QMU8e7
3. H. Zheng, H. Yu, Y. Hao, Y. Wu and S. Li, "Rumor Detection Based on Improved Transformer," 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML), 2021, pp. 247-253
4. Wei Wang, Yuchen Qiu, Shichang Xuan, Wu Yang, "Early Rumor Detection Based on Deep Recurrent Q-Learning", Security and Communication Networks, vol. 2021, Article ID 5569064, 13 pages, 2021.
5. Z. Guo, K. Yu, A. Jolfaei, A. K. Bashir, A. O. Almagrabi and N. Kumar, "A Fuzzy Detection System for Rumors through Explainable Adaptive Learning," 2021 in IEEE Transactions on Fuzzy Systems, doi: 10.1109/TFUZZ.2021.3052109.
6. Kumar, A., Bhatia, M.P.S. & Sangwan, S.R. Rumour detection using deep learning and filter-

wrapper feature selection in benchmark twitter dataset. Multimed Tools Appl (2021).
7. Salminen, J., Hopf, M., Chowdhury, S.A. (2020). Developing an online hate classifier for multiple social media platforms. Hum. Cent. Comput. Inf. Sci.
8. Salminen J, (2018). Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In: Proceedings of the international AAAI conference on web and social media (ICWSM)
9. Castelle M. (2018). The linguistic ideologies of deep abusive language classification. In: Proceedings of the 2nd workshop on abusive language online (ALW2), Brussels; 2018
10. Watanabe H.(2018). Hate speech on twitter: a pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE Access.
11. Lee H-S.(2018) An abusive text detection system based on enhanced abusive and non-abusive word lists. Decis Support Syst
12. Pitsilis GK, (2018). Detecting offensive language in tweets using deep learning. arXiv preprint arXiv:1801.04433
13. Mathew.(2019). Thou shalt not hate: countering online hate speech. In: Proceedings of the 13th international AAAI conference on web and social media (ICWSM-2019). Munich
14. Qian J, (2018). Leveraging intra-user and inter-user representation learning for automated hate speech detection. In: Proceedings of the conference of the north american chapter of the association for computational linguistics: human language technologies, volume 2 (Short Papers), New Orleans;
15. Kshirsagar R, (2018). Predictive embeddings for hate speech detection on twitter. arXiv preprint arXiv:1809.10644
16. Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and Monitoring Hate Speech in Twitter. Sensors (Basel, Switzerland), 19(21), 4654.
17. Y. Zhou, Y. Yang, H. Liu, X. Liu and N. Savage,(2020). "Deep Learning Based Fusion Approach for Hate Speech Detection," in IEEE Access, vol. 8, pp. 128923-128929
18. M. Mozafari, R. Farahbakhsh, and N. Crespi.(2019).‘‘A BERT-based transfer learning approach for hate speech detection in online social media,’’ in Int. Conf. Complex Netw
19. Mondal, Mainack & Silva, Leandro & Benevenuto, Fabrício. (2017). A Measurement Study of Hate Speech in Social Media. 85-94.
20. Alshalan, Raghad & Al-Khalifa, Hend & AlSaeed, Duaa & Al-Baity, Heyam & Alshalan, Shahad. (2020). Detection of Hate Speech in COVID-19–Related Tweets in the Arab Region: Deep Learning and Topic Modeling Approach. Journal of Medical Internet Research.
21. P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao.(2020). A Framework for Hate Speech Detection Using Deep Convolutional Neural Network. in IEEE Access, vol. 8, pp. 204951-204962, 2020, doi: 10.1109/ACCESS.2020.3037073
22. Unsvåg, Elise & Gambäck, Björn. (2018). The Effects of User Features on Twitter Hate Speech Detection.
23. Berglind, Tor & Pelzer, Björn & Kaati, Lisa. (2019). Levels of hate in online environments. 842-847
24. M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. M. Belding. Hate lingo .(2018). A target-based linguistic analysis of hate speech in social media. CoRR, abs/1804.04257, 2018
25. F. Weiss, M. Mendoza and E. Milios, "Time series classification for rumor detection," 11th International Conference of Pattern Recognition Systems (ICPRS 2021), 2021, pp. 176-181
26. J. Han, S. Wu, and X. Liu.(2019). Identifying and categorizing offensive language in social media. in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 652–656
27. R. Yan, D. Li, W. Wu, D. -Z. Du and Y. Wang(2020). Minimizing Influence of Rumors by Blockers on Social Networks: Algorithms and Analysis. in IEEE Transactions on Network Science and Engineering, vol. 7, no. 3, pp. 1067-1078, 1.
28. Tong, W. Wu, L. Guo, D. Li, C. Liu, B. Liu.(2020). An efficient randomized algorithm for rumor blocking in online social networks. IEEE Trans. Netw. Sci. Eng., vol. 7, no. 2, pp. 845-854
29. S. Ye, J. Wang and H. Fan.(2020). Minimize Social Network Rumors Based on Rumor Path Tree," in IEEE Access, vol. 8, pp. 167620-167630
30. Huang, Da-wen & Yang, Lu-Xing & Yang, Xiaofan & Tang, Yuan & Bi, Jichao. (2020).

Defending against Online Social Network Rumors through Optimal Control Approach. Discrete Dynamics in Nature and Society.
31. Wang, X. Wang, F. Hao, G. Min, and L. Wang.(2019). Efficient coupling diffusion of positive and negative information in online social networks. IEEE Transactions on Network and Service Management, vol. 16, no. 13, pp. 1226–1239.
32. Yin, Fulian & Lv, Jiahui & Zhang, Xiaojian & Xia, Xinyu & Wu, Jianhong. (2020). COVID-19 information propagation dynamics in the Chinese Sina-microblog. Mathematical Biosciences and Engineering.
33. Y. Zan, DSIR double-rumors spreading model in complex networks, Chaos Soliton.(2018).Fract., 110, 191-202.