Indian Engineering Journal
ISSN: 2984-9047
Vol 4 Issue 1 (2026)

Building a Safer Social Network: Machine Learning Algorithm for Blocking Digital Bullying Words

A. Sharfudeen[1], Dr. R. Sankarasubramanian[2]

[1]Research Scholar, PG & Research Department of Computer Science, Erode Arts & Science College (Autonomous), Erode

[2]Principal, Erode Arts & Science College (Autonomous), Erode

Abstract

A significant societal issue, particularly among youngsters, is cyberbullying."The deliberate, repeated, and hostile use of information technology to harm or harass other people" is the definition of cyberbullying. It has increased in popularity since the creation of social media platforms like Twitter and Face book. As a result, social media academics are placing more and more emphasis on the automatic identification of messages that engage in cyberbullying. The use of rules and norms, human moderators, and blacklists based on offensive language are some of the traditional strategies for combating cyberbullying. In order to automatically identify cyberbullying behaviors, a principled learning framework must be created. A flexible rule-based system that enables users to tailor the filtering criteria to be factual to their walls and a soft classifier powered by machine learning that automatically labels messages in content-based filtering help achieve this. Server can learn the words and save them in the database based on this filtering. Before the message can be exchanged, the server might review the words at the moment of transmission and prohibit.

Keywords: cyberbullying, human moderators, blacklists, deliberate, filtering

Introduction

Overview

The need to elicit public opinion about predefined topics is widespread in the social sciences, government and business. Traditional survey-based methods are being partly replaced by social media datamining but their potential and limitations are poorly understood. This article investigates this issue by introducing and critically evaluating a systematic social media analytics strategy to gain insights about a topic from social network. The results of an investigation into sets of dance style videos show that it is possible to identify plausible patterns of subtopic difference, gender and sentiment. The analysis also points to the generic limitations of social media analytics that derive from their fundamentally exploratory multi-method nature. The majority of the handlers in these networks are under the age of thirty. Researchers have taken advantage of the vast amounts of data available on numerous social networking sites and undertaken extensive research in a variety of fields. Sentiment Analysis is a popular field of study that utilizes a lot of data from social media.

Problem Statement

Detecting cyberbullying words involves identifying and categorizing language used in online communication that is intended to harass, intimidate, or harm others. This is a critical task in ensuring the safety and well-being of individuals in online communities, especially on social media platforms. The challenge lies in developing algorithms that can accurately differentiate between normal, non-threatening language and potentially harmful or abusive content. This requires a nuanced understanding of language context, including slang, sarcasm, and cultural nuances, to avoid false positives. Furthermore, the system must be able to adapt to evolving forms of cyber bullying and be scalable to handle the vast amount of text data generated online daily. By effectively detecting cyberbullying words, online platforms can creates after environments, protect users from harm, and foster positive interactions among users.

Exisiting System

In content-based filtering each user is assumed to operate independently. As a result, a content-based filtering system selects information items based on the correlation between the content of the items and the user preferences as opposed to a collaborative filtering system that chooses items based on the correlation between people with similar preferences. While electronic mail was the original domain of early work on information filtering, subsequent papers have addressed diversified domains including newswire articles, Internet "news "articles, and broader network resources. Documents processed in content-based filtering are mostly textual in nature and this makes content-based filtering close to text classification. The activity of filtering can be modeled, infact, as a case of single label, binary classification, partitioning incoming documents in to relevant and non-relevant categories. More complex filtering systems include multi-label text categorization automatically labeling messages into partial thematic categories. Content-based filtering is mainly based on the use of the ML paradigm according to which a classifier is automatically induced by learning from a set of pre-classified examples. A remarkable variety of related work has recently appeared, which differ for the adopted feature extraction methods, model learning, and collection of samples. The feature extraction procedure maps text into a compact representation of its content and is uniformly applied to training and generalization phases. Several experiments prove that BagofWords (BoW) approaches yield good performance and prevail in general over more sophisticated text representation that may have superior semantics but lower statistical quality.

Disadvantages

- There is no approach for filter unwanted messages in social networks
- Can't deal with posts that are associated with images
- Difficult analyze short text tags.
- Automatic blocking can't be implemented

Proposed System

On-line Social Networks (OSNs) are today one of the most popular interactive medium to communicate, share and disseminate a considerable amount of human life information. One fundamental issue into day On-lineSocialNetworks (OSNs) is to give users the ability to control the messages posted on their own private space to avoid that unwanted content is displayed. Up to now OSNs provide little support to this requirement. To fill the gap, in this paper, propose a system allowing OSN users to have a direct control on the messages posted on their walls. This is achieved through a flexible rule-based system, that allows users to customize the filtering criteria to be applied to their walls, and a Machine Learning based soft classifier automatically labeling messages in support of content-based filtering. Deep learning (DL) is used as text categorization techniques to automatically assign each short text message with in a set of categories based on its content. The major efforts in building a robust back propagation algorithm concentrate in the extraction and selection of a set characterizing and discriminating features. Here, a database of the categorized words is built and it is used to check the words if it has any indecent words. If the message consists of any vulgar words, then they will be sent to the Blacklists to filter out those words from the message. Finally, the message without the indecent words will be posted in the user's wall on the result of the content-based-filtering technique. A system automatically filters s unwanted messages using the blacklists on the basis of both message content and the message creator relationships and characteristics. Major difference include , a different semantics for filtering rules to better fit the considered domain, to help the users Filtering Rules(FRs) specification, the extension of the set of features considered in the classification process.

Advantages

- Short texts are classified based BPNN approach.
- And also blocked malicious users.
- System uses deep learning classifier to enforce customizable content dependent rules
- Also eliminate the unwanted links in OSN pages

System Architecture

A system architecture or systems architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. System architecture can comprise system components, the externally visible properties of those components, the relationships (e.g. the behavior) between them. It can provide a plan from which products can be procured, and systems developed, that will work together to implement the overall system. There have been efforts to formalize languages to describe system architecture; collectively these are called architecture description languages (ADLs).

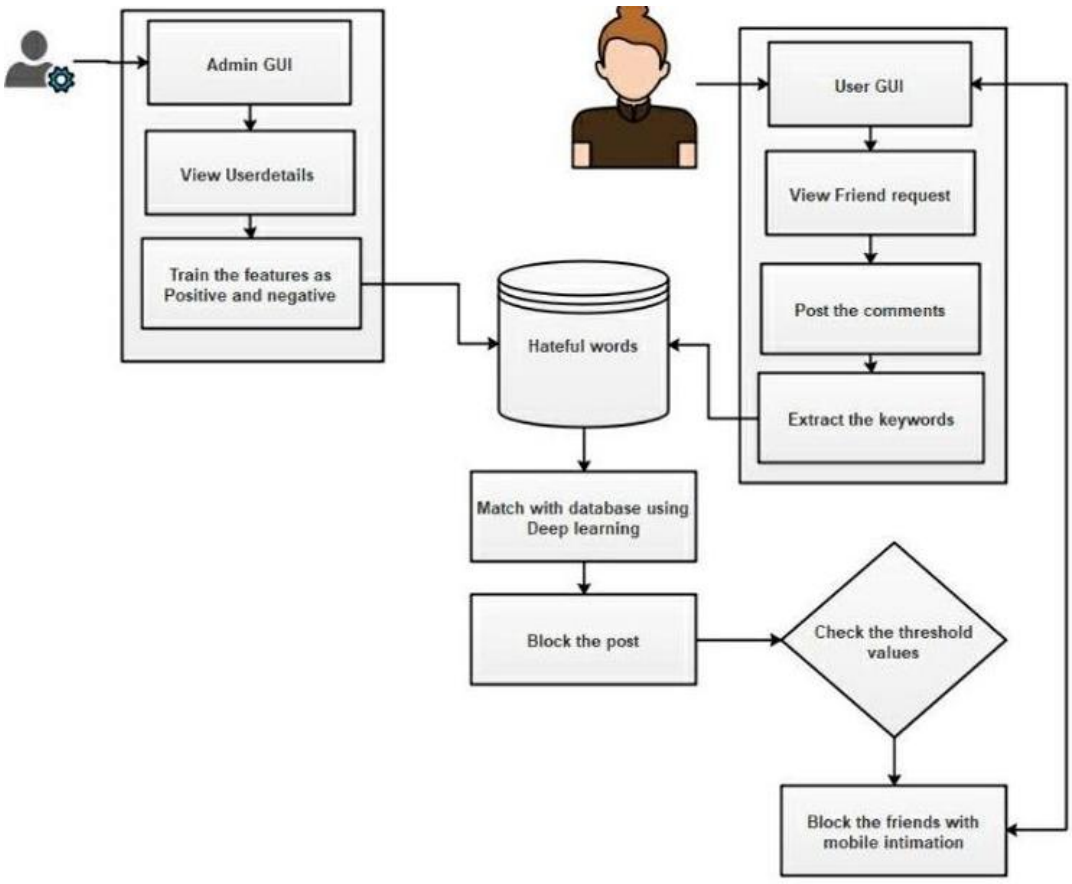

Fig 1.1 Architecture Diagram

Conclusion And Future Enhancement Conclusion

I demonstrated a solution to filter unwanted messages from OSN walls. The system uses a DL soft classifier to enforce a content-dependent filtered rules system that may be customized. The extraction and selection of a set

of characterizing 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 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. 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

Future Enhancement

I plan to use similar strategies to infer BL rules and FRs in the future. I 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.