ISSN: 2366-1313

# OFFENSIVE LANGUAGE DETECTION USING MACHINE LEARNING TECHNIQUES ON SOCIAL MEDIA

[1]ALAKUNTA. SWATHI, [2]G. RAMASWAMY

[1]M.Tech student dept of CSE, Malineni Lakshmaiah Women's Engineering College, Pulladigunta, Guntur.

[2]Head of Department of CSE, Malineni Lakshmaiah Women's Engineering College, Pulladigunta, Guntur

**Abstract**: Offensive communications have invaded social media content. One of the most effective solutions to cope with this problem is using computational techniques to discriminate offensive content. Moreover, social media users are from linguistically different communities. In this study, we propose a model for text classification consisting of modular cleaning phase and tokenizer, three embedding methods, and eight classifiers. Our experiments show a promising result for detection of offensive language on our dataset obtained from Twitter. Considering hyperparameter optimization, three methods of AdaBoost, SVM and MLP had highest average of F1-score on popular embedding method of TF-IDF.

**Keywords**: cyberbullying; adolescent safety; offensive languages; social media.

## I. INTRODUCTION

Text classification is the process of classifying textual documents into predefined categories based on their content. Classifying text is the automated assignment of natural language texts to predefined categories. Text classification is the primary requirement of text retrieval systems, which retrieve texts in response to a user query, to extract some knowledge and text understanding systems, which transform text in some way such as producing summaries, answering questions, making decisions or extracting data. Text mining has become one of the trendy fields in technology area that has been incorporated in several research fields such as computational linguistics, Information Retrieval (IR) and data mining. Natural Language Processing (NLP) techniques were used to extract knowledge from the textual data that is written by human beings. Text mining reads an unstructured form of data to provide meaningful information patterns in a shortest time period [1]. Now a days, social networking sites are a great source data generation as most of the people in today's world use these sites in their daily lives to keep connected to each other. Social networking websites create new ways for engaging people belonging to different communities [2]. Social networks allow users to communicate with people exhibiting

ISSN: 2366-1313

different moral and social values. The websites provide a very powerful medium for communication among individuals that leads to mutual learning and sharing of valuable knowledge. On social media it becomes a common practice to not write a sentence with correct grammar and spelling. This practice may lead to different kinds of ambiguities like lexical, syntactic and semantic and due to this type of unclear data, it is hard to find out the actual data order. Therefore, extracting logical patterns with accurate information from such unstructured form of data is a critical task for performing analysis [3]. Social network analysis applications have experienced tremendous advances from past few years due in part to increasing trends towards users interacting with each other on the internet. Social networks are organized as graphs, and the data on social networks takes on the form of massive streams, which can be mined for various purposes. Social Network Text Analysis from the post covers an important era in the social network analytics field. This edited volume, contributed by prominent researchers in this field that helps in presenting a wide selection of topics on social network data mining such as Structural Properties of Social Networks, Algorithms for Structural Discovery and Content Analysis in Social Networks [4]. With the rise of social media, people obtain and share different types of information updates on a 24/7 basis. Social media includes social networking sites and blogs where people can easily connect with each other. Social media has been mainly defined as "the many relatively inexpensive and widely accessible electronic tools that facilitate anyone and anytime access information, collaborate on a common effort, or build relationship". Many research areas have tried to gain valuable insights from these large volumes of freely available user generated content. The research areas for e-Commerce, intelligent transportation systems, smart cities, Cyber Crime, etc. are no exception. However, extracting meaningful and actionable knowledge from user generated content is a complex task. As each social media service has its own data collection formats and constraints. The volume of messages posts produced becomes overwhelming for automatic processing and mining [5]. Along with this, social media texts are usually short, informal, with a lot of abbreviations, jargon, slang leads to unstructuredness. As per the above mentions application of social media platform now a day's social media has been the important part of one's life and plays a vital role in transforming people's life style. Since the emergence of these social networking sites like Twitter, Facebook and Instagram as key tools for

ISSN: 2366-1313

news, journalists and their organizations have performed a high-wire act. These sites have become a day-to-day routine for the people. However, one does not need to look very far to experience the darker side of social media use. News reports of cyber bullying, violence, criminal activity, and suicide fueled by social media is shocking and troubling. Social networks are an inherent part of today's Internet and used by more than a billion people worldwide. They allow people to share ideas and interact with other people, from old friends to strangers. This interaction reveals a lot of information, often including personal information visible to anyone who wants to view it. Children are also growing around and surrounded by mobile device and uses interactive social networking sites. So, it becomes necessary to avoid this drawback and improve the detection of violence posting on the social media.

## II. REVIEW OF LITERATURE

### Offensive language detection

As of late, identifying cyberbullying, aggression, hate speech, toxic comments, and offensive language in social media receives much attention from the researcher's community. Several public datasets are available to train machine classifiers for those assignments. However, there are no standard benchmark corpora or training sets that can be combined to obtain more robust classification systems. Kumar et al. (2018) presented the report and findings of the shared task on aggression identification. The provided dataset contains 15,000 annotated Facebook posts and comments in English and Hindi. The goal was to discriminate between three classes: non-aggressive, covertly aggressive, and overly aggressive. The toxic comment classification was an open competition at Kaggle. Various methods were evaluated for this task on a dataset containing users with comments from Wikipedia. These comments are organized into six classes: toxic, severe toxic, obscene, threat, insult, identity hate. Concerning hate speech identification, Davidson et al. (2017) presented a recent hate speech detection dataset with over 24,000 English tweets belonging to three categories: non-offensive, hate speech, and profanity. Mandl et al. (2019) reported the shared tasks on offensive language identification where three datasets were developed from Twitter and Facebook and made available for Hindi, German, and English. Moreover, Zampieri et al., 2019, Zampieri et al., 2020 presented several offensive language detection results in several languages obtained by teams of SemEval competition.

### Multilingual text classification

ISSN: 2366-1313

Multilingual text classification is an emerging field in text classification. However, not many previous works have been realized in this area. Early, Lee et al. (2006) presented a multilingual text categorization method using the latent semantic indexing technique. This method consists of performing multiple monolingual approaches on English and Chinese datasets. In another work, Prajapati et al. (2009) introduced an approach relying on the translation of documents to universal language and then performed the classification. They incorporated the knowledge using WordNET to map terms to concepts then classify text using linear classifier Rocchio and probabilistic Naïve Bayes and K-Nearest Neighbor (KNN). Amini et al. (2010) investigated MTC by combining two semi-supervised learning techniques, including co-regularization and consensus-based self-training. They trained different monolingual classifiers on the Reuters Corpus Volume 1 and 2 (RCV1/RCV2) containing five different languages: English, German, French, Italian, and Spanish. The authors validated their method using six classification methods: Boost, co-regularized boosting, boosting with self-training, Support Vector Machine (SVM) with self-training, co-regularization + self-training, and boosting with full self-training. Bentaallah and Malki (2014) compared two WordNet-based approaches for multilingual text categorization. The first relied on machine translation to directly access WordNet and used a disambiguation strategy to consider only the most common meaning of the term. While the second excluded the translation and explored the WordNet associated with each language. Mittal and Dhyani (2015) addressed the multilingual text classification based on N-gram techniques. They studied MTC in Spanish, Italian, and English languages. They proceeded by predicting the language of a document and used Naïve Bayes in the classification phase. More recently, Kapila and Satvika (2016) addressed the problem of MTC on Hindi and English languages using different machine learning algorithms, including SVM, KNN, Decision Tree, Self-Organizing Map, and Genetic Algorithms. They improved the method accuracy by employing various feature selection methods.

Recently, deep neural networks and contextual embeddings were proposed in the text classification domain for English (Liu and Guo, 2019, etc.).

In short, despite the considerable amount of work on cross-lingual text classification, the MTC is almost neglected and few studies were proposed using classical techniques such as SVM and KNN. In

ISSN: 2366-1313

addition, research on the offensive language detection field had been prospected only from a monolingual perspective. Other studies investigated recent techniques as mBERT but only in the cross-lingual area. A new approach is therefore needed to study the MTC in the offensive language detection field from deep learning aspects using promising transfer learning techniques as BERT.

## III. RESEARCH METHODOLOGY

In this study, we propose a modular text classification pipeline consisting of modular cleaning phase and tokenizer, three embedding methods, and eight classifiers. The experiment done in this study is based on Twitter, and a dataset was optimized effectively. Although we do not claim that our framework would perform well on all social media platforms, it could provide future research direction to guide academic and industry researchers. The broader impact of this paper can be related to the systematically investigation of detecting online harassment on social media platforms. Moreover, due to social media platforms' inherent features, it is impossible to generalize a model for all the platforms. For example, shows that training a classifier on Reddit is more challenging than Gab because the average length of posts.

This section briefly discusses the steps taken for cleaning and preparing the dataset and also con ducting the experiments. Furthermore, Fig. 1 shows a graphical overview of these steps, discussed in the following.

### A) Data Preparation

Data Preparation is the first step for training bi nary classifiers. The strategies for data preparation, which need to be carefully conducted, are described as below:

- **Basic cleaning methods**: We need to clean the data as (i) extracting the pure text from the dataset, removing duplicates, and NaNs (ii) transforming to lowercase (iii) expanding the abbreviations.

- **Slangs:** Given the micro-blogging style of Twitter, using slangs are typical. Slangs bring difficulties to text mining approaches, especially for those emerging lately and thus do not have an updated entry in any dictionaries. So, we plan to transform the text into a canonical form using the reference dictionary1 for slangs and abbreviations.

- **Removing methods**: Using hashtags, user references, links, and emojis are typical on social media platforms. Therefore, preprocessing the data and selectively removing the typical pat terns are essential to normalize the text.

ISSN: 2366-1313

• **TF-IDF:** One way to represent words into vectors is to count the occurrence of words seen in the whole documents. One caveat of this method is the overemphasizing the frequent words in the dataset. In contrast with the word counting method, TF-IDF distributes the weight of frequent words by their relative frequency.

• **Word2Vec:** The word2vec method takes a corpus of text as input and returns word vectors as output. There are two model architectures to produce a distributed representation of words. The continuous bag-of-words (CBOW) architecture predicts the current word based on the context(window size), and the Skip-gram predicts surrounding words(defined window) given the current word.

• **FastText:** FastText represents a low dimensional vector text that is generated by summing vectors corresponding to the words in the text. Neural Network is being used in FastText for word embedding. FastText model is often compared to other deep learning classifiers with a higher speed and accuracy for training and evaluation.

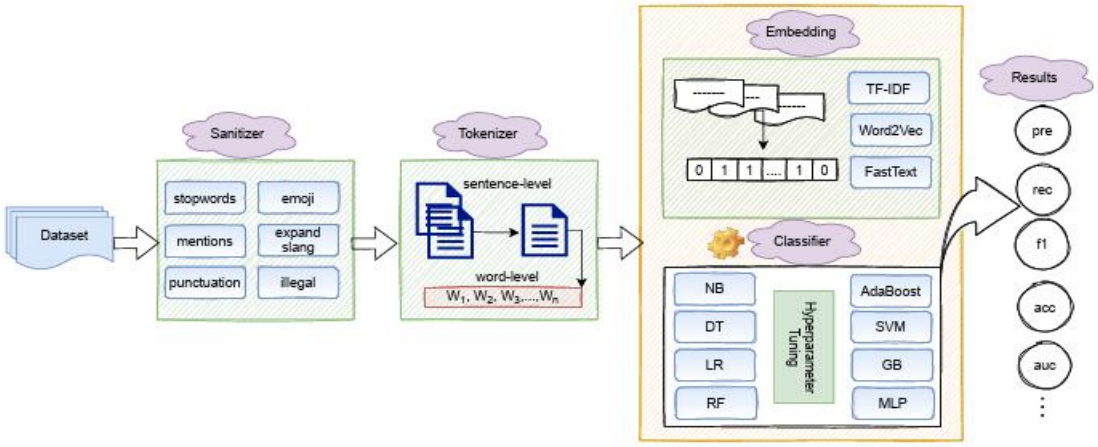


Fig.1 The modular experimental setting with the flow of data from dataset to results.

### B) Using Text Mining Techniques to Detect Online Offensive Contents

Offensive language identification in social media is a difficult task because the textual contents in such environment is often unstructured, informal, and even misspelled. While defensive methods adopted by current social media are not sufficient, researchers have studied intelligent ways to identify offensive contents using text mining approach. Implementing text mining techniques to

ISSN: 2366-1313

analyze online data requires the following phases: 1) data acquisition and preprocess, 2) feature extraction, and 3) classification. The major challenges of using text mining to detect offensive contents lie on the feature selection phrase, which will be elaborated in the following sections.

### C) Message-level Feature Extraction

Most offensive content detection research extracts two kinds of features: lexical and syntactic features. Lexical features treat each word and phrase as an entity. Word patterns such as appearance of certain keywords and their frequencies are often used to represent the language model. Early research used Bag-of-Words (BoW) in offensiveness detection. The BoW approach treats a text as an unordered collection of words and disregards the syntactic and semantic information. However, using BoW approach alone not only yields low accuracy in subtle offensive language detection, but also brings in a high false positive rate especially during heated arguments, defensive reactions to others' offensive posts, and even conversations between close friends. N-gram approach is considered as an improved approach in that it brings words nearby context information into consideration to detect offensive contents. N-grams represent subsequences of N continuous words in texts. Bi-gram and Tri-gram are the most popular N grams used in text mining. However, N-gram suffers from difficulty in exploring related words separated by long distances in texts. Simply increasing N can alleviate the problem but will slow down system processing speed and bring in more false positives. Syntactic features: Although lexical features perform well in detecting offensive entities, without considering the syntactical structure of the whole sentence, they fail to distinguish sentences' offensiveness which contain same words but in different orders. Therefore, to consider syntactical features in sentences, natural language parsers are introduced to parse sentences on grammatical structures before feature selection. Equipping with a parser can help avoid selecting un-related word sets as features in offensiveness detection

### D) User-level Offensiveness Detection

Most contemporary research on detecting online offensive languages only focus on sentence-level and message-level constructs. Since no detection technique is 100% accurate, if users keep connecting with the sources of offensive contents (e.g., online users or websites), they are at high risk of continuously exposure to offensive contents. However, user-level detection is a more challenging task and studies associated with the user level of analysis

ISSN: 2366-1313

are largely missing. There are some limited efforts at the user level.

### E) Machine learning algorithms

NaiveBayes (NB) and SVM—are used to perform the classification, and 10-fold cross validation was conducted in this experiment. To fully evaluate the effectiveness of users' sentence offensiveness value (LSF), style features, structure features and content specific features for user offensiveness estimation, we fed them sequentially into the classifiers, and get the result in Fig.3. The "Strong+Weak" means simply uses offensive words as the base feature to detect offensive user. Similarly, "LSF" means the sentence offensiveness value generated by LSF is used as the base feature.

## IV. EXPERIMENTAL RESULTS

This section describes several experiments we conducted to examine LSF on detecting offensiveness languages in social media. Dataset Description The experimental dataset, retrieved from Youtube comment boards, is a selection of text comments from postings in reaction to the top 18 videos. Classification of the videos includes thirteen categories: Music, Autos, Comedies, Educations, Entertainments, Films, Gaming, Style, News, Nonprofits, Animals, Sciences, and Sports. Each text comment includes a user id, a timestamp and text content. The user id identifies the author who posted the comment, the timestamp records when the comment was posted and the text content contained a user's comments. The dataset includes comments from 2,175,474 distinct users.

Pre-processing Before feeding the dataset to the classifier, an automatic pre-processing procedure assembles the comments for each user and chunks them into sentences. For each sentence in the sample dataset, an automatic spelling and grammar correction process precedes introduction of the sample dataset to the classifier. With the help of WordNet corpus and spell-correction algorithm2, correction of spelling and grammar mistakes in the raw sentences occurs by tasks such as deleting repeated letters in words, deleting meaningless symbols, splitting long words, transposing substituted letters, and replacing the incorrect and missing letters in words. As a result, words missing letters, such as "speling," are corrected to "spelling"; misspelled words, such as "korrect," change to "correct." Experiment Settings in Sentence Offensive Prediction The experiment compares six approaches in sentence offensive prediction: a) Bag-of-words (BoW): The BoW approach disregards grammar and word order and detects offensive sentences by checking

ISSN: 2366-1313

whether or not they contain both user identifiers and offensive words. This approach also acts as a benchmark. b) 2-gram: The N-gram approach detects offensive sentences by selecting all sequences of n words in a given sentence and checking whether or not the sequences include both user identifiers and offensive words. In this approach, N equals to 2, it also acts as a benchmark. c) 3-gram: N-gram approach, selecting all sequences of 3 words in a given sentence. It also acts as a benchmark. d) 5-gram: N-gram approach, selecting all sequences of 5 words in a given sentence. It also acts as a benchmark.

## V. Evaluation Metrics

In our experiments, standard evaluation metrics for classification in sentiment analysis (i.e., precision, recall, and f-score) are used to evaluate the performance of LSF. In particular, precision presents the percent of identified posts that are truly offensive messages. Recall measures the overall classification correctness, which represents the percent of actual offensive messages posts that are correctly identified. False positive (FP) rate represents the percent of identified posts that are not truly offensive messages. False negative (FN) rate represents the percent of actual offensive messages posts that are unidentified. F-score represents the weighted harmonic mean of precision and recall, which is defined as:

$$f-score=\frac{2(precision\times recall)}{precision+recall}$$

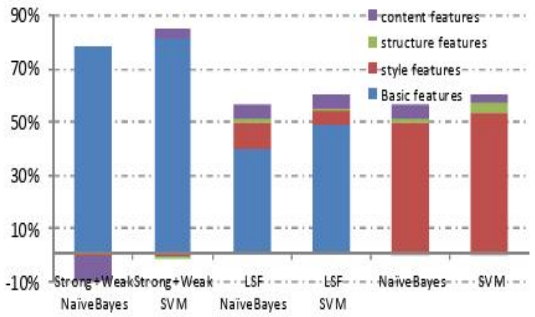


Fig.2 F-score for different feature sets using NB and SVM

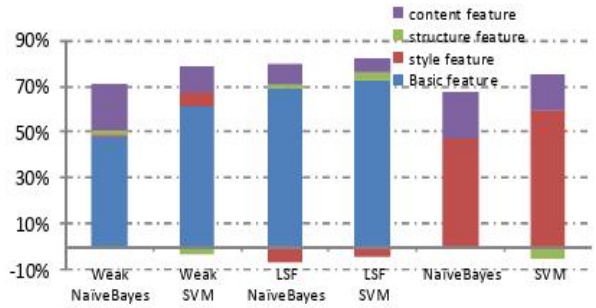


Fig.3 F-score for different feature sets using NB and SVM (without strongly offensive words)

## V. CONCLUSION

In this study, we investigate existing text-mining methods in detecting offensive contents for protecting adolescent online safety. In this work, we propose a modular text classification pipeline on social media datasets focusing on Twitter. Our proposed approach is to leverage a modular

ISSN: 2366-1313

development that allows easy use for combining different text classification components. This paper's main contribution is that it presents a new modular text classification pipeline to facilitate bench marking by conducting a detailed analytical study of the best-performing approaches, features, and embeddings reported by the state-of-the-art.

## REFERENCES


1. P. Hajibabaee, F. Pourkamali-Anaraki, and M. Hariri Ardebili, "An empirical evaluation of the t-sne algorithm for data visualization in structural engineering," in 2021 IEEE International Conference on Machine Learning and Applications. IEEE, 2021.
2. S. Zad, M. Heidari, J. H. J. Jones, and O. Uzuner, "Emotion detection of textual data: An interdisciplinary survey," in 2021 IEEE World AI IoT Congress (AIIoT), 2021, pp. 0255–0261.
3. S. Zad, M. Heidari, J. H. Jones, and O. Uzuner, "A survey on concept-level sentiment analysis techniques of textual data," in 2021 IEEE World AI IoT Congress (AIIoT), 2021, pp. 0285–0291.
4. M. Heidari, S. Zad, B. Berlin, and S. Rafatirad, "Ontology creation model based on attention mechanism for a spe cific business domain," in 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp. 1–5.
5. M. Heidari, S. Zad, and S. Rafatirad, "Ensemble of super vised and unsupervised learning models to predict a prof itable business decision," in 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp.
6. A. Esmaeilzadeh, M. Heidari, R. Abdolazimi, P. Ha jibabaee, and M. Malekzadeh, "Efficient large scale nlp feature engineering with apache spark," in 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2022.
7. R. Abdolazimi, M. Heidari, A. Esmaeilzadeh, and H. Naderi, "Mapreduce preprocess of big graphs for rapid connected components detection," in 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2022.
8. M. Malekzadeh, P. Hajibabaee, M. Heidari, and B. Berlin, "Review of deep learning methods for automated sleep staging," in 2022 IEEE 12th Annual Computing and Com munication Workshop and Conference (CCWC). IEEE, 2022.

ISSN: 2366-1313

9. A. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, "Offensive language detection using multi-level classification," Advances in Artificial Intelligence, vol. 6085/2010, pp. 16-27, 2010.
10. Mahmud, Ahmed, Kazi Zubair, and Khan, Mumit "Detecting flames and insults in text," in Proc. of 6th International Conference on Natural Language Processing (ICON' 08), 2008.
11. D. Yin, Z. Xue, L. Hong, and B. Davison, "Detection of harassment on Web 2.0," in the Content Analysis in the Web 2.0 Workshop, 2009.
12. Z. Xu and S. Zhu, "Filtering offensive language in online communities using grammatical relations," in Proceedings of The Seventh Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS'10), 2010