HATE SPEECH DETECTION USING SENTIMENT AND EMOTIONAL ANALYSIS

[1] M. RAJ KUMAR, [2] B. PRADEEP KUMAR, [3] V. KALYANI

[123]Assistant Professor, Department of IOT, Sri Indu College of Engineering and Technology, Hyderabad, Telangana-501510

ABSTRACT

Social media provides users with an online platform to express themselves freely; yet, when users make unpleasant and opinionated comments that target certain people or communities, this may lead to hostility towards them. Due to the widespread condemnation of obesity (fatness), a lot of fat-shaming material has been put online. The project titled "Framework for Detection and Integration of Unstructured Data of Hate Speech on Facebook Using Sentiment and Emotion Analysis" aims to develop a software framework using Java for detecting and integrating unstructured data of hate speech on Facebook. The proposed framework employs sentiment and emotion analysis techniques to identify and categorize hate speech, and further integrates the data into structured formats for further analysis. The framework provides a user-friendly interface for data input and analysis, and employs sentence classification for enhancing the accuracy of hate speech detection. The project aims to contribute to the growing need for automated tools to detect and combat online hate speech, and to assist researchers and policy makers in analyzing trends and patterns in online hate speech.

INTRODUCTION

Hate speech is a significant problem on social media platforms, with potentially severe consequences for individuals and society as a whole. The rise of social media has facilitated the spread of hateful and discriminatory messages, which can fuel prejudice, bigotry, and even violence. To address this issue, we propose a framework called FADOHS, which stands for "Framework for Automatic Detection Of Hate Speech." FADOHS aims to detect hate speech on Facebook pages and integrate unstructured data through sentiment and emotion analysis [1].

The proposed framework consists of four stages: data collection and preprocessing, sentiment and emotion analysis, hate speech clustering, and evaluation. In the first stage, the system collects the data from Facebook pages and preprocesses it to remove noise and irrelevant information. In the second stage, the system performs sentiment and emotion analysis to understand the underlying emotions and sentiments of the text. In the third stage, the system clusters the text based on the degree of hate speech expressed. Finally, in the fourth stage, the system evaluates the effectiveness of different strategies for data analysis and natural language processing [2].

The system provides two types of applications, one that uses a dataset and one that allows users to post comments. The admin part of the application allows for monitoring of sentiment analysis and other results, which can help to identify potential issues early and take appropriate actions to address them [3].

Overall, the FADOHS system offers a powerful tool for detecting and addressing hate speech on social media platforms. By identifying the most significant factors from the unstructured data, the system can effectively cluster information into different groups according to the degree of hatred being expressed. The system has the potential to make a significant contribution towards creating a more respectful and inclusive online community [4].

LITERATURE SURVEY

Existing Problem

The authors of ``Hate Me, Hate Me Not: Hate Speech on Facebook" have proposed several classification methods to distinguish among different types of hate speech. More specifically, they leverage morpho-syntactic features, sentiment polarity, and word-embedded lexicons to design and implement two classifiers for Italian. Their framework utilizes support vector machines (SVMs) and long short-term memory (LSTM) networks. This study was premised on the concept outlined in Del Vigna et al.'s study and our understanding of hate speech [5].

In the existing system first they recognize a set of pages from American-based websites, known to discuss controversial topics such as immigration, race, and religion. We use these ``seeds", or Facebook IDs, to crawl the Facebook graphs and construct a small network using the ``follow" relationship. Leveraging graph analysis techniques, we identify the most influential pages spreading hate speech and crawl their latest posts and comments. We then apply sentiment and emotion analysis algorithms to recognize posts with highly negative tones, specifically those suspected of instigating hatred. Finally, we convert each post into a document by concatenating all comments, and using the K-means algorithm to create clusters of posts based on the topics they discuss [6].

Disadvantages of Existing System

- Limited Scope: The system only focuses on American-based websites, which limits its ability to detect hate speech in other regions.
- Bias: The system uses a predefined set of controversial topics, which can lead to biases and misses out on newer or less-discussed topics that may also contain hate speech
- False Positives: The sentiment and emotion analysis algorithms used to identify posts with highly negative tones may not always accurately distinguish between hate speech and legitimate expressions of opinion, leading to false positives.
- False Negatives: The system's reliance on the "follow" relationship may result in the omission of influential pages that do not have a large number of followers or likes, resulting in false negatives [7].
- Lack of Context: The system's focus on individual posts and comments may lead to a lack of context, making it difficult to understand the intent behind a particular post or comment.
- Overreliance on Machine Learning: While machine learning algorithms can be useful in identifying patterns and clusters, they may also miss out on subtle nuances and contextual cues that human reviewers would be able to pick up on.

SURVEY 1: Racism, hate speech, and social media: A systematic review and Critique, AUTHORS: A. Matamoros-Fernández and J. Farkas

Departing from Jessie Daniels's 2013 review of scholarship on race and racism online, this article maps and discusses recent developments in the study of racism and hate speech in the subfield of social media research. Systematically examining 104 articles, we address three research questions: Which geographical contexts, platforms, and methods do researchers engage with in studies of racism and hate speech on social media? To what extent does scholarship draw on critical race perspectives to interrogate how systemic racism is (re)produced on social media? What are the primary methodological and ethical challenges of the field? The article finds a lack of geographical and platform diversity, an absence of researchers' reflexive dialogue with their object of study, and little engagement with critical race perspectives to unpack racism on social media. There is a need for more thorough interrogations of how user practices and platform politics co-shape contemporary racisms [8].

SURVEY 2: Hate me, Hate me not: Hate speech detection on Facebook, AUTHORS: F. Del Vigna, A. Cimino, F. Dell-TOrletta, M. Petrocchi, and M. Tesconi

While favoring communications and easing information sharing, Social Network Sites are also used to launch harmful campaigns against specific groups and individuals. Cyber bullism, incitement to self-harm practices, sexual predation are just some of the severe effects of massive online offensives. Moreover, attacks can be carried out against groups of victims and can degenerate in physical violence [9].

In this work, we aim at containing and preventing the alarming diffusion of such hate campaigns. Using Facebook as a benchmark, we consider the textual content of comments appeared on a set of public Italian pages. We first propose a variety of hate categories to distinguish the kind of hate. Crawled comments are then annotated by up to five distinct human annotators, according to the defined taxonomy [10]. Leveraging morpho-syntactical features, sentiment polarity and word embedding lexicons, we design and implement two classifiers for the Italian language, based on different learning algorithms: the first based on Support Vector Machines (SVM) and the second on a particular Recurrent Neural Network named Long Short Term Memory (LSTM). We test these two learning algorithms in order to verify their classification performances on the task of hate speech recognition. The results show the effectiveness of the two classification approaches tested over the first manually annotated Italian Hate Speech Corpus of social media text.

SURVEY 3: The K-means algorithm: A comprehensive survey and performance evaluation,AUTHORS: M. Ahmed, R. Seraj, and S. M. S. Islam

The k-means clustering algorithm is considered one of the most powerful and popular data mining algorithms in the research community [11]. However, despite its popularity, the algorithm has certain limitations, including problems associated with random initialization of the centroids which leads to unexpected convergence. Additionally, such a clustering algorithm requires the number of clusters to be defined beforehand, which is responsible for different cluster shapes and outlier effects. A fundamental problem of the k-means algorithm is its inability to handle various data types. This paper provides a structured and synoptic overview of research conducted on the k-means algorithm to overcome such shortcomings. Variants of the k-means algorithms including their recent developments are discussed, where their effectiveness is investigated based on the experimental analysis of a variety of datasets. The detailed experimental analysis along with a thorough comparison among different k-means clustering algorithms differentiates our work compared to other existing survey papers. Furthermore, it outlines a clear and thorough understanding of the k-means algorithm along with its different research directions [12].

SURVEY 4: BuzzFace: A news veracity dataset with Facebook user commentary and egos.,AUTHORS: G. C. Santia and J. R. Williams

Veracity assessment of news and social bot detection have become two of the most pressing issues for social media platforms, yet current gold-standard data are limited. This paper presents a leap forward in the development of a sizeable and feature rich gold-standard dataset [13]. The dataset was built by using a collection of news items posted to Facebook by nine news outlets during September 2016, which were annotated for veracity by BuzzFeed. These articles were refined beyond binary annotation to the four categories: mostly true, mostly false, mixture of true and false, and no factual content. Our contribution integrates data on Facebook comments and reactions publicly available on the platform's Graph API, and provides tailored tools for accessing news article web content. The features of the accessed articles include body text, images, links, Facebook plugin comments, Disqus plugin comments, and embedded tweets. Embedded tweets provide a potent possible avenue for expansion across social media platforms. Upon development, this utility yielded over 1.6 million text items, making it over 400 times larger than the current gold-standard. The resulting dataset—BuzzFace—is presently the most extensive created, and allows for more robust machine learning applications to news veracity assessment and social bot detection than ever before [14].

SURVEY 5: A path-based model for emotion abstraction on Facebook using sentiment analysis and taxonomy knowledge, AUTHORS: V. Franzoni, Y. Li, and P. Mengoni

Each term in a short text can potentially convey emotional meaning. Facebook comments and shared posts often convey human biases, which play a pivotal role in information spreading and content consumption. Such bias is at the basis of human-generated content, and capable of conveying contexts which shape the opinion of users through the social media flow of information. Starting from the observation that a separation in topic clusters, i.e. sub-contexts, spontaneously occur if evaluated by human common sense, this work introduces a process for automated extraction of sub-context in Facebook. Basing on emotional abstraction and valence, the automated extraction is exploited through a class of path-based semantic similarity measures and sentiment analysis. Experimental results are obtained using validated clustering techniques on such features, on the domain of information security, over a sample of over 9 million page users. An additional expert evaluation of clusters in tag clouds confirms that the proposed automated algorithm for emotional abstraction clusters Facebook comments compatibly with human common sense. The baseline methods rely on the robust notion of collective concept similarity [15].

PROPOSED SYSTEM

In this proposed system we aim to create a framework (FADOHS) that can first detect hate speech on Facebook and then integrate unstructured data through clustering using sentiment and emotion analysis. It identifies the most significant factors from the unstructured data of posts and comments on Facebook pages that allegedly promote hate speech. We will conduct experiments to measure the effectiveness of different strategies for data analysis and natural language processing by implementing four stages of FADOHS. Ultimately, we can demonstrate the strong effectiveness of hate-speech clustering using hybrid methods of data analysis and natural language processing to identify and categorize information into different groups according to the degree of hatred being expressed [15].

In our proposed system we develop 2 types of application where the first one use dataset and the second one using comments posted by the user. In the first type the facebook comments which are referred from kaggle website, are uploaded into the system. Then the preprocessing is done and each and every comment given in the dataset is processed one by one and the Sentiment Score is calculated and the Sentiment type is predicted for the each dataset record. Additionally it computes the Emotion Score, Emotion Type (Happiness, Sadness, Anger, Fear, Disgust, Surprise), Hate Word and Label the dataset record completely. Finally a static Graph is plotted with the results which we received [16].

In the second type we develop an application with two entities: Admin and user. Where user can post the comments and in the admin part the Sentiment Analysis and other results will be monitored. In the admin part mainly we predict the Sentiment Score, Sentiment type, Emotion score, Emotion type and Hate Word.

Advantages of Proposed System

- Efficient prototype: The proposed system can efficiently detect hate speech which is crucial for maintaining a safe and respectful online community. The proposed system is an efficient prototype for detecting the hate speech in Online social media networks like facebook or twitter.
- Integration of unstructured data: The system can integrate unstructured data using sentiment and emotion analysis, which allows for a more comprehensive understanding of the data and can reveal important insights.
- Identification of significant factors: FADOHS can identify the most significant factors from the unstructured data of posts and comments on that allegedly promote hate speech. This can help to focus efforts on addressing the most critical issues [17].
- Strong effectiveness of hate-speech clustering: FADOHS can effectively cluster information into different groups according to the degree of hatred being expressed. This can help to prioritize and target resources towards the most severe cases of hate speech.
- Two types of applications: The system provides two types of applications, one that uses a dataset and one that allows users to post comments.
- Admin monitoring: The admin part of the application allows for monitoring of sentiment analysis and other results, which can help to identify potential issues early and take appropriate actions to address them.
-

IMPLEMENTATION

Input Dataset Module:

In the first module we develop our proposed system to accept the input dataset which is referred from kaggle containing facebook comments of users. This dataset doesn’t contain any label. It takes the comments as input and uses natural language processing techniques to identify language patterns that are indicative of hate speech. Once identified, the module flags the comments as potentially containing hate speech.

Preprocessing Module:

This module is responsible for cleaning and preparing the data for analysis. It removes irrelevant information and transforms the data into a format that can be easily analyzed by other modules.

Sentiment Analysis Module:

This module uses natural language processing techniques to analyze the sentiment of Facebook posts and comments. It calculates a sentiment score for each post and comment, as well as predicting the sentiment type (positive, negative, or neutral).

Emotion Analysis Module:

This module uses natural language processing techniques to analyze the emotions expressed in Facebook posts and comments. It calculates an emotion score for each post and comment, as well as predicting the emotion type (happiness, sadness, anger, fear, disgust, or surprise).

Hate Speech Clustering Module:

This module uses hybrid methods of data analysis and natural language processing to identify and categorize information into different groups according to the degree of hatred being expressed. It clusters similar comments together and identifies the most significant factors from the unstructured data of posts and comments on Facebook pages that allegedly promote hate speech [18].

Static Graph Module:

This module is responsible for plotting static graphs with the results obtained from the analysis. It generates a visual representation of the normal comment and hate speech comment in it.

User Interface Module:

This module provides a user-friendly interface for users to interact with the system. It allows users to upload Facebook comments or post new comments and provides real-time results on the sentiment score, emotion score, emotion type, and hate word.

Admin Monitoring Module:

This module is responsible for monitoring and analyzing the sentiment score, sentiment type, emotion score, emotion type, and hate word of the comments posted by users. It provides insights and analytics.

Algorithm Description

Support Vector Machines (SVM):

SVM is a popular algorithm used for binary classification tasks. It can be trained on sentiment and emotional features extracted from text to distinguish between hate speech and non-hate speech instances. SVM seeks to find an optimal hyperplane that separates the two classes based on the feature vectors.

Naive Bayes:

Naive Bayes is a probabilistic classification algorithm commonly used for text classification tasks. It works on the assumption of independence between features. In the context of hate speech detection, Naive Bayes can be trained using sentiment and emotional features to classify text instances as hate speech or non-hate speech.

Random Forest:

Random Forest is an ensemble learning algorithm that combines multiple decision trees to make predictions. Each decision tree is trained on a random subset of features, and the final prediction is based on the majority vote of the individual trees. Random Forest can be trained using sentiment and emotional features to detect hate speech.

Recurrent Neural Networks (RNN):

RNNs are neural network architectures that can process sequential data, making them suitable for analyzing text. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are popular variants of RNNs. By training an RNN on sentiment and emotional features extracted from text, it can learn the sequential patterns and capture the nuanced context in hate speech detection.

Convolutional Neural Networks (CNN):

CNNs are commonly used for image recognition tasks but can also be applied to text classification. In the context of hate speech detection, a CNN can be trained on sentiment and emotional features represented as text embedding’s. By applying convolutional layers to capture local patterns, CNNs can learn to differentiate between hate speech and non-hate speech instances.

Transformer-based Models:

Transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers), have revolutionized natural language processing tasks. BERT can be fine-tuned on sentiment and emotional features for hate speech detection. It can capture contextual information and semantic relationships in text, leading to improved performance.

These algorithms can be combined with sentiment analysis techniques, such as lexicon-based approaches or machine learning-based models, to extract sentiment features. Additionally, emotion recognition models, such as the Ekman's six basic emotions or more complex models like Affectiva or EmoNet, can be utilized to extract emotional features from text.

It's worth noting that the choice of algorithm depends on factors like the size and nature of the dataset, available computational resources, and desired performance. Experimentation and fine-tuning are often necessary to identify the most effective algorithm for a specific hate speech detection task.

RESULTS

Dataset

For both stages of the intermediate and the target task, we use the HateXplain dataset. It contains 20,148 items collected from Twitter and Gab. Every item consists of one English sentence with its own ID and annotations about labels for its category, target groups, and rationales, which are annotated by two or three annotators. Based on the IDs, the dataset is split into 8:1:1 for training, validation, and test phases. Following the permanent split provided by the dataset, the models can't reference any test data during the training phases of all stages.

The evaluation is according to the metrics of HateXplain, which are classified into three types: performance-based, bias-based, and explainability based. The performance-based metrics measure the detection performance in distinguishing among three classes (i.e., hate speech, offensive, and normal). Accuracy, macro F1 score, and AUROC score are used as the metrics. The bias-based metrics evaluate how biased the model is for specific expressions or profanities easily assumed to be hateful. HateXplain follows AUC-based metrics developed by Borkan et al. (2019). The model classifies the data into 'toxic'– hateful and offensive–and 'non-toxic'–normal.

For evaluating the model's prediction results, the data are separated into four subsets: D+ g , D− g , D+, and D−. The target group labels are considered standard for dividing data into subgroups. The notations with g denote the data of a specific subgroup among the subgroups, and the notations without g are the remaining data. + and − mean that the data are toxic and non-toxic, respectively. Based on these subsets, three AUC metrics are calculated. Subgroup AUC is to evaluate how biased the model is to the context of each target group: AUC(D− g + D+ g). The higher the score, the less biased the model is with its prediction of a certain social group. BPSN (Background Positive, Subgroup Negative) AUC measures the model's false-positive rates regarding the target groups: AUC(D++D− g). The higher the score is, the less a model is likely to confuse non-toxic sentences whose target is the specific subgroup and toxic sentences whose target is one of the other groups. BNSP (Background Negative, Subgroup Positive) AUC measures the model's false-negative rates regarding the target groups: AUC (D−+D+ g). The higher the score is, the less the model is likely to confuse non-toxic sentences whose target is the specific group and toxic sentences whose target is one of the other groups.

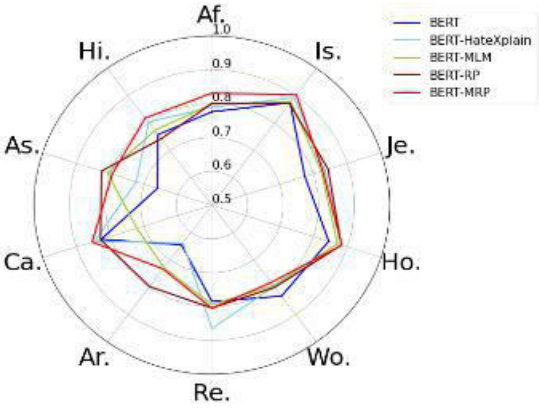

Fig.1.-Graph

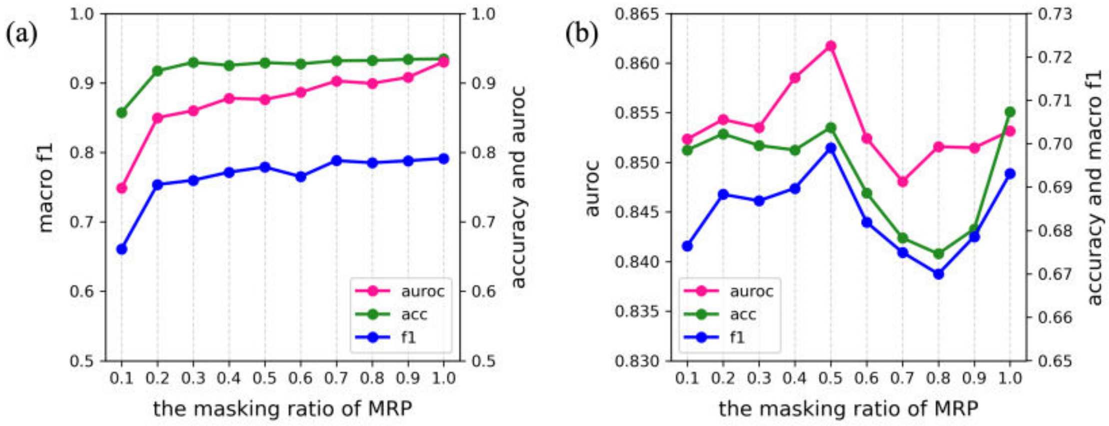

Fig 2.Classification test scores of the proposed models according to masking ratio in MRP. (a) is for token classification after training on MRP in the first stage, and (b) is for hate speech detection in the final stage. The case of masking 100% of tokens is the same as BERT-RP.

CONCLUSION

In conclusion, the proposed system, FADOHS, offers a comprehensive framework for detecting hate speech on Facebook and integrating unstructured data through sentiment and emotion analysis. By identifying the most significant factors from the unstructured data, the system can effectively cluster information into different groups according to the degree of hatred being expressed. This can help to prioritize and target resources towards the most severe cases of hate speech. The system provides two types of applications, one that uses a dataset and one that allows users to post comments, which ensures that the system can be used in a variety of contexts and can be adapted to different user needs. The admin part of the application allows for monitoring of sentiment analysis and other results, which can help to identify potential issues early and take appropriate actions to address them. Overall, the FADOHS system offers a powerful tool for detecting and addressing hate speech on social media platforms. The system has the potential to make a significant contribution towards creating a more respectful and inclusive online community.

REFERENCES

[1] Ahmed .M, Seraj,.R and Islam. S. M. S., (2020) ‘The K-means algorithm: A comprehensive survey and performance evaluation,'

[2] Chinnasamy.S and Manaf,. N. A. (2018) ’Social media as political hatred mode is the general election,'

[3] CNBC. (2020). ‘Facebook's Artificial Intelligence Still Has Trouble Finding Hate Speech but it finds a lot of nudity.’ Accessed (May 11, 2018).

[4] Del Vigna .F, Cimino, A, Dell-TOrletta.F, Petrocchi, . M and Tesconi. M, (2017) `Hate me, hate me not: Hate speech detection on Facebook,' in Proc. 1st Italian Conf. Cybersecur. (ITASEC), Venice, Italy

[5] Facebook. (2020). ‘Community Standards Home.’Accessed: (May 11,2018).[Online]. Available: https://www.facebook.com/communitystandards/.

[6] Fortune. (2018). ‘Facebook Removed 2.5 Million Pieces Hate Speech 1st Quarter.’ Accessed: (Jul. 16, 2018.) [Online].Available:https://fortune.com/2018/05/15/facebook-hate-speech-removals/.

[7] ILGA. (2018). ‘Hate Crime & Hate Speech.’Accessed: (May 6, 2018). Available:https://www.ilga-europe.org/what-we-do/ouradvocacy-work/hate-crime-hate-speech

[8] LV. Z .Liu. T, Benediktsson,J.A. and H. Du, (2019)`Novel land cover change detection method based on K-means clustering and adaptive majority voting using bitemporal remote sensing images,' IEEE Access, vol. 7, pp. 34425-34437, 2019.

[9] K. Bhargavi. An Effective Study on Data Science Approach to Cybercrime Underground Economy Data. Journal of Engineering, Computing and Architecture.2020;p.148.

[10] [21] M. Kiran Kumar , S. Jessica Saritha. AN EFFICIENT APPROACH TO QUERY REFORMULATION IN WEB SEARCH, International Journal of Research in Engineering and Technology. 2015;p.172

[11] K BALAKRISHNA,M NAGA SESHUDU,A SANDEEP. Providing Privacy for Numeric Range SQL Queries Using Two-Cloud Architecture. International Journal of Scientific Research and Review. 2018;p.39

[12] K BALA KRISHNA, M NAGASESHUDU. An Effective Way of Processing Big Data by Using Hierarchically Distributed Data Matrix. International Journal of Research.2019;p.1628

[13] P.Padma, Vadapalli Gopi,. Detection of Cyber anomaly Using Fuzzy Neural networks. Journal of Engineering Sciences.2020;p.48.

[14] Kiran Kumar, M., Kranthi Kumar, S., Kalpana, E., Srikanth, D., & Saikumar, K. (2022). A Novel Implementation of Linux Based Android Platform for Client and Server. In A Fusion of Artificial Intelligence and Internet of Things for Emerging Cyber Systems (pp. 151-170). Springer, Cham.

[15] Kumar, M. Kiran, and Pankaj Kawad Kar. "A Study on Privacy Preserving in Big Data Mining Using Fuzzy Logic Approach." *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 11.3 (2020): 2108-2116.

[16] M. Kiran Kumar and Dr. Pankaj Kawad Kar. “Implementation of Novel Association Rule Hiding Algorithm Using FLA with Privacy Preserving in Big Data Mining”. Design Engineering (2023): 15852-15862

[17] K. APARNA, G. MURALI. ANNOTATING SEARCH RESULTS FROM WEB DATABASE USING IN-TEXT PREFIX/SUFFIX ANNOTATOR, International Journal of Research in Engineering and Technology. 2015;p.16.

[18] Matamoros-Fernández .A and Farkas,.J, (2021) `Racism, hate speech, and social A systematic review and critique,’Telev New Media,(vol.22,NO.2,PP.205224,Feb 2021)