International Journal of System Dynamics Applications
Volume 11 • Issue 2

BERT-BU12 Hate Speech Detection Using Bidirectional Encoder-Decoder

Shailja Gupta, Manav Rachna University, India
Manpreet Kaur, Manav Rachna University, India
Sachin Lakra, Manav Rachna University, India

ABSTRACT

Transfer learning models have been known to exhibit good results in the area of text classification for question-answering, summarization, and next word prediction, but these learning models have not been extensively used for the problem of hate speech detection. The authors anticipate that these networks may give better results in another task of text classification (i.e., hate speech detection). This paper introduces a novel method of hate speech detection based on the concept of attention networks using the BERT attention model. The authors have conducted exhaustive experiments and evaluation over publicly available datasets using various evaluation metrics (precision, recall, and F1 score). They show that the model outperforms all the state-of-the-art methods by almost 4%. They have also discussed in detail the technical challenges faced during the implementation of the proposed model.

KEYWORDS

Attention Networks, Classification, FusedAdam, Gelu, Hate, Machine Learning, Transfer Learning, Uncased

INTRODUCTION

The right to speak and the right to express oneself freely are two of the various rights provided by the constitution of countries. People have been enjoying these rights by expressing their sentiments, opinions and their feelings with each other. Modern technology provides humans with social networking sites and microblogging sites to understand each other's culture and emotions even while living in various parts of a country or a world. However, people have also started misusing these platforms by trying to oppose the opinions or thoughts of other users by using abusive language, offensive words, and aggressive sentences on these platforms, as part of their communication. These platforms have also been used in recent times by religious groups, political parties and bullies to oppose others and improve their image among the general public for their own interest by posting hateful, offensive and abusive contents to spoil the image of opposing parties or groups. The younger generation which is tech-savvy and has not developed the understanding of worldly ways, are highly affected by reading and viewing such content.

According to statistics related to Hate Crime, (2019), there have been 103,379 hate crimes recorded in the year 2018-19 in England and Wales, where the majority have been race-related (76%), 56% of hate crimes recorded by police have been for public offenses and (36%) have involved violence. 5% of these crimes have been recorded as criminal damage and arson. A campaign advisor of a non-profit organization has reported that 73% of people with learning disabilities and autism have

DOI: 10.4018/IJSDA.20220701.oa4

experienced hate crime. Based on Hate Crime Statistics, (2018), the statistics collected by the FBI reported 7036 hate incidents involving 8646 victims, where 59.6% of hate crime has been reported under the categories of race, ethnicity, and ancestry bias, 0.7% of hate crimes reported have been gender-related, while the contribution of hate crimes against individuals with disabilities has been reported as 2.1%. 2.2% of hate crimes have been found to be related to gender identity, 16.7% of hate crimes have been found to be related to sexual orientation while hate crimes falling into the category of relational border constitute 18.7%.

Social networking sites are also gaining a bad reputation due to the presence of such content. There are many challenges faced in implementing hate speech detection by researchers in the field of developing automated hate speech detection methods, which make it difficult to assess an individual's contribution towards the problem. The reasons for the challenges in the hate speech detection problem are varying definitions of hate speech, limitation of data or content availability for the training and testing of these systems, casual approach for framing of the sentences, lack of grammar correctness, syntactic structure and comparative evaluation among the datasets.

For these reasons, governmental and social networking sites are trying to find solutions for reducing and removing hateful content from these platforms. Deriving from an article of the Council on Foreign Relations & United Nations Strategy and Plan of action on hate speech, (2019), social media agencies are investing hundreds of millions of Euros, along with time, and staff known as content moderators to combat the issue of hate speech detection by manually reviewing content present online and by detecting material that is not fit to be viewed. The basic problem of the detection of hate speech has been the understanding of the definition of hate speech as it can vary from person to person. The authors have attempted to understand the definition of hate speech by understanding its different terminologies.

Hate Speech

Hate can be expressed in many forms. It is difficult to identify if a part of a sentence contains hateful content or not, merely by reading it. The understanding of hate in hateful sentences is important and has been explained by different sources like ILGA, Facebook, YouTube, Twitter and other European countries, which are responsible for maintaining a code of conduct. Twitter, (2019); Nobata et al. (2016) & ILGA, (2016) has termed "hate" as words that incite discrimination, hostility, violence and lead to threatening or direct attack towards a person, people or a group of people based on certain actual or perceived attributes like age, sexual orientation, race, disability, gender, ethnicity, religious affiliations, disease, national origin, veteran status or gender identity. Social networking platforms like Facebook differentiate hateful from not hateful content by allowing content like standup comedy, jokes, lyrics of songs that might be considered as an attempt to express hateful words among others, but perceived as the bad taste of the authors or speakers. The presence of hateful content that criticizes a nation on its views is also considered as non-hateful but if the hateful comments or content are for a certain community or a group of people, it is considered as hate, as stated by YouTube, (2019). In brief as termed by Fortuna & Nunes, (2018), "hate" has specific targets and is used to promote violence or hate. The only purpose of "hate" is to attack or diminish a particular group of people.

Till now the problem of hate speech has been tacked using various deep learning models but still a lot of scope can be seen in improving the performance of the model. The rest of this paper is outlined as follows. Section Related Work discusses the literature survey carried out for the task of hate speech detection. The methodology applied for the task of hate speech detection has been discussed in Methodology Section. The Methodology section discusses the libraries used in the experiments conducted, the pre-processing of the dataset, the fine-tuning of the BERT model for classification followed by training, compilation, and optimization of our model. Section Experiment Conducted elaborates the experimental settings and discusses different comparative models that have been considered while evaluating the proposed model. The experimental results obtained from our proposed model, based on different evaluation metrics on four publicly available datasets have been