Detection of categorized abusive text in Social Media for English and dictionary-based search for Assamese language using CNN Application

Suhangee Dawo[1], Nayan Jyoti Kalita[2]

[1]Department of Royal School of Information Technology, Royal Global University, Guwahati, Assam-785001, India,

[2]Department of Computer Science and Engineering, Royal Global University, Guwahati, Assam-785001, India,

---***---

Abstract -In today's world Social Networking sites likeFacebook, Twitter, Reddit etc. plays a really significant role. Since people of various cultures and academic backgrounds share their ideas and perspectives on social media therefore detection of toxic language in user generated online content has become the foremost increasing issue in recent years.

Therefore, during this paper, we have experiment by selecting some random comments from Facebook for both English and Assamese language for developing a corpus from Facebook comments. We have experiment those comments using deep learning approach and identify what proportion of offensive the text contains, by categorizing as "toxic", "severe_toxic", "obscene", "threat", "insult", "identity_hate". Also, our system detects the abusive comment based on the abusive words in the dictionary.

Key Words: Natural Language Processing, DeepLearning, Convolutional Neural Network, Offensive Language Detection, Facebook comments, slang words

1.INTRODUCTION

Today, almost everyone features a social profile online, where they share all their personal information for anyone to view without giving it a second thought. Online interactions like VoIP calls, text chats, sharing ideas/opinions on a public forum are more prevalent in recent times. While there are useful conversations, similarly many hates and abuse online interactions are equally available. In a study we found that nearly 1 in 5 children have daily faced hate and abuse threats online. The most disturbing fact is that 1 in 4 children have been a victim of cyberbullying. This is clear indication that these online interactions should be controlled to eliminate or a minimum to attenuate online harassment. The textual contents on online social media mainly face book are highly unstructured, informal, and sometimes misspelled, so existing research on message-level offensive language detection cannot accurately detect those offensive contents. Although a person could recognize the useless annoying texts among the useful ones, it's not a simple task for computer programs. It has become a difficult challenge to verify whether the comment is an offensive or not offensive due to the ever-growing number of messages. The only way we can solve this issue without compromising individual privacy, is by automating the method of understanding the language to detect potential abuse during a comment and reporting users if they're found to be misbehaving. Since our task is to solve the problem of abuse detection in social media, so we havechosen Facebook as our domain which is the most widely used online social media platform.

Abuse Word	Comment
suwaseleka	BJP r **suwaseleka**. **Suwaseleka**kam nokoriba.
hubidhabadi, sarthopor, luvia	Nojoke etia asomiya buli kobo laz lagibo dhorise...kio ba! Hosai manuh eman **hubidhabadi, sarthapor, luvia** keneke hobo pare...!

Table1: Sample representation of data in Assamese comments

Abuse Word	Comment
cheapass	What a **cheapass**!
shameless	**Shameless**dirtypolitician What a **shame** for horse trades, they must apologize to the nation.
bloody, fool	**Bloodyfool** is talking shit!

Table2: Sample representation of data in English comments

1.1 Reason to choose Facebook

- The reason we chose Facebook is because, it is the easiest platform to express opinions about various posts as comments which is sometimes valuable.
- From a study we found that 900 million users are monthly active on face book, so it's obvious that the number of users is increasing day by day hence the diversity of opinion is also increasing.
- Another reason is that it reaches to people very fast.

2. LITERATURE REVIEW

Dawei Yin and his colleagues [1] explored a context-based approach back in 2009 for abuse classification. They analyzed content features, sentiment features and context features of a comment and documented that a supervised-learning approach using an SVM classifier with n-grams performs better than conventional methods using TF-IDF.

Another approach is inspired from commercial rule-based spam filtering using blacklists. This work is seen in the work of S.O. Sood, et.al in 2012 [2]. They have used blacklists along with an edit distance metric and showed that their approach was better for online profanity detection compared to previous approaches that has been done.

In a recent publication Yahoo [3] seems to automatically moderate online abuses. Yahoo seems to train a classifier using supervised machine learning techniques for using a combination of parser, lexical and syntactic features.

Google Jigsaw [4] recently published a paper that used data from the Wikipedia Detox project. This paper is based on abuse classification where they discuss the effectiveness of logistic regression and multi-layer perceptron. This paper also compares their approach with a human baseline. The Methods discussed in this paper is actually put to use in Google's Perspective API – an API that takes in a piece of text and returns if the text is an abuse. It is been found that all the published works mentioned above seems to focus more on the data through progressive feature selection. Less effort has been found to explore deep learning techniques to detect abuses in social media comments. Deep learning methods frequently require little to no feature engineering. However, it seems that there has been some significant work done in the past that employs deep learning for text but for completely different problems.

Ji Young Lee and his colleague in the year 2016 [5] worked on building a sequential short-text classifier. Their classifier employed the use of a recurrent neural network. They proved that their model achieves state-of-the-art results on three different datasets for dialog act prediction.

Kim Yoon, in 2014, [6] explored convolutional neural networks for sentence classification. It has been found that his architecture proved to be effective for quite a lot of sentence classification tasks including sentiment analysis. He tested four CNN variations, and showed that CNN models could outperform previous approaches for several classification tasks. He tested the classification problems by taking robust datasets.

Ji Ho Park and Pascale Fung [7] explored a two-step approach of performing classification on abusive language and then classifying into specific types and they compared it with one step approach of doing onemulticlass classification for detecting racist and sexist language on social media platform. They proposed Hybrid CNN which takes both character and word features as input.

Imon

Imon Banerjee et al.2018 [8] proposed two distinct deep learning models CNN Word-Glove and Domain phrase attention-based hierarchical neural network (DPA-HNN), for synthesizing information on pulmonary emboli (PE) from clinical thoracic CT free-text radiology reports. They trained those models on Stanford dataset and are tested on four major healthcare centers dataset. They performed comparative analysis on deep learning models against the current state-of -the-art PEFinder as well as with traditional machine learning models SVM and Adaboost with bag-of-words features. Their work proposed experimental insight on the proficiency of CNN and RNN to automatize the analysis of unstructured imaging reports.

One of the recent works on NLP is seen in 2019 by Shah Zaib and Muhammad Asif and Maha Arooj [9]. They performed an experiment based on word wise and sentence wise tokenization with the help of existing sentiment lexicons. They have collected a dataset consisting of comments and replies by users on Face book to perform descriptive analysis. And concluded that sentence wise approach performed better than word wise approach. These are some of the common mainstream uses of CNN for sentence classification today.

2.1 Challenges Faced

Detecting an abusive language in social media is a difficult problem. Especially in face book data, the local netizens who post comments with informal language makes the researchers have to perform special techniques in normalizing the data. Detecting an abusive language in Assamese language in Facebook becomes more difficult because many netizens in Assam use abusive words in a foreign language in their conversations, both in the context of jokes or to curse someone indirectly. The use of abusive words in a foreign language is not only in the formal form, but informal ones. For example, many Assam netizens type 'Fak yu!' to say 'Fuck you!'. Not only using abusive words in foreign language, many Assam netizens also usually use abusive words in their local language. The examples of abusive words in Assamese local language that often used by Assam netizens in their conversation are kukur (assamese language, means dog), burbok (assamese language, means idiot) etc.

We aim to solve this problem by solving the most important and difficult part in the aforementioned process, understanding a comment and telling if it is safe or not. This requires a great deal of natural language processing / understanding and a significant amount of machine learning and deep learning approaches. An efficient implementation of such a system takes during a piece of text (a comment) as input and produces a binary label, “toxic”, “severe_toxic”, “obscene”, “threat”, “insult”, “identity_hate”. As output.

3. METHODOLOGY

The main aim of this thesis is to explore abusive threats of users in comments by manually creating a corpus of comments collected from Facebook and to preprocess (slang words) the collected data to build up a proficient calculation for felling investigation and predicting the sentiment of comments as ["toxic", "severe_toxic", "obscene", "threat", "insult", "identity_hate"]. Our work takes a more recent, relevant and manually annotated dataset of online user comments from Facebook post and explore a deep learning architecture. We also craft a convolutional layerto capture higher level representations. Very less amount of work has been done to detect abuses on these higher-level representations. For example, “I love to play cricket” can be chunked as [“I”, “love”, “to”, “play”, “cricket”]. We used Convolutional Neural Network (CNN) as a baseline to get the result of our experiment.

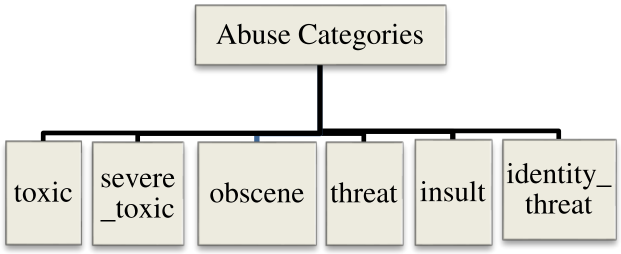


Figure1: Representation of basic abuse categories

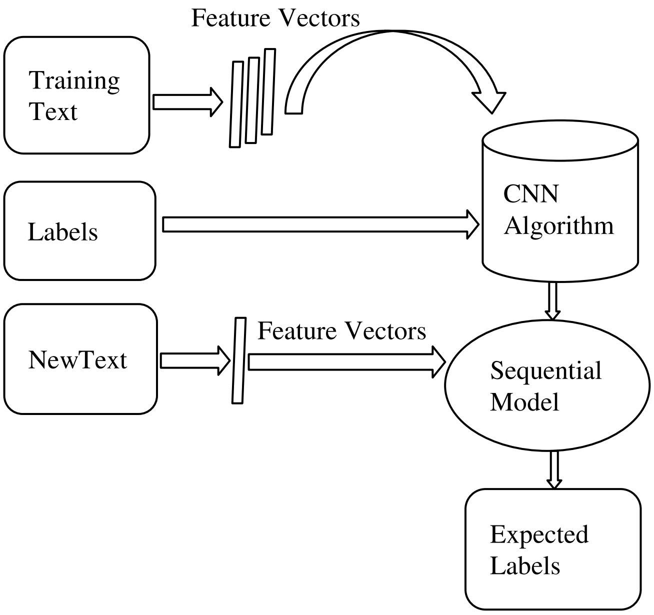


Figure2: Block Diagram of the Process Flow

3.1 Data Collection:

The process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes is referred to as data collection. In this work we have collected data from Facebook by creating a Facebook user account and then created a corpus manually by collecting comments of both English and Assamese from Facebook post of various fields (Groups, Page, User Account etc.). A rich lexicon dictionary is created for enhanced processing of the algorithm. The goal for all the data collection is to tackle the problem of an online abuse detection and categorize the data using applications of CNN.

Script	**No. of comments**
Assamese	2,000
English	4,000
Total	6,000

Table3: Text Statistics in corpus

While collecting data from Facebook various pages and posts were identified and crawled. It includes pages and posts of the below mentioned type:

- Political Parties like BJP, CONGRESS, RSS etc.
- News Websites like REPUBLIC TV, ABP News, NDTV etc.
- Personal User Account of Various Posts.
- Bollywood Pages etc.

After the collection of data, it was labelled as [“toxic”, “severe_toxic”, “obscene”, “threat”, “insult”, “identity_hate”].

3.2 Data Processing:

One of the most important tasks of pre-processing is to filter out useless and unwanted data. In natural language processing (NLP), useless or unwanted words are always considered as stop words. We would never want these useless words to take up space in our database, or take up valuable processing time. For this, we can remove them easily by storing a list of words that we consider to as stop words. In python, NLTK (Natural Language Toolkit) has a list of stop words stored in 16 different languages. To remove this stop word, we have used NLTK library to import those stop words. And modified the list by adding words of our choice in the text file in the stop words directory. Before processing a natural language, we need to identify the words that constitute a string of characters. That is why tokenization is considered as the basic step towards proceeding with NLP (text data). This step is important because the meaning of the text could be easily interpreted by analyzing the words present in the text. Considering the string: “Burn them in hell.” We get [‘Burn’, ‘them’, ‘in’, hell’]. There are various uses of doing this. We use this tokenized form to,

- Count the number of words in the text
- Count the frequency of the word, that is, the number of times a particular word is present.

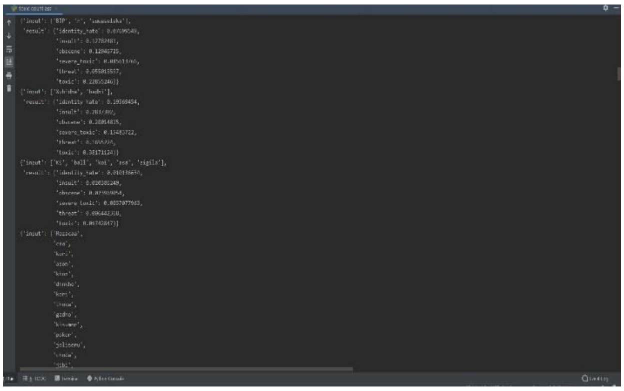
Wikipedia is the largest opensource online encyclopedia where more than 200 different languages are found, it is the best source for Natural Language Processing Task (NLP). In this work, we use 1 million-word vectors trained on Wikipedia using fastText (T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin.2018 Advances in Pre-Training Distributed Word Representations). These word vectors are in binary format. fastText library makes the word representation and sentence classification efficient. The models are trained on CBOW method using ngram features.

4. IMPLEMENTATION

ForImplementation we used Python programming language. It is a general-purpose interpreted, interactive, and high-level programming language which supports modules and packages to encourage code reusability. PyCharm is an integrated development environment (IDE) used as the front end and Python 3 as the backend. To tokenize the entire data, we used Keras preprocessing text Tokenizerand labeled the data as “toxic", "severe_toxic", "obscene", "threat", "insult", "identity_hate". We divided the data into test set and train set and feed to the convolutional layers. This layer comprises of neuron that scans the input for pattern and perform necessary operation (conv1d, max pooling, concatenate, dropout, dense, fully connected layer). During the training phase CNN automatically learns the value of it filters based on the task we want to perform. Our proposed model ‘Sequential Model’ performs the prediction and gives the output. Once the Sequential model is built, it behaves like a Functional API model.

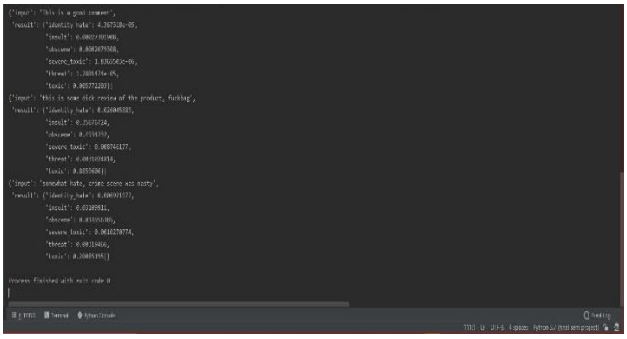
5. RESULT AND DISCUSSION

For this work, we have manually collected a Corpus from Facebook post and created two slang words dictionaries for assamese and English words. We installed the necessary testing libraries and input it into the program. Training of the data set is done by providing the processed data to the training system. For each sentence, we compare it with abusive words in the dictionary. The sentences are then checked as ["toxic", "severe_toxic", "obscene", "threat", "insult", "identity_hate"]. Finally, based on the count, we get the result of how much proportion of abuse the comment contains, by detecting the slang words present in the sentence. The result shows both for Assamese and English Sentences.



```
{'input': ['BJP', '[illegible]', '[illegible]'],
 'result': {'identity_hate': 0.07699549,
            'insult': 0.12782481,
            'obscene': 0.12846725,
            'severe_toxic': 0.0456117[illegible],
            'threat': 0.055813537,
            'toxic': 0.22855246}}
{'input': ['Xohidhe', 'hadhi'],
 'result': {'identity_hate': 0.19569454,
            'insult': 0.2817[illegible],
            'obscene': 0.28814835,
            'severe_toxic': 0.13483722,
            'threat': 0.1655224,
            'toxic': 0.38171124}}
{'input': ['Ki', 'ball', 'koi', 'asa', 'zigila'],
 'result': {'identity_hate': 0.010136694,
            'insult': 0.020381249,
            'obscene': 0.0[illegible],
            'severe_toxic': 0.0037077963,
            'threat': 0.006442318,
            'toxic': 0.06742847}}
{'input': ['Mozacaa',
           'cea',
           'kori',
           'ason',
           'kinn',
           'dinshe',
           'kori',
           '[illegible]',
           'gadho',
           'kinumne',
           'pokar',
           'jolisoru',
           'chala',
           'jibi',
```

Figure3: Output for Assamese comments



```
{'input': 'This is a good comment',
 'result': {'identity_hate': 4.367318e-05,
            'insult': 0.00022401900,
            'obscene': 0.0002079308,
            'severe_toxic': 1.0365503e-06,
            'threat': 1.2881474e-05,
            'toxic': 0.005772203}}
{'input': 'this is some dick review of the product, fuckbag',
 'result': {'identity_hate': 0.020045889,
            'insult': 0.35676724,
            'obscene': 0.4934232,
            'severe_toxic': 0.008746177,
            'threat': 0.0031024814,
            'toxic': 0.8859686}}
{'input': 'somewhat hate, crime scene was nasty',
 'result': {'identity_hate': 0.006921977,
            'insult': 0.03109911,
            'obscene': 0.010956406,
            'severe_toxic': 0.0010270774,
            'threat': 0.0019466,
            'toxic': 0.20085195}}

Process finished with exit code 0
```

Figure4: Output for English comments

6. CONCLUSION AND FUTURE WORK

Natural Language Processing is one of the emerging fields that is mainly used in many application areas. Its scope is increasing day by day. This project aimed at categorizing Facebook comments according to a new set of selected categories as ["toxic", "severe_toxic", "obscene", "threat", "insult", "identity_hate"] and detects the Assamese and English slang words from the sentence that incites violence towards people. A dictionary of slang words is created for enhanced processing of the algorithm. We conclude that stop words are not always meaningless, as they play a major role in improving the performance of some classification.

As a future extension of this work, similar dataset for other language can be built to explore depth into Assamese language besides English. NLP in deep learning using CNN application can be applied to further more datasets for better analysis. Also, various categories, and other preprocessing approaches based on other features can be proposed. The accuracy of the algorithm can be checked by collecting the comments from different blogs and sites such as INSTAGRAM, REDDIT and apply different types of classifiers on the dataset and their accuracy can be compared to know which classifier is helpful for achieving better efficiency.

REFERENCES

[1] Yin D., Xue Z., Hong L., Davison B. D., Kontostathis A., Edwards L., « Detection of harassment on Web 2.0 », WWW Workshop: Content Analysis in the WEB 2.0, p. 1-7, 2009.

[2] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. Proceedings of the Content Analysis in the WEB 2 (2009), 1–7.

[3] S. Sood, J. Antin, and E. Churchill. Profanity uses in online communities. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 1481--1490. ACM, 2012.

[4] Kim Y. Convolutional Neural Networks forSentence Classification. 2014; Zhang Y, Wallace B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. arXiv preprint arXiv:151003820. 2015; PMID: 463165.

[5] Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. arXiv preprint arXiv:1603.03827 (2016)

[6] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In Proceedings of TRAC.

[7] Hao Chen, Susan Mckeever, and Sarah Jane Delany. 2017. Presenting a La-belled Dataset for Real-Time Detection of Abusive User Posts. In Proceedings of WI '17, Leipzig, Germany,August 23-26, 2017, 7 pages.

[8] I Banerjee, Y Ling, MC Chen, SA Hasan, CP Langlotz… - Artificial intelligence in medicine, 2019.Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification.

[9] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and RiteshKumar 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In Proceedings of SemEval.