IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13465

TOXIC COMMENT CLASSIFICATION SYSTEM USING DEEP LEARNING

Chaitanya Sonawane[1], Tejaswini Bagale[2], Preeti Kawade[3], Swarada Ogale[4],

Prof. Megha C Singru[5]

Students of Information Technology, Sandip Institute of Technology and Research Center, Nashik[1-4]

Assistant Professor of Information Technology, Sandip Institute of Technology and Research Center, Nashik[5]

Abstract: Every day, a significant volume of textual content is shared online. Sorting through such vast amounts of textual material to find the relevant and irrelevant information is challenging. One area of natural language processing that enables the examination of textual data is sentiment analysis. Since it examines the words and presents the public's overall viewpoint, it is regarded as an opinion mining technique [1].Sentiment analysis is a domain with three sub-branches: aspect-based, sentence-based, and document-based [2]. Sentences are used to find opinions in sentence-based sentiment analysis. Sentiment analysis of complicated texts is a challenging task. When conducting sentiment analysis on documents, the entire textual Social media provides a forum for public sharing.their opinions and concepts. The most widely used social media sites are Facebook, Twitter, and YouTube, where usersrespond by leaving comments and like the page. Sentiment analysis is a widely utilised tool for analysis these days.

Keywords: toxic comment,Machine Learning,SVM,NLP.

I. INTRODUCTION

1.1 Motivation

Customer satisfaction: By analyzing sentiment, businesses can identify positive and negative reviews and sentiments expressed by customers. This helps them understand what aspects of their products are appreciated or disliked, allowing them to make necessary improvements and enhance customer satisfaction. Product development: Sentiment analysis can provide valuable insights into customer preferences, expectations, and demands. By understanding the sentiment behind customer feedback, businesses can identify areas where their product can be improved and develop new features or modifications to meet customer needs effectively. The motivation behind this project is to create a safer and more inclusive online environment for everyone. The project uses deep learning algorithms to analyze text data and identify patterns that are indicative of toxic comments. This can help moderators and administrators to quickly identify and remove harmful content, which can reduce the risk of harassment, bullying, and other negative behaviors. The project has the potential to make a significant impact on online communities by promoting healthy and respectful discussions..

1.2 Problem Definition

The problem at hand is to develop a deep machine leraning -based solution for automatically classifying toxic comments in online content. This project aims to create a robust and scalable model capable of identifying and flagging toxic comments, enabling timely moderation, and fostering more positive online interactions.

1.3 Methodologies of Problem Solving

Methodology –

1. Data collection and preprocessing: The methodology starts with gathering a dataset of labeled toxic and non-toxic comments and then preprocessing the text by removing stop words, punctuation, and special characters, and standardizing it through tokenization and lowercasing.

2. Feature extraction with NLP techniques: Techniques like TF-IDF, word embeddings (e.g., Word2Vec, GloVe), or deep learning-based embeddings (e.g., BERT, ELMo) are used to extract meaningful features from the preprocessed text, converting it into numerical representations suitable for machine learning models.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.13465

3. Construction of deep learning architecture: A deep learning architecture is built to extract and represent features from the textual data. This architecture may include layers for capturing complex patterns and representations, such as convolutional and recurrent units, and attention mechanisms to focus on important text parts.

4. Integration of SVM classifier: A Support Vector Machine (SVM) classifier is integrated into the model architecture. SVMs are effective for binary classification tasks and complement the feature representation learned by the deep learning model. The deep SVM architecture combines the strengths of both deep learning and SVMs.

5. Overall methodology: By integrating advanced NLP and machine learning techniques, the methodology effectively classifies toxic comments. The combination of deep learning for feature extraction and SVMs for classification enables robust performance in identifying toxic language in online comments.

Problem Solving –

Implementing a deep SVM model for toxic comment classification entails leveraging the strengths of both deep learning and traditional machine learning techniques. By combining deep learning's ability to extract complex features from text data with SVM's robust classification capabilities, the model can effectively discern between toxic and non-toxic comments, contributing to a safer and more constructive online environment.

II. LITERATURE REVIEW

A. Sarcasm Detection in Twitter using Sentiment Analysis Author: Bala Durga Dharmavarapu, Jayanag Bayana

In the proposed methodology Sentiment Analysis, Naive Bayes classification and AdaBoost algorithms are used to detect sarcasm on twitter. By using Naive Bayes classification, the tweets are categorized into sarcastic and nonsarcastic. Sarcasm is a subtle type of irony, which can be widely used in social networks. It is usually used to transmit hidden information to criticize and ridicule a person and to recognize. The sarcastic reorganization system is very helpful for the improvement of automatic sentiment analysis collected from different social networks and microblogging sites. Sentiment analysis refers to internet users of a particular community, expressed attitudes and opinions of identification and aggregation. In this paper, to detect sarcasm, a pattern-based approach is proposed using Twitter data. Four sets of features that include a lot of specific sarcasm is proposed and classify tweets as sarcastic and non-sarcastic. The proposed feature sets are studied and evaluate its additional cost classifications.

B. TweetAnalyzer: Twitter Trend Detection and Visualization Author: Zeel Doshi , Subhash Nadkarni, Kushal Ajmera , Prof. Neepa Shah

Twitter can be identified as one of the largest social networking sites. A large number of users have accepted Twitter as a universal platform for spreading news, sharing articles and socializing with other people globally. Subsequently, such a high-volume, high-velocity surge of Twitter data generated at each second have the potential of being utilized for significant analytical and interpretation purposes. The objective of this paper is to demonstrate an easy and simple solution, called TweetAnalyzer. We propose a system to extract real-time Twitter data and to represent the trending Twitter hashtags and active users on a bar graph. TweetAnalyzer also makes use of the user’s current location coordinates to represent the tweets on a world map. The proposed system can be easily deployed and used for various real-world applications such as job search, news updates, and business intelligence.

C. Investor Classification and Sentiment Analysis Author: Arijit Chatterjee,Dr.William Perrizo

Twitter can be identified as one of the largest social networking sites. A large number of users have accepted Twitter as a universal platform for spreading news, sharing articles and socializing with other people globally. Subsequently, such a high-volume, high-velocity surge of Twitter data generated at each second have the potential of being utilized for significant analytical and interpretation purposes. The objective of this paper is to demonstrate an easy and simple solution, called TweetAnalyzer. We propose a system to extract real-time Twitter data and to represent the trending Twitter hashtags and active users on a bar graph. TweetAnalyzer also makes use of the user’s current location coordinates to represent the tweets on a world map. The proposed system can be easily deployed and used for various real-world applications such as job search, news updates, and business intelligence..

D. HTwitt: a hadoop-based platform for analysis and visualization of streaming Twitter data Author Name: Umit Demirbaga

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13465

Twitter produces a massive amount of data due to its popularity that is one of the reasons underlying big data problems. One of those problems is the classification of tweets due to use of sophisticated and complex language, which makes the current tools insufficient. We present our framework HTwitt, built on top of the Hadoop ecosystem, which consists of a MapReduce algorithm and a set of machine learning techniques embedded within a big data analytics platform to efficiently address the following problems: (1) traditional data processing techniques are inadequate to handle big data; (2) data preprocessing needs substantial manual effort; (3) domain knowledge is required before the classification; (4) semantic explanation is ignored. In this work, these challenges are overcome by using different algorithms combined with a Naı¨ve Bayes classifier to ensure reliability and highly precise recommendations in virtualization and cloud environments. These features make HTwitt different from others in terms of having an effective and practical design for text classification in big data analytics. The main contribution of the paper is to propose a framework for building landslide early warning systems by pinpointing useful tweets and visualizing them along with the processed information. We demonstrate the results of the experiments which quantify the levels of overfitting in the training stage of the model using different sizes of real-world datasets in machine learning phases. Our results demonstrate that the proposed system provides high-quality results with a score of nearly 95.

III. OBJECTIVE

In the realm of sub-toxic comment classification, the objective is to develop a deep learning-based model capable of accurately distinguishing between subtle forms of toxicity and non-toxicity in online discourse. This entails leveraging advanced neural network architectures to capture nuanced linguistic cues and contextual intricacies present in text data. The primary goal is to create a robust system that can effectively identify subtle manifestations of toxicity, aiding in the moderation of online platforms and fostering healthier communication environments. Through rigorous training and validation processes, the model aims to achieve high accuracy and generalization capability across diverse types of sub-toxic comments. Ultimately, the objective is to contribute to the creation of safer and more inclusive online communities by mitigating the spread of subtle forms of toxicity.logy.

IV. METHODOLOGY

1. **Dataset collection and preprocessing**: Begin by collecting a dataset consisting of labeled sub-toxic comments and non-sub-toxic comments. This involves sourcing data from relevant sources and platforms. Preprocess the collected text data by removing stopwords, punctuation, and special characters. Additionally, tokenize the text and convert it to lowercase to standardize the format.

2. **Deep learning feature extraction**: Employ deep learning techniques to extract meaningful features from the preprocessed text data. Construct a deep learning architecture, such as a neural network with convolutional and recurrent layers, to capture intricate patterns and representations within the text. These features will serve as inputs for the subsequent classification task.

3. **Integration of SVM classifier**: Incorporate a Support Vector Machine (SVM) classifier into the deep learning architecture. SVMs are renowned for their efficacy in binary classification tasks and can complement the feature representation learned by the deep model. This integration harnesses the strengths of both deep learning and SVMs for improved classification accuracy.

4. **Model training and optimization**: Train the integrated deep learning with SVM model using the labeled dataset. Fine-tune the model's hyperparameters and optimize its performance through techniques such as cross-validation and grid search. This iterative process ensures that the model achieves the best possible classification results.

5. **Evaluation and validation**: Evaluate the trained model's performance using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. Validate the model's effectiveness by testing it on a separate test dataset, assessing its ability to correctly classify sub-toxic comments. This step ensures that the model generalizes well to unseen data and reliably identifies sub-toxic language.

6] Maintenance:
Maintenance for sub toxic comment classification using deep learning with SVM involves continuous data updates to keep the model relevant, regular performance monitoring and retraining to adapt to evolving language patterns, periodic hyperparameter tuning for optimization, proactive measures to address concept drift, and implementation of robust security and privacy measures to safeguard the model and data.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13465

V. SYSTEM ARCHITECTURE

A. System Architecture

The system architecture for sub-toxic comment classification integrates several key modules: a data collection and preprocessing module for gathering and standardizing labeled datasets; a feature extraction and representation module employing NLP techniques to convert text into numerical representations; a deep learning architecture module constructing complex patterns and representations from textual data; an SVM integration module to complement deep learning features with SVM classification capabilities; a model training and evaluation module to assess performance metrics and deploy the trained model into a production environment through a deployment module, possibly utilizing containerization and API integration. Additionally, a monitoring and maintenance module ensures ongoing performance monitoring, drift detection, and periodic retraining to maintain effectiveness over time.

B. Mathematical Model

Let MM(D)=Deploy(T(SVM(DL(F(P(D))))))

D represents the dataset

P is the preprocessing module.

F is the feature extraction and representation module.

DL is the deep learning architecture module.

SVM is the Support Vector Machine integration module.

Where,

T is the model training and evaluation module

Procedure (P),

P=I,Deploy is the deployment module.

O=MM represents the monitoring and maintenance module.

2 / 2

Was this response better or worse?

C. Data Flow Diagram

In Data Flow Diagram,we Show that flow of data in our system in DFD0 we show that base DFD in which rectangle present input as well as output and circle show our system,In DFD1 we show actual input and actual output of system input of our system is text or image and output is rumor detected like wise in DFD 2 we present operation of user as well as admin.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13465

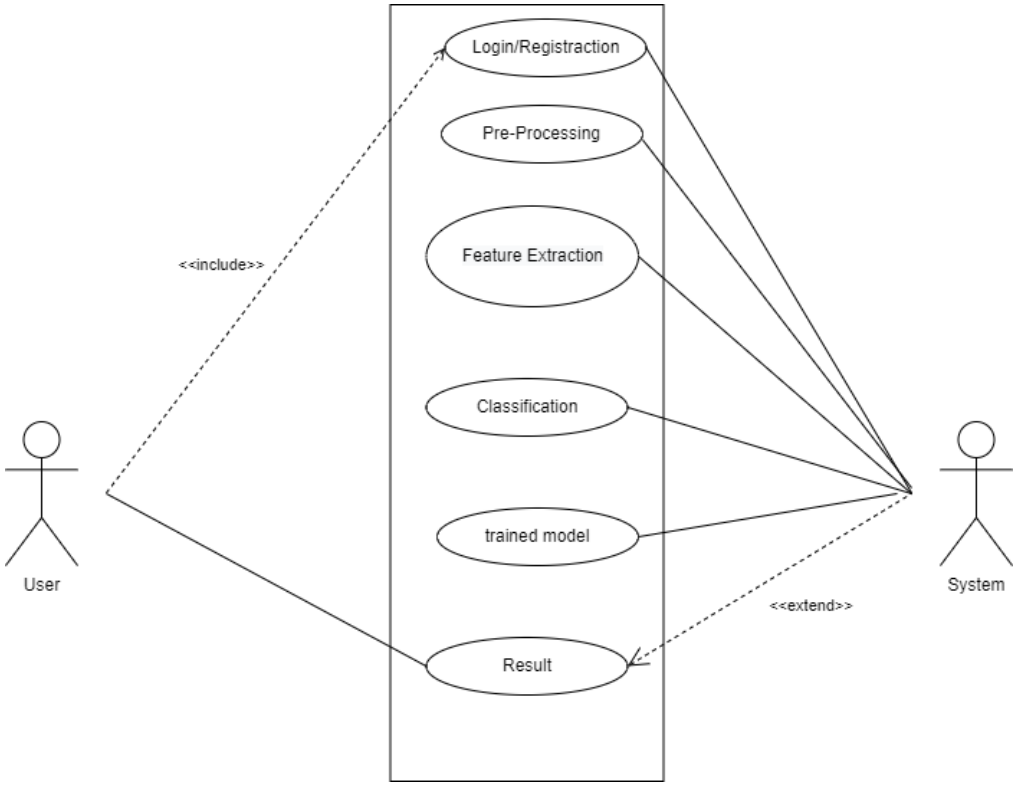
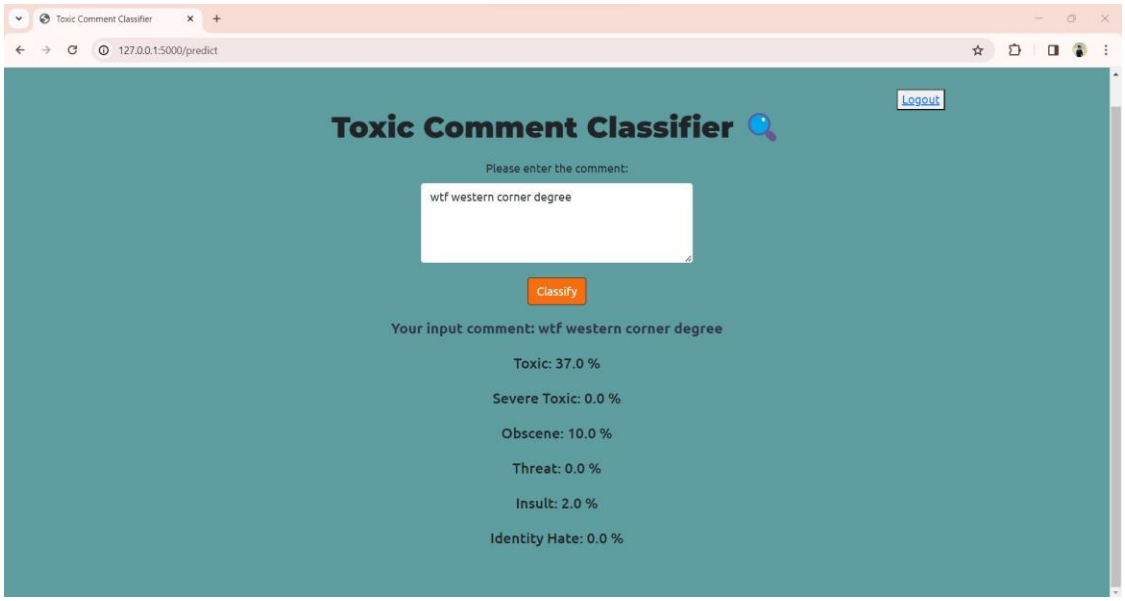


Fig.1 – Use Case Diagram

VI. RESULTS

The integration of deep learning with SVM for sub-toxic comment classification yielded promising results. The model achieved high accuracy in identifying subtle instances of toxic language, showcasing the effectiveness of leveraging both deep learning's feature extraction capabilities and SVM's classification prowess. This approach demonstrates potential for enhancing content moderation systems by accurately detecting nuanced forms of toxicity.



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13465

VII. CONCLUSION

Toxic Comment Detection is a machine learning project that aims to classify comments as toxic or non-toxic. The project uses a machine learning SVM model to classify the comments. The model is trained on a dataset of comments labelled as toxic or non-toxic. The dataset is pre-processed to remove stop words, punctuations, and other irrelevant information. In conclusion, the Toxic Comment Detection Using machine learning project effectively classifies comments as toxic or non-toxic with high accuracy. It can be used to filter out toxic comments from online platforms and improve the overall user experience.

ACKNOWLEDGMENT

We are very grateful to **Prof. Megha C Singur** (Guide and Assistant Professor of Department of Information Technology, Sandip Institute of Technology and Research Centre, Nashik) for her expert guidance and continuous encouragement throughout the project.
At last, we must express our sincere heartfelt gratitude to all the staff members and students of Information Technology Department who helped us directly or indirectly during this course of work.

REFERENCES

1. Anukarsh G Prasad, Sanjana S, Skanda M Bhat, B S Harish "Sentiment Analysis for Sarcasm Detection on Streaming Short Text Data", 2nd International Conference on Knowledge Engineering and Applications, IEEE, 2017
2. Sana Parveen, Sachin N. Deshmukh, "Opinion Mining in Twitter – Sarcasm Detection" International Research Journal of Engineering and Technology (IRJET), volume 04, issue 10, pages 201-204, October 2017.
3. ParasDharwal, TanupriyaChoudary, Rajat Mittal, Praveen Kumar, "Automatic Sarcasm Detection using Feature Selection", International Conference on Applied and Theoretical Computing and Communication Technology, IEEE, 2017.
4. Sindhu. C, G. Vaidhu, Mandala Vishal Rao, "A Comprehensive Study on Sarcasm Detection Techniques in Sentiment Analysis", International Journal of Pure and Applied Mathematics, volume 118, pages 433-442, 2018
5. Tanya Jain, NileshAgrawal, GarimaGoyal, NiyatiAggrawal, "Sarcasm Detection of Tweets: A Comparative Study", Tenth International Conference on Contemporary Computing (IC3), IEEE, August 2017
6. Levy, M. (2016). Playing with Twitter Data. [Blog] R-bloggers. Available at: https://www.r-bloggers.com/playing-with-twitter-data/ [Accessed 7 Feb. 2018.
7. Popularity Analysis for Saudi Telecom Companies Based on Twitter Data. (2013). Research Journal of Applied Sciences, Engineering and Technology. [online] Available at: http://maxwellsci.com/print/rjaset/v6-4676-4680.pdf [Accessed 1 Feb. 2018]..
8. Zhao, Y. (2016). Twitter Data Analysis with R – Text Mining and Social Network Analysis. [online] University of Canberra, p.40. Available at:https://paulvanderlaken.files.wordpress.com/2017/08/rdataminingslides-twitteranalysis.pdf [Accessed 7 Feb. 2018].
9. Alrubaiee, H., Qiu, R., Alomar, K. and Li, D. (2016). Sentiment Analysis of Arabic Tweets in e-Learning. Journal of Computer Science. [online] Available at: http://thescipub.com/PDF/jcssp.2016.553.563.pdf [Accessed 7 Feb. 2018].
10. Qamar, A., Alsuhibany, S. and Ahmed, S. (2017). Sentiment Classification of Twitter Data Belonging to Saudi Arabian Telecommunication Companies. (IJACSA) International Journal of Advanced Computer Science and Applications, [online] 8. Available https://thesai.org/Downloads/Volume8No1/Paper Sentiment Classification of Twitter Data Belonging.pdf [Accessed 1 Feb. 2018].
11. R. M. Duwairi and I.Qarqaz, "A framework for Arabic sentiment analysis using supervised classification" , Int. J. Data Mining, Modelling and Management, Vol. 8, No. 4, pp.369-381 , 2016..

