Auto-Off ID: Automatic Detection of Offensive Language in Social Media

Auto-OffID: Automatic Detection of Offensive Language in Social Media

Geetha R, Karthika S, Chaluvadi Jwala Sowmika and Janani Bharathi M

Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India

E-mail: geethkaajal@gmail.com

Abstract. As the popularity of social media grows, computer-mediated anonymity allows users to engage in activities that they would not do in real life. This makes users vulnerable to abuse through Internet platforms. Due to the enormous number of social media data, it is not possible to manually filter out the overflow of abusive content in online communities and social networking sites. The research work proposes a multi-level classification model that deploys various machine and deep learning models to effectively identify offensive content in a tweet. The proposed Auto-OffID system is designed to build a system that classifies tweets as offensive or non-offensive; filters out and classifies offensive tweets as either targeted or non-targeted; filters out targeted tweets and identify mentions of individuals and organizations who have been bullied. The study is supported by the text analysis features with lexicon features using LIWC, POS tags for primary and secondary users, Twitter Tag Scores (TTS). This system is evaluated using a diverse choice of machine learning and deep learning models from which it is proved that C-LSTM outperform with an accuracy of 91.72% for offensive language identification; LDA + Logistic Regression training with SVM accuracy of 90.87% for offensive tweet classification.

1. Introduction

Every day, billions of users connect over the Online Social Networks (OSN). With the emergence of internet-based social media applications such as Facebook, Twitter, Instagram, and Snapchat provide public contact across digital channels with a significant shift over the past decade. These OSN have unique characteristics which give them an advantage to different users who wish to engage in some ways. The use of OSN is growing rapidly to share confidential information across various applications that help users get in regular communication with one another without considering vulnerabilities of cyber security. Twitter is one of the significant social networks, a 2006 launched micro-blogging application that enabled individuals to interact through a brief discussion of 280 characters short sentences.

Twitter is not only a simple method of exchanging personal thoughts and experiences, but also a rapidly growing platform that enables people to exchange experiences at the mass level, such as elections, conferences and, sadly, even tragedies. These interactions can lead to some risky results when it comes to infusing different kinds of security threats into OSN. Social media posts can involve sharing certain forms of violent or offensive material that can ultimately lead to threats like cyber-bullying.

Cyber-bullying is an online assault focused on intentionally offending, threatening, humiliating, or online harassment. Because slurs, rumors and disinformation can be circulated to a wide audience immediately, cyber-bullying on OSN turn-out as a traumatic experience for the people affected. The most important factor that gives chances for cyber bullying and offensive talks is the availability of personal and private data in cyber space leading to privacy issues and regrets as discussed by [1]. To avoid inappropriate actions to social media, online communities, OSN sites and technology providers have invested extensively in initiatives to counter offensive language. Twitter provides access to data for analyzing sentiments, opinions [2] and its impacts on business and society for researchers. One of the most successful tactics to resolve this problem is to use the computational methods to identify offenses, aggressive actions, and hateful speech in user generated contents like micro-blogs, posts, comments, etc.,

Henceforth, this study proposed a novel system called Auto-OffID, especially designed to detect unpleasant social conversations regarding harmful and offensive content via tweets. Auto-OffID uses a hierarchical methodology to distinguish tweets as offensive or not, their type and the organizations and individuals who have been bullied.

The objective of the proposed Auto-OffID method is to create a systematic classification model for classifying tweets as follows:

1. offensive or non-offensive.
2. filter offensive tweets and classifies them as either targeted or non-target.
3. filter targeted tweets and identify the organizations and individuals who have been bullied.

Outline: The section 2 briefs about the related research works carried out in offensive content analysis. The section 3 explains the system architecture and the planned approach to implement the hierarchical classification of offensive content as various tasks. The section 4 presents the results for the Auto-OffID system with sample data and its inferences and the performance metrics. Finally, the section 5 concludes the research work with future works planned by the authors.

2. Related Works

The task of detecting hate speech in Spanish and English tweets was performed in [3] where hate speech against women and immigrants was analysed. The task is structured as two sub tasks of categorization: a large binary subtask to identify the presence of hateful speech, and a fine-grained aspect dedicated to distinguishing additional features on hateful contents such as violent attitude and harassment, to discern if the hate speech is against a person or perhaps a community. This research proved that a trained SVM model with RBF kernel gave better F1-scores when compared to Neural Network models and, more specifically, Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTMs) networks. But it worked with a limitation of consuming only the provided data, exploiting sentence embeddings from Googles Universal Sentence Encoder.

The research work [4] followed the standard pre-processing for feature extraction like unigram, bigram and trigram where each is weighted by its TF-IDF to capture information about the syntactic structures based on which the Penn Part-of-Speech (POS) tags were constructed. The authors used modified Flesch-Kincaid Grade Level and Flesch Reading Ease scores to capture the quality of each tweet and the Sentiment lexicon is used to assign a sentiment score to each tweet. Finally, logistic regression with L2 regularization is used as the final model for classification.

In [5], the authors attempt to understand the bullying by analyzing the abusive content. Depending on hate speech and cyber-bullying, a conceptual framework is proposed to capture key similarities and differences between offensive types from which the implications for data annotation and feature construction are optimized. The authors of [6] have been able to maximize the strength of sentiment analysis for detecting bullying on Twitter. The evaluation was carried by designing a tweet extractor component over the open-source libraries like Twitter4J and Twitter Streaming API. This novel component will continuously crawl the public timeline of Twitter in hunt of tweets comprising

the words of their interest. The Naive Bayes (NB) classifier was extended to perform classification with the LingPipe toolkit and an accuracy of nearly 70% was achieved.

The authors of [7] have deployed a system to detect negative internet experiences through text messages as well as images in terms of abusive information. A novel automated system was built by combining the visual word bag (BoVW) model and the local binary pattern (LBP) with the SVM classifier to identify the abusive image in messages. The detection of violence in text messages included model Bag-of -Words (BoW) with an NB classifier. Finally, the binary system is used to categorize content which include abusive language by evaluating the results gathered by the classification of both the text and images.

The authors of [8] deployed a model that collects user information and text information from Twitter and attempts to establish an auto-detection model of cyber-bullying tweets based on the text, readability, sentiment score, and other user information to predict the tweets with harassment and ridicule cyberbullying. Three data mining techniques, k-nearest neighbours (KNN), support vector machine (SVM), and decision tree are used in this system to detect cyberbullying tweets and select the best performance model for cyberbullying prediction. This system contributes to predicting the bully tweets through the models timely which avoid harmful posts appearing on the social network service and helps the manager of the website to remove the vicious posts as soon as possible to prevent the proliferation of cyberbullying.

3. System Design

The objective of Auto-OffID system was split into three major tasks namely, Task A – Offensive Language Identification, Task B – Offense Type Categorization and Task C – Offense Language Target Recognition. The initial phase of data collection was performed by the SEMEVAL which is discussed in section 3.1 and the data pre-processing module is described in section 3.2. The proposed system architecture of Auto-OffID is presented in Figure 1.

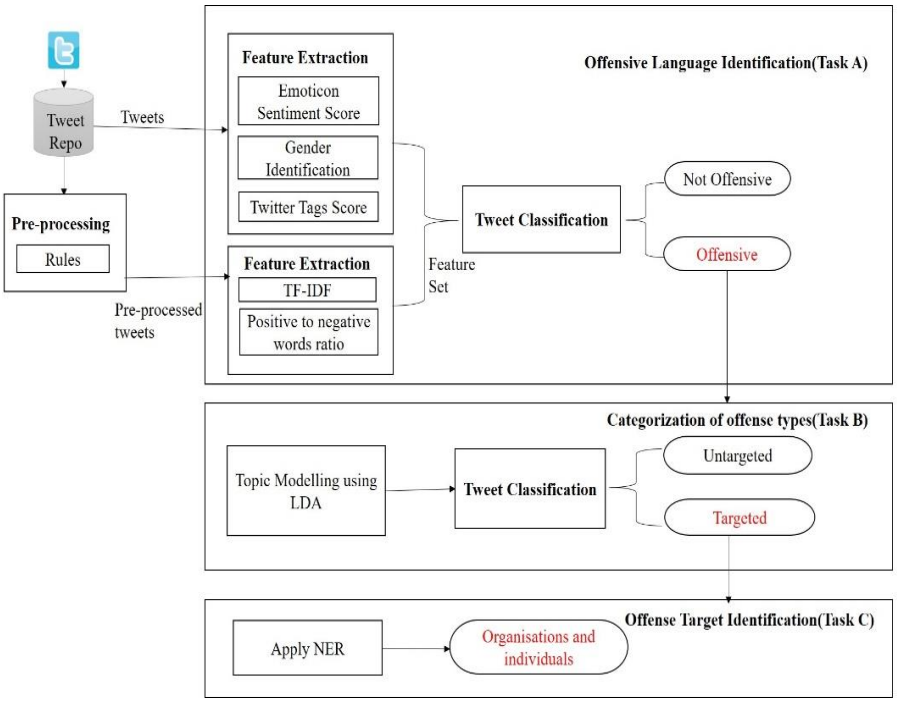

Figure 1. The proposed architecture of Auto-OffID System

3.1. Dataset Description

The dataset (OLID) used in the project references the dataset used in [9]. The dataset encompasses 13,240 tweets and is manually annotated for each of the three Task (A, B and C) as shown in Table 1. The rules and assumptions on which the annotation process is performed is described below for each task.

- Task A: Offensive Language Identification - Level A distinguishes between the kinds:

– Non-offensive (NOT): Post containing non-offensive or profane messages,
– Offensive (OFF): Post containing some degree of offensive language (profanity) or deliberate violation that could be covered up. This comprises of insults, threats and posts that include profane or swear words.

- Task B: Offense Type Categorization - Level B categorizes the kinds of offense:

– Targeted Insult (TIN): posts involving an individual, community or other person face an insult / threat,
– Untargeted (UNT): posts that contain untargeted swearing and profanity. Posts with overall curse words that are not targeted but contain inappropriate language.

- Task C: Offensive Language Target Recognition - Level C classifies insult/threatening targets:

– Individual (IND): individual targeted posts; may be a renowned person, a named person, or an undisclosed participant. Insults and individual-targeted threats are very often defined as cyber-bullying,
– Group (GRP): Posts trying to target a group of individuals assumed to be united by reason of the same ethnic background, gender or sexuality, political orientation, religious beliefs, and perhaps other common characteristics,
– Other (OTH): The target for offensive contents need not necessarily apply to any of the earlier two categories (e.g., an organization, a circumstance, an incident / event, or a question).

Table 1. Sample tweets with their labels for each task of the annotation schema.

S.No	Tweet	Task A	Task B	Task C
1	@USER He is so generous with his offers.	NOT	-	-
2	I'M FREEEEE!!!! WORST EXPERIENCE OF MY FUCKING LIFE	OFF	UNT	-
3	@USER Fuk this fat cock sucker	OFF	TIN	IND
4	@USER Figures! What is wrong with these idiots? Thank God for @USER	OFF	TIN	GRP

3.2. Data Pre-processing

Twitter data may be imperfect, unclear, and noisy, tends to result in inaccurate mining results. Data pre-processing is a proven method of resolving such issues. It translates raw data into a comprehensible format. It also filters out useless data. Various steps involved in it are data cleaning, integration, feature reduction, and transformation. The Auto-OffID system built a library to automate data pre-processing with the rules mentioned in Table 2.

Table 2. A list of pre-processing rules followed by Auto-OffID System.

Rule No	Rules for Pre-processing	Description
1	Convert to lowercase	Transform tweets into lowercase
2	Stop-word removal	English stop-word list provided by NLTK is used to remove stop words from tweets like a, and, the, etc.,
3	Punctuation removal	Remove punctuations like !()-[]{};:'•"\,<>.?$%&^*£_~+ from tweets
4	URL and User Mentions replacement	Replace URL links and @USER from tweets with 'url' and '@username' respectively
5	Date and Time Replacement	Replace date and time from tweets with 'ddmmyyyy' and 'hhmmss' respectively
6	Number replacement	Replace numbers from tweets with 'num'
7	Hashtag replacement	Replace hashtags like #something from tweets with 'hashtag'
8	Emoticon removal	Remove emojis from tweets
9	Stripping whitespaces	Remove additional whitespaces

3.3. Task A – Offensive Language Identification

The Auto-OffID system extracts the content and contextual features for learning the presence of offensive language in a tweet. Therefore, from the pre-processed tweets, the contextual features are extracted based on trigrams - TF-IDF and content features are studied by the novel measures like emoticon sentiment score, positive to negative words ratio, TTS, gender identification. The process involved in measuring each feature is depicted in Figure 2 and discussed below.

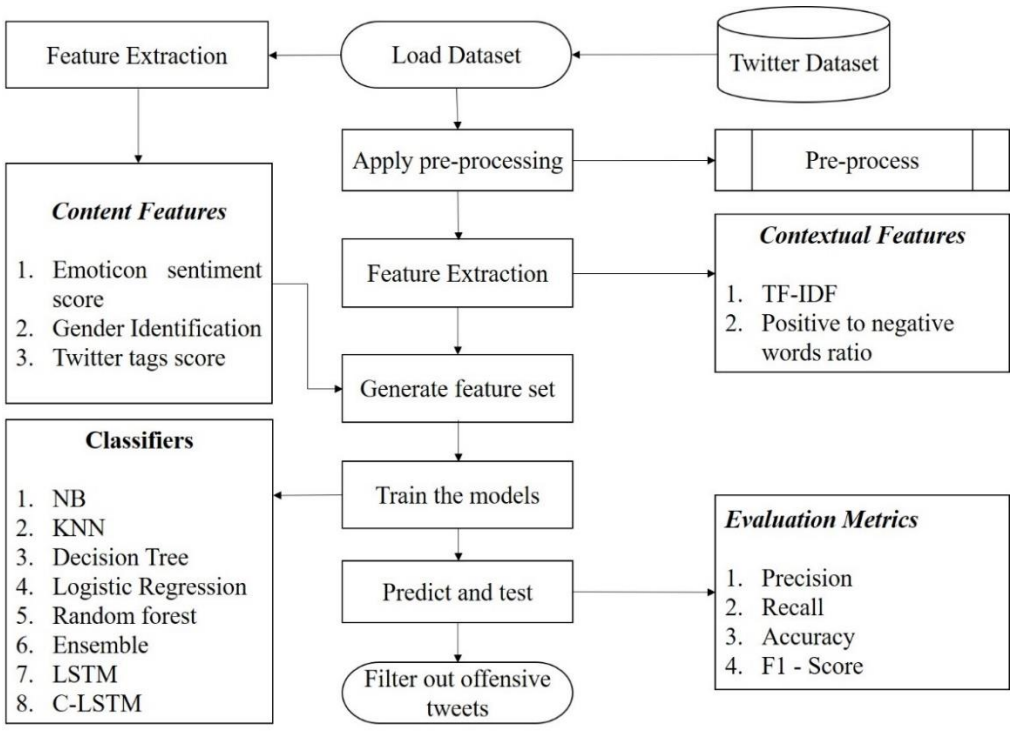

Figure 2. The process flow of Task A – Offensive Language Identification

3.3.1. Emoticon Sentiment Analyzer

The Auto-OffID system used Vader's most efficient sentiment analyser to construct a generalized valence-based human-cured gold standard sentiment lexicon. This will help to identify the polarity of the emoticons in tweets. The authors have chosen Vader for sentiment analysis for the following reasons like, due to its context-awareness, it outperformed the Emosent Analyser; it worked well on our dataset as well as on social media text; no training data is required; it does not suffer from speed performance trade-offs.

3.3.2. Positive to Negative Words Ratio

The identification of lexical features adapted in [10] seemed to be the most efficient lexicon between SentiWordNet, WordNet-Affect, MPQA and SenticNet for the twitter data set. In accordance with the statistics, MPQA was identified as the best lexical resource for the dataset. The MPQA Subjectivity lexicon with 8,222 words that were labelled based on subjective expressions was collected from a variety of sources. This vocabulary contains a collection of words labelled with polarity (positive, negative, neutral) and intensity (weak, strong) with their POS tags that can be used to calculate the positive-negative word ratio.

3.3.3. Gender Identification

A POS tagger usually processes a word sequence and attaches a part of the speech tag to each word. The authors have decided to use the SpaCy POS Tagger is more time efficient. SpaCy also provides word support vectors, whereas NLTK is not.

3.3.4. Twitter Tag Score (TTS)

Twitter Tag Score is the defined as the ratio computed from the occurrences of emoticons, user mentions and hashtags in a tweet as shown in Equation 1.

$$TTS = \frac{Sum(hasEmoticon(t),\ hasMention(t),\ hasHashtag(t))}{Count(tags)} \quad (1)$$

where, t - tweet instance
hasEmoticon(t) = 1 if emoticon is present and 0 otherwise
hasUsermention(t) = 1 if user mention is present and 0 otherwise
hasHashtag(t) = 1 if hashtag is present and 0 otherwise

These sets of features extracted are used to train ML models. Predicted classes by each model undergoes a process of hard voting to decide the final class of each of the tweets of the testing set.

3.3.5. Trigram TF-IDF

In the process of contextual information retrieval, the Auto-OffID uses TF-IDF features built with trigrams. It is a measure which reflects the quantifiable relevance intended to reflect degree of relevance a word holds in a collection or corpus with respect to a document. The term frequency (TF) in TF-IDF measures the frequency of a word w in a tweet whereas the inverse document Frequency (IDF) refers to the IDF the inverse of the document frequency which measures the informativeness of word w in the whole corpus.

3.3.6. Ensemble Classifier

The Auto-OffID System adopts a hard-voting approach in which each individual classifier will consider voting for a class and win the majority. In mathematical terms, the ensemble's predicted target label is the pattern in which independently predicted labels are distributed. For instance, consider that the Auto-OffID system uses 3 classifiers (1, 2, 3) and two classes (A, B), and after training, predict of class should be performed at a single point. Suppose the classifier 1 predicts class A, classifier 2 predicts class B and classifier 3 predicts class B. According to the hard-voting

approach, it could be inferred that 2 out of 3 classifiers predict class B, so class B is the ensemble decision.

3.4. Task B – Offensive Type Categorization

Tweets classified as offensive by the Offence Language Identification Module are further classified to determine whether a tweet is targeted or not. Auto-OffID adopts Latent Dirichlet Allocation (LDA) to generate a set of topics to identify the types of offenses and related keywords. LDA is a "probabilistic topic model" of a set of composites with parts of textual contents. It is mostly applied during natural language processing (NLP) and topic modelling tasks. As far as the topic modelling is concerned, the composites are twitter posts, and the parts are words and / or phrases (words / phrases in length refer to n-grams) for analysis.

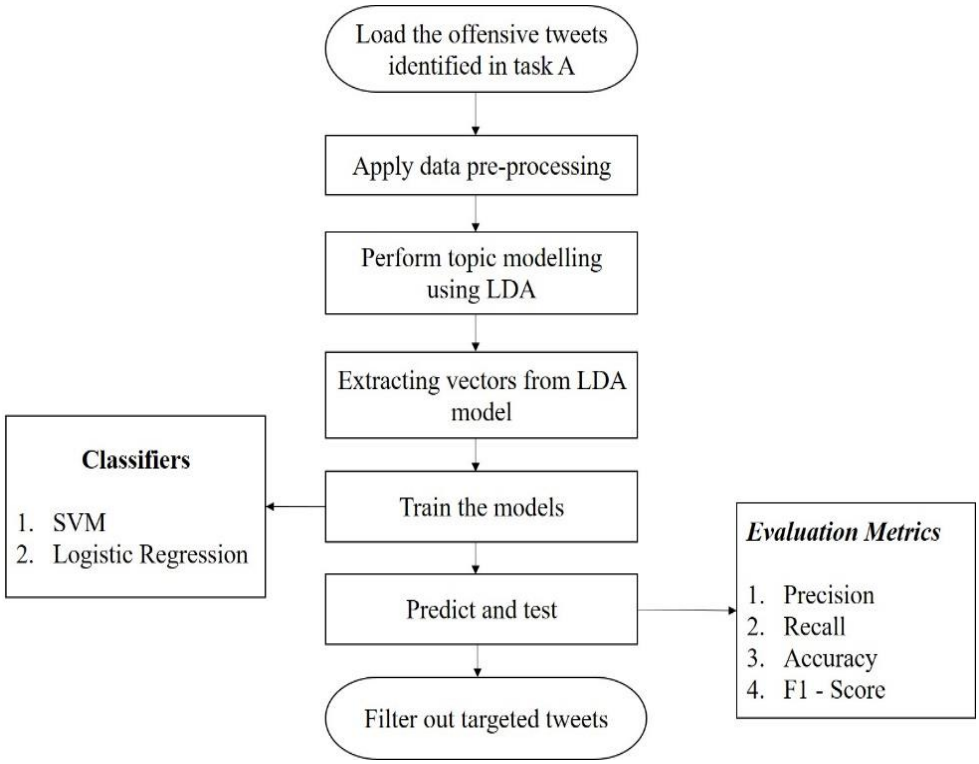

Figure 3. The process flow of Task B – Offensive Type Classification

The main objective of the LDA is to map all offensive tweets to themes in such a way that the words in each tweet are mostly captured by themes. The extracted feature vectors of offensive tweets are used to train and test using various classification, regression, and deep learning models. The topics that need to be modelled in by Auto-OffID system are targeted and untargeted and hence the topic count was fixed at 2. Top 10 words in each topic were found with assigned weights after application of LDA.

3.5. Task C – Offensive Target Identification

The task of offensive target identification is performed by analyzing only the targeted tweets in which the individuals and organizations targeted are identified using Named Entity Recognition (NER)

techniques. The Auto-OffID used the Stanford NER Tagger to discover the targeted individuals and organizations [11].

NER is a methodology for extracting information that intends to identify and categorize named entities in text into predetermined categories including names of persons, organizations, times, location of places, quantitative and monetary values, percentage distribution, etc. In our project, Stanford NER was used for identifying the names of the targeted organizations and individuals in the input set.

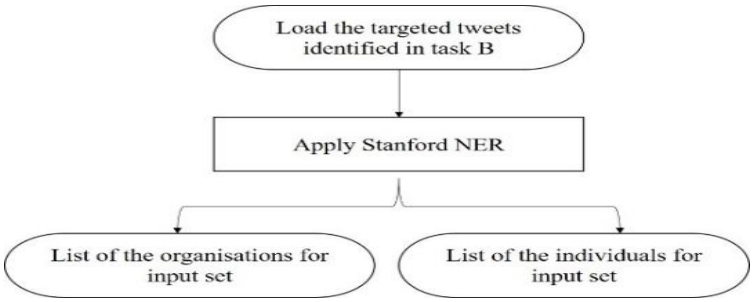

Figure 4. The process flow of Task C – Offensive Target Identification

4. Results and Discussions

The Auto-OffID system evaluated the resultant features based on the performance metrics such as True Positive (TP), True Negative (TN), False Negative (FN), False Positive (FP), accuracy, precision, recall, f1 score that reflects the performance of the system. The dataset described in Section 3.1 is divided into training and testing data in the ratio of 80: 20. The accuracy, precision, recall and f1 score can be computed as given in Equations 2,3,4 and 5.

$$Precision = TP/(TP + FP) \quad (2)$$

$$Recall = TP/(TP + FN) \quad (3)$$

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad (4)$$

$$F1 - Score = (2 * Precision * Recall)/(Precision + Recall) \quad (5)$$

4.1. Dataset Pre-processing

All the pre-processing rules discussed in the section 3.2 are applied on the tweets and a sample result of the pre-processing rules is presented in Table 3.

Table 3. The transformation stages of a tweet for each pre-processing rules.

Sample Tweet	@USER Love ya girl!! Never change!! Your perfect as you are! Check her out www.diva_smiles.com 100 wishes🙋🌹💋 #MuchLove	
Rule No	Pre-processing Rule	Resultant Tweet Data
1	Convert to lowercase	@user love ya girl!! Never change!! Your perfect as you are!check her out www.diva_smiles.com 100 wishes🙋🌹💋 #muchlove
2	Stop word removal	"['@user', 'love', 'ya', 'girl!!', 'never', 'change!!','your','are!','check','her','out', 'www.diva_smiles.com','100','wishes'🙋🌹💋',' #muchlove ']"

3	Punctuation removal	"['@user', 'love', 'ya', 'girl', 'never', 'change','your','are','check','her','out', 'www.diva_smiles.com','100','wishes'🙋🌹💋',' #muchlove ']"
4	URL and User Mentions replacement	"['@username', 'love', 'ya', 'girl', 'never', 'change','your','are','check','her','out','url','100' ,'wishes'🙋🌹💋',' #muchlove ']"
5	Date and Time Replacement	"['@username', 'love', 'ya', 'girl', 'never', 'change','your','are','check','her','out','url','100' ,'wishes',🙋🌹💋',' #muchlove ']"
6	Number replacement	"['@username', 'love', 'ya', 'girl', 'never', 'change','your','are','check','her','out','url','num ','wishes',🙋🌹💋''' #muchlove ']"
7	Hashtag replacement	"['@username', 'love', 'ya', 'girl', 'never', 'change','your','are','check','her','out','url','num ','wishes',🙋🌹💋','hashtag']"
8	Emoticon removal	"['@username', 'love', 'ya', 'girl', 'never', 'change','your','are','check','her','out','url','num ','wishes',hashtag']"
9	Stripping whitespaces	"['@username', 'love', 'ya', 'girl', 'never', 'change','your','are','check','her','out','url','num', 'wishes,'hashtag']"

4.2. Emoticon Sentiment Score

The Auto-OffID system used Vader Sentiment Analyzer for extracting the sentiment of the emoticons used in the tweets. The Vader provides four scoring namely positive, negative, compound, and neutral for analyzing a tweet instance. The positive, negative, and neutral scores are the percentage of text falling within these categories. The compound value is a measure which quantifies the sum of all the normalized lexicon ratings between -1 (extremely negative) and +1 (extremely positive). A tweet is of positive sentiment if the compound score is >= 0.05, neutral sentiment if it lies be-tween -0.05 and 0.05 and of negative sentiment if it is <= -0.05.

```
mysterious and I still think you are😅--- {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
None
@USER I'm sooo happy you are a Gator!! 🐊💙🐊💙🐊 {'neg': 0.0, 'neu': 0.62, 'pos': 0.38, 'compound': 0.6467}
None
@USER BITCH I FUCKING FELT THAT SHIT😤💀-- {'neg': 0.432, 'neu': 0.568, 'pos': 0.0, 'compound': -0.5859}
None
☹️---------------------------------- {'neg': 0.706, 'neu': 0.294, 'pos': 0.0, 'compound': -0.34}
None
```

Figure 5. Illustration for Emoticon Sentiment Analyzer with Emoticon Sentiment Scores

According to the dataset and chosen classifier properties, negative values cannot be used as a feature for training few models. So, the compound score reported by VADER is normalized to a

positive range as follows. Thus, if the compound score is negative (negative sentiment, i.e., <= -0.05) then the tweet is assigned a value 0. In case of the positive score (>= 0.05), the same score is taken as the feature value. If there occurs no emoji in the tweet instance, then the system assigns a value of 1. The neutral sentiment is also assumed as positive sentiment. A set of emoticon sentiment score predicted for the dataset in Auto-OffID system is displayed in Figure 5.

4.3. Positive to Negative Word Ratio

The Auto-OffID used MPQA Lexicon to identify the positive and negative words in the given tweet by performing set intersection with MPQA. The count of the positive and negative words in the tweet is computed and used to calculate positive/negative words ratio. The Auto-OffID system followed a simple logic that if there is no negative word in a tweet, it is assigned a value of 1. Thus, in the presence of both positive and negative words, the ratio is calculated as shown in Table 4.

Table 4. Positive to Negative words ratio for two sample tweets

S.No	Tweet	Positive – Negative Ratio
1	mysterious and I still think you are	1.0 (Absence of negative words)
2	@USER Please get out of my sight! Die Bitch	0.5 (Please is a Positive word. Die and bitch are negative. Hence ½ =0.5)

4.4. Gender Identification

The Auto-OffID system used POS tagger is used to identify the gender-based mentions of “he” and “she”. The occurrences of personal pronouns are counted to determine if the gender targeted is male or female as shown in Table 5.

Table 5. Gender Identification for sample tweets

S.No	Tweet	Gender
1	@USER She is sooooo pretty!!!!!	Female
2	@USER John erry is a loser and a traitor to our country. He doesn’t understand that we voted his party out bec we hated the iran deal. He is now working for iran against our gov.	Male
3	@USER @USER He is Satan	Male
4	@USER Please get out of my sight!Die Bitch	Neutral (Default – Male)

4.5. Twitter Tag Score (TTS)

The Auto-OffID system proposed a novel scoring to identify the targets and tags in a tweet and utilize it as an effective feature for offensive language identification. The computed TTS for a set of sample tweets is listed in Table 6.

Table 6. Twitter Tags Score (TTS) for sample tweets

S.No	Tweet	TTS
1	@USER Love ya girl!! Never change!! Your perfect as you are!Check her out www.diva_smiles.com 100 wishes🙆🌹💋 #MuchLove (NOT)	0.6 (HasUsermention=1, HasEmoticon =1, HasHashtag= 1) Total occurrences = 3/5 (1 usermention+ 3 emoticons + 1 hashtag) =5

2	@USER John erry is a loser and a traitor to our country. He doesn't understand that we voted his party out bec we hated the iran deal. He is now working for iran against our gov.	1 (HasUsermention=1) Total occurrences =1
3	@USER @USER He is Satan	0.5 (HasUsermention=1) Total occurrences =2
4	@USER Please get out of my sight!Die Bitch	1 (HasUsermention=1) Total occurrences =1

4.6. Task C – Offensive Target Identification

The Auto-OffID system identified targets like person and organization based on the user mentions available in the tweets. A sample result analysis will serve as a reference for Task C as presented in Table 7.

Table 7. Offensive Target Identified by Auto-OffID system using NER approach

Tweet	Targets Identified by NER
@USER If Jamie Oliver fucks with my £3 meal deals at Tesco I'll kill the cunt	Jamie Oliver – person Tesco – organization

4.7. Performance Metric for Auto-OffID Tasks

The Auto-OffID system implemented various state-of-art classifiers to evaluate the suitability of models in predicting offensive content. The task wise results for various models are presented in Figure 6 and 7. The maximum accuracy yielded by ML models was around 60-70%. In case of unstructured data as in Auto-OffID dataset and assumptions, it would be more pragmatic to approach the problem statement with reinforcement learning. To justify the assumption, the authors decided to explore DL models. With epoch count of 3 and softmax activation function, LSTM model outperformed ML models with an accuracy of 87.34%.

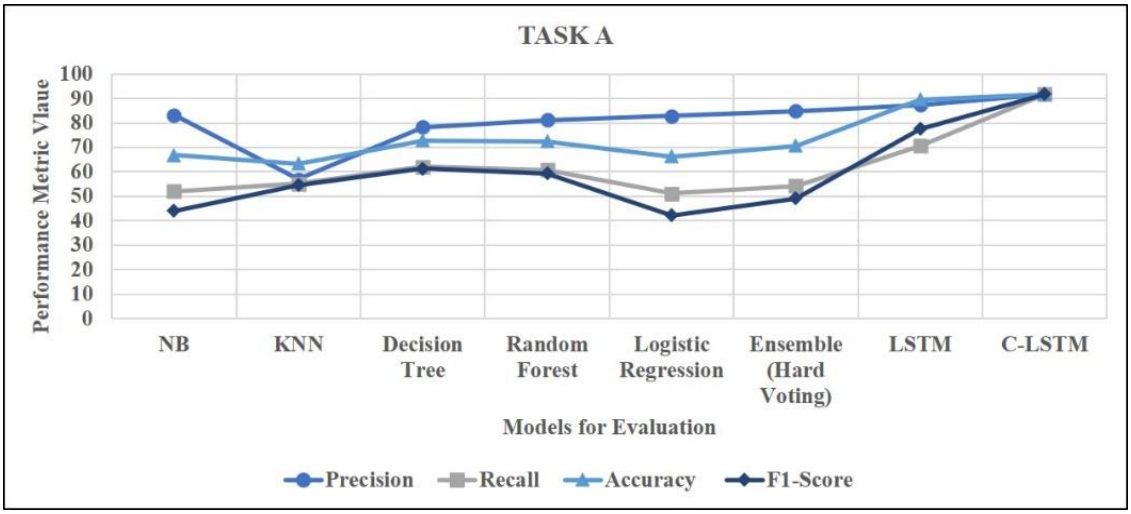

Figure 6. Performance of various models for Task A – Offensive Language Identification

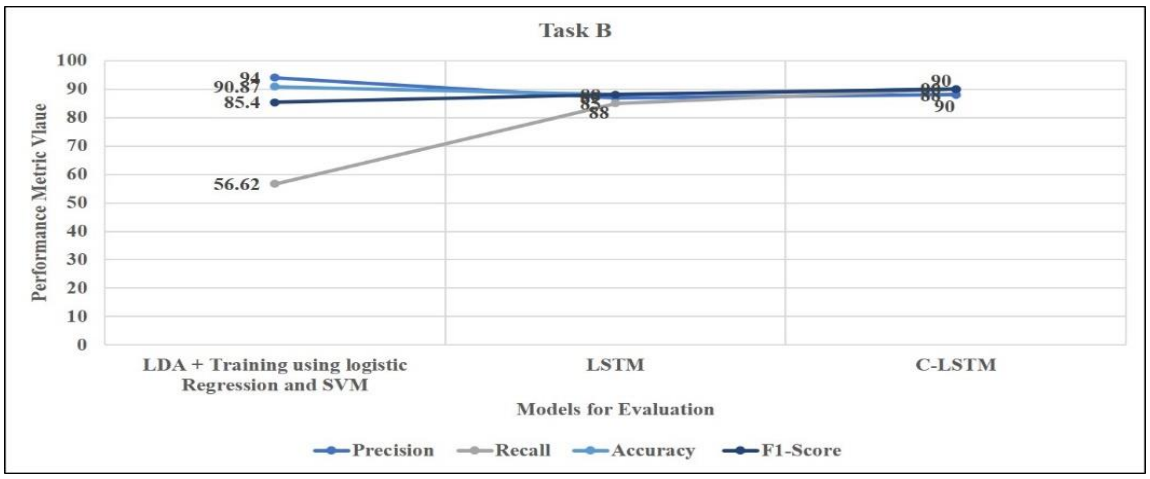

Figure 7. Performance of various models for Task B – Offensive Type Categorization

As the increase in the epoch count did not improve the accuracy loss significantly, Auto-OffID tried another DL model (C-LSTM). The offensive language identification is effectively performed by LSTM and C-LSTM models with significant results that are found to be better than that of machine learning models deployed in ensemble. The C-LSTM provided the best results among all the other models with a F1-score of 91.72. In Task B, topic modelling using LDA rendered the best results in terms of accuracy. But the authors decided to explore deep learning models to improve recall as LDA resulted in poor recall. C-LSTM provided the best results by faring well in all evaluation metrics.

The top 20 words that were identified to be present in offensive tweets in each levels of hierarchical classification are presented in Figure 8.

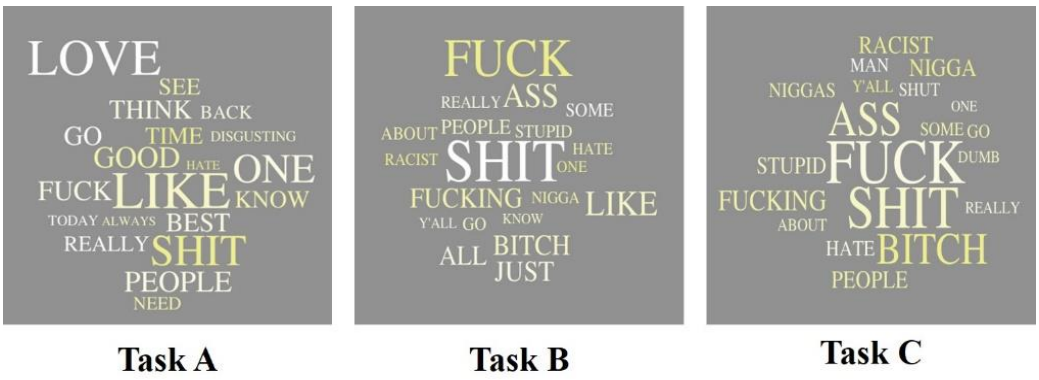

Figure 8. The most frequently used offensive words in the tweets at various levels of analysis

5. Conclusion

The Auto-OffID system was able to influence the potentiality of sentiment analysis by the proposed hierarchical classification approach in bullying detection on Twitter messages. The presented system experimented an ensemble method (hard voting) by deploying a variety of ML classifiers as voters to reduce the outliers and misclassifications in the process of offensive content identification. The state-of-art classifiers were close to 70% accurate. Based on an intuitive assumption that reinforcement learning could perform better in case of unstructured dataset such as [9, 12], it was found truly supportive. The authors empirically found that DL models performed better for our dataset and were able to nearly achieve a 90% accuracy. Therefore, the Auto-OffID System achieved an accuracy of 91.72% by C-LSTM for Task A – Offensive Language Identification and 90.87% accuracy from LDA

+ Logistic Regression on SVM for Task B - Offensive Target Categorization. It may however require a long-term synchronous system that could execute for a longer period to extract real time data from Twitter provided handling the constrained accesses to user tweets. The task of identifying potential bullies and bullying activities is difficult in Twitter as the tweets can be instantly removed by users even before it can be fully tracked by the analyser.

The limitation of this study is that only text content was analysed from the UGC. Apart from instant text messaging, cyber-bullying acts using audio, video, and images can also cause the victims to step forward towards some serious life-threatening course of action. Individuals has been bullied with disgusting or vicious audios or videos of any kind. As the upcoming work, the authors are aiming to adopt audio and video message identification techniques for cyber-bullying. To enhance the optimal detection of bullying content present in tweets, the research work could be extended by incorporating user-based context and sentiment related knowledge.

References

[1] Geetha R, Karthika S, Kumaraguru P. 'Will I Regret for This Tweet?'—Twitter User's Behavior Analysis System for Private Data Disclosure. *The Computer Journal*. 2020 May 9.

[2] Geetha R, Rekha P, Karthika S. Twitter opinion mining and boosting using sentiment analysis. In *2018 International Conference on Computer, Communication, and Signal Processing (ICCCSP)* 2018 Feb 22 (pp. 1-4). IEEE.

[3] Basile V, Bosco C, Fersini E, Debora N, Patti V, Pardo FM, Rosso P, Sanguinetti M. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation* 2019 (pp. 54-63). Association for Computational Linguistics.

[4] Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media* 2017 May 3 (Vol. 11, No. 1).

[5] Waseem Z, Davidson T, Warmsley D, Weber I. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint* arXiv:1705.09899. 2017 May 28.

[6] Sanchez H, Kumar S. Twitter bullying detection. ser. *NSDI*. 2011;12(2011):15.

[7] Kansara KB, Shekokar NM. A framework for cyberbullying detection in social network. *International Journal of Current Engineering and Technology*. 2015;5(1):494-8.

[8] Lee PJ, Hu YH, Chen K, Tarn JM, Cheng LE. Cyberbullying Detection on Social Network Services. *PACIS*. 2018;61.

[9] OffensEval 2019, https://sites.google.com/site/offensevalsharedtask/home

[10] Musto C, Semeraro G, Polignano M. A Comparison of Lexicon-based Approaches for Sentiment Analysis of Microblog Posts. *InDART@* AI* IA 2014 Dec 10 (pp. 59-68).

[11] Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R. Predicting the type and target of offensive posts in social media. *arXiv preprint* arXiv:1902.09666. 2019 Feb 25.

[12] Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*. 2018 Jul 20;13(3):55-75.