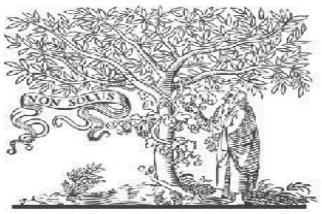
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

Link : https://ijiemr.org/downloads/Volume-12/Issue-05

10.48047/IJIEMR/V12/ISSUE05/17

Title **FADOHS Framework for Detection and Integration of Unstructured Data of Hate Speech on Facebook Using Sentiment and Emotion Analysis**

Volume 12, Issue 05, Pages: 157-167

Paper Authors

Dr. N Swapna, Mohammed Fahad, Mohd Sumair Ali, Mohd Zubair Ahmed, Yasar Ali Ahemad Khan

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

FADOHS Framework for Detection and Integration of Unstructured Data of Hate Speech on Facebook Using Sentiment and Emotion Analysis

1. **Dr. N Swapna** Associate professor & Head of the Department, MTech(Phd), Department of CSE, (Vijay Rural Engineering College(VREC)) swapnanaralas@gmail.com

2. **Mohammed Fahad**, BTech, Department of CSE, (Vijay Rural Engineering College(VREC)) mohammedfahad7532@gmail.com

3. **Mohd Sumair Ali**, BTech, Department of CSE, (Vijay Rural Engineering College(VREC)) mohammedsumairali91@gmail.com

4. **Mohd Zubair Ahmed,** BTech, Department of CSE, (Vijay Rural Engineering College(VREC)) mohammedzubair6786@gmail.com

5. **Yasar Ali Ahemad Khan,** BTech, Department of CSE, (Vijay Rural Engineering College(VREC)) yasarali4454@gmail.com

ABSTRACT: Hate speech is a kind of communication that targets an individual or a group based on their race, ethnicity, religion, sexual orientation, or other characteristics. Although it may be conveyed in a variety of ways, both online and offline, the growing popularity of social media has expanded both its usage and intensity significantly. As a result, the goal of this study is to find and analyse unstructured data from chosen social media postings that attempt to promote hatred in the comment sections. To address this problem, we offer FADOHS, a new framework that combines data analysis and natural language processing methodologies to alert all social media providers to the prevalence of hatred on social media. We examine recent posts and comments on these sites using sentiment and emotion analysis algorithms. Posts suspected of containing dehumanising language will be screened before being sent into the clustering algorithm for further analysis. According to the experimental findings, the suggested FADOHS framework outperforms the state-of-the-art technique by around 10% in terms of accuracy, recall, and F1 scores.

Keywords *– Emotion recognition, clustering algorithm, sentiment analysis, data mining.*

1. INTRODUCTION

Facebook CEO Mark Zuckerberg previously said, "Hate speech and bigotry have no place on

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Facebook." [1]. Despite the fact that Facebook has used different artificial intelligence (AI) tools to combat hate speech on its site, several concerns remain. "For hate speech, our technology still does not operate properly, therefore it has to be reviewed by our review staff," the business claimed when they provided data on the crackdown on hate speech. In the first quarter of 2018, we eliminated 2.5 million pieces of hate speech, 38% of which were reported by our system." [2]. The most recurring issue in this effort is very difficult to answer with AI alone: What exactly constitutes hate speech? This subject generates ongoing debate, with numerous definitions of hate speech being offered; for example, "Hate speech is public utterances that disseminate, provoke, advocate, or excuse hatred, prejudice, or hostility against a certain group." [3] and "We define hate speech as a direct assault on someone based on protected characteristics such as race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and significant sickness or disability." [4].

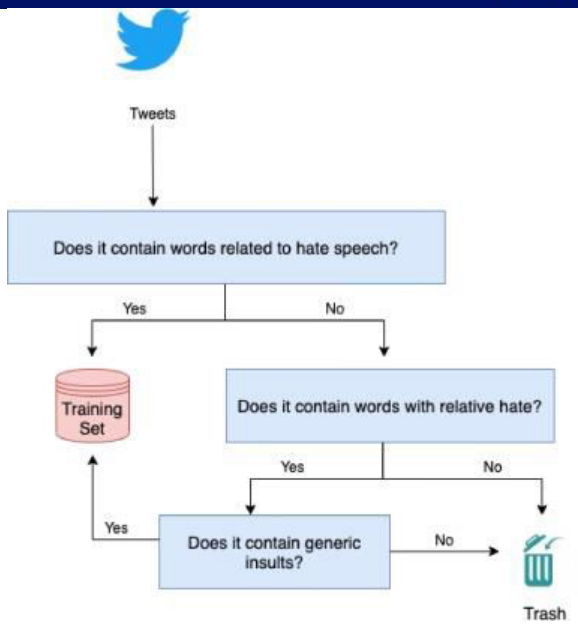


Fig.1: Example figure

Facebook recognised that the issue stems from the fact that AI is not yet smart enough to tell if someone is spreading hatred or merely recounting an experience [5]. Hate speech, according to Sara Chinnasamy and Norain Abdul Manaf, may also be fostered in subtle ways, such as presenting sensitive themes to evoke hate remarks [6]. According to Anat Ben-David and Ariadna MatamorosFernandez, despite Facebook's attempts, hate comments persist. According to the authors, many people show their latent anger via harsh messages or comments. These postings, which are not detected by Facebook algorithms, are widespread on the network. The authors also found that, despite regulations and attempts to

counteract it, overt hate speech and covert discriminatory actions are nonetheless prevalent on Facebook [7]. After we've defined hate speech, we may create a framework for investigating it. The authors of "Hate Me, Hate Me Not: Hate Speech on Facebook" [8] offered many categorization systems to differentiate between various forms of hate speech. They propose and implement two classifiers for Italian using morpho-syntactic characteristics, sentiment polarity, and word-embedded lexicons. Support vector machines (SVMs) and long short-term memory (LSTM) networks are used in their framework. The notion presented in Del Vigna et alresearch .'s and our knowledge of hate speech served as the foundation for this investigation. We looked at early methods for detecting hate speech on Facebook, particularly covert speech in the comments area of postings on hot themes.

2. LITERATURE REVIEW

Racism, hate speech, and social media: A systematic review and critique:

This article maps and explores current advancements in the study of racism and hate speech in the domain of social media research, departing from Jessie Daniels' 2013 assessment of scholarship on race and racism online. We address three research topics by conducting a systematic review of 104 articles: In studies of racism and hate speech on social media, which geographical settings, platforms, and approaches do researchers use? How can scholarship use critical race views to investigate how systematic racism is (re)produced on social media? What are the field's key methodological and ethical challenges? To unravel racism on social media, the paper discovers a lack of regional and platform diversity, a lack of researchers' reflective interaction with their object of study, and insufficient engagement with critical race views. More in-depth examinations of how user behaviours and platform politics co-shape current racisms are required.

Hate me, hate me not: Hate speech detection on Facebook

While Social Network Sites facilitate communication and information exchange, they are sometimes used to start damaging campaigns against certain organisations and people. Cyberbullying, encouragement to self-harm, and sexual predation are only a few of the serious consequences of enormous internet offensives. Furthermore, assaults against groups of victims might develop into physical violence. The goal of this endeavour is to restrict and prevent the worrisome spread of such hate campaigns. Using Facebook as a model, we examine the linguistic content of comments posted on a collection of

public Italian sites. To differentiate the kind of hatred, we first suggest a number of hate types. According to the stated taxonomy, crawled comments are then annotated by up to five separate human annotators. We propose and implement two classifiers for the Italian language that use morpho-syntactical characteristics, sentiment polarity, and word embedding lexicons. The first is based on Support Vector Machines (SVM), and the second on a specific Recurrent Neural Network called Long Short Term Memory (LSTM). We put these two learning algorithms to the test in order to validate their classification performance on the hate speech identification assignment. The findings demonstrate the efficacy of the two classification methods evaluated on the first manually annotated Italian Hate Speech Corpus of social media material.

The K-means algorithm: A comprehensive survey and performance evaluation

In the scientific community, the k-means clustering method is regarded as one of the most powerful and widely used data mining techniques. Despite its popularity, the technique has certain drawbacks, including issues with random centroids initialization, which leads to unexpected convergence. Furthermore, such a clustering technique necessitates the pre-definition of the number of clusters, which is accountable for variable cluster forms and outlier effects. The inability of the k-means algorithm to accommodate different data formats is a basic issue. This article presents an organised and synoptic summary of research on the k-means method to address such deficiencies. Variants of the k-means algorithms are reviewed, including recent advances, and their usefulness is explored through experimental analysis of a range of datasets. The rigorous experimental analysis, as well as the full comparison of several k-means clustering methods, distinguishes our study from other previous survey publications. Furthermore, it provides a clear and comprehensive overview of the k-means algorithm as well as its many research paths.

Student Engagement Level in e-Learning Environment: Clustering Using K-means

Several obstacles confront e-learning platforms and procedures, including the notion of customising the e-learning experience and keeping students motivated and interested. This effort is part of a bigger project that will use a range of machine learning approaches to address these two difficulties. To that aim, this article suggests using the k-means algorithm to cluster students based on 12 engagement measures classified as interaction-related and effort-related. Quantitative analysis is used to identify

pupils who are not engaged and may need assistance. Two-level, three-level, and five-level clustering models are explored. The dataset under consideration is the students' event log from a second-year undergraduate Science course taught in a hybrid manner at a North American institution. MATLAB is used to convert the event log and create a new dataset containing the metrics under consideration. The analysis of experimental findings demonstrates that, among the interaction-related and effort-related metrics evaluated, the number of logins and average time to submit assignments are the most reflective of the students' involvement level. Furthermore, using the silhouette coefficient as a performance indicator, it is shown that the two-level model has the greatest cluster separation performance. The three-level approach, on the other hand, performs similarly while better recognising kids with low involvement levels.

Novel land cover change detection method based on K-means clustering and adaptive majority voting using bitemporal remote sensing images

The detection of land cover change (LCCD) using bitemporal remote sensing pictures has become a hot issue in the world of remote sensing. Despite the fact that several approaches have been advocated in recent decades, improvements in usability and performance of these methods have remained important. A unique LCCD strategy based on the combination of k-means clustering and adaptive majority voting (kmeans AMV) techniques is presented in this work. Three key strategies comprise the proposed k-means AMV method. To begin, an adaptive zone surrounding a central pixel is formed by identifying the spectral similarity between the centre pixel and its eight nearby pixels in order to use contextual information in an adaptable way. Second, after the adaptive area expansion has ended, the k-means clustering approach is used to identify the label of each pixel inside the adaptive region. Finally, an existing AMV approach is employed to improve the label of the adaptive region's core pixel. The label of each pixel in the change magnitude image (CMI) may be improved and the binary change detection map can be formed when the CMI is scanned and processed in this way. Three photographic scenes from various land cover change events are modified to assess the efficacy and performance of the proposed k-means AMV technique. The findings reveal that the suggested k-means AMV strategy outperforms the other widely used approaches in terms of detection accuracy and visual performance.

3. METHODOLOGY

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Hate speech is a kind of communication that targets an individual or a group based on their race, ethnicity, religion, sexual orientation, or other characteristics. Although it may be conveyed in a variety of ways, both online and offline, the growing popularity of social media has expanded both its usage and intensity significantly. As a result, the goal of this study is to find and analyse unstructured data from chosen social media postings that attempt to promote hatred in the comment sections.

Disadvantages:

1. The growing popularity of social media has expanded both its usage and severity significantly.
2. To identify and analyse unstructured material from chosen social media postings with the intent of spreading hatred in the comment area.

To address this problem, we offer FADOHS, a new framework that combines data analysis and natural language processing methodologies to alert all social media providers to the prevalence of hatred on social media. We examine recent posts and comments on these sites using sentiment and emotion analysis algorithms. Posts suspected of containing dehumanising language will be screened before being sent into the clustering algorithm for further analysis. According to the experimental findings, the suggested FADOHS framework outperforms the state-of-the-art technique by around 10% in terms of accuracy, recall, and F1 scores.

Advantages:

1. The suggested methodology offers a unique approach to clustering posts and comments, recognising hotly debated issues that create hate speech, and identifying hate speech.
2. This study exemplifies the use of unstructured data, such as Facebook postings, in conjunction with a framework for effective analysis.

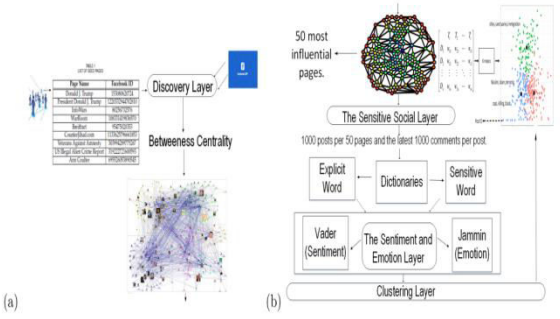


Fig.2: System architecture

MODULES:

To carry out the aforementioned project, we created the modules listed below.

- Data exploration: we will put data into the system using this module.

- Processing: we will read data for processing using this module.
- Splitting data into train and test: Using this module, data will be separated into train and test models.
- GPT2 - Random Forest - SVM - Voting Classifier - MLP + RF + SVN - LSTM - LSTM with SVM Compiler - CNN + LSTM with SVM Compiler. Calculated algorithm accuracy.
- User registration and login: Using this module will result in registration and login.
- User input: Using this module will provide input for prediction
- Prediction: the final projected value will be presented

4. IMPLEMENTATION

ALGORITHMS:

Random Forest: A Random Forest Method is a supervised machine learning algorithm that is widely used in Machine Learning for Classification and Regression issues. We know that a forest is made up of many trees, and the more trees there are, the more vigorous the forest is. Random Forest is a supervised machine learning technique that develops and merges several decision trees to form a "forest." It may be used in R and Python for classification and regression tasks.

SVM: Support Vector Machine (SVM) is a supervised machine learning technique that may be used for classification and regression. Though we call them regression issues, they are best suited for categorization. The SVM algorithm's goal is to identify a hyperplane in an N-dimensional space that clearly classifies the input points. When there is a clear margin of distinction between classes, SVM performs rather well. SVM is more successful in high-dimensional domains and uses less memory. SVM is useful when the dimensions are more than the number of samples.

Voting classifier: A voting classifier is a machine learning estimator that trains numerous base models or estimators and predicts based on the results of each base estimator. Aggregating criteria may be coupled voting decisions for each estimator output. The Voting Classifier is a form of Ensemble Learning that may be both homogeneous and heterogeneous, meaning that the basic classifiers can be of the same or different type. As previously stated, this form of ensemble may also be used as an extension of bagging (e.g. Random Forest).

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

LSTM: LSTM is an abbreviation for Long-Short Term Memory. In terms of memory, LSTM is a sort of recurrent neural network that outperforms standard recurrent neural networks. LSTMs perform far better when it comes to learning specific patterns.

CNN: A CNN is a kind of network architecture for deep learning algorithms that is primarily utilised for image recognition and pixel data processing jobs. There are different forms of neural networks in deep learning, but CNNs are the network design of choice for identifying and recognising things. CNNs, in general, perform better with data that has a spatial link. The CNN input is typically two-dimensional, in the form of a field or matrix, but it may be altered to one-dimensional, enabling it to construct an internal representation of a one-dimensional sequence.

5. EXPERIMENTAL RESULTS

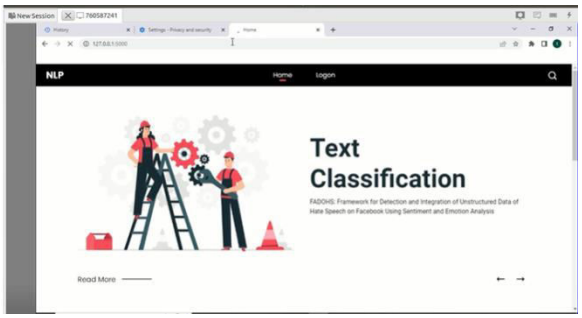


Fig.3: Home screen

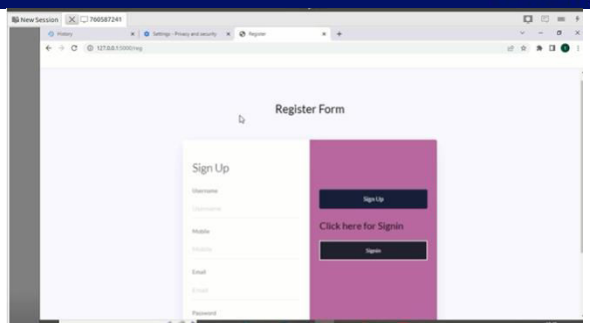


Fig.4: User registration

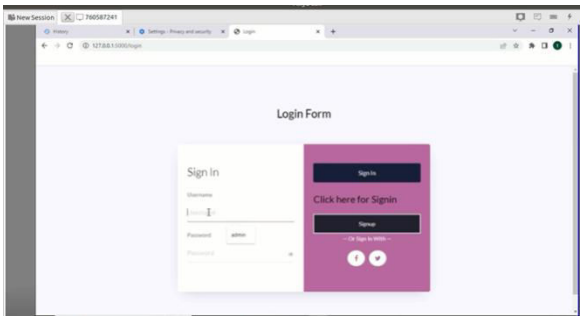


Fig.5: User login

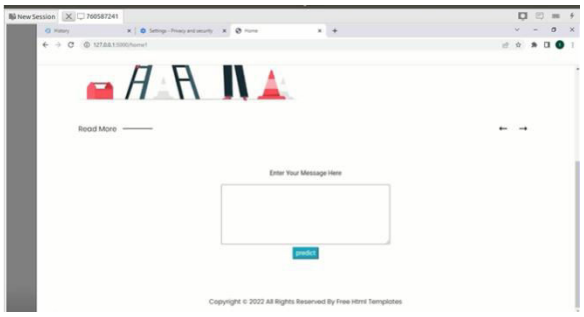


Fig.6: Main page

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

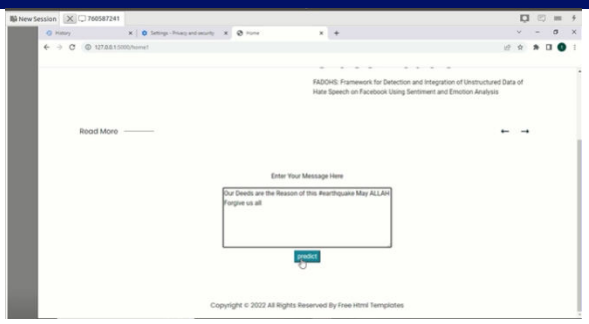


Fig.7: User input

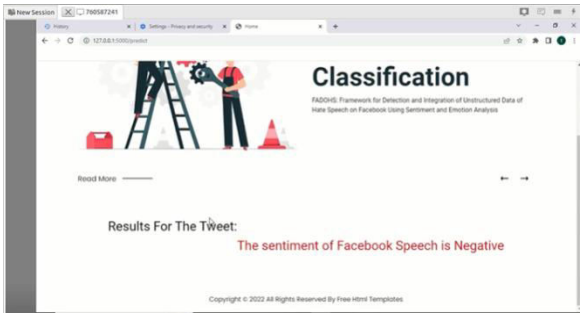


Fig.8: Prediction result

6. CONCLUSION

In this work, we offer FADOHS, which discovers and combines unstructured data from Facebook pages that purportedly encourage hate speech, allowing us to identify the typical subjects covered. This was initially difficult since non-personal sites and accounts on Facebook tend to avoid using overly explicit language in messages in order to avoid getting deleted from the network or receiving criticism. Nonetheless, many sites manage to arouse unpleasant feelings and seem to encourage hate speech among their followers by addressing sensitive themes while using relatively benign wording. The suggested methodology offers a unique approach to clustering articles and comments, recognising hotly debated issues that create hate speech, and identifying hate speech. FADOHS clusters and analyses postings that may include hate speech by combining graph analysis, dictionaries, sentiment/emotion analysis, and clustering algorithms. To appropriately handle the hate speech problem, we begin our investigation with a select collection of sites known to cover sensitive themes that may provoke hate comments. Using graph analysis, we may create three tiers of direct social graphs and identify significant sites based on this study. We identify posts with a particular amount of hostility in the comments using preset vocabulary, sentiment, and emotion analysis. The results led us to believe that unstructured data may be recognised and incorporated from hate speech-promoting websites. The next critical step seeks to categorise these data, which is accomplished by using the K-means clustering method, followed by testing various settings to uncover coherent clusters of subjects. We next manually examine the postings that fit into each category and give each cluster a manual label. We may infer that both variables are matched by comparing the manual label to the cluster centroids, showing the success of our technique. Our findings show

that a tiny collection of seeds may identify multiple sites supposedly encouraging hate speech and associated themes. This study exemplifies how to take unstructured data, such as Facebook postings, and apply a framework for useful analysis. According to the experimental findings, the suggested FADOHS framework outperforms the state-of-the-art technique by around 10% in terms of accuracy, recall, and F1 scores. In future research, we want to use our approach not just on remarks but also on their answers to more correctly identify persons accused of propagating hate speech. Long-term advantages might be incredibly useful since this could identify cyberbullies and cyberterrorists. We would also want to do a more in-depth investigation of the emotion filtering and grouping data in order to determine the most reliable arrangement for improving outcomes.

REFERENCES

[1] Zuckerberg Refugee Crisis: Hate Speech Has, Place Facebook, Street Guardian, Honolulu, HI, USA, 2010.

[2] 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/.

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

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

[5] CNBC. (2020). Facebook's Artificial Intelligence Still Has Trouble Finding Hate Speech—But it Finds a Lot of Nudity. Accessed: May 11, 2018. [Online]. Available: https://www.cnbc.com/2018/05/15/facebook artificial-intelligence-still-finds-it-hard-to-identify-hate-speech.html

[6] S. Chinnasamy and N. A. Manaf, ''Social media as political hatred mode in Ts 2018 general election,'' in SHS Web Conf., vol. 53, 2018, p. 2005.

[7] A. Matamoros-Fernández and J. Farkas, ''Racism, hate speech, and social media: A systematic review and critique,'' Telev. New Media, vol. 22, no. 2, pp. 205–224, Feb. 2021.

[8] F. Del Vigna, A. Cimino, F. Dell-TOrletta, M. Petrocchi, and M. Tesconi, ''Hate me, hate me not: Hate speech detection on Facebook,'' in Proc. 1st Italian Conf. Cybersecur. (ITASEC), Venice, Italy, 2017, pp. 86–95.

[9] M. Ahmed, R. Seraj, and S. M. S. Islam, ''The K-means algorithm: A comprehensive survey and performance evaluation,'' Electronics, vol. 9, no. 8, p. 1295, Aug. 2020.

[10] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, ''Student engagement level in an e-Learning environment: Clustering using K-means,'' Amer. J. Distance Educ., vol. 34, no. 2, pp. 137–156, Apr. 2020.