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

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁒ Impact Factor 7.39 ⁒ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11680

Cyberbullying Detection

Vaishnavi K[1], Prof. Pallavi N[2], Prof. Padmini C[3]

[1]Student, Department of Computer Science and Engineering, Atria Institute of Technology, Bangalore, India

[2,3]Assistant Professor, Department of Computer Science and Engineering, Atria Institute of Technology,

Bangalore, India

Abstract: As a side effect of increasingly popular social media, cyberbullying has emerged as a seriousproblem afflicting children, adolescents and young adults. Since the textual contents on online social media area highly unstructured, informal, and Often misspelled, existing research on message level offensive language detection cannotaccurately detect offensive contents. Here we design a framework called Lexical Syntactic Feature (LSF) architecture to detectoffensive contents and identify potential offensive users in social media. We distinguished the contribution of profanities and obscenities in determining offensive content and introduce hand authoring syntactic rule in identifying name calling harassments. In particular we incorporated a user's writing style, structure and specific cyberbullyingcontents as features to predict the users capability to send out offensive content. Results from the experiments shows that the LSF framework performed significantly betterthan existing methods in offensive content detection.

Keywords: Cyberbullying, Offensive ,Lexical syntactic feature, detection.

I.INTRODUCTION

with the rapid growth of social media users especially adolescents are spending significant amount of time on various social networking sites to connect with others to share information and to pursue common interests it has been found that 70 of teens use social media sites on a daily basis and nearly one in four teens hit their favorite social media sites 10 or more times a day 19 of teens report that someone has written or posted mean or embarrassing things about them on social networking sites as adolescents are more likely to be negatively affected by biased and harmful contents than adults detecting online offensive contents to protect adolescents online safety becomes an urgent task to address concerns on children access to offensive contents over internet administrators of social media often manually review online contents to detect and delete offensive materials however manually reviewing are labor intensive and time consuming.

Some automated content filtering software packages such as Appen, Eblaster, iambigbrother, and Internet Security Suite are designed to detect and filter inappropriate online content. Most of them simply blocked websites and paragraphs that contained dirty words. Not only do these word-based approaches affect the readability and usability of your website, they also can't detect subtle offensive messages. To address these limitations, we propose a lexical syntax function-based (LSF) language model to effectively detect offensive languages on social media and protect adolescents. LSF looks not only at the message, but also at the person who posted the message and its posting pattern. LSF can be implemented as a client-side application for individuals and groups who are concerned about the online safety of young people. Users can also adjust the threshold for acceptable levels of offensive content.

II.LITERATURE REVIEW

Maral Dadvar et al. [19] proposed offensive speech recognition on social media is a daunting task because the text content in such an environment is unstructured, unofficial, and often misspelled.

researchers have considered smart ways to identify offensive content using a text mining approach. Implementing text mining techniques for analyzing online data requires the following phases:

1) Data collection and preprocessing,

2) Feature extraction, and

3) Classification.

The main challenge in detecting offensive content using text mining is in the feature selection phase. This will be discussed in more detail in the next section.

a) Message-level feature extraction: Most offensive content detection studies extract two types of features: lexical features and syntactic features. The lexical feature treats each word and sentence as one unit. Word patterns, such as the appearance and frequency of specific keywords, are often used to represent a language model. For example, a bag of words (BoW) and n-gram.

b) User-level aggression detection: Most of the latest research on offensive online language detection focuses only

DOI: 10.17148/IJARCCE.2022.11680

on sentence-level and message-level composition. There is no 100% accurate detection method. However, user-level detection is a more difficult task.

Kontostathis et al[18] proposed rule-based communication model for tracking and classifying online predators. Pendar uses a phrase function with a machine learning classifier to distinguish between victims and predators in an online chat environment. Pazienza and Tudorache suggest using user profiling features such as online presence and conversation to detect positive discussions. However, the above survey did not include user information such as stylistic style, user posting trends, or reputation to improve detection rates.

Kasturi DewiVarathan et al. [16] presented popular online social networking sites employ multiple mechanisms to screen for inappropriate content. For example, you can enable YouTube safe mode to hide all comments, including offensive language, from users. However, pre-checked content will still be displayed. The user simply clicks Text Comment and the derogatory term is replaced with an asterisk. Facebook allows users to add comma-separated keywords to the Moderation Blacklist. When people insert edited keywords into page posts or commands, the content is automatically identified as spam and is reviewed. The Twitter client "Tweetie 1.3" has been rejected by the Apple Company for allowing users to display vulgar expressions in their tweets.

In general, the most popular social media uses a simple dictionary-based approach to filter offensive content. These lexicons are either predefined (such as YouTube) or compiled by you (such as Facebook).

For adolescents who often lack cognitive awareness of risk, these approaches are less effective at preventing exposure to offensive content. Therefore, to protect young people from the potential for exposure to vulgar, pornographic and hateful words, parents use more specialized software and discomfort to protect young people from the efficient detection of offensive content techniques. Requires efficient detection of content detection technology.

III.PROPOSED METHODOLOGY

To successfully detect hazardous languages on social media and protect teenagers, we propose a lexical syntax function-based (LSF) language model. LSF not only sees the message, but also the individual who posted it and the trend of publishing. For individuals and organization's concerned about young people's online safety, LSF can be implemented as a client-side application. The threshold for acceptable levels of offensive content can also be adjusted by users. In terms of precision, recall, and F-score, experimental results reveal that algorithms for forecasting SPF phrase vulnerability and evaluating user vulnerability outperform classical learning-based approaches. Indicates. It also has fast processing rates for efficient social media use.

A. Design Framework

To tackle these issues, we developed a Lexical Syntactic Feature (LSF)-based framework for detecting offensive content and identifying offensive users on social media. The framework comprises two processes for detecting offensiveness. The goal of Phase 1 is to detect offensiveness at the phrase level. On the user level, Phase 2 determines offensiveness.

Pre-processing and two primary components make up the system: sentence offensiveness prediction and user offensiveness prediction.

Users' interactions are chunked into posts, then sentences, at the pre-processing step.

During sentence offensiveness prediction, each phrase's offensiveness is determined by two factors: the offensiveness of its terms (using lexical characteristics) and the context (using syntactical features). i.e., parse texts grammatically into dependency sets to capture all forms of connections between words.

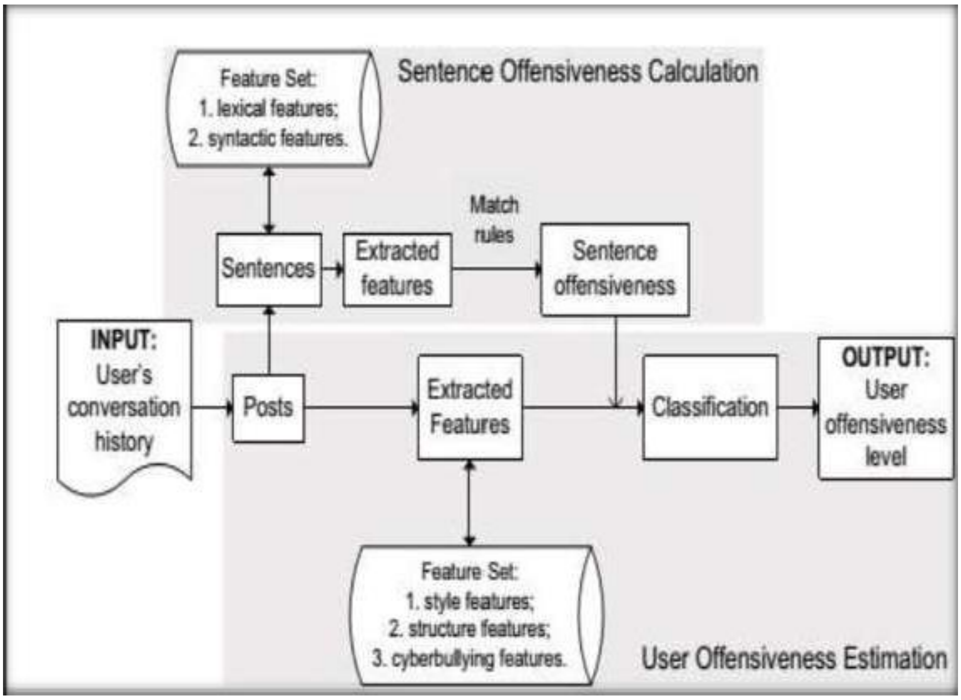


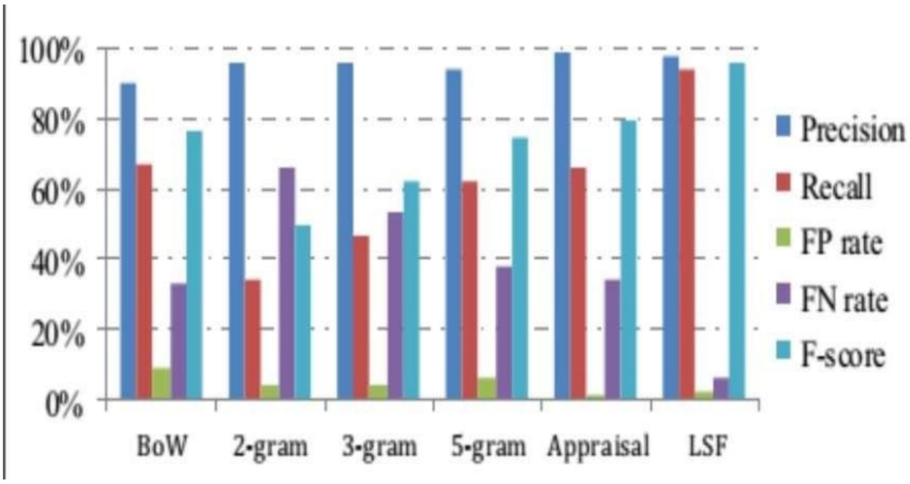
Figure1. Data Flow Diagram

B. Experiments

1. Sentence Offensiveness Calculation

Based on objectionable word lexicons and sentence syntactic structures, a novel approach of sentence level analysis is offered. To begin, we create two offensive word dictionaries with varying degrees of offensiveness. Second, the syntactic intensifier idea is created to modify the level of offensiveness of words dependent on their context. Finally, an offensiveness value is calculated for each sentence by adding the words offensiveness' together. Mathematically
The offensiveness of each offensive word, W in a phrase, is indicated by the letter S.
If W is a really insulting term, use option A1.
A2 = Ow if W is a mildly objectionable term, 0 if not

The significance of intensifier Iw can be calculated as k j=1 dj for the offending word w. As a result, the offensiveness value of a sentence, s, becomes a linear combination of the offensiveness of words. Os = ∑ Ow Iw
b1 if user identified
dj =b2 if is an offensive word
0 otherwise
Figure2. Sentence Offensiveness Calculation



DOI: 10.17148/IJARCCE.2022.11680

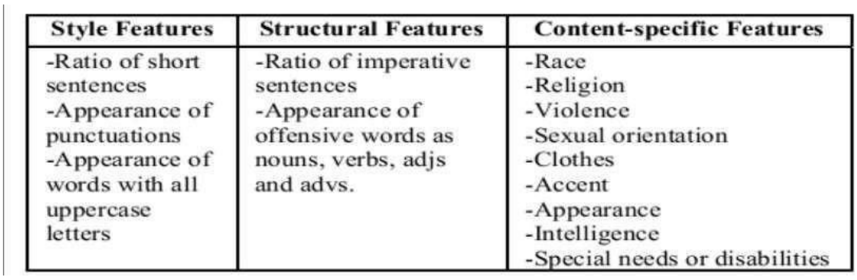
Because many of the objectionable sentences are imperatives, which eliminate all user identification, none of the baseline techniques achieves a recall rate higher than 70%, as shown in the above Fig. The BoW techniques had the highest recall rate (66) among the baseline approaches. BoW, on the other hand, has a high false positive rate.

2. User Offensive Estimation

Given a user, u,they retrieve his/her conversation history which contains several posts {P1,…………., Pm} and each post contains sentences { s1,……..., sn}. Sentences offensiveness Values are dentoted as {Os1, , Osn}.The original offenesiveness value of post p, $Op=\sum Os$.The offensiveness value of modified posts can be presented as, $Op\rightarrow s$.

So the final post offensiveness Op of post p can be calculated as $Op=\max(Op,Op\rightarrow s)$. Hence, the offensiveness value of Ou,can be presented as $Ou=1/m\sum Op$.

Classification is performed using machine learning techniques such as Naive Bayes (NB) and SVM with 10-Ford cross-validation. To fully assess the usefulness of the user's sentence vulnerability score (LSF), style features, structural features, and content-specific features for estimating user vulnerability, the data is sequenced into classifiers. Supplied and the results were obtained.



Style Features	Structural Features	Content-specific Features
-Ratio of short sentences -Appearance of punctuations -Appearance of words with all uppercase letters	-Ratio of imperative sentences -Appearance of offensive words as nouns, verbs, adjs and advs.	-Race -Religion -Violence -Sexual orientation -Clothes -Accent -Appearance -Intelligence -Special needs or disabilities

Figure3. Features Classification

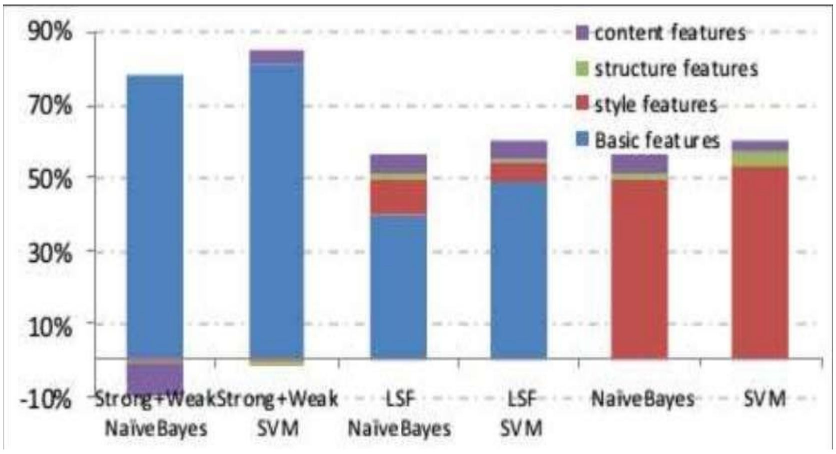


Figure4.User Offensiveness Estimation-with presence of Strongly offensiveness words

As seen in the figure, offensive words and user language traits do not compensate for each other to boost detection rates, showing that they are not mutually incompatible. Classifiers that include user language features, on the other hand, have a higher detection rate than those that just use LSF.

IV.CONCLUSION

In this study, an existing text mining method to detect offensive content to protect the online safety of young people, a vocabulary syntax function (LSF) proposed to identify offensive content on social media. Think of an approach and how to predict how users will find sending content unpleasant. .. There are several contributions to this research. First, the concept of offensive online content is concretely conceptualized, further distinguishing the contribution of derogatory / profane and profane words to the determination of offensive content, and identifying harassment by hand by attribution. Second, improved the traditional machine leaning methods by not only using lexical features to detect offensive

DOI: 10.17148/IJARCCE.2022.11680

languages, but also incorporating style features, structure features and context-specific features to better predict a user's potentially to send out offensive content in social media.

Experimental result shows that the LSF sentences offensiveness prediction and user offensiveness estimate algorithm's outperform traditional learning based approaches in terms of precision, recall and f-score. If also achieves high processing speed for effective deployment in social media.

V.REFERENCES

[1] Abeele, M.V. and De Cock, R. (2013). Cyberbullying by Mobile Phone among Adolescents: The Role of Gender and Peer Group Status. Communications, 38(1), p.107-118

[2] Al Mazari, A. (2013). Cyber-bullying taxonomies: Definition, forms, consequences and mitigation strategies. IN: International Conference on Computer Science and Information Technology (CSIT). 5th . Amman. March 27 – 28, 2013. IEEE, 126-133

[3] Argamon, S., Koppel, M., Fine, J. and Shimoni, A. R. (2003) Gen der, Genre, And Writing Style In Formal Written Texts. Text In terdisciplinary Journal for the Study of Discourse, 23, p.321-3