IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

HARNESSING TEXT CLASSIFICATION USING TEXTMINING AND DEEPLEARING TOSAFE GUARD ONLINE DISCOUSE

[1]E. JonesMerlin,[2]T. Bhuvaneshwari, [3]S. Priyadharshini ,[4]V. Ruba,[5]S. Subasri
[1]Assistant Professor, [2,3,4,5]Student
[1]Computer Science and Engineering,
[1]Sri Ramakrishna College Of Engineering, Perambalur, TamilNadu, India

Abstract : Everyone is entitled to the right to free speech. However, under the guise of freedom of speech, this privilege is misused to discriminate and harm others. This prejudice is called hate speech. A clear definition of hate speech is speech that expresses hatred against an individual or group of people based on characteristics such as race, religion, ethnicity, gender, national origin, disability, or sexual orientation. It can be communicated by voice, writing, gesture, or exhibition when someone is attacked due of the group to which they belong. Hate speech has been more prevalent both offline and online in recent years. The social media and other online platforms play a significant role in the development and dissemination of hateful information, which eventually fuels hate crimes. The growing use of social media and information sharing has brought about significant advantages for humanity. But this has also given rise to a number of issues, such as the disseminating and sharing of hate speech messages. This project's goal is to examine comments on social networks using Natural Language Processing (NLP) and a Deep Learning method called Valence Aware Dictionary for Sentiment Reasoning (VADER) method. In order to identify the text as positive or negative, VADER are used to extract the keywords from user-generated content. If it's negative, immediately block the comments in accordance with the user's preferences and block the friends in accordance with pre-established threshold values

IndexTerms:UsersOnlineDiscourse,safeguard,TextMining -

I.INTRODUCTION

Data mining is the computing process of discovering patterns in large data sets. involving methods at the intersection of machine learning, statistics, and database system. It is an interdisciplinary subfield of computer science. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets.

It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records, unusual records, and dependence. This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analysis.

II.RELATED WORK

In content-based filtering each user is assumed to operate independently. As a result, a content-based filtering device selects data gadgets based totally on the correlation between the content material of the objects and the consumer preferences as adversarial to a collaborative filtering machine that chooses gadgets based totally on the correlation between humans with comparable preferences. While digital mail was once the authentic area of early work on statistics filtering, subsequent papers have addressed varied domains inclusive of newswire articles, Internet "news" articles, and broader community resources. Documents processed in content-based filtering are ordinarily textual in nature More complex filtering systems include multi-label text categorization automatically labeling messages Online Social Networks (OSNs) are now one of the most popular interactive mediums for communicating, sharing, and disseminating a significant amount of human life data. Deep learning (DL) is a text classification technique that uses machine learning to assign each brief text message to one of several categories based on its content. which is then used to take a look at the phrases for any inappropriate words. If the conversation carries any vulgar terms, the message will be submitted to the Blacklists, which will filter these phrases out. Finally, as a result of the content-based-filtering technique, A system uses blacklists to automatically filter unwanted messages to assist users with Filtering Rules (FRs) formulation, and the extension of the collection of features examined in the classification process are some of the major differences. Finally predict the friends who are posted continues unwanted messages on user pages with alert system

III.METHODOLOGY

3.1 Framework Constructions

A social networking service is an online platform that people use to build social networks with other people who share similar personal or career interests, activities, real-life connections. In this module we can create the interface for admin and user. User can login to the machine and view the buddy request. The person can share the photographs to friends. The range and evolving vary of stand-alone and built-in social networking offerings in the on-line area introduces a task of definition.

3.2 Read Comments

social media is turning into an crucial phase of lifestyles on-line as social web sites and purposes proliferate. Most standard on line media encompass social components, such as remark fields for users. In business, social media is used to market products, promote brands, and join to modern-day clients and foster new business. In this module, we can remark in on-line social network. Comment in the structure of text. The textual content may also be uni-gram, bi-gram and multi grams. This module is used to get the enter from social users. Comments may also be a number of types such as hyperlinks or texts or quick texts. Comments are study and ship to server page

3.3 Classification

In this module, we sketch an automatic system, known as Filtered Wall (FW), in a position to filter undesirable messages from OSN person walls. The structure in assist of OSN offerings is a three-tier structure social media is turning into an crucial phase of lifestyles on-line as social web sites and purposes proliferate. Most standard on line media encompass social components, such as remark fields for users. The first layer many times targets to grant the simple OSN functionalities. The main efforts in constructing a sturdy lower back propagation neural community (BPNN) are targeted in the extraction and resolution of a set of characterizing and discriminant features. In order to specify and implement these constraints, we make use of the textual content classification. From VADER factor of view, we method the project by means of defining a hierarchical two-level approach assuming that it is higher to perceive and do away with "neutral" sentences, then classify "non-neutral" sentences via the category of hobby alternatively of doing the whole lot in one step. Finally, the supported SNA may also require an extra layer for their wished graphical consumer interfaces (GUIs). The main efforts in constructing a sturdy lower back propagation neural community (BPNN) are Centre in the extraction and resolution of a set of characterizing and discriminant points

3.4 Rules Implementation

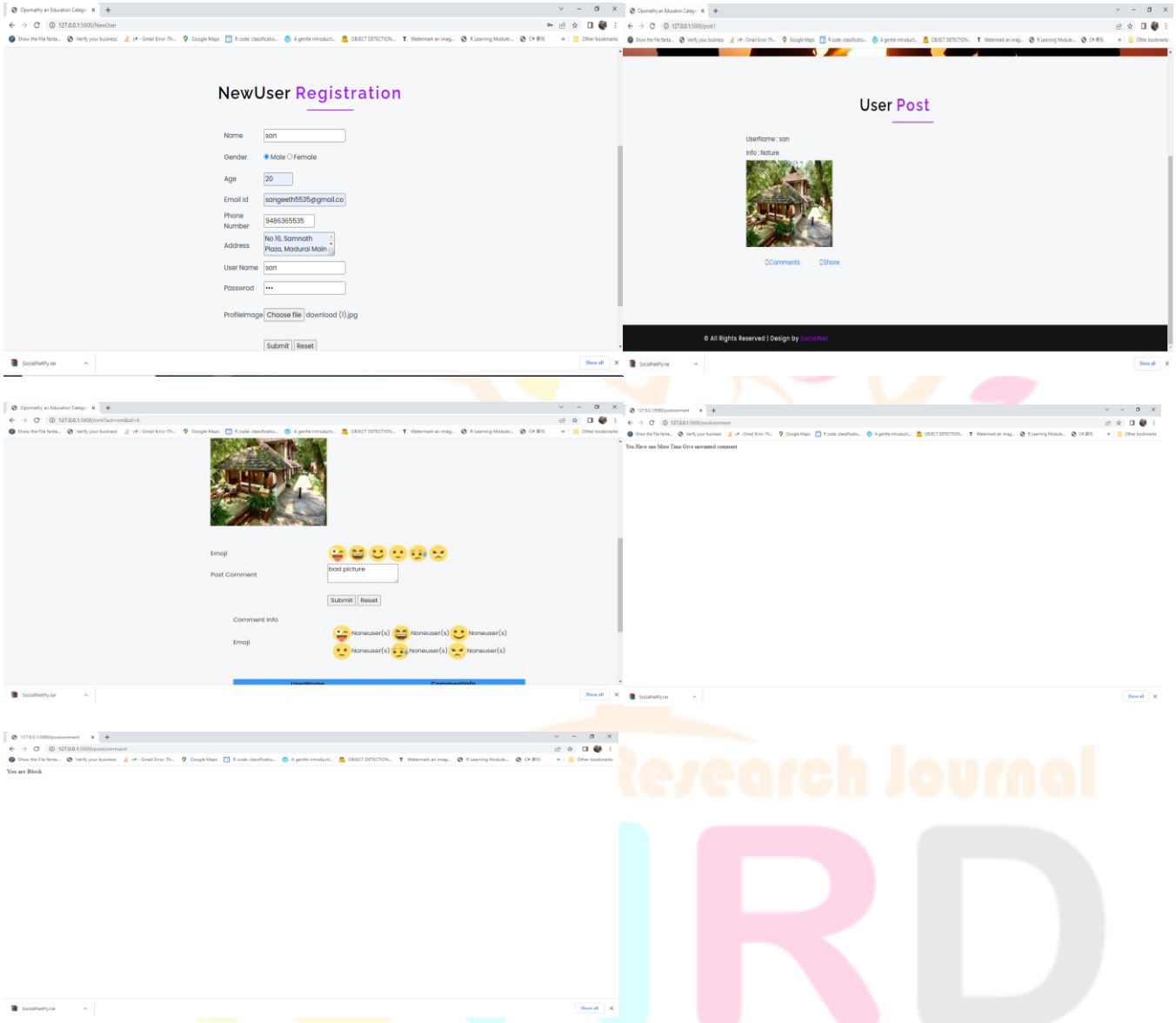
The filtering guidelines must enable customers to nation constraints on message creators. Thus, creators on which a filtering rule applies need to be chosen on the foundation of a number of extraordinary criteria; one of the most applicable is with the aid of imposing prerequisites on person profile's attributes. In such a way it is, for instance, viable to outline policies making use of solely to younger creators, to creators with a given religious/ political view, or to creators that we consider are now not specialist in a given discipline (e.g., through posing constraints on the work attribute of consumer profile). This capability filtering regulations figuring out messages in accordance to constraints on their contents. And block the customers who are publish the terrible remarks extra than 5 instances and additionally ship cell intimation to customers at the time offline

3.5 Alert System

BL'S are without delay managed with the aid of the system, which need to be in a position to decide who the customers are inserted in the BL and figure out when the person retention in the BL is finished. To enhance flexibility, this data is in the machine through a set of rules; the guidelines on BL. Rules are generated through server for placing threshold values. Alert ship to person web page and warning for 4 instances and then block the person and ship SMS to person always posting bad feedback Next

IV.RESULTS AND DISCUSSION

One of the most important issues in today's On-line Social Networks (OSNs) is giving users the opportunity to control the messages posted on their own private area in order to prevent the appearance of undesired content. Until date, OSNs have been unable to meet this condition. To fill the void, we propose in this paper a mechanism that gives OSN users direct control over the messages put on their walls. This is accomplished via a flexible rule-based system that allows users to design the filtering criteria that will be applied to their walls, as well as a Machine Learning.



It makes use of techniques at the intersection of synthetic intelligence, computer learning, statistics, and systems. The ordinary purpose of the statistics mining technique is to extract data from a statistics set and seriously change it into an comprehensible shape for in addition use Aside from the uncooked evaluation step, it includes database and facts administration aspects, records pre-processing, mannequin and inference considerations, interestingness metrics, complexity considerations, post-processing of found structures, visualization, and online updating.

The extraction and selection of a set of characterizing and discriminating features is the focus of most efforts in developing a robust back propagation algorithm. A database of categorized terms is created here, which is then used to check the words .A system uses blacklists to automatically filter unwanted messages based on both message content and message creator relationships and characteristics. A revised semantics for filtering rules to better fit the considered domain, to assist users with Filtering Rules (FRs) formulation, and the extension of the collection of features examined in the classification process are some of the major differences. Finally predict the friends who are posted continues unwanted messages on user pages with alert system.

STEPS INVOLVED IN THIS PROJECT

Step 1: Collect comments or text data with samples of both hateful and non-hateful language.
Step 2: Implement natural language processing for tokenizing. Tokenize text-based reviews as single terms, Analyze unigrams, bigrams, and n-grams
Step 3: Remove stop words, analyze stemming words, and remove special characters Finally, extract key phrases, Analyze extended words that can be substituted with right word
Step 4: Eliminate the stop words, stemming words and extract key terms based on text mining approach
Step 5: Classification can be including VADER algorithm to label comments
Step 6: Each token is looked up in the sentiment lexicon, which contains words annotated with pre-defined sentiment scores. These scores range from -1 (most negative) to 1 (most positive).
Step 7: VADER calculates the sentiment intensity of each token based on its sentiment score. The sentiment scores for all tokens in the text are aggregated to obtain an overall compound sentiment score.
Step 8: Based on the compound sentiment score, VADER classifies the text into positive, negative, or neutral categories. The threshold values for classification may be adjusted depending on the application
Step 9: SMS alert at the time of posting negative comments in social network page Block the friends who continuously post the negative common

V.CONCLUSION

We demonstrated a solution to filter unwanted messages from OSN walls in this project. The system uses a DL soft classifier to enforce a content-dependent filtered rules system that may be customised. The extraction and selection of a set of characterising and discriminant features are the most time-consuming aspects of developing a robust short text classifier. Furthermore, the handling of BLs improves the system's versatility in terms of filtering choices. This project is the initial step in a larger one. The early promising results we've seen with the classification technique encourage us to keep working on other projects aimed at improving classification quality.

The DL soft classifier is used in this system to filter out undesirable signals. BL is used to increase the filtering system's flexibility. We'll create a mechanism that takes a more comprehensive approach to determining when a user should be added to the BL. In addition to classification features, the system includes a strong rule layer that uses a flexible language to construct Filtering Rules (FRs), which allow users to decide which information should not be displayed on their walls. FRs can enable a wide range of filtering criteria that can be combined and tailored to meet the needs of the user. FRs leverage user profiles, user relationships, and the output of the DL classification process to specify the filtering criteria that will be used.

VI.FUTURE WORK

We plan to use similar strategies to infer Block list rules and Filtering rules in the future. We can enhance the framework in the future to implement this approach in a variety of languages with higher accuracy. Also included is the semi-supervised technique to unlabeled data analysis. Unlabeled comments such as images can be filter

VII. REFERENCES

[1] Akuma, Stephen, Tyosar Labem, and Isaac Terngu Adom. "Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets." International Journal of Information Technology 14.7 (2022): 3629-3635.
[2] Alkomah, Fatimah, and Xiaogang Ma. "A literature review of textual hate speech detection methods and datasets." Information 13.6 (2022): 273.
[3] Aluru, Sai Saketh, et al. "Deep learning models for multilingual hate speech detection." arXiv preprint arXiv:2004.06465 (2020).
[4] Antypas, Dimosthenis, and Jose Camacho-Collados. "Robust hate speech detection in social media: A cross-dataset empirical evaluation." arXiv preprint arXiv:2307.01680 (2023).
[5] Cao, Rui, Roy Ka-Wei Lee, and Tuan-Anh Hoang. "DeepHate: Hate speech detection via multi-faceted text representations." Proceedings of the 12th ACM Conference on Web Science. 2020
[6] Fortuna, Paula, et al. "Directions for NLP Practices Applied to Online Hate Speech Detection." Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022.
[7] Gravano, Agustín, et al. "Assessing the Impact of Contextual Information in Hate Speech Detection." IEEE Access, vol. 11, pp. 30575-30590, 2023, do: 10.1109/ACCESS. 2023.3258973. (2023).
[8] Khan, Shakir, et al. "BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection." Journal of King Saud University-Computer and Information Sciences 34.7 (2022): 4335-4344.
[9] Khan, Shakir, et al. "HCovBi-caps: hate speech detection using convolutional and Bi-directional gated recurrent unit with Capsule network." IEEE Access 10 (2022): 7881-7894.
[10] Malik, Jitendra Singh, Guansong Pang, and Anton van den Hengel. "Deep learning for hate speech detection: a comparative study." arXiv preprint arXiv:2202.09517 (2022).
[11] Mosca, Edoardo, Maximilian Wich, and Georg Groh. "Understanding and interpreting the impact of user context in hate speech detection." Proceedings of the Ninth International Workshop on Natural Language Processing for social media. 2021.

[12] Mozafari, Marzieh, Reza Farahbakhsh, and Noel Crespi. "A BERT-based transfer learning approach for hate speech detection in online social media." Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEXNETWORKS2019 8. Springer International Publishing.
[13] Mullah, Nanlir Sallau, and Wan Mohd Nazmee Wan Zainon. "Advances in machine learning algorithms for hate speech detection in social media: a review." IEEE Access 9 (2021): 88364-88376.
[14] Patil, Hrushikesh, Abhishek Velankar, and Raviraj Joshi. "L3cube-mahahate: A tweet-based marathi hate speech detection dataset and Bert models." Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022). 2022.
[15] Rabiul Awal, Md, et al. "AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection." arXiv e-prints (2021): arXiv-2103.
[16] Roy, Pradeep Kumar, et al. "A framework for hate speech detection using deep convolutional neural network." IEEE Access 8 (2020): 204951-204962.
[17] Sabat, Benet Oriol, Cristian Canton Ferrer, and Xavier Giro-i-Nieto. "Hate speech in pixels: Detection of offensive memes towards automatic moderation." arXiv preprint arXiv:1910.02334 (2019).
[18] Toraman, Cagri, Furkan Şahinuç, and Eyup Halit Yilmaz. "Large-scale hate speech detection with cross-domain transfer." arXiv preprint arXiv:2203.01111 (2022).
[19] Velankar, Abhishek, Hrushikesh Patil, and Raviraj Joshi. "A review of challenges in machine learning based automated hate speech detection." arXiv preprint arXiv:2209.05294 (2022).
[20] Zhou, Xianbing, et al. "Hate speech detection based on sentiment knowledge sharing." Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021.