International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 03 | Mar 2022 www.irjet.net p-ISSN: 2395-0072

An Analytical Survey on Hate Speech Recognition through NLP and Deep Learning

Sagar Mujumale, *computer dept., KJCOEMR, Pune*

Bogiri Nagaraju, *computer dept., KJCOEMR, Pune (Guide)*

***Abstract* — Hate speech it is one of the most problematic occurrences that have been gaining a lot of traction in the recent years. The growth of hate speech and hate related crimes have been inexplicably interlinked which have been the main concern of violent crimes that have been happening in the recent years. Due to the introduction of the internet platform hate speech has been gaining traction in a lot of different media rather than just text. There are various audios, videos and other media that have been circulated on the online platforms for the purpose of inciting hatred towards a particular community or a group of people. Most of the online social networks have been trying very hard to combat this problem but due to the complex nature of hate speech and its recognition it has been an uphill task of late. Therefore to provide solution to this kind of problem and to elevate the concerns surrounding hate speech this survey paper analyses past works on hate speech recognition to achieve our methodology for the same. The methodology achieved in this analysis will be detailed for the in the upcoming publications and research articles on this topic.**

Keywords: *Fuzzy Artificial Neural Networks, Natural Language Processing, Bag of Words, Entropy Estimation.*

I INTRODUCTION

Communication is one of the most integral part of the day to day activities of a human being. There is a universal need for socialization has been observed and well documented in humans and other primates. The socialization is almost entirely based on communication between individuals. The main aim of the socialization is the promotion of healthy conversation that enables effective improvement in the mood and the realization of social connections. The lack of communication or socialization can be extremely detrimental to the overall wellbeing of an individual which can lead to a lot of mental health problems.

There are a wide variety of techniques that are used to socialize or communicate with one another, but one of the most common methods is through the use of speech or language. This is one of the most potent and the most preferred forms of socialization that is prevalent across the world. This can be noticed as the large selection of varied forms of languages that have been evolved and being used all over the globe. There have been insightful storytelling and other forms of communication over the course human evolution that have resulted in the development of language that we see and use today.

In the recent years, there has been a proliferation of the information or knowledge at a rapid pace which has been catalyzed through the internet paradigm. The rapid expansion has led to the realization of a varied variety of different and unique implementations. These services have been swiftly achieving new grounds and are facilitating an increased convenience in all walks of life, including socialization. The socialization has been transferred online through the use of online social networks that have been getting increasingly popular in the past few years. There is an influx of new users to these online platforms every single day which has swelled their user base considerably.

These services for socialization provide the users to interact and socialize with their friends and followers through sharing text messages and media such as images, video, etc. The paradigm of social networks allows the users to interact with their followers and share their thoughts which will be conveyed to them. Internet users are drawn to social network sites and tweeting services more than any other type of content. Services like Facebook, Instagram, and Twitter are becoming increasingly popular amongst individuals of all ages, ethnicities, and pursuits. Their contents are growing rapidly, making them a fascinating example of so-called big data. Big data has piqued the interest of researchers who are interested in automating the study of people's ideas and the organization or dispersion of users in organizations, among other things.

Social media is widely utilized to share many types of material. People frequently utilize social media to communicate their thoughts and ideas. Whereas these platforms provide a platform for discussion to explore and express their ideas, the sheer volume of postings, remarks, and conversations makes it nearly difficult to maintain control over the quality. Furthermore, because of the diversity of origins, customs, and beliefs, many people use violent and abusive words while conversing with others from varied ethnicities.

Social media has changed in popularity as a fantastic way to express one's sentiments and sentiments. However, under the guise of freedom of expression, growing adoption of social media has led in the proliferation of hate propaganda. Denying that social media is highly quick, open, free, and easy to use, it is also highly susceptible owing to its rapid expansion. It becomes a vehicle for miscreants to disseminate various sorts of hatred or prejudice speech directed at some other community. Hate speech is defined as a rhetoric that is potentially hurtful to a person's or group's sentiments and may promote to violence or lack of compassion, as well as illogical and inhuman action.

Hate speech, which is illegal, has increased as a result of the growth of online social media. Hate speech and racial discrimination are linked, and there is evidence that hate offences are on the rise. As the problem of hateful speech gains momentum, several government-sponsored measures are being implemented.

This literature survey paper segregates the section 2 for the evaluation of the past work in the configuration of a literature survey, and finally, section 3 provides the conclusion and the future work.

II RELATED WORKS

A pattern-based technique to identifying hate speech on Twitter has been proposed by H. Watanabe et al. [1]. The authors construct a set of parameters to maximize the pattern collection by extracting patterns from the training set pragmatically. They also describe a technique for identifying hate speech that accumulates words and phrases pragmatically indicating hatred and offense and mixes them with patterns and other sentiment-based traits. The suggested collections of unigrams and trends will be used as already-built dictionaries for future hate speech identification studies. To differentiate between hostile tweets and those that are simply offensive, they separate tweets into three categories (rather than just two).

K. A. Qureshi et al. recognized the difficult challenge of multi-class automated hate speech, and text classification is handled with significantly better results when the fundamental difficulties are identified first. There are 10 separate binary classified datasets comprised of different hate speech categories [2]. Using a set of thorough, well-defined rules, experts annotated each dataset with a high degree of agreement across annotators. The datasets were well-balanced and thorough. They were also given grammatical nuance. The compilation of such a dataset was completed as a necessary need for filling the gap in the field. Following the compilation of high-quality datasets, a list of effective, commonly used, and recommended text mining features were compiled from relevant text mining research.

L. H. Son et al. introduced sAtt-BLSTM convNet, a hybrid of soft attention-depend bidirectional long short-term memory (sAtt-BLSTM) and convolution neural network (convNet), to detect sarcasm in the brief text (tweets). The network was trained using semantic word embeddings and pragmatic auxiliary features [3]. When analyzed to the baseline models, the proposed model has the greatest classification accuracy for both datasets. The usage of mash-up languages and innovative vocabulary with complex architectures add to the difficulty of automated sarcasm detection and highlight certain outstanding concerns.

P. K. Roy et al. utilize a deep convolutional neural network to disclose hate speech on Twitter. LR, RF, NB, SVM, DT, GB, and KNN were first employed to detect HS-related messages on Twitter, with features extracted utilizing the tf-idf approach. However, using a 3:1 train-test dataset, the strongest ML model, SVM, was only able to properly predict 53% of HS tweets. The imbalanced dataset may be to blame for the poor forecast of HS tweets; as a result, the model is biased towards NHS tweets prediction because it has the bulk of cases. With the fixed partitioned dataset, deep learning-based CNN, LSTM, and their combinations of C-LSTM models produce comparable results [4]. On a given partitioned of train-test, none of the models predicted the HS tweets with adequate accuracy, regardless of whether they were typical machine learning-based models or deep learning-based models.

O. Oriola et al. gathered an English corpus of South African tweets to detect offensive and hate speech [5]. Because the tweets contained varied signals from South African languages, the corpus was annotated by multilingual annotators. After tokenization and preprocessing, four unique feature sets and their combinations were retrieved from the tweets. To classify tweets as hate speech, offensive speech, or free speech, researchers used three types of improved machine learning models: hyper-parameter

optimization, ensemble, and multi-tier meta-learning on distinct machine learning algorithms like Logistic Regression, Support Vector Machine, Gradient Boosting, and Random Forest.

Using embedding representation of words and DNNs, Ashwin Geet d'Sa et al. looked at the multiclass classification of hate speech. The categorization was done on a Twitter data collection using a three-class classification system: hatred, offensive, and neither. For hate speech categorization, they presented feature-based and finetuning techniques [6]. A sequence of word embeddings is utilized as input for the classifiers in the feature-based technique. They looked into fastText embedding and the BERT embedding as word embeddings. The performances of these two forms of embeddings are nearly equal within the scope of a feature-based approach.

For the first time, M. Mozafari et al. looked at the possibility of a meta-learning technique as a viable solution to the problem of few-shot learning in cross-lingual hate speech and offensive language detection tasks [7]. To that purpose, the authors prepared two benchmark datasets for cross-lingual hate speech and offensive language classification tasks using a wide range of publically available datasets comprising hateful and offensive content from many languages. To train a model that can generalize fast to a new language with a few labeled data (k samples per class), the authors used a meta-learning technique depending on optimization-based and metric-based methodologies (MAML and Proto-MAML). The results show that meta learning-depend models beat transfer learning-based models in the majority of scenarios, with Proto-MAML being the top-performing model in terms of identifying hostile or offensive text with a little amount of labeled data.

Y. Zhou et al. presented the principles of three types of text classification methods, ELMo, BERT, and CNN, and utilized them to detect hate speech, then enhanced performance by fusion from two perspectives: fusion of ELMo, BERT, and CNN classification results, and fusion of three CNN classifiers with different parameters. The findings demonstrated that fusion processing can aid in the identification of hate speech [8].

Two novel optimization-based techniques to solving the HSD problem in online social networks are proposed by C. Baydogan et al. For the first time in the research, the most modern metaheuristic algorithms, ALO and MFO, were adopted to solve the HSD issue. To compare the performance of the suggested metaheuristic-based techniques, researchers employed eight different supervised machine learning algorithms, SSO, and state-of-the-art TSA. The pre-processing step was accomplished utilizing NLP approaches for the specified real-world problems linked to HSD. The feature extraction was done by combining the BoW+TF+Word2Vec techniques [9]. Then, to tackle HSD issues, twelve distinct algorithms competed. Except for one dataset, the customized ALO algorithm yielded the greatest accuracy, precision, sensitivity, and f-score values in the research.

M. Z. Ali et al. prepared a complete data collection by obtaining Urdu language tweets and having them categorized on aspect and emotion levels by qualified linguists [10]. There is no current data collection of Urdu hate speech with aspect and emotion levels annotated. The authors used state-of-the-art approaches to solve the three most common difficulties in machine learning-based sentiment analysis, namely sparsity, dimensionality, and class skew, and noticed the performance increase over the baseline model. The classifier was trained using two machine learning algorithms: SVM and Multinomial Nave Bayes. The authors employed dynamic stop words filtering to reduce sparsity, a variable global feature selection scheme (VGFSS) to reduce dimensionality, and synthetic minority oversampling to reduce class imbalance (SMOTE).

F. M. Plaza-Del-Arco et al. focus on HS detection in Spanish corpora and propose an MTL model that incorporates related tasks like polarity and emotion classification. Experiments on two benchmark corpora demonstrate the usefulness of their suggested strategy in outperforming a STLBETO model and obtaining state-of-the-art results [11]. The results of the proposed model, as well as a comprehensive knowledge transfer study from SA, reveal that polarity and emotion classification tasks aid the MTL model in successfully classifying HS by utilizing emotional information. The linked impacts of emotional knowledge and HS open the door to new approaches to developing NLP systems in other fields where polarity and emotion may play a key role.

H. S. Alatawi et al. utilized Deep learning to investigate domain-specific and agnostic word embedding (BiLSTM). The findings suggest that this technique is effective in combating white nationalist hate speech [12]. The BERT model has also demonstrated that it is the most up-to-date solution for this issue. The findings of the experiment demonstrate that BERT surpasses domain-specific strategy by 4 points; however, the domain-specific technique can recognize purposefully misspelled words and popular lingo from the hate community, but the BERT model fails to do so because it is trained on Wikipedia and literature. Some datasets in the experiments are unbalanced to imitate real-world data, while others are balanced to evaluate the model's performance under ideal conditions.

A. Rodriguez et al. introduced FADOHS, a proposal that discovers and integrates unstructured data from Facebook pages that purportedly encourage hate speech to identify the most common subjects addressed [13]. This problem was initially difficult to solve since non-personal Facebook pages and accounts tend to avoid using blatantly explicit phrases in postings to risk getting deleted from the platform or receiving negative feedback. Nonetheless, many sites manage to elicit unpleasant feelings and appear to promote hate speech among their followers by addressing sensitive themes while using relatively innocuous wording. The suggested methodology offers a unique approach to clustering posts and comments, as well as recognizing and identifying hate speech generated by widely discussed themes.

V. -I. Ilie et al. discuss research on ContextAware Misinformation Detection Using Deep Learning Architectures. For multi-class classification, they use two text preprocessing pipelines (Lemma and Aggressive Text Preprocessing), three context-aware word embeddings (Word2Vec, FastText, and GloVe), and ten Neural Networks [14]. Context-aware word embeddings are either pre-trained on the dataset (Generic Word Embeddings) or custom trained on the dataset (Specific Word Embeddings). Depending on these findings, they suggest a preprocessing and classification pipeline. The dataset utilized for the experimental validation comprises 100 000 news articles categorized as either true or false.

Racist remarks on social media sites such as Twitter are becoming more common, and they should be automatically identified and blocked to prevent their dissemination. E. Lee et al. approaches racism detection from a sentiment analysis standpoint, detecting racist tweets by recognizing negative feelings [15]. Deep learning is supplemented by the ensemble technique, in which GRU, CNN, and RNN are stacked to build the GCR-NN model, to achieve high-performance sentiment analysis. A huge dataset gathered from Twitter and annotated with TextBlob is utilized for experimentation with machine learning, deep learning, and the suggested GCR-NN model. Racist statements were found in 31.49 % of the 169,999 tweets collected. Results reveal that deep learning models perform much better than machine learning models, with the suggested GCR-NN achieving an average accuracy score of 0.98 in sentiment analysis for positive, negative, and neutral categories.

III CONCLUSION AND FUTURE SCOPE

This survey paper highlights the past researchers that have been performed for the purpose of hate speech recognition on online social networks and their dissemination in various media formats. Hate speech has been increasing rapidly across the world with large numbers of people coming online to spread their hatred and intolerance towards a particular group of people. This has become a major concern where hate speech has become a potent catalyst for the purpose of inciting violence and achieving nefarious goals and intentions of political nature in a gruesome manner. Therefore the need of the hour is for the recognition and the prevention of hate speech and its proliferation on social media in different formats such as video images and audio etc. For this purpose the indepth analysis of previous researches have been useful in identification and the realization of our methodology which will be detailed in the future.

REFERENCES

[1] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," in IEEE Access, vol. 6, pp. 13825-13835, 2018, DOI: 10.1109/ACCESS.2018.2806394.

[2] K. A. Qureshi and M. Sabih, "Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text," in IEEE Access, vol. 9, pp. 109465-109477, 2021, DOI: 10.1109/ACCESS.2021.3101977.

[3] L. H. Son, A. Kumar, S. R. Sangwan, A. Arora, A. Nayyar, and M. Abdel-Basset, "Sarcasm Detection Using Soft Attention-Based Bidirectional Long Short-Term Memory Model With Convolution Network," in IEEE Access, vol. 7, pp. 23319-23328, 2019, DOI: 10.1109/ACCESS.2019.2899260.

[4] P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, "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.

[5] O. Oriola and E. Kotzé, "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets," in IEEE Access, vol. 8, pp. 21496-21509, 2020, DOI: 10.1109/ACCESS.2020.2968173.

[6] Ashwin Geet d'Sa, Irina Illina, and Dominique Fohr, " Classification of Hate Speech Using Deep Neural Networks" in HAL Open Access, HAL Id: hal-03101938 https://hal.archives-ouvertes.fr/hal-03101938.

[7] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Cross-Lingual Few-Shot Hate Speech and Offensive Language Detection Using Meta-Learning," in IEEE Access, vol. 10, pp. 14880-14896, 2022, DOI: 10.1109/ACCESS.2022.3147588.

[8] Y. Zhou, Y. Yang, H. Liu, X. Liu, and N. Savage, "Deep Learning-Based Fusion Approach for Hate Speech Detection," in IEEE Access, vol. 8, pp. 128923-128929, 2020, DOI: 10.1109/ACCESS.2020.3009244.

[9] C. Baydogan and B. Alatas, "Metaheuristic Ant Lion and Moth Flame Optimization-Based Novel Approach for Automatic Detection of Hate Speech in Online Social Networks," in IEEE Access, vol. 9, pp. 110047-110062, 2021, DOI: 10.1109/ACCESS.2021.3102277.

[10] M. Z. Ali, Ehsan-Ul-Haq, S. Rauf, K. Javed, and S. Hussain, "Improving Hate Speech Detection of Urdu Tweets Using Sentiment Analysis," in IEEE Access, vol. 9, pp. 84296-84305, 2021, DOI: 10.1109/ACCESS.2021.3087827.

[11] F. M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López and M. T. Martín-Valdivia, "A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis," in IEEE Access, vol. 9, pp. 112478-112489, 2021, DOI: 10.1109/ACCESS.2021.3103697.

[12] H. S. Alatawi, A. M. Alhothali, and K. M. Moria, "Detecting White Supremacist Hate Speech Using Domain-Specific Word Embedding With Deep Learning and BERT," in IEEE Access, vol. 9, pp. 106363-106374, 2021, DOI: 10.1109/ACCESS.2021.3100435.

[13] A. Rodriguez, Y. -L. Chen and C. Argueta, "FADOHS: Framework for Detection and Integration of Unstructured Data of Hate Speech on Facebook Using Sentiment and Emotion Analysis," in IEEE Access, vol. 10, pp. 22400-22419, 2022, DOI: 10.1109/ACCESS.2022.3151098.

[14] V. -I. Ilie, C. -O. Truică, E. -S. Apostol and A. Paschke, "Context-Aware Misinformation Detection: A Benchmark of Deep Learning Architectures Using Word Embeddings," in IEEE Access, vol. 9, pp. 162122-162146, 2021, DOI: 10.1109/ACCESS.2021.3132502.

[15] E. Lee, F. Rustam, P. B. Washington, F. E. Barakaz, W. Aljedaani, and I. Ashraf, "Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model," in IEEE Access, vol. 10, pp. 9717-9728, 2022, DOI: 10.1109/ACCESS.2022.3144266.