Hate Speech Detection As A Chrome Extension

Dhar Padma Patanjali
K A Meghashree

Literature Survey

Detecting hate speech is very important in various fields like controversial event finding, reporting and blocking users that spread hate and misinformation amongst the masses. The complexity of the task is very challenging as we need to take care of natural language constructs.There are many baseline methods for detecting hate speech.

The most naive methods simply suggest classifying words as good or bad. So if anyone uses a word/words that are included in the bad word dictionary, they would be informed of it. There is presently a google chrome extension working mostly on this principle called Hate Free. The developers have maintained bad words and other words which if appear along with it may be hurtful to some section of the society.

We read some technical papers about various hate speech detection techniques. They are as follows.

1. **HateMiners: Detecting Hate Speech Against Women**

In this paper, the authors have classified posts as misogynistic in twitter data set by generating sentence embeddings which generated a vector representation of the text. TFIDF vectors were generated on the preprocessed text followed by BoWv(Bag of words vector) for each tweet. These were passed through a logistic regression classifier to classify tweets as misogynistic.

2. **Hate Speech Classification in Social Media Using Emotional Analysis**

In this paper, the authors have presented a combination of lexicon-based and machine learning approaches to predict hate speech contained in a text, using an emotional approach through sentiment analysis.
Using the emotional information contained in text helps to increase the accuracy on hate speech detection. This analysis still has limitations that lead to exciting future research directions. It is reasonable to question the definition of hateful content, in the sense that it is not clear what is the threshold a published text shared in social media has to violate to be considered hateful due to the subjectivity of the definition of hate-speech.

3. **Identification of Hate Speech in Social Media**

The authors in this paper have spoken about various models that can be used to detect hate speech including various lexical analysis model in which machine is fed patterns of languages to detect hate words and negative polarity words. The second approach they looked into was various machine learning techniques. Logistic regression with L2 regularisation and 10 fold cross validation was used for this purpose. They also looked into hybrid methods which used the combination of 3 different classification features (Maximum Entropy, SVM and Random Forest). They however found that the best performance was obtained on using the Naive Bayes classifier with TFIDF features. Thus, they used both supervised as well as unsupervised methods to get results for detection of hate speech.

4. **Hate speech detection: Challenges and solutions**

As online content continues to grow, so does the spread of hate speech. This paper identifies and examines the challenges faced by online automatic approaches for hate speech detection in text. Among these difficulties are subtleties in language, differing definitions on what constitutes hate speech, and limitations of data availability for training and testing of these systems. Furthermore, many recent approaches suffer from an interpretability problem—that is, it can be difficult to understand why the systems make the decisions that they do. The paper proposes a multi-view SVM approach that achieves near state-of-the-art performance, while being simpler and producing more easily interpretable decisions than neural methods. Both technical and practical challenges for this task are discussed as well.

References

1. HateMiners: Detecting Hate Speech Against Women
   https://arxiv.org/abs/1812.06700

2. Hate Speech Classification in Social Media Using Emotional Analysis
   https://ieeexplore.ieee.org/document/8575590

3. Identification of Hate Speech in Social Media
   https://ieeexplore.ieee.org/document/8615517

4. Hate speech detection: Challenges and solutions
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6701757/