# Predicting the Type and Target of Offensive Posts in Social Media

Mohaimenul Islam
*Computer Science and Engineering*
*BRAC University*
mohaimenul.islam.moon@g.bracu.ac.bd

Mst. Sabrina Mobassira
*Computer Science and Engineering*
*BRAC University*
mst.sabrina.mobassira@g.bracu.ac.bd

Sharafat Kabir Shajid
*Computer Science and Engineering*
*BRAC University*
sharafat.kabir.shajid1@g.bracu.ac.bd

Nuzhat Hasina
*Computer Science and Engineering*
*BRAC University*
nuzhat.hasina@g.bracu.ac.bd

***Abstract*—One increasingly pressing issue on social media is the abundance of offensive content. This paper offers a hierarchical classification method for determining the kind and intended audience of inappropriate and offensive social media posts. The first suggested strategy specifies the kind of transgression (such as hate speech, cyberbullying, or obscenity), and then determines who the offender is. (For instance, targeted, untargeted, and other insults). The method is assessed using the Dataset for Offensive Language Identification. (OLID), a recently created dataset including tweets flagged as objectionable with a two-layer annotation approach that is finely tuned. The method attains cutting-edge outcomes on OLID, surpassing earlier approaches by a substantial amount.**

***Index Terms*—Offensive language, Machine learning, Natural language processing, social media, Sentiment analysis, Text classification, Spam filtering, Hate speech, Racism, Sexism, Discrimination, Bias, Ethics**

## I. Introduction

The spread of objectionable content on social media is a difficult and complicated issue. This content may be harmful to people individually and in groups, and it may also be a factor in the propagation of prejudice and hatred. Effective techniques are required to detect and remove offensive content from social media, but this is a challenging undertaking because of the variety of offensive content and the ambiguity of language. This paper proposes a hierarchical classification method for identifying the nature and intended audience of inflammatory social media messages. The suggested method begins by classifying the offense (such as hate speech, cyberbullying, or vulgarity) and then classifies the offense's target (such as targeted, untargeted). The method's effectiveness is assessed on the recently created Offensive Language Identification Dataset (OLID), which consists of tweets that have been fine-tuned using a two-layer annotation system. On OLID, the method produces state-of-the-art results, significantly outperforming earlier approaches. The suggested method is a useful tool for locating and reducing objectionable social media information. Researchers can use it to examine the characteristics and effects of offensive content on the internet, and social media sites can use it to flag and remove offensive messages. Our suggested strategy has a hierarchical structure and has two smaller tasks: Sub-task A: Offensive language identification as offensive (OFF) or not offensive (NOT) • Sub-task B: Automatic categorization of offense types as either targeted insult (TIN) or untargeted (UNT) Our paper was influenced by Zampieri et al.'s work on OLID Dataset [1]

## II. Background Study and Related Works

### A. Literature Review

Supervised learning techniques are widely used to identify aggressive or abusive language in text. Supervised learning techniques are widely used to identify the aggressive or abusive language in the text, such as "Smokey" developed by Spertus [**?**], More recent approaches are based on GloVe [**?**] and RNNs [**?**]. Singh et al. [**?**] used CNNs and LSTM models to identify aggressive language in social media text from multilingual speakers who sometimes use code-mixing, in a single sentence. Mishra et al. [**?**] employed graph mining techniques, particularly node2vec, to capture user community structures and homophily in a Twitter network, integrating these learned node representations with text-based features to enhance the detection of abusive content. While many existing approaches to detecting offensive language use terms like "cyberbullying," "abusive," and "hate speech" interchangeably, it is important to distinguish between these terms. Each term has a specific definition and target, and proper categorization of offensive language is crucial for organizations and government entities to make informed mitigation efforts.

## III. Methodology

We first performed text pre-processing steps, such as removing non-alphanumeric characters, hashtags, URLs, and direct @ mentions. This is because these characters and symbols can be irrelevant to the task of offensive language detection and can make the text more difficult to process.

For each sub-task, We split the training data into a train and validation set, with a 67-33 split. This means that 67% of the data was used to train the classifiers, and the remaining 33% was used to evaluate the performance of the classifiers.

We trained the classifiers on the train set and tuned the model by looking at the performance of the validation set. This is a process of adjusting the hyperparameters of the classifier to improve its performance.

Finally, We reported the results and performance for each classifier on the holdout test set. The holdout test set is a set of data that was not used to train or tune the classifiers. This is used to get an unbiased estimate of the performance of the classifiers.

The different classifiers we experimented with are:

NBSVM: This is a simple yet powerful approach to text classification that combines a linear model such as SVM (or logistic regression) with Bayesian probabilities. Naive Bayes: This is a statistical classifier that assumes that the presence or absence of a particular feature is independent of the presence or absence of other features. This makes it a relatively simple and efficient classifier to train. Support vector machines (SVMs): These are a type of machine learning model that can be used for classification and regression tasks. SVMs work by finding the hyperplane that best separates the two classes of data.

LSTM: This is a recurrent neural network that can learn long-term dependencies between words. Recurrent neural networks: These are a type of neural network that can process sequences of data. This makes them well-suited for tasks such as natural language processing, where the order of the words is important. Long short-term memory (LSTM): This is a type of recurrent neural network that is specifically designed to learn long-term dependencies. This makes them well-suited for tasks such as machine translation, where the meaning of a sentence can depend on words that are far apart.

CNN-Text: This is a convolutional neural network that can learn local patterns in text. Convolutional neural networks: These are a type of neural network that can learn local patterns in data. This makes them well-suited for tasks such as image classification, where the features of an image are often local. Text: This refers to the text that is being classified.

## IV. Dataset

The data used in this work was released as part of SemEval 2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval) (9). Offensive Language Identification Dataset (OLID) dataset, is a large collection of English tweets annotated using a hierarchical three-layer annotation model to distinguish between whether the language is offensive or not (A), its type (B), and its target (C). It contains 14,100 annotated tweets divided into a training partition of 13,240 tweets and a testing partition of 860 tweets. Each level is described in more detail in the following subsections.

Level A: • Not Offensive (NOT): Posts that do not contain offense or profanity;

• Offensive (OFF): Posts containing any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words

Level B: • Targeted Insult (TIN): Posts containing insult/threat to an individual, a group, or others;

• Untargeted (UNT): Posts containing nontargeted profanity and swearing. Posts with general profanity are not targeted, but they contain non-acceptable language.

OLID contains 14,100 annotate tweets. It has been used as the official dataset for OffensEval: Identifying and Categorizing Offensive Language in Social Media. The dataset was annotated using crowdsourcing. The gold labels were assigned taking the agreement of three annotators into consideration. No correction has been carried out on the crowdsourcing annotations. Twitter user mentions were substituted by @USER and URLs have been substitute by URL.

OLID is annotated using a hierarchical annotation. Each instance contains up to 3 labels each corresponding to one of the following levels:

- Level (or sub-task) A: Offensive language identification;

- Level (or sub-task) B: Automatic categorization of offense types;

Whenever a label is not given, a value NULL is inserted (e.g. INSTANCE NOT NULL NULL)

Level A: Offensive language identification

- (NOT) Not Offensive - This post does not contain offense or profanity. - (OFF) Offensive - This post contains offensive language or a targeted (veiled or direct) offense

In our annotation, we label a post as offensive (OFF) if it contains any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct.

(B) Level B: Automatic categorization of offense types

- (TIN) Targeted Insult and Threats - A post containing an insult or threat to an individual, a group, or others (see categories in sub-task C). - (UNT) Untargeted - A post containing non-targeted profanity and swearing.

Posts containing general profanity are not targeted, but they contain non-acceptable language.

## V. Exploratory Data Analysis

The data used in this study was released as part of the SemEval 2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). The Offensive Language Identification Dataset (OLID) is a large collection of English tweets that have been annotated using a hierarchical three-layer annotation model. This model distinguishes between whether the language is offensive or not (Level A), its type (Level B), and the categorization of the targets of insults/threats (Level C). As Level C is not in our project scope, this data has been discarded in this paper. The dataset contains 14,100 annotated tweets, which are divided into a training set of 13,240 tweets and a test set of 860 tweets. Each level of the annotation model is described in more detail in the following subsections.

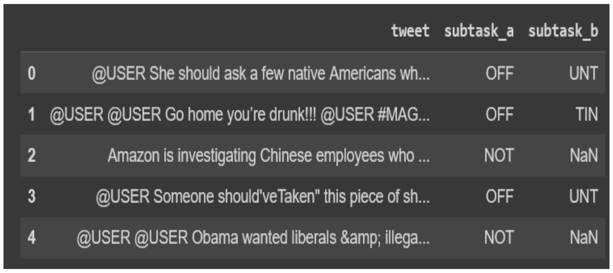
Level A: Offensive or not

The first level of the annotation model classifies tweets as either offensive or not offensive. Offensive tweets are those that are intended to cause harm or offense to others. They may contain insults, threats, or other forms of harmful language.

Level B: Type of offense

The second level of the annotation model classifies offensive tweets into one of four types:

Targeted insult: A tweet that insults or threatens an individual or group of people. Untargeted profanity: A tweet that contains profanity but does not target anyone in particular. Other: A tweet that is offensive but does not fit into any of the other categories. Not offensive: A tweet that is not offensive.



|  | tweet | subtask_a | subtask_b |
|---|---|---|---|
| 0 | @USER She should ask a few native Americans wh... | OFF | UNT |
| 1 | @USER @USER Go home you're drunk!!! @USER #MAG... | OFF | TIN |
| 2 | Amazon is investigating Chinese employees who ... | NOT | NaN |
| 3 | @USER Someone should'veTaken" this piece of sh... | OFF | UNT |
| 4 | @USER @USER Obama wanted liberals &amp; illega... | NOT | NaN |

Fig. 1. Five tweets from the OLID dataset, with their labels for each level and final categorization.

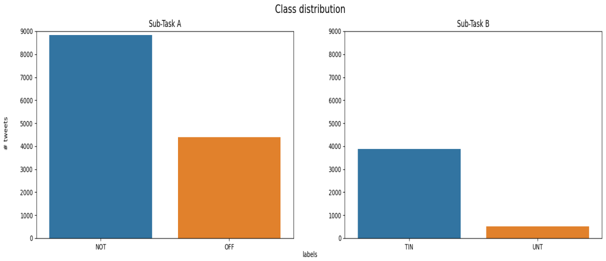


Fig. 2. Class Distribution from the data set.

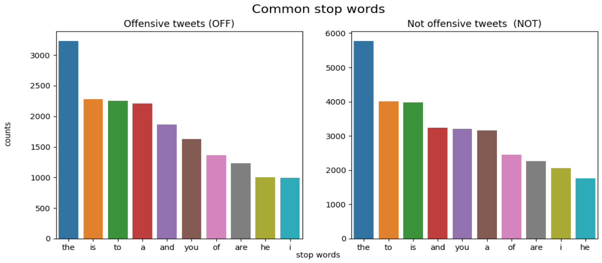


Fig. 3. Common Stop Words.

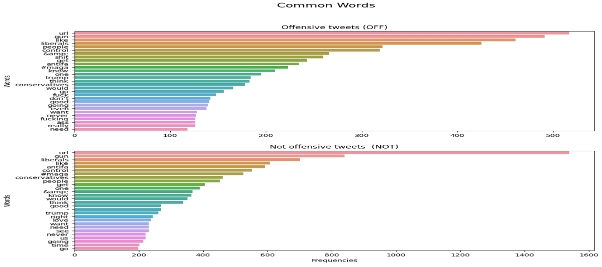


Fig. 4. common words.