e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 12, December 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712053

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media

E. Raju[1], Ragula Saketh[2], O.K.V.Krishna[3], P.G.Kushanu[4]

Assistant Professor, Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India[1]

Student, Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India[2,3,4]

ABSTRACT: This study reviews machine learning (ML) methods and algorithms for detecting hate speech on social media (SM). The problem of hate speech is typically modeled as a text categorization task. In this study, we used machine learning techniques to analyze the fundamental baseline elements of hate speech classification. Data collection and exploration, feature extraction, dimensionality reduction, classifier selection and training, and model evaluation are the five fundamental baseline components that were examined. Over time, the machine learning algorithms used to detect hate speech have improved. In the literature, various performance metrics and new datasets have been presented. In order to keep researchers up to date on these developments in automatic hate speech detection, a thorough and modern state-of-the-art is required. This study offers three contributions. Initially, to provide readers with the essential details regarding the crucial procedures in hate speech detection through machine learning algorithms. The second step is to critically assess each method's advantages and disadvantages in order to let researchers make decisions about which algorithm to use. Finally, some unresolved issues and research limitations were noted. The various ML technique variations, such as deep learning, ensemble approaches, and traditional ML, were covered. The results of this study will be extremely beneficial to both professionals and researchers.

I. INTRODUCTION

Social media networks (SMNs) are the fastest means of communication as messages are sent and received almost instantaneously. SMNs are the primary media for perpetrating hate speeches nowadays. In line with this, cyber-hate crime has grown significantly in the last few decades. More researches are being conducted to curb with the rising cases of hate speechesin social media (SM). Different calls have been made to SM providers to filter each comment before allowing it into the public domain . The impacts of hate crimes are already overwhelming due to widespread adoption of SM and the anonymity enjoyed by the online users . In this era of big data, it is time-consuming and difficult to manually process and classifymassive quantities of text data. Furthermore, human factors like competence and fatigue can readily affect how accurately manualtext is classified. It is advantageous to automate the text classification procedures using machine learning (ML) techniques in order to produce more accurate and objective findings. For the detection of hate speech, machine learning (ML) algorithms have advanced significantly from conventional ML, ensemble, and deep learning (DL) approaches. Owing to the remarkable progress in natural language processing (NLP), a number of machine learning techniques have produced better results.

II. LITERATURE SURVEY

Mohammed S.Albarrak(2020) Reducing information asymmetry between investors and a firm can have an impact on the cost of equity, especially in an environment or times of uncertainty. New technologies can potentially help disseminate corporate financial information, reducing such asymmetries. In this paper we analyse firms' dissemination decisions using Twitter, developing a comprehensive measure ofthe amount of financial information that a company makes available to investors (iDisc) from a big data of firms' tweets(1,197,208 tweets). Using a sample of 4131 firm-year observations for 791 non-financial firms listed on the US NASDAQ stock exchange over the period 2009–2015, we find evidence that iDisc significantly reduces the cost of equity. These results are pronounced for less visible firms which are relatively small in size, have a low analystfollowing and a small number of investors. Highly visible firms are less likely to benefit from iDisc in influencing their cost of equity as other communication channels may have widely disseminated their financial information. Our investigations encourage managers to consider the benefits of directly spreading a firm's

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

financial information to stakeholders and potential investors using social media in order to reduce firm equity premium (COE).

Sundong Kim et al(2020) In social networks, user attention affects the user's decision-making, resulting in a performance alteration of the recommendation systems. Existing systems make recommendations mainly according to users' preferences with a particular focus on items. However, the significance of users' attention and the difference in the influence of different usersand items are often ignored. Thus, this paper proposes an attention-based multi-layer friend recommendation model to mitigate information overload in social networks. We first constructed the basic user and item matrix via convolutional neural networks (CNN). Then, we obtained user preferences by using the relationships between users and items, which were later inputted intoour model to learn the preferences between friends. The error performance of the proposed method was compared with the traditional solutions based on collaborative filtering. A comprehensive performance evaluation was also conducted using large-scale real-world datasets collected from three popular locationbased social networks. The experimental results revealed that our proposal outperforms the traditional methods in terms of recommendation performance.

Qian Li(2020)Text classification is the most fundamental and essential task in natural language processing. The last decade has seen a surge of research in this area due to the unprecedented success of deep learning. Numerous methods, datasets, and evaluation metrics have been proposed in the literature, raising the need for a comprehensive and updated survey. This paper fills the gap by reviewing the state of the art approaches from 1961 to 2020, focusing on models from shallow to deep learning. We create a taxonomy for text classification according to the text involved and the models used for feature extraction and classification. We then discuss each of these categories in detail, dealing with both the technical developments and benchmark datasets that support tests of predictions. A comprehensive comparison between different techniques, as well as identifying the pros and cons of various evaluation metrics are also provided in this survey. Finally, we conclude by summarizing key implications, future research directions, and the challenges facing the research area.

Qusay Al-Maatouk(2020)The purpose of this article was to reduce the dissimilarities in the literature regarding the use of social media for training and its impact on students' academic performance in higher education institutions. The main method of data collection for task-technology fit (TTF) and the technology acceptance model (TAM) was a questionnaire survey. This research hypothesizes that TTF applied to social media for learning will affect technology, task, and social characteristics that in turn improve students' satisfaction and students' academic performance. It also posits that the behavioral intent to use social media for learning will affect comprehension efficiency, ease of use, and enjoyment, all of which also improve students' satisfaction and students' academic performance. The data collection questionnaire was conducted with 162 students familiar with social media. Quantitative structural equation modeling was employed to analyze the results. A significant relationship was found between technology, task, and social features with TTF for utilizing social media for academic purposes, all of which fostered student enjoyment and improved outcomes. Similarly, a clear relationship was found between comprehension efficiency, ease of use, and enjoyment with behavioural intentions to utilize social media for academic purposes that positively affected satisfaction and achievement. Therefore, the study indicates that TTF and behavioral intentions to use social media improve the active learning of students and enable them to efficiently share knowledge, information, and discussions. We recommend that students utilize social media in pursuit of their educational goals. Educators should also be persuaded to incorporate social media into their classes at higher education institutions.

III. EXISTING SYSTEM

In Existing System, In order to identify hate speech in Arabic tweets, this article sought to examine a number of neural network models based on convolutional neural networks (CNN) and recurrent neural networks (RNN). On the task of Arabichate voice detection, it also assessed the latest language representation model bidirectional encoder representations from transformers (BERT). In order to do our tests, we first created a brand-new dataset of hate speech with 9316 annotated tweets. Next, we evaluated four models—CNN, CNN + GRU, CNN + GRU, and BERT—through a series of tests on two datasets.

© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712053

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. PROPOSED SYSTEM

In proposed system, This article reviewed advances made so far in automatic hate speech detection in social media. Hate speech as a societal problem is an old research area in the arts and humanities, however, it is still a new research area in the computing domain. This study found out that there is more research work in hate speech detection using classical ML than ensemble and deep learning techniques. That means researchers can explore more on hate speech detection using ensemble and deep learning methods.

V. SYSTEM ARCHITECTURE

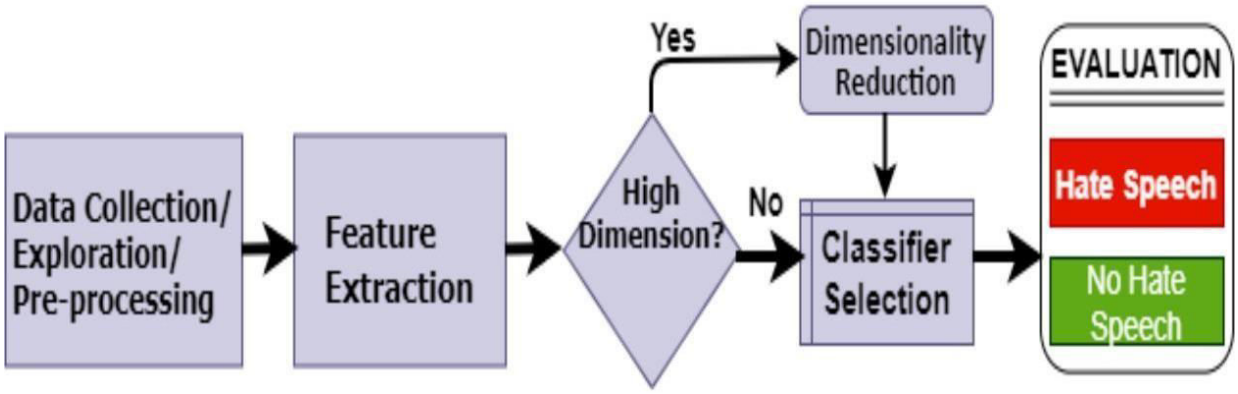

Figure A:- SYSTEM ARCHITECTURE

VI. METHODOLOGY

The system is organized into key modules, each designed to handle distinct aspects of the fake user identification on social network. The modules are as follows:

- Dataset
- Pre-Processing
- Splitting
- Apply Algorithm
- Visualization
- Accuracy

Modules:-

1. Dataset:

A data set is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question.

2. **Pre-Processing:**

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3. **Splitting:**

Data splitting is the act of partitioning available data into. two portions, usually for cross- validatory purposes. one portion of the data is used to develop a predictive model. and the other to evaluate the model's performance.

- Training Data: Used for train the model or given as input to the to the learning model
- Testing Data: Used for test the model or given as input to the model for prediction.

4. Apply Algorithm:

In this we are using support vector machine algorithm to predict accuracy. It is a non- probabilistic supervised machine learning approaches used for classification and regression. It assigns a new data member to one of two possible classes. It defines a hyper plane that separates n-dimensional data into two classes.

5. Visualization :

Visualization is a technique that uses an array of static and interactive visuals within a specific context to help people understand and make sense of large amounts of data. The data is often displayed in a story format that visualizes patterns, trends and correlations that may otherwise go unnoticed.

6. Accuracy:

Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

VII. IMPLEMENTATION

Algorithm:-

1. Import all necessary libraries for Flask, TensorFlow, scikit-learn, and text processing. Set up the environment for GPU memory usage and enable dynamic growth to optimize GPU allocation. Initialize the Flask application.
2. Load the pre-trained models using pickle. Also, load and preprocess the dataset dataset/hate_tweets.csv for hate speech detection.
3. Define a helper function dummy(token) that returns the token without modifications. Additionally, implement a function to convert prediction output into descriptive class names and another function to predict class probabilities for the given text.
4. Set up Flask routes:
   - Define routes for static pages like home (/, /first, /home), login (/login), and dataset upload (/upload).
   - Implement the /preview route to allow users to upload and preview datasets. The uploaded dataset is read using pandas and displayed on a web page.
5. Create the /predict route to handle text classification requests. For this:
   - Receive the input text from a web form.
   - Preprocess the text using the defined pipeline, ensuring invalid or empty text is handled gracefully.
   - Use the loaded model to predict the text's classification and explain the results with LIME (TextExplainer). Configure LIME parameters to create random samples based on the input text.
   - Display the top-2 predictions and their explanations to the user, along with relevant visualizations such as word clouds.
6. Start the Flask application by running it on port 5002.

Experimental Outcome

1. Generate markers

This uses the aruco module in OpenCV to create and save distinct ArUco markers. It starts by setting the marker size to 400 pixels and defining a predetermined vocabulary of 4x4 markers. After that, the code generates each marker by iterating through the marker IDs and stores them as PNG files in the markers directory.

© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712053

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

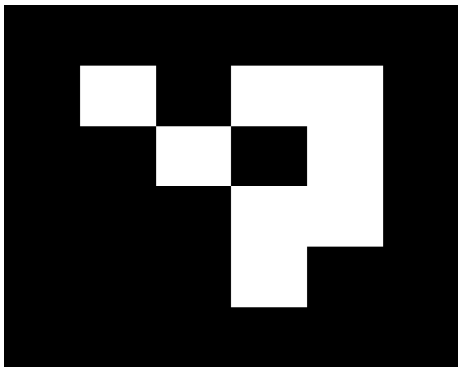

Figure:- GERATED MARKER

2. Marker detection

Real-time ArUco marker detection from a webcam feed is accomplished using OpenCV. It first establishes detection parameters and imports apredetermined dictionary of 4x4 markers (DICT_4X4_250). Aruco.detectMarkers() is used to detect ArUco markers after the video frame has been collected and converted to grayscale in a loop. The marker ID appears in the upper-right corner, and the corners of any markers that are located are drawn on the frame.

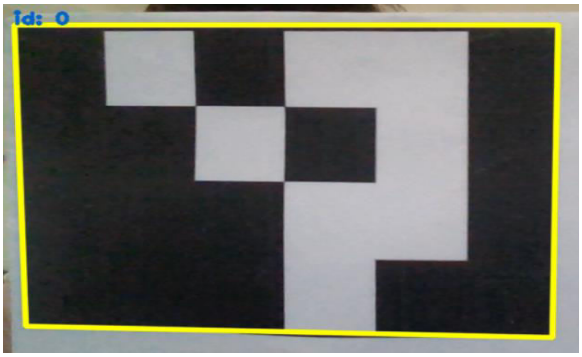

Figure:- MARKER DETECTION

3. Image augmentation

This code performs image augmentation on a webcam feed by overlaying different images onto detected ArUco markers. It first loads a list of images and detects markers in real-time from the video feed. When a marker is detected, the image augmentation function is used to warp and overlay the corresponding image onto the marker's position on the frame using homograph.

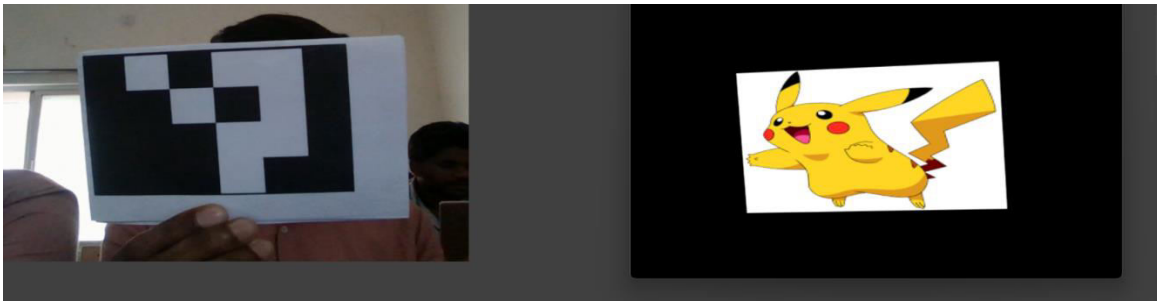

Figure:- IMAGE AUGMENTATION

© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712053

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

4. Distance Estimation

It uses a webcam feed to identify ArUco markers in real time, determines their pose, and shows extra data like the marker's ID and distance from the camera. First, it loads camera calibration information from a file, including the distortion coefficients and intrinsic matrix. The code determines the distance from the camera, computes the rotation and translation vectors for each detected marker using aruco estimate Pose Single Markers(), and then displays the marker's pose (axes) on the video frame.

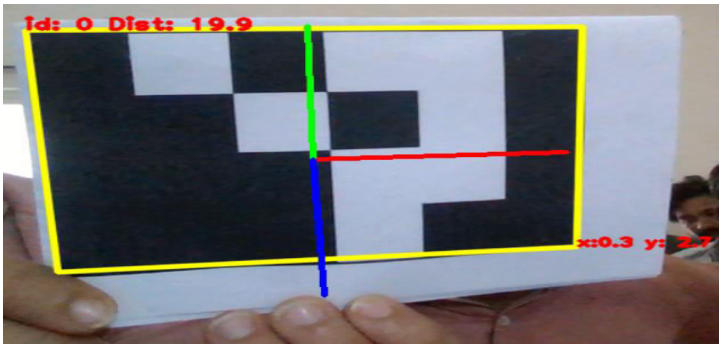

Figure:- DISTANCE ESTIMATION

VIII. CONCLUSION

This article reviews advancements in automatic hate speech detection on social media, a relatively new area in computing compared to its long history in the arts and humanities. It highlights the dominance of classical machine learning approaches over ensemble and deep learning methods, suggesting the latter as areas for further exploration. The study discusses the strengths and weaknesses of various techniques and identifies challenges such as cultural variations, data sparsity, imbalanced datasets, and limited data availability.

The research emphasizes the need for region-specific approaches, noting that hate speech variables differ by context. For example, terms like "419" and factors such as marital or health status in Nigeria remain unaddressed in current models. Aimed at newcomers, this review offers guidance on text classification using machine learning and outlines open challenges to inspire future research.

IX. FUTURE ENHANCEMENT

In future work, it is essential to encourage and support the application of machine learning (ML) for automatic hate speech detection on social media. A critical area needing more attention is the consideration of hate speech variables specific to each country or region, as these can vary significantly. For instance, in Nigeria, factors like marital status and health status are often used as hate speech variables, yet they remain unaddressed in existing studies. Additionally, unique symbols and terms, such as "419" to signify unethical behavior in Nigeria, have not been captured by current state-of-the-art methods.

This research review is primarily intended for newcomers to the field of hate speech classification on social media. It offers a step-by-step guide for conducting text classification tasks using ML and highlights key challenges in the domain, providing a foundation for further research.

REFERENCES

[1] M. S. Albarrak, M. Elnahass, S. Papagiannidis, and A. Salama, ‘‘The effect of Twitter dissemination on cost of equity: A big data approach,’’ Int. J. Inf. Manage., vol. 50, pp. 1– 16, Feb. 2020.
[2] C. Cai, H. Xu, J. Wan, B. Zhou, and X. Xie, ‘‘An attention-based friend recommendation model in social network,’’ Comput., Mater. Continua, vol. 65, no. 3, pp. 2475–2488, 2020.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[3] 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,’’ IEEE Access, vol. 6, pp. 13825–13835, 2018.
[4] P. Fortuna and S. Nunes, ‘‘A survey on automatic detection of hate speech in text,’’ ACM Comput. Surv., vol. 51, no. 4, pp. 1–30, Sep. 2018.
[5] A. Guterres, ‘‘United nations strategy and plan of action on hate speech,’’ United Nations, New York, NY, USA, Tech. Rep., 2019.
[6] Q. Li et al., A Survey on Text Classification: From Shallow to Deep Learning, vol. 37, no. 4. New York, NY, USA: Cornell Univ. Library, 2020.
[7] Q. Al-Maatouk, M. S. Othman, A. Aldraiweesh, U. Alturki, W. M. Al-Rahmi, and A.A. Aljeraiwi, ‘‘Task-technology fit and technology acceptance model application to structure and evaluate the adoption of social media in academia,’’ IEEE Access, vol. 8, pp. 78427–78440, 2020.
[8] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, ‘‘Text classification algorithms: A survey,’’ Information, vol. 10, no. 4, pp. 1–68, 2019.
[9] T. Davidson, D. Warmsley, M. Macy, and I. Weber, ‘‘Automated hate speech detection and the problem of offensive language,’’ in Proc. 11th Int. Conf. Web Soc. Media (ICWSM), 2017, pp. 512–515.
[10] Z. Waseem and D. Hovy, ‘‘Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter,’’ in Proc. NAACL Student Res. Workshop, 2016, pp. 88–93.
[11] P. Burnap and M. L. Williams, ‘‘Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making,’’ Policy Internet, vol. 7, no. 2, pp. 223–242, Jun. 2015.
[12] S. S. Bodrunova, A. Litvinenko, I. Blekanov, and D. Nepiyushchikh, ‘‘Constructive aggression? Multiple roles of aggressive content in political discourse on Russian YouTube,’’ Media Commun., vol. 9, no. 1, pp. 181–194,Feb. 2021.
[13] F. Tulkens, ‘‘The hate factor in political speech. Where do responsibilities lie?’’ Polish Ministry Admin. Digitization Council Eur., Warsaw, Poland, Tech. Rep., 2013.
[14] R. Slonje, P. K. Smith, and A. Frisén, ‘‘The nature of cyberbullying, and strategies for prevention,’’ Comput. Hum. Behav., vol. 29, no. 1, pp. 26–32, Jan. 2013.
[15] M. A. Al-Garadi, M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H. A. Khattak, and A. Gani, ‘‘Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges,’’ IEEE Access, vol. 7, pp. 70701–70718, 2019.
[16] M. Stegman and M. Loftin, ‘‘An essential role for down payment assistance in closing America’s racial homeownership and wealth gaps the price of the homeownership gap,’’ Urban Inst., Washington, DC, USA, Tech. Rep., 2021.
[17] R. Alshalan and H. Al-Khalifa, ‘‘A deep learning approach for automatic hate speech detection in the Saudi Twittersphere,’’ Appl. Sci., vol. 10, no. 23, pp. 1–16, 2020.
[18] A. Al-Hassan and H. Al-Dossari, ‘‘Detection of hate speech in social networks: A survey on multilingual corpus,’’ in Proc. Comput. Sci. Inf. Technol. (CS IT), Feb. 2019, pp. 83–100.
[19] A. Schmidt and M. Wiegand, ‘‘A survey on hate speech detection using natural language processing,’’ in Proc. 5th Int. Workshop Natural Lang. Process. Social Media, 2017, pp. 1–10.
[20] A. Alrehili, ‘‘Automatic hate speech detection on social media: A brief survey,’’ in Proc. IEEE/ACS 16th Int. Conf. Comput. Syst. Appl. (AICCSA), Nov. 2019, pp. 1–6.
[21] A. Rodriguez, C. Argueta, and Y.-L. Chen, ‘‘Automatic detection of hate speech on Facebook using sentiment and emotion analysis,’’ in Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC), Feb. 2019, pp. 169–174.
[22] G. Weir, K. Owoeye, A. Oberacker, and H. Alshahrani, ‘‘Cloud-based textualanalysis as a basis for document classification,’’ in Proc. Int. Conf. High Perform.Comput. Simul. (HPCS), Jul. 2018, pp. 629–633.
[23] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, ‘‘Antisocial behavior in online discussion communities,’’ in Proc. 9th Int. Conf. Web Soc. Media (ICWSM), 2015, pp. 61–70, 2015.
[24] T. Granskogen and J. A. Gulla, ‘‘Fake news detection: Network data from social media used to predict fakes,’’ in Proc. CEUR Workshop, vol. 2041, no. 1, 2017, pp. 59– 66.
[25] L. Tamburino, G. Bravo, Y. Clough, and K. A. Nicholas, ‘‘From population to production: 50 years of scientific literature on how to feed the world,’’ Global Food Secur., vol. 24, Mar. 2020, Art. no. 100346.
[26] V. S. Raleigh, ‘‘Trends in world population: How will the millennium compare with the past,’’ Hum. Reprod. Update, vol. 5, no. 5, pp. 500–505, 1999.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com