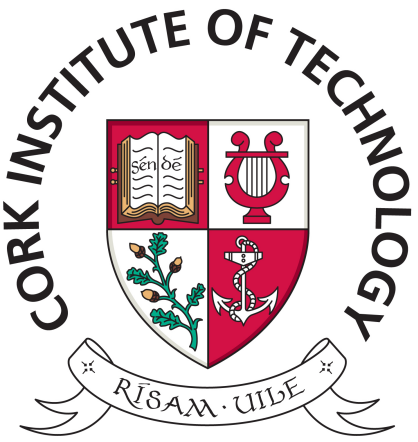


# Detection of Offensive Language in Social Media Posts

by
Sidharth Mehra

This Interim Report has been submitted in partial fulfillment for the module AI Research Project

in the
Faculty of Engineering and Science
Department of Computer Science

March 8, 2020

# Declaration of Authorship

I, Sidharth Mehra, declare that this thesis titled, Detection of Offensive Language in Social Media Posts and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for an masters degree at Cork Institute of Technology.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at Cork Institiute of Technology or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this project report is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- I understand that my project documentation may be stored in the library at CIT, and may be referenced by others in the future.

Signed: _____

Date: _____

# *Abstract*

Posting offensive or abusive content on the social media have been a serious concern in the recent years. This has created a lot of problems because of the huge popularity and usage of the social media sites like Facebook and Twitter. The scope of our work lies in predicting whether the posted content is offensive or not. We would be using various Machine learning algorithms like Logistic Regression, Support Vector Machines, Naive Bayes and Random Forest along with the Deep learning architectures like CNN, BiLSTM, RNN and BERT to develop our predictive model. The main motivation lies in the fact that our model will automate the quick detection of the posted offensive content so as to facilitate the relevant actions and moderation on these offensive posts. We would be using the publicly available benchmark dataset OLID (Offensive Language Identification Dataset) [15] and performing the comparative analysis of various Machine learning algorithms. Main focus is to also predict the type and the target of the offensive content which has not been studied much in the previous work. We would at the same time select the various text pre-processing and vectorization techniques to eventually develop a strong model by extracting the important features. Moreover the top performing machine learning models will undergo hyper-parameter optimization to further escalate the model detection performance. To evaluate the performance of our model, we will be using per class precision, recall and F1-score along with the macro average. Finally a real time system could be deployed on various social media platforms to detect and analyze the offensive post content and taking the appropriate action in order to normalize the behaviour on these sites and the society.

# *Acknowledgements*

I would like to thank my professor Dr.Mohammed Hasanuzzaman, for guiding me through out this project phase of the preparation of this Interim Report and also for providing the extreme valuable suggestions to proceed further in this AI research project.

# Contents

# Chapter 1

# Introduction

This introduction chapter mainly focuses on three things. Firstly we will discuss about the motivation behind choosing this particular topic for our AI research project. Secondly we give a brief idea of our research objectives explaining what we are going to contribute to this particular problem area. Then at the last we present the structure of this interim report representing what content is there in which chapter.

Increase in the usage of social media sites likes Facebook and Twitter have given the crowd a great platform to express their opinions/feelings for the individual, groups or events happening around them or in society. This digital media has become a great resource to share the information and also gives the full freedom of speech to everyone on the platform. This feature of the social media to express something openly to the world have created the major problems for these online businesses and negatively impacted the well being of the societal decorum. With the gaining popularity of these platforms; there also comes the negative part along with its benefits. There are increasing cases of the abuse or offense on the social media like Hate speech, Cyber-bullying, Aggression or general Profanity. It is very much important to understand that this behaviour can not only immensely affect the life of an individual or a group but could be suicidal in some cases; adversely hampering the mental health of the victim/s. This increasing negative situation on the internet has created a huge demand for these social media platforms to automate the task of detecting the objectionable content for the human moderation. This effort by the numerous studies as evident by the chapter 2 of

this report has taken the support the Artificial Intelligence, Machine Learning and Natural Language Processing to develop the systems that are able to detect these types of offensive content so that the appropriate action could be taken as quickly as possible.

The research objective of this work is to develop an accurate system that is able to distinguish the offensive content from the clean or non-offensive ones. For this purpose we would be using the recently created novel dataset OLID which is discussed in great detail in chapter 3. We aim to apply various Machine Learning and Deep Learning techniques with the different combination of the text featurization methods as there is not much work published on this dataset. Hyper-parameter optimization and dealing with class imbalance issue would be also the major focus of this study to escalate the model detection accuracy on the unseen tweets.

This document is **structured** as follows: **Chapter 2** presents the existing solutions as well as the background information available on the offensive language detection. **Chapter 3** presents the proposed work flow towards predicting the nature of the social media posts; the algorithms to be employed and the process used for training and tuning the best Machine learning models.**Chapter 4** closes the document with a summary.

# Chapter 2

# Literature Review

In this section we would be critically analysing the few academic papers that revolve around the detection of offensive language in social media using machine learning. We would be extracting the summaries of these papers along with their main characteristics which are defined with the proper structure described in the later part of this chapter. We will also determine the common and different points between these papers which will help us to extract the structure and the context of the problem solution. After analysing these papers that are formulated according to the below structure, we will develop a similar structure to illustrate the context of our problem solution. We will also compare the existing methods and try to identify that is there any clear best solution or are there any trade-offs. The main characteristics of these papers are structured as the individual sections of this chapter following a chronological order.

## 2.1 Motivation and Challenges

The motivation, challenges and the direction of mitigation are mentioned below for each of the related works

- Paper [4] was one of the first significant works that was related with detecting cyber-bullying on the social media with the help of NLP. Earlier this problem was studied mostly by the psychiatrists and social scientists. Authors believed that anonymity and lack of meaningful supervision has put out this topic into more attention to computational linguists. They moreover believed that the appropriate action on this content through NLP could

prevent the tragic outcomes of bullying on social media platforms. Authors modeled this problem into 2 parts. First one is to determine whether the given comment is sensitive or not making it a binary classification task. And if the comment is sensitive then classify it as the comment revolving around one of the areas in sexuality, race  culture or intelligence which made this as multi-class classification subtask.

- The very strong baseline paper [14] aimed to tackle the serious issue of cyber-bullying on social media platforms using NLP techniques so that the further investigation on the same domain could be explored in future by the NLP community with the help of their baseline results. They moreover formulated the bullying on the social platforms into 4 major NLP tasks namely text classification, role labeling, sentiment analysis and topic modeling. They defined the bullying traces as the posts by the individuals who have been a victim of bullying. Objective of subtask A was to distinguish the bullying episodes from the non-bullying traces in the dataset examples. The subtask B dealt with the role labelling of the tweets into Author's role and Person mention role which can be thought of a binary text classification task. The other thing in this subtask is to further sequentially tag these roles into one of the 5 categories which are Accuser (A), bully (B), reporter (R), victim (V) and other (0). Subtask C was related with classifying the sentiment of the bullying episode to understand their motivation. This task was a binary task of classifying the tweet as teasing or not. Subtask D was related with latent content modelling which extracts the main topics to better understand the bullying traces.

- The authors in [2] realized the need of automating and improving the task of detecting the bullying content on the social media. Authors clearly outlined that the previous work has only investigated the bullying detection only from individual comments. They strongly felt that taking user characteristics and profile information could definitely yield in better model for characterizing the sensitive content.

- Authors in [8] believed that they were among the first one to investigate the topic of hate speech detection on social media particularly when it comes to racist comments which were proliferating hugely on these online platforms

because of freedom of speech and the features of re-tweeting. They analysed the racism against the black as they constituted 25% of the whole twitter population at that time and they were the highest one to face the racism on the platform. They formulated the problem as the binary classification task to determine whether the tweet is racist or nonracist.

- The motivation of authors in the work [1] was triggered by the murder of Drummer Lee Rigby in Woolwich, London, UK. They observed and believed that immediately after such events there is a potential opportunity of spreading the hate speech on online platforms. They formulated this problem as the binary classification task of predicting whether the tweet is hateful/ antagonistic or not with a focus on race, ethnicity or religion. They derived their features from the context of each tweet and experimented with various probabilistic, rule-based, and spatial-based classifiers with a voted ensemble meta-classifier. They believe that their contribution is closely related to the effective decision and policy making.

- In [5], the authors took the problem of detecting hate speech on social media using user comments. The motivation of their work was developed because of increasing cases of hate on the online platforms causing the decline in the online business and the user experience. They aim to develop a low dimensional representation of comments using neural language models which they would be sending to classification algorithms. They addressed the problem of high dimensionality and sparsity due to previous work of BOW model and aimed to achieve accurate predictive models for the domain.

- Authors in [11] were the first one up till 2016 to focus on the problem of detecting the offensive language on online platforms rather than just focussing on the specific type of abuse such as cyber-bullying or hate-speech as evident in the past years. They brought out the attention to the drawbacks of regular expression and black-lists in the extreme cases of hate speech which are more subtle and less obvious than regular ones. They tend to outperform the state of the art deep learning approach to build a predictive model and also created a new corpus of user comments as a unique one of its kind.

- Authors in [3] raised the challenge of separating the hate speech with other types of offensive content as the previously used lexical methods failed to

distinguish between the different offense types. It is important to accurately identify the hate speech particularly from the other offensive content as both tend to have the different implications on the society and the individuals. They worked with twitter data that was labelled into 3 categories namely hate speech, offensive language and neither of these two for which they trained multi-class classifier.

- In [9] the authors extended the main challenge from [3] to distinguish the hate speech from general profanity or offense to come up with a lexical baseline for the same.

- The work presented in [12] aimed to identify and empirically analyse the relationship; particularly similarities and differences between the different subtasks that come under the offensive language detection. They also proposed a topology that could help the researchers for data annotation and feature construction.

- Authors in [7] believed that the offensive content is increasing at an unprecedented rate on the online platforms which have potential to hamper the mental lives of individual and groups leading to a negative effective in the society. They believed that the manual moderation is practically impossible with this rapid generation of the data. Therefore they introduced this problem at TRAC workshop in 2018 and modelled the problem as the multi-class classification task of differentiating the posts and the comments into Overtly Aggressive, Covertly Aggressive, and Non-aggressive texts. They also identified the important challenges in the shared task mainly revolving around the annotation and the language issues.

- The related work presented in [10] also aims to distinguish hate speech from the regular type of offensive content / profanity which has not been studied much in the previous work. They used the n-grams, skip-grams and clustering based techniques for the feature representations. They analysed that the previous work formulated the task of hate speech detection as the binary classification problem in which the systems are likely to misclassify the instances as non-offensive that did not contain any objectionable words but are actually offensive in a semantic and deeper sense. They also focussed on ensemble classifiers rather than the single classifiers which were mainly

used in the past studies. The work considered that the opinion variation of the annotators is also an issue for this problem which needs to be resolved.

- The authors in [13] dealt with the topic of offensive language on the German tweets which was announced as the shared task in GermEval 2018 competition. The problem was formulated into 2 subtasks. First one is the binary classification task to differentiate between the offensive and non-offensive German tweets. The second one is the further classification of offensive tweets into profanity, insult, abuse and other.

- Authors in [6] believed that the text with high toxicity on the internet can cause personal attacks, bullying behaviours, threatening and harassment. They utilized the Wikipedia dataset released under Kaggle competition of toxic comment classification which was modelled as the 6-class classification problem. They employed CNN for the learning the structure of words in the document due to the advances in hardware and cloud computing.

- The work in [15] considered the problem of the offensive language detection on the social media as the whole rather than focussing on the specific type of abusive content on the web as evident by the previous studies. They came up with the new dataset with three level of annotation scheme providing an opportunity for the researchers to delve deeper into the topic. Subtask A is to determine whether the post is offensive or not. Subtask B further identifies the type of the offense content whether it is targeted or untargeted. Subtask C then categorizes the target of the offensive post into either one of 3 classes namely individual, group and other.

## 2.2 Datasets Involved in Previous Work

- Related work [4] used the labeled corpus of 4500 YouTube video comments that were scrapped from web. The sensitive examples from the dataset were labeled into one of the 3 categories namely sexuality, race  culture and intelligence.

- In paper [14], the authors used the sampled version containing labeled 1762 tweets from the TREC corpus developed in 2011.

- At that time as no dataset for the bullying detection was publically available so authors in [2] scrapped the comments from the top 3 videos of different categories found in YouTube movies. The final dataset contained 4626 user comments consisting of 3856 distinct users and were manually labeled as bullying and non-bullying. Comment history of each user was also recorded and on an average the dataset contained 54 comments per user profile.

- Authors in [8] prepared a balanced dataset from the twitter accounts that contained 24582 carefully labeled tweets by annotators as racist or nonracist.

- The authors in [1] collected the data from twitter API within the 2 week window after the murder of the drummer as they believed that majority of the hate comments bound to happen in the 2 weeks' time and gradually decline. Therefore they wanted to capture this immediate reaction. A total of 450,000 tweets were captured during the study window and labeled as 'Yes' or 'No' meaning that the comment is hateful or not.

- In [5], the authors prepared the largest dataset available till 2015 for hate speech detection for over 6 months and named as WWW-2015 dataset. They collected the comments from the Yahoo Finance website containing a total of 951,736 user comments out of which only 5% were hate speech comments and rest 95% of the comments were clean without the use of any offensive language. This was therefore modelled as the binary classification task as well.

- The dataset in [11] was the extended version of the one used in [6]. It contained the user comments from the Yahoo finance as well as the Yahoo news adding the diversity to it because of presence of all types of abusive language like hate speech, profanity or derogatory language. They marked each user comment as the "abusive" or "clean" making it a binary classification problem. In the news dataset only 16% of the data was abusive and in the finance dataset only 7% of the comments were abusive.

- Authors in [3] prepared the dataset from twitter API that contained 24,802 tweets each of them labeled as hate speech (5%), offensive (76%) or neither of these two (19%).

- The authors in [9] used the open source dataset containing 14,509 English

tweets which are labeled as HATE (16%), OFFENSIVE (33%), or OK (51%) reflecting the absence of any offensive content.

- The dataset in [7] contained roughly 12,000 training comments from Facebook. For the development they provided the participants with 3000 comments in English and Hindi language respectively. Each comment was annotated with a label indicating the levels of aggression namely Overtly Aggressive (OAG), Covertly Aggressive (CAG) and Non-Aggressive (NAG) making it a multi-class classification task. The test set contained 916 and 970 comments in English and Hindi respectively.

- The authors in [10] used the same dataset as in [9] making it a 3-class classification problem of distinguishing between hate speech and general profanity.

- The dataset in [13] contained over 8500 annotated tweets which were labeled as OFFENSE (33.7%) and OTHER (66.3%). Each offensive tweet was further annotated as ABUSE (20.4%), INSULT (11.9%), PROFANITY (1.4%) and OTHER (66.3%).

- Authors in [6] used a dataset having Wikipedia comments such that each comment is annotated into either of 6 classes namely toxic, severe toxic, obscene, threat, insult and identity hate.

- In the work [15], the authors came up with a new dataset OLID for the offensive language detection in social media. At level A, each tweet was assigned a label as Not Offensive (NOT) or Offensive (OFF). At level B, the offensive tweets were annotated by either of 2 labels namely Targeted Insult (TIN) and Untargeted (UNT). At level C, the targeted offensive tweets were given an either of 3 labels namely Individual (IND), Group (GRP), Other (OTH).

## 2.3 Existing Methodology

- Authors in [4] applied the standard pre-processing techniques like stemming, and removal of stop-words and unimportant characters from the textual data. They experimented with three classifiers Naïve Bayes, SVM and Decision Tree with 10-fold cross validation. They divided their feature space

into general features and label specific features. For the general features they used TF-IDF and for the specific features they used unigrams and bigrams.

- In [14] the authors trained the model to detect and analyze the cyber-bullying on the social media platforms. They divided their work into 4 subtasks. For the subtask A, after the regular pre-processing techniques they applied three techniques for the featurization namely unigram, unigram+bigram and POS colored unigram+bigram. For the classification task they experimented with Naive Bayes, SVM with linear kernel, SVM with RBF kernel and Logistic Regression with 5 fold cross validation using the WEKA implementation. In subtask B, for the author's role they used the same classifiers as the subtask A with 10-fold cross validation and also tuned the best model jointly with 5-fold cross validation with grid search CV. For categorizing the person mention role into respective categories they used the named entity recognition and trained the linear CRF and SVM respectively with 10-fold cross validation. In subtask C, they used the same feature representation, classifiers and parameter tuning as for the previous 2 subtasks with 10-fold cross validation. They used LDA as well as its variational inference implementation as their exploratory tool to discover the relevant topics from the bullying trace in the subtask D.

- Authors in [2] used three feature sets to train the cyber-bullying classifier which were content-based, cyber-bullying based and user-based features. For the pre-processing they removed all the stop words and applied stemming to their dataset. They trained a Support Vector Machine to classify the bullying comments and non-bullying comments with 10-fold cross validation.

- To deal with the problem of hate against black community on twitter the authors in [8] trained a Naïve Bayes classifier to able to classify the new tweet as racist or nonracist. They pre-processed the dataset by eliminating the URL's, mentions, stop words and punctuation along with lowercasing and replacing the wrong spellings with the correct ones. Authors found that 86% of the tweets that were racist only because they contained the offensive words so they preferred unigram model to featurize the training data.

- Work in [1] was a binary classification task of predicting whether the comment is hateful or not. They followed a pipeline starting from Data collection and annotation, Feature selection, Data pre-processing, Feature preparation and finally Model selection. They realized that the offensive words from the tweet could be the important features so they utilized the frequency of occurring of unigram and bigram. As the offensive tweet contain the certain instances following a particular pattern therefore for the extraction of typed dependencies within the tweet text they employed a Stanford lexical parser along with a context free lexical parsing model which represented the syntactic grammatical relationship in a sentence that are used a important features for the classifier. They came out with more common sense type of reasoning approach for this feature extraction phase. For the pre-processing phase they followed a generalized pipeline of tokenization, lowercase conversion, removal of stop words and alphanumeric characters, stemming. To preserve the context of words and the surrounding they employed unigrams to trigrams. They experimented with the 2 approaches of n-grams and collection of derogatory or hateful terms to check the contribution of other terms in determining the strong predictors. They ran a Bayesian Logistic Regression using all the typed dependencies features and came up with the vector representation of the tweet containing list of ngrams that included words, typed dependencies or combination of both. They used the three classifiers Bayesian Logistic Regression, Random Forest Decision Tree and Support Vector Machine for this binary classification task. They also employed the meta voting ensemble classifier made from these classifiers.

- After applying the standard pre-processing techniques, they [5] divided their work into two parts for the detection of hate speech from the user comments. First they employed paragraph2vec to learn the distributed representation of comments and words using the neural language model of the continuous BOW (CBOW). This produced a low dimensional embeddings where the semantically similar comments resided in the same part of the space. Secondly a logistic regression classifier was trained on these embeddings to classify the type of user comment as hateful or clean.

- Authors in [11] used the Vowpal Wabbit's regression model to measure the different aspect of the user comments using NLP features. They divided

their features into 4 categories which were N-grams, Linguistics, Syntactic and Distributional Semantics. Due to noise found in the data they performed some mild-preprocessing for the first three features but did not performed any normalization for the fourth feature.

- In [3], the pre-processing part was undertaken as to convert the tweet into lowercase and performed stemming through the porter stemmer. After that they featurized the tweets as weighted TF-IDF unigrams, bigrams and trigrams followed by the construction of the POS tagging using NLTK. They used Flesch-Kincaid Grade Level and Flesch Reading Ease scores to capture the quality of each tweet and also assigned the sentiment scores to each of the tweet. For the hashtags, mentions, retweets and URL's, they included binary and count indicators and for the number of characters, words and syllables, they included features. They tried various models in Scikit-learn like Logistic regression, Naïve Bayes, Random Forest, Decision Tree and Linear SVM to train the model using 5-fold cross validation along with L1 regularization to reduce the dimensionality of the text data. They also performed the grid search parameter tuning to find the optimal parameters.

- Authors in [9] used a LIBLINEAR SVM implementation for this multi-class classification task which has proven to be very effective on Native language identification and temporal text identification. For the features they used character n-grams, word n-grams and word skip-grams.

- In [12], the authors defined their topology based on the prior work in the field of detection of different types of abusive language. They considered a 2 fold approach where the first aspect is to analyse the target of the abuse and another aspect is to analyse the degree to which the abuse is explicit. They also laid the implications of this topology on the annotation and the modeling of this problem. They suggested that the data annotation strategies should be dependent on the type of the abuse that is intended to be identified. On the other hand to select the most relevant features for the modeling, it is important to identify whether the abuse is directed, generalized, explicit or implicit.

- In TRAC workshop proceedings [7], there were a total of 30 teams who submitted their systems for English and Hindi Language. Participants applied

various techniques like LSTM, CNN, SVM, BiLSTM, Logistic Regression, Random Forest and many more to classify the English and Hindi Facebook comments.

- The participants of the shared task of GermEval [13] used tokenization, POS-tagging, lemmatization and stemming and parsing as the methods for tweets pre-processing. They used SVM, Logistic Regression, Naïve Bayes, CNN, LSTM, GRU and the combination for the classification of the tweets.

- The authors in [6] compared word embeddings and CNN against the BOW approach with the classifiers such as SVM, Naïve Bayes, k-NN and LDA that were applied on the Document Term Matrix.

- Related work as evident in [15] applied various machine learning and deep learning techniques for each of the subtask within the problem domain. They first applied linear SVM trained on word unigrams followed by BiLSTM with the softmax activation function in the final layer with FastText embeddings. Finally they also experimented with CNN on this dataset.

## 2.4 Results Analysis and Evaluation Metrics

- In [4] the authors achieved the best accuracy of 80.2% with the help of rule based JRIP whereas kappa measure was highest for SVM with the average value of 0.75.

- In paper [14], for the subtask A the best model came out to be Linear SVM with a combination of unigrams and bigrams achieving an F-measure of 0.77. For the subtask B they achieved the cross validation accuracy of 61% with SVM linear along with the combination of unigram and bigram for the author's role. For the person mention role linear CRF outperforms the SVM achieving the accuracy of 87% and F-measure of 0.47. For the subtask C the best validation accuracy of 89% is achieved by linear SVM. For the subtask D, LDA discovered 5 topics from the bullying post namely feeling, suicide, family, school, verbal bullying and physical bullying.

- The results from the work in [2]clearly indicated that the detection accuracy was boosted upon adding the bullying specific features and the user context.

The model achieved the F-measure of 0.64 when they used all the three feature space to train the model.

- Authors in [8] achieved a 10-fold cross validation accuracy of 76% using the Naïve Bayes classifier and an error rate of 24%.

- Results as evident from [1] indicated that the most efficient features proved to be a combination of ngram typed dependencies and hated terms. All the classifiers BLR, SVM, RFDT and Ensemble performed equally well on the test set with an F-measure of 0.77

- In [5] the authors achieved the Area Under Curve (AUC) value of 0.80 with the paragraph2vec representation trained on a logistic regression classifier.

- In [11] the authors achieved the best F-measure of 0.79 on the finance data and 0.81 on the news data when trained considering all of the features rather than the selective features.

- Results from the work in [3] indicated that Logistic regression with L2 regularization performed significantly better than the other models with the F1 measure of 0.90.

- In [9] , the authors achieved a good accuracy of 78% with the character 4-gram model with Linear SVM evaluated using the stratified 10-fold cross validation.

- In the shared task of TRAC [7], the best ranked team used LSTM with the data augmentation strategy. They used a combination of CNN and RNN on the surprize twitter dataset for the feature representation. The team performed the spelling correction, emojis conversion and sentiment score computation as the part of text pre-processing and achieved F-score of 0.64 for both the Hindi and English on the Facebook test set. They also managed to come up as the team that significantly performed better on the twitter dataset despite being trained on the Facebook dataset achieving the F-measure of 0.60 and 0.50 for English and Hindi respectively.

- The results from GermEval [13] displayed that the top performing systems came out to be CNN and LSTM for the both the shared tasks. They achieved

an average macro F1 score of 76.77% and 52.71 % for the binary-class coarse-grained subtask A and multi-class fine-grained subtask B respectively.

- Authors in [6], for the 6-class classification problem of toxic comments into respective categories achieved a descent accuracy of 91.2% with CNN.

- In [15], as the dataset is fairly imbalanced for each of the levels so the authors used per class precision, recall and F1-score with the weighted average to compute the performance of each of the models. For the detection of the offensive language, CNN came out to be the best model with macro average F1 score of 0.80. For the categorization of the offensive language, CNN was the best model with the average macro F1 score of 0.69. In the identification of the target of the offensive posts, both CNN and BiLSTM performed equally well achieving an F1-macro of 0.47.

## 2.5 Conclusion Analysis

In this section we would be critically analyzing the above discussed related work on the detection of abusive/ offensive language on the various online platforms. This will eventually lead and form a strong basis to our contribution in the problem area through this research project.

- The work presented in [2], [4], [14] focussed on the detection of **Cyber-Bullying** on the social media platform like Youtube and Twitter. Some of them modeled this as the binary classification problem of a content being sensitive or not while one work took the user characteristics and profile information into the account. One work also formulated the problem as sentiment analysis and topic modeling.

- Related work [6] focussed mainly on the detection and classification of **Toxic Comments** into 6 respective categories namely toxic, severe toxic, obscene, threat, insult and identity hate.

- Majority of the related work exists in the detection of the **Hate Speech** because of its large presence on the social media when compared to the other types of the offensive content on these platforms. The work presented in [1], [5], [8] are carried out with the same objective. Some formulated the problem as the binary classification of distinguishing racist content from the

non-racist one, while some of them posed it as differentiating the hateful comments from non-hateful/clean ones based on race, ethnicity and religion. Work presented in [3], [9], [10] primarily contributed in distinguishing the hate speech from general profanity and modeled the problem as 3-class classification task with the annotation of the instances as Hate, Offensive or Clean.

- The work in [7] was the identification of the **Aggression** on the social media particularly the English and Hindi language comments of Facebook and Twitter. The contribution formulated the idea as the 3-class classification task of classifying the posts into 3 categories namely non-aggressive, covertly aggressive and overtly aggressive.

- Related work presented in [11], [13], [15] focussed on the detection of **Offensive language** rather than the specific type of offence or the abuse on these platforms. Some modeled the problem as the binary classification task of categorizing the content as abusive or clean while other focussed on the further multi class classification of the detected offensive German tweets as the profane, insult or abuse. The most recent work came out with the very suitable dataset for the detection of offensive language with the 3 levels of annotation scheme helpful in identifying the type and the target of the offensive posts for an in-depth analysis and moderation.

- Upon analyzing each of the above work we can reach to the conclusion that the previous studies revolving around the detection of abusive/offensive content on the social media platforms can be classified into 5 main categories. These were mainly Cyber-bullying detection, Hate speech detection, Aggression identification, Toxic comment classification and the last but not least is the detection of overall Offensive content that includes all these previous types.

- We strongly believe that OLID dataset [15] aims to capture the differences and similarities between the pre-existing datasets revolving around the above mentioned 5 tasks of offensive language detection. This dataset also treats the problem as the whole because it covers all these 5 aspects which was not evident in the previous studies. It also at the same time enable the identification of the type and the target of the offensive tweets

for the further analysis.

- As the OLID dataset was recently released in 2019, therefore it opens up a new opportunity for the researchers to explore this novel dataset for the further improvement in the field. We also aim to contribute and come with the different techniques in order to make significant improvements after the work carried in [15]

# Chapter 3

# Proposal

The **conclusion analysis** from the Literature Review chapter has formed a basis about the existing work and as well as gives us the insight about the pre-existing datasets for the offensive language detection. This section represents the direction that we will undertake for this research project. We have already defined the research objectives in the previous chapters; here we will discuss the **techniques** to achieve these objectives. Therefore we would be presenting our **contribution** in terms of design decisions, algorithms and methods, towards the accurate detection of the offensive language on the social media.

## 3.1 Dataset Description

To carry out the research project we would be using the publicly available benchmark dataset [15] which is named as OLID (Offensive Language Identification Dataset). This dataset generalizes the task of detecting the all types of offensive content that was defined in the literature chapter using its 3 level annotation scheme. This types of dataset is very first of its kind and opens up the new opportunity to explore it even further which we also aim to do in our research project.

This OLID dataset annotate each tweet instance/post with the 3 level annotation scheme meaning that each instance is labeled with 3 corresponding type of labels. The first level denotes whether the tweet is offensive (OFF) or not (NOT). The second level identifies the type of the offensive tweet which could be targeted insult (TIN) or untargeted (UNT). The third level identifies the target of the offensive post categorized mainly into individual (IND), group (GRP) or other (OTH).

OLID is a collection of 14,100 annotated tweets obtained using the twitter API. The training dataset consists of 13240 annotated tweets while the test partition contained about 860 tweets. The whole problem can be formulated as the 3 sub-tasks of detecting the ofensive content, identifying the type of the offensive content and lastly to categorize the target of the targeted offensive content.

The distribution of OLID dataset is reflected with the table 3.1 and the insights from the table for the training set is depicted below in the hierarchical order of annotation.

| A | B | C | Train | Test | Total |
|---|---|---|---|---|---|
| OFF | TIN | IND | 2,407 | 100 | 2,507 |
| OFF | TIN | OTH | 395 | 35 | 430 |
| OFF | TIN | GRP | 1,074 | 78 | 1,152 |
| OFF | UNT | — | 524 | 27 | 551 |
| NOT | — | — | 8,840 | 620 | 9,460 |
| **All** | | | 13,240 | 860 | 14,100 |

Figure 3.1: Distribution of 3 levels of annotation in OLID

1. **For sub-task A** – The distribution of the each of the labels in the training set for this subtask A is below
   - OFF – 4480 tweets (33.23%)
   - NOT- 8840 tweets (66.76%)
2. **For sub-task B** – The distribution of the each of the labels in the training set for this subtask B is below
   - TIN- 3876 (88.09%)
   - UNT- 524 (11.90%)
3. **For sub-task C** – The distribution of the each of the labels in the training set for this subtask C is below
   - IND – 2407 (62.10%)
   - GRP – 1074 (27.70%)
   - OTH – 395 (10.19%)

## 3.2 Proposed Workflow

We would be following an exhaustive Machine learning lifecycle for Natural Language Processing tasks to proceed with our research project of detecting and analyzing the offensive content from the social media. This entire pipeline would be broken down into the individual steps which are briefly depicted below

1. **Data Cleaning:** Here we would be cleaning our data to achieve the data deduplication as it can hamper our predictive models. We would be also removing any kind of retweets that may occur in the data.

2. **Text Data Pre-processing:** In this step we would be processing our training tweets instances. Here we would be employing the standard techniques such as tokenization, stemming, lemmatization, stop word removal, emojis encoding and lower case conversion.

3. **Exploratory Data Analysis:** This step will help us summarize the main characteristics of our OLID dataset with the help of visual methods. EDA is for seeing what data can tell us beyond formal modelling or hypothesis testing task. We will understand the text data in terms of the frequency and the type of the words occurring in the offensive and clean tweets. We will look into some visual plots like box-plots, scatter-plots and correlation matrix along with the word-clouds that will help us to make more sense out of the data.

4. **Text Featurization:** This step is very important for a valid input to many Machine learning and Deep learning algorithms as they cannot directly process the text as it is about the mathematical calculations that goes behind fitting our encoded textual data into some machine learning algorithm. We would be employing various techniques like unigram, bigram, word2vec, BOW, skip grams to vectorize our training data. We would also be exploring the fast-text embeddings.

5. **Model Building:** Here we would be employing the non-neural machine learning algorithms such as Naive Bayes, SVM and Logistic Regression . Also we would be applying the ensemble techniques on the dataset like Ada-boost, Random Forest and Gradient Boosting. These techniques combine the results from the number of learning algorithms to obtain a better

predictive performance than any single machine learning algorithm. Also we would be exploring with the deep learning models like CNN, BiLSTM and BERT as they are the state of the art in many NLP tasks. We would be training these models on the standard parameters on the training set using the 10-fold cross validation.

6. **Model Evaluation:** Most widely used evaluation metrics for the binary classification problem like ours are confusion metrics, f1-score, precision, recall and ROC area. We would be then choosing the best performing models based on these metrics.

7. **Hyper-Parameter Optimization:** We will then tune the parameters of the best models with the help of the techniques like Grid Search CV and Randomized Search CV. We will compare each of these techniques based on the performance and present the comparative results.

# Chapter 4

# Summary

The main objective of this interim report was to develop a pathway in order to successfully complete the AI research project. This report starts with an introduction where we discussed about the research objectives and the importance of detecting the offensive content on the social media platforms. We gradually developed a direction by studying and reviewing various academic papers that were available for this task. Based on the critical analysis of the literature review, we reached to the conclusion to take up the exploration on the recently created OLID dataset [15] which seems to be very appropriate for this problem domain. Moreover this dataset also considers the problem as the whole rather than just focussing on the specific type of the abuse on social media.As not much work is carried on this dataset, we propose a comparative analysis of various Machine learning and Deep learning algorithms to solve the different subtasks that the dataset offers. The proposal chapter describes the dataset in full detail and also put some light on our proposed workflow fitting in the NLP Machine learning pipeline with the minute details of the techniques that are to be used for carrying out this research project.

# Bibliography


[1] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.

[2] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer, 2013.

[3] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*, 2017.

[4] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. In *fifth international AAAI conference on weblogs and social media*, 2011.

[5] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30, 2015.

[6] Spiros V Georgakopoulos, Sotiris K Tasoulis, Aristidis G Vrahatis, and Vassilis P Plagianakos. Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, pages 1–6, 2018.

[7] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, 2018.

[8] Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*, 2013.

[9] Shervin Malmasi and Marcos Zampieri. Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*, 2017.

[10] Shervin Malmasi and Marcos Zampieri. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202, 2018.

[11] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153, 2016.

[12] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, 2017.

[13] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. Overview of the germeval 2018 shared task on the identification of offensive language. 2018.

[14] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666. Association for Computational Linguistics, 2012.

[15] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, 2019.