Comparative analysis of bangla online abusive comment classification

Health & Medicine, Addiction

As internet access rises, harassment through internet or other electronic media have expanded rapidly. In spite of the fact that there are a number of methods available to flag and report toxic and abusive comments online in English, there’s none in Bangla language. In this work, we are not only focusing to detect obscene and abusive remarks but also threats, hate speeches, racial slurs in Bangla. We used n-gram with TF-IDF variants and other different classifiers, LSTM to classify toxic comments and furthermore created a dataset of Bangla user-generated offensive comments, which itself is first of its kind. Keywords: abusive remarks, bangla online harassment, hate speech, NLP, toxic commentsI.

Introduction

Everyday people deal with online user-generated content in social media comments, public forums, online discussion group etc. Abusive and toxic comments on these forums are a major problem which hampers effective online conversations. Online ridicules can also take a huge toll on users mental health.

Bangladesh, specifically, has been confronting this issue as far back as the web turned out to be broadly available [1]. It’s been nothing unanticipated that by and large the young in Bangladesh has tumbled to the trap of online induction due to the genuine mitigation of social foundation. This unites using slangs and savage language on the web. Therefore, web tormenting is rising exponentially as an expansive number of user belonging to this portion.

An abusive comment can be of various types, for instance, it might be obscene, hostile, racist, xenophobic, threatening, misogynistic. Women, in particular, are the biggest victim of cyber-bullying in Bangladesh [2][16]. Next, to women, school children are the frequent target of cyber bullies. [15] With the number of internet users rising in Bangladesh [20] and other Bangla speaking regions [21] it is now essential to ensure quality online conversation. To make an online conversation more civil, different machine learning and deep learning algorithms have been used to detect and flag abusive comments in various languages. Automatically detecting toxic comments can also save the hassle of manually banning certain users after a report. Unfortunately, there is none developed in Bangla in spite of the fact that it is the seventh most spoken language in the world [3]. Abusive comments in Bangla can be categorized in several ways:

1. Comments with profanity, obscene words.
2. Racial slurs, extremist comment against certain religious group, derogatory remarks
3. Threatening or hostile comments
4. Crude commentsIn this paper we implemented several algorithms to figure out best performing one to determine abusive bangla comments.

Related Work

Several works have been done to classify online abusive comments in different languages. The vast majority of these works have been done in the English dialect. Reynolds, Kelly, April Kontostathis, and Lynne Edwards have

used a language-based model of identifying cyberbullying. Nobota, Chikashi et. al have performed different syntactic and embedding features and combined them with standard NLP features which outperformed deep learning model.

They also used character n-gram to tackle noisy and obfuscated words with characters. [7] One recent work done by Mohammed, Fahim showed that preprocessing data for toxic comment classification does not improve the overall performance. [5] Che, Hao, Susan McKeever, and Sarah Jane Delany found that in FastText classifier with pre-trained word embeddings does not necessarily outperform standard neural network.[8] X. Zhang, J. Zhao, and Y. Lecun, compared character level CNN against traditional models such n-grams with TF-IDF variants, deep learning models etc. on different datasets to classify text.[9] Georgakopoulos, Spiros V., Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos compared CNN against the traditional bag-of-words approach which was effective in classification. Pedro M. Sosa on other hand combined CNN and LSTM to determine sentiment analysis in twitter dataset which performed 2. 7%-8. 5% better than regular models.[10] Dinakar, Karthik, Roi Reichart, and Henry Lieberman showed that binary classifiers outperformed multiclass classifier when it comes to detecting online cyberbullying. Interesting work has been done by Hosseini, Hossein, Sreeram Kannan, Baosen Zhang, and Radha Poovendran where they presented that certain bastardization of words can deceive automated online toxicity detection. Chu, Theodora, Kylie Jue, and Max Wang compared among LSTM, CNN with word embeddings and CNN character embeddings, and

came to the conclusion that CNN with character embedding performed the best.[14]

Although some works have been done in Bangla sentiment analysis and phrase level polarity identification [18] to our knowledge, no work has been done before to classify and detect abusive comments in Bangla Language.

Challenges

Bangla abusive comment detection is challenging due to many reasons. Although several works have been done to classify abusive comments in other languages before, none was done for Bangla. The lack of available dataset is assumed to be the main reason for it. It’s very common to use code-mixed language in social media in Indian sub-continent.[19] People in online forum post both Bengali and English-Bengali code-mixed comments, which is phonetically writing Bangla in English alphabets. A word in Bangla can be spelled in several ways in English alphabet. Bangla alphabet also contains alphabets with similar sound which leads to multiple spelling for same words other than correct one.