**MIT School of Engineering**

**Department of Computer Science and Engineering**

**Project Synopsis**

**Group ID:14**

**Project Title: HATE SPEECH DETECTION SYSTEM.**

**Group Members:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Enrollment Number** | **Roll No.** | **Name of student** | **Email Id** | **Contact Number** |
| **MITU22BTCS0772** | **2223178** | **Shivam Khode** | **shivamkhode04@gmail.com** | **9764210397** |
| **MITU22BTCS0514** | **2223163** | **Omkar More** | **moreomkar609@gmail.com** | **7720893298** |
| **MITU22BTCS0432** | **2223158** | **Mayank Modi** | **mayankmodi7126@gmail.com** | **8839600663** |

**Problem Statement:** Develop an accurate and adaptable hate speech detection system capable of identifying offensive and discriminatory language in real-time online textual content, while minimizing false positives and false negatives

**Abstract:** In the wake of a polarizing election, social media is laden with hateful content. Context accompanying a hate speech text is useful for identifying hate speech, which however has been largely overlooked in existing datasets and hate speech detection models. We provide an annotated corpus of hate speech with context information well kept. Then we propose two types of supervised hate speech detection models that incorporate context information, a logistic regression model with context features and a neural network model with learning components for context. Further, to address various limitations of supervised hate speech classification methods including corpus bias and huge cost of annotation, we propose a weakly supervised two-path bootstrapping approach for online hate speech detection by leveraging large-scale unlabeled data. This system significantly outperforms hate speech detection systems that are trained in a supervised manner using manually annotated data. Applying this model on a large quantity of tweets collected before, after, and on election day reveals motivations and patterns of inflammatory language.

**Literature Survey :**

1. Linguistic Rule-Based Approach

In 2014, C. J. Hutto et al. proposed an approach to classify sentiment using VADER, which is a rule-based approach [18]. At first, they created a list of lexical features that are highly sensitive to the sentiment of social media posts. After then they combined that list of lexical features with five general rules that encapsulate syntactical and grammatical rules for presenting sentiment intensity. At last, they have found that VADER performed 96% accuracy using the rule-based model on Twitter sentiments. Dennis Gitariet al. in 2015 proposed a method to identify the Sentiment Analysis of the Social Media Text using the Rule-based method [19]. In this work, They categorized the hate speech problem into three fields religion, nationality, and race. The main objective of this paper is to develop a classification model that employs sentiment analysis. The developed model not only detects subjective sentences but also classifies and ranks the polarity of sentiment phrases. After then they relate the semantic and subjective features with hate speech. Finally, they achieved 71.55 % precision using the lexicon-based approach.

1. Supervised Learning Approach

Fatahillah et al. (2017) used Naive Bayes Classifier Algorithm to detect hate speech on Instagram using the k-nearest neighbor classifier [20]. They collected the data set using Twitter API from Twitter and annotated those data set manually. After preprocessing and feature engineering phase, they applied the Naive Bayes Classifier algorithm and found 93% of accuracy. M. Ali Fauzi et al. (2018) proposed an approach to identify hate speech using a set of supervised learning algorithms [21]. They ensembled five different classification algorithms, including K-Nearest Neighbours, Random Forest, Naive Bayes, Support Vector Machine, and Maximum Entropy. They collected the data set using Twitter API and annotated those data set manually. In preprocessing phase, They employed tokenization, filtering, stemming, and term weighting methods. They utilized the bag of words features with TFIDF techniques. The naive Bayes algorithm performed best with 78.3 % of accuracy among all the other five stand-alone classifiers. In 2019, P. Sari et al. proposed an approach to detect hate speech using logistic regression on Twitter. [22] They collected the data from Twitter and employed Case Folding, Tokenizing, Filtering, and Stemming methods in preprocessing phase. After Pre-processing, the TF-IDF technique is used for vectorization. After Feature engineering, the Logistic regression algorithm has been applied, and they have found 84% of accuracy. In 2020, Oluwafemi Oriola et al. proposed an approach to detect offensive speech on tweeter [5]. The author collected the data set using Twitter API and annotated those data set into two sections, free speech „FS‟ and hate speech „HT.‟ In preprocessing phase, they removed special characters, emojis, punctuations, symbols, hashtags, stopwords to clean the data. In the feature engineering phase, they employed the TF-IDF technique to transform the text into feature vectors. After applying an optimized support vector machine with n-gram, they have found 89.4% of accuracy. In 2020, Annisa Briliani et al. proposed an approach to identify hate speech on Instagram using the knearest neighbor classifier [23]. They collected the data set using Instagram API from Instagram and annotated those data set manually. They divided the dataset into 2 labels, namely zero and one. In preprocessing phase, they cleaned the data and employed the TF-IDF technique in the feature engineering phase. After then, they applied the k-nearest neighbor algorithm and found 98.13% of accuracy.

1. Unsupervised Learning Approach

Rui Zhao et al. (2015) proposed an approach to detect cyberbullying using Semantic-Enhanced Marginalized Denoising Auto-Encoder [24]. They used two sources of data set. The first source is Twitter, and the second source is Myspace. Twitter data was collected through Twitter stream API, and Myspace data was collected using the web crawling technique. They have achieved 84.9 % accuracy using smSDA for the Twitter dataset, and they have got 89.7% of accuracy with smSDA with the MySpace dataset. Axel Rodríguez et al. (2019) proposed an approach to detect hate speech content using sentiment analysis on Facebook [25]. They used Graph API to extract the post and comments from Facebook. To remove the unrelated texts VADER and JAMMIN were used. In preprocessing phase, they filtered out all unnecessary stopwords or symbols. Preprocessed documents converted into the vector using TFIDF. The resulting matrix is passed to the k-means clustering algorithm as an input matrix. The most negative articles and responses were collected using sentiment and emotion analysis. Sylvia Jaki et al. (2019) demonstrated an approach to detect hate speech content using unsupervised learning on Twitter [26]. They collected over 50,00 data set using Twitter API. They used NLP techniques to group the words into similar clusters. They computed three clusters of the top 250 most biassed terms using spherical k-means clustering and skip-grams. As a result, they have got an 84.21% F1 score. Michele Di Capua et al. (2019) proposed an approach to detect cyberbullying using unsupervised learning [27]. They collected over 54,000 data set from YouTube and Annotated all data sets manually. The GHSOM network algorithm was implemented using the SOM-Toolbox-2 platform. They trained and tested GHSOM using a K-fold method with K = 10. As a result, they have got 64% of accuracy.

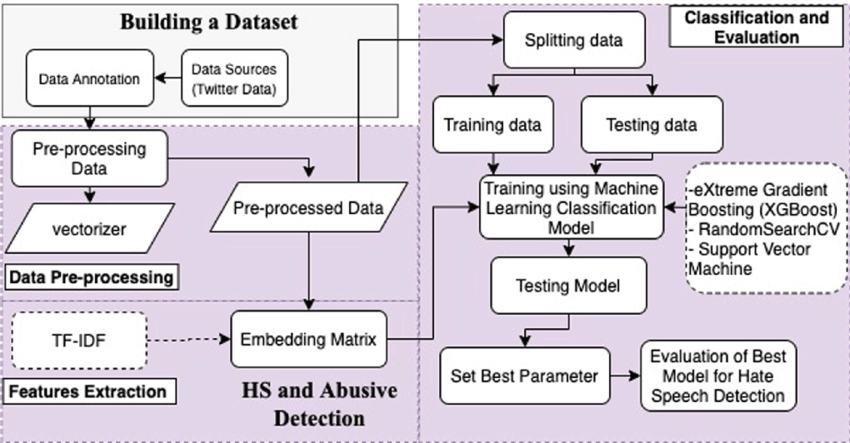
1. Deep Learning Approaches

Hugo Rosa et al. (2018) proposed an approach to detect cyberbullying using deep learning [28]. In this paper, the training and testing data set was collected from Kaggle. At first, they initiated CNN, which holds a certain similarity to the issue of cyberbullying. It starts with a single-layer CNN and continues with a completely linked layer with a dropout of 0.5 and softmax performance. Then they combined CNN-DNN-LSTM to achieve maximum accuracy. They employed TFIDF for vector representation. They achieved 64.9% precision with google embeddings. Tin Van Huynh et al. (2019) proposed an approach to detect hate speech using Bi-GRU-CNN-LSTM Model [29]. In this paper, they collected data from Twitter and categorized their data into three labels (OFFENSIVE, HATE, and CLEAN). After cleaning the data, they implemented three neural network models such as BiGRU-LSTM-CNN, Bi-GRU-CNN, and TextCNN to identify hate speech. They achieved a 70.57% of F1 score as a result. Gambäck et al. (2019) utilized a deep learning algorithm to detect hate speech on Twitter [30]. In this paper, they collected data from Twitter and divided the data set into four categories(sexism, racism, combined(sexism and racism), and non-hate-speech). They employed four CNN models that were trained with character n-gram, word2vec, random vectors combined(word2vec and character n-gram). The author utilized a 10-fold technique to improve the accuracy of the model. Among all four models, word2vec based CNN model performed well with a 78.3% of F-score.

1. Hybrid based Approach

Viviana Patti et al. (2019) proposed a Hybrid based approach to detect hate speech [31]. In this paper, they employed two models. In their first model, they implemented a linear support vector classifier (LSVC), and in the second model, they employed a long short-term memory (LSTM) neural model with word embedding. They concatenated 17 categories, such as HurtLex, with two types, namely LSVC and LSTM. Joint learning with a multilingual word embedding model, including HurtLex, performed best with 68.7% of F1-score. Safa Alsafari et al. (2020) proposed a Hate speech detection model for Arabic social media [32]. In this paper, they collected the data set using Twitter search API, and the data set is categorized into four classes (Religious, Nationality, Gender, and Ethnicity). They cleaned the data set in preprocessing phase by removing unnecessary words such as URLs, punctuations, symbols, tags, and stopwords. They implemented a three-class classification with CNN and Bert to achieve 75.51% of the F1-score.frequent validation or on demand validation - both can generate considerable, often unnecessary, network traffic and the latter reduces much of the latency gains offered by caching. The viable alternative in such circumstances is resource-driven invalidation where the server invokes a callback on the cache to inform it whenever an update has occurred [7][8]. Although this solution involves the server maintaining knowledge of its caches there will be applications which are willing to accept these memory costs in preference to the communication costs of pollingbased invalidation.

**Proposed System (Block Diagram):**



**Conclusion:**

The Hate Speech Detection Project aims to harness the power of machine learning and natural language processing to create a safer online environment. By detecting and mitigating hate speech, this project contributes to promoting tolerance, diversity, and inclusivity in the digital age.

**References:**

[1] Z. Waseem and D. Hovy, “Hateful symbols or hateful people? predictive features for hate speech detection on twitter,” in Proceedings of NAACL-HLT, pp. 88–93, Association for Computational Linguistics, 2016.

[2] Z. Waseem, “Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter,” in Proceedings of the first workshop on NLP and computational social science, pp. 138–142, 2016.

[3] E. Wulczyn, N. Thain, and L. Dixon, “Ex machina: Personal attacks seen at scale,” arXiv preprint arXiv:1610.08914, 2016.

[4] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, “Measuring the reliability of hate speech annotations: The case of the european refugee crisis,” arXiv preprint arXiv:1701.08118, 2017.

[5] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, “Abusive language detection in online user content,” in Proceedings of the 25th International Conference on World Wide Web, pp. 145–153, International World Wide Web Conferences Steering Committee, 2016.

[6] I. Kwok and Y. Wang, “Locate the hate: Detecting tweets against blacks.,” in AAAI, Association for the Advancement of Artificial Intelligence, 2013.

[**https://deliverypdf.ssrn.com/delivery.php?ID=600090118106119104065094074096126069101037019087067074070003086094098122010080073124050026032061006034030081072126097066065116029029009051050010070012127023023075022046085012009127025121095112096076075014115083082113084103078107071086075099000003104064&EXT=pdf&INDEX=TRUE**](https://deliverypdf.ssrn.com/delivery.php?ID=600090118106119104065094074096126069101037019087067074070003086094098122010080073124050026032061006034030081072126097066065116029029009051050010070012127023023075022046085012009127025121095112096076075014115083082113084103078107071086075099000003104064&EXT=pdf&INDEX=TRUE)

[**https://www.sciencedirect.com/science/article/pii/S1110016823007238**](https://www.sciencedirect.com/science/article/pii/S1110016823007238)

[**https://www.mdpi.com/2078-2489/13/6/273**](https://www.mdpi.com/2078-2489/13/6/273)