Republic of Iraq
Ministry of Higher Education
and Scientific Research
University of Diyala
College of Sciences
Department of Computer
Science

Identifying Offensive Posts in Social Media using SVM and CNN

A Thesis

Submitted to the Department of Computer Science/ College of Science/ University of Diyala

In Partial Fulfillment of the Requirements for the Degree of MASTER in Computer Science

By

Waleed Molan Salih

Supervised By

Professor

Naji M. Sahib

2021 AD ***IRAQ*** ***1442 AH***

بسم الله الرحمن الرحيم

﴿ نَرْفَعُ دَرَجَاتٍ مَّن نَّشَاءُ وَفَوْقَ كُلِّ ذِي عِلْمٍ عَلِيمٌ ﴾

صدق الله العظيم

سورة يوسف

الاية (76)

Acknowledgments

*First of all, praise is to **Allah** the lord of the whole creation, on all the blessing was the help in achieving this research to its end.*

*I wish to express my thanks to my supervisors **professor Naji Mutter Sahib** for supervising this research and for the generosity, patience, and continuous guidance throughout the work. It has been my good fortune to have the advice and guidance from him. My thanks to the academic and administrative staff at the Department of the computer sciences\University of Diyala for their hospitality and generosity.*

I would like to express my gratitude to my family for their great support and continuous encouragement.

Waleed Molan Salih

Dedication

To...
The soul of my father and my brothers (Muhammad and Saadoun), may God have mercy on them

My dear mother

my family

my sincere thanks and gratitude to Professor.Naji M. Sahib for his good guidance and valuable instructions, as he had a great impact on the research reaching this image
All our distinguished teachers those who paved the way for our science and knowledge
Those who taught me how to stand firmly on the ground
To all of those who have received advice and support
I present to you a summary of my scientific efforts

Waleed Molan Salih

Abstract

Social media has become a part of our lives. This platform is used by billions of users as a communication device and as a data source in real-time and it has become huge in people's popularity. Online Social Networking (OSN) such as Twitter, Facebook, and Instagram are the most effective venues for free expressions of people of all ages.

With all of the easy-to-use and benefits technology that have emerged, there have also been negative consequences. Cybercriminals make use of this information and utilize social media to perform various types of cybercrimes, such as cyberbullying. Cyberbullying is a type of harassment carried out using digital technology. It's a global issue that's just getting worse. If a text has racist slurs, assaults or condemns any religious or community position, or stimulates criminal activity, it is considered threatening or abusive.

As manual filters take some time and can lead to human annotators suffering from post-traumatic stress disorders, a lot of research has been done to automate the process. The work is frequently modeled on a supervised classification issue in which algorithms are trained on posts that are noticed in about offensive or abuse content.

In the proposed work, the main focus was on examining the effects of private regulation on hate speech on social media by using a variety of algorithms to achieve this goal, including Random Forest (RF), Support Vector Machine (SVM), Ada boost, Bagging, and Convolution Neural Network (CNN). The five different algorithms on two datasets (Twitter) are applied to detect hate speech and compare their accuracy.

The best accuracy value overall classification algorithms were obtained with new dataset and Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction methods, the best accuracy value of the SVM classifier =0.9648, and for the RF classifier is 0.9598, and the best accuracy value of the AdaBoost classifier is 0.9476, and for the Bagging classifier is 0.9534, and finally, best accuracy value of the CNN classifier is 0.9572.

And finally, in the new dataset, the best results were obtained for the SVM algorithm, with an accuracy of 0.9648, and in the old dataset, the best results for the Bagging algorithm were obtained with an accuracy of 0.9224.

Contents

List of Figures

List of Tables

List of Abbreviatiations

Abbreviations	Meaning
Adaboost	Adaptive Boosting
AUROC	Area Under the Receiver Operating Characteristic Curve
Bagging	Bootstrap Aggregation, is a simple and very powerful ensemble method.
BiRNN	Bidirectional Recurrent Neural Networks
C-LSTM	Convolutional-Long Short Term Memory
CNN	Convolution Neural Network
CSV	comma-separated values
DCNN	Deep Convolution Neural Network
Doc2Vec	A set of language modeling and feature learning

	techniques called word embeddings became increasingly popular for NLP tasks
DS	Decision Stumps
GRU	Gated Recurrent Unit
HS	Hate Speech
IOTLML	Identify Offensive Tweet Language based on Machine Learning
KNN	k-Nearest Neighbors
LSTM	Long Short Term Memory
ML	Machine Learning
NHS	Non-Hate Speech
NLP	Natural Language Processing
OOB	Out-Of-Band
OSH	Optimal Separating Hyperplane
OSN	Online Social Networking
RF	Random Forest
RNN	Recurrent Neural Network
SVM	Suport Vector Machine
T	Threshold
TF-IDF	Term Frequency–Inverse Document Frequency
Twitter API	Twitter Application Programming Interface
Word2Vec	is an algorithm used to produce distributed representations of words

Chapter One

General Introduction

Chapter one

General Introduction

1.1 Introduction

The Internet era has had a huge impact on the world since it has brought people together from all over the world. Now, without ever meeting them personally, people can talk and establish connections with others. Social media has become a part of our lives [1].

Social media is an ideal platform for exchanging words and ideas as well as disseminating the most up-to-date information. Recent news can be obtained at a breakneck speed and in the blink of an eye [2]. With all of the benefits and easy-to-use technology that have emerged, there have also been negative consequences. Cybercriminals make use of this information and utilize social media to perform various types of cybercrimes, such as cyberbullying. Cyberbullying is a type of harassment carried out using digital technology. It's a global issue that's just getting worse. If a text has racist slurs, assaults or condemns any religious or community position, or stimulates criminal activity, it is considered threatening or abusive [3].

It has been established that victims of cyberbullying become dangerously afraid and may have violent revenge fantasies or even suicidal thoughts as a result of the bullying. They are depressed, have low self-esteem, and are anxious. Cyberbullying is worse than physical bullying because it occurs behind the scenes and at all hours of the day and night. Even the bully's tweets or comments don't go away; they linger with the victim for a long time and have a mental impact on them. It's almost like ragging, but it happens in front of tens of thousands of mutual friends, and the scars last as long as the messages do. The victims are humiliated to an unimaginable degree as a result of the nasty and humiliating messages. The necessity for analysis to provide findings that empower victims, strengthen

public campaigns, and discourage abusers is highlighted by Simons [4]. Likewise, the employment of remedial measures to prevent online harassment against Muslims or women by social network operators, such as Twitter or Facebook urged by Barlow and Awan [5].

As manual filters take some time and can lead to human annotators suffering from post-traumatic stress disorders, a lot of research has been done to automate the process. The work is frequently modeled on a supervised classification issue in which algorithms are trained on posts that are noticed about to offensive or abuse content [6].

In this study, Random Forest, SVM, Ada boost, Bagging, and CNN methods are being used for detecting hate speech. Random forest (RF) classification is a group algorithm in which each group tree consists of a sample of replacements from the training pack (i.e. a bootstrap sample). Instead of allowing each classification vote for one class, the Random Forest Classifier combines classification by averaging its probabilistic prediction [7]. SVMs are supervised learning algorithms for regression and classification that operate well in high-dimensional domains. The text classification tasks SVM classifications demonstrate good performance [8]. Robert Schapire and Yoav Freund proposed the Ada-boost, as one of the collaborative boosting classifiers in 1996. It combines many classifiers to improve the accuracy of the classifier. AdaBoost is an iterative ensembles approach. The AdaBoost classification generates a powerful classification by merging many low-performance classification systems, leading to a very precise classification [9]. Bagging is an ensemble method that trains each classifier for a random redistribution of the training set, each training set of the classification is formed by random drawing, replacement N examples — where N is the size of the first training set; in the resulting training set many of the original instances can be repeated, while others can remain. Every individual classifier in the group is produced with a separate random

selection of the training set [10]. One of the biggest issues in natural language processing was text classification. The development of profound learning is a prominent option for the neural network (CNN). The initially proposed CNNs for images nonetheless suffer several of critical challenges in text handling [11].

1.2 Related Works

Many kinds of researches and studies are dealing with the detection and hate speech and abusive languages, among which we include the following:

- **Del Vigna et al. 2017** [12] attempted to halt and curtail the worrying spread of such hate campaigns. They evaluate the content of the linguistic comments that occurred on a collection of pages of public Italian, using Facebook as a benchmark. To separate the types of hate, they first suggest a range of hate categories. Following that, up to five different human annotators annotate the crawled comments according to the established taxonomy. They design and implement two classifiers for the Italian language, based on distinct learning algorithms, using sentiment polarity, word embedding lexicons, and morpho-syntactical characteristics. The first is relying on Support Vector Machines (SVM), while the second is based on a specific Recurrent Neural Network termed Long Short Term Memory (LSTM). On the job of hate speech recognition, these two learning algorithms were put to the test to see how well they classified. The results demonstrate that the two classification approaches tested in the first Italian Hate Speech Corpus handwritten text in social media are efficient. Obtain results of F-score of about 0.72.

- **Albadi et al. 2018** [13] Proposed a system to handle the problem of recognizing discourse on Arabic Twitter that promotes religious intolerance. Describe how the initial public Arabic dataset annotated for the goal of the detection of religious hate and the first Arabic vocabulary consisting of phrases often used in religious discourses and values that signify their polarity and strength was produced. Then, using deep learning, n-gram, and lexicon-based approaches, created several categorization models. Following that, a detailed comparison of the performance of various models on a newly unseen dataset is presented. They conclude that with pre-trained word embeddings and simplistic recurrent neural network (RNN) architecture of Gated recurrent unit (GRU)scan adequately detect religious hate speech with AUROC of 0.84.

- **Ruwandika et al. 2018** [14] To complete the objective of automatically identified hate speech, supervised and unsupervised machine learning algorithms of five models were created and utilizing. A local English text collection was used for the experiment. For the purposes of this experiment, hate speech is defined as the use of language to insult or spread hatred toward a group or individual based on social status, gender, race, or religion. The task of hate speech identification was then compared using both supervised and unsupervised learning algorithms with various feature types. With an F-score of 0.719, the Naive Bayes classifier with Tf-idf features outperformed all other supervised and unsupervised models. The KMeans clustering model, out of all five, performed the worst in practically every case. This could explain why the problem of

identifying online hate speech has been framed as a supervised learning activity.

- **L. Jiang et al. 2019** [15] According to multiple methods, they tried to find out which method has the best accuracy of detecting hate speech from tweets. The major innovation of this article is that they used different ratios of data to compare with multiple methods at the same time. They used the Dataset named Hate speech dataset published on Kaggle tiled. For this Dataset, two CSV files are present in the downloadable folder referring to the training and testing set respectively. In this case, roughly less than 10,000 unique labeled values (tweets data) are present. To add useful information to their model, they append it to the end of the other dataset. And I always used the other Dataset called Twitter Sentiment Analysis datasets published on Kaggle. For this Dataset, there are also has two files present in the downloadable folder referring to the training and testing set. All of the datasets have the labeled dataset of not racist/sexist and racist/sexist even not so more. As a result, good performance is obtained by using machine learning when data is small. Good results can be obtained by using deep learning when they used more data for the experiments. Using BiRNN can get the best results, compared with other methods they used. Even if this method is superior to other models, they have to consider the type of data set in the future. The results show the good performance is obtained by using when data is small, and good results can be obtained by using deep learning when we use more data for our experiments.

- **Sandaruwan et al. 2019** [16] they Proposed machine learning-based and lexicon-based techniques for automatical detection of offensive

speeches and Sinhala hate shared on social media. With the lexicon generation process, the lexicon-based methodology was launched and the corpus-based lexicon provided 0.763 accuracy for offensive, hate, and neutral speech identification. A 3000 commentary body has been developed, which is equally spread across hate, offenses, and neutrality. The approach to machine learning has begun. Using this corpus of remark, feature groups and models for Sinhala hate speech detection might be identified in the best way. The highest recall value was 0.84 with an accuracy of 0.9233 using Multinomial Nave Bayes.

- **R. Shah et al. 2020** [17] study aims to put forward ideas regarding cyber-bullying detection on the social media platform Twitter. Their work involved finding the best approach and best classifier which will accurately detect bully tweets. Pre-processing of data has two steps: Collection of data and Cleaning of data. The very first and basic step is the collection of data that is done in two ways. The Twitter API was accessed and tweets were extracted, the rest of the tweets were obtained from the Kaggle dataset. The dataset was divided into training and testing data. The tweets of the training data were labeled by the values 0 and 1. The bully tweets were represented by value 1 and the non-bully tweets were represented by value 0. The test data was not labeled. The next step was cleaning the data. The outcome of this study is that whichever tweet is a bully tweet is represented by the value 1, thus all the bully tweets are detected. The Twitter dataset is equally distributed into a bully and non-bully tweets and fed to different machine learning models. The logistic regression classifier provides an accurate classification of the bully and non-bully tweets with the precision of 0.91, recall 0.94, and F1-score 0.93. This work

will help curb cyber-bullying so that the users can stay at bay from victimization.

- **P. K. Roy et al. 2020** [18] proposed an automated system is developed using the Deep Convolutional Neural Network (DCNN). The dataset used for this study is taken from Kaggle.com. It was prepared by collecting the tweets from Twitter. The dataset description was missing on the uploaded webpage however, by manual inspection during the research they found that the dataset contained English written tweets only. The other tweets languages were not considered for this case. The developed dataset contained a total of 31,962 English written tweets, of which 29,720 tweets (0.9298) are Non-Hate Speech (NHS) and the remaining 2,242 tweets (0.702) are Hate Speech (HS) related tweets. The proposed DCNN model utilizes the tweet text with GloVe embedding vector to capture the tweets' semantics with the help of convolution operation and achieved the precision, recall, and F1-score value as 0.97, 0.88, 0.92 respectively for the best case and outperformed the existing models. They also tested other deep neural network-based models such as Long Short-Term Memory (LSTM), and Convolutional-LSTM (C-LSTM) network for the same and found the DCNN model is a better choice for this research.

- **T. T. Han et al. 2020** [19] they tried to the problem of spreading hate speech over the social network by autonomously detection the posts/tweets of the network users. They performed the pre-processing of language context using NLP tools and then exploit a deep learning model called bidirectional recurrent neural network (Bi-RNN) to detect if the tweets are vulnerable to hate speech or not. The system is

then implemented according to the proposed architecture and tested with the popular Twitter dataset for analysis of hate speech. The experimental works are executed and measured with evaluation metrics called precision, error rate, and processing time. The proposed Bi-RNN model Precision results as follow Error, Rate, Processing Time for training 0.93, 0.16, 450 seconds, and for Testing 0.91, 0.19, 78 seconds respectively.

- **Senarath et al. 2020** [20] This article gives a new empirical investigation with different semantic characteristics on social media tasks in hate speech classification. In particular, they give broad empirical analyses where they examine the characteristics of the corpus-based semantic vector space model representation, the neural word embedding for distributional semantics, and declarative knowledge patterns from the external domain semantics knowledge base. In contrast to the situation of a single type of feature representation, their experimental results suggest that combining varied feature representations improves the effectiveness of hateful behavior classification. Results of two major Twitter datasets for the detection of hate speech demonstrated a constant performance boost for the classification models based on hybrid characteristics (F1 score gain up to 0.30). The implementation of the proposed method used to combine various representations of the features to help the improve ment of the monitoring systems to human behavior.

- **Mubarak et al. 2020** [21] They introduced a strategy for creating an offending dataset that is free of the subject, dialect, or target bias. The largest Arabic dataset with particular tags for vulgarity and hate speech has been created. They examined the data to see which

themes, dialects, and gender are most related to offensive tweets, as well as how Arabic speakers use offensive language. They created precise standards for categorizing tweets as clean or offensive, including special tags for obscene tweets and hate speech. 10,000 tweets have been tagged Finally, using SVM algorithms, run a huge battery of tests to get strong results (F1=0.797) on the dataset.

1.3 Problem Statement

Hate speech has been available for a very long time in many communities. Some people believe racism should be suppressed in democratic society. The problem arises when an individual or group of people use terms that they believe to be protected racist language and some people hear and comprehend the same words believes it is hate speech. One effective strategy to avoid such scenarios is definitely to be wary of tweeting and having a possibly objectionable content control mechanism. This project aims to put forward ideas regarding hate speech detection on the social media platform twitter. By looking at most of the previous studies, several studies have been focused on detecting offensive language in social media. However, some of these studies have data volume problems, and others an accuracy need to be improved.

1.4 Aim of The Thesis

The proposed identification of the offensive posts in the social media system has several goals to achieve them, as follows:

1. Proposed an Automatic System for Identify Offensive Tweet Language based on Machine Learning and Deep CNN (IOTLML - DCNN) to classify offensive Twitter messages automatically in two classes: offensive and clean with a high level of precision.

2. Using two Twitter databases and getting important features from them to denotes offensive words, based on two feature extraction methods (count vectorizer and term frequency-inverse document frequency (TF-IDF))
3. Employ four of the most popular machine learning algorithms in the suggestion system which are SVM, Random Forest (RF), Ada Boost, and Bagging classifier
4. Conducting an investigation study to compare the results of the proposed system obtained from machine learning and deep learning.

1.5 Thesis Organization

Besides this chapter, the remaining parts of this thesis include the following chapters:

Chapter Two: Theoretical Background

In this chapter, the linguistic aspect of the tools, techniques, and algorithms that will be applied in designing and implementing our system is presented.

Chapter Three: **The Proposed System**

The suggested method for detecting hate speech and abusive languages is described in this chapter, along with a full discussion of the tools and techniques utilized in classification.

Chapter Four: Experimental Results and Evaluation

The findings acquired from the installation of the suggested system are presented in this chapter, as well as the analysis and discussion of the results, their testing, and comparisons with previous studies.

Chapter Five: Conclusions and Suggestions for Future work

In this chapter, a set of conclusions obtained from the design and implementation of the proposed system are presented.