Hate Speech on Twitter A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection.

ABSTRACT

With the rapid growth of social networks and microblogging websites, communication between people from different cultural and psychological backgrounds became more direct, resulting in more and more “cyber” conflicts between these people. Consequently, hate speech is used more and more, to the point where it became a serious problem invading these open spaces. Hate speech refers

to the use of aggressive, violent or offensive language, targeting a specific group of people sharing a common property, whether this property is their gender (i.e., sexism), their ethnic group or race (i.e., racism) or their believes and religion, etc. While most of the online social networks and microblogging websites forbid the use of hate speech, the size of these networks and websites makes it almost impossible to control all of their content.Therefore, arises the necessity to detect such speech automatically and filter any content that presents hateful language or language inciting to hatred. In this paper we propose an approach to detect hate expressions on Twitter. Our approach is based on unigrams and patterns that are automatically collected from the training set. These patterns and unigrams are later used, among others, as features to train a machine learning algorithm. Our experiments on a test set composed of 2 010 tweets show that our approach reaches an accuracy equal to 87.4% on detecting whether a tweet is offensive or not (binary classification), and an accuracy equal to 78.4% on detecting whether a tweet is hateful, offensive or clean (ternary classification).

**EXISTING SYSTEM**

* Gitari et al. [15] extracted sentences from some major “hate sites” in United States. They annotated each of the sentences into one of three classes: “strongly hateful (SH)”, “weakly hateful (WH)”, and “non-hateful (NH)”. They used semantic features and grammatical patterns features, run the classification on a test set and obtained an F1-score equal to 65.12%.
* Nobata et al. [16] used lexicon features, n-gram features, linguistic features, syntactic features,pretrained features, “word2vec” features and “comment2vec” features to perform the classification task into two classes, and obtained an accuracy equal to 90%.
* Kwok et al. [17] targeted the detection of hateful tweets against black people. They used unigram features which gave an accuracy equal to 76% for the task of binary classification. Obviously, the focus on the hate speech toward a specific gender, ethnic group, race or other makes the collected unigrams related to that specific group. Therefore, the built dictionary of unigrams cannot be reused to detect hate speech towards other groups with the same efficiency.
* Burnap et al. [3] used typed dependencies (i.e., the relation between words) along with bag of words (BoW) features to distinguish hate speech utterances from clean speech ones.

**Disadvantages**

* There is no Sentiment-based Features detection to find Hate Speech on Twitter.
* There is no Binary Classification to categories All Hateful Messages.

**PROPOSED SYSTEM**

1) The system proposes a pattern-based approach to detect hate speech on Twitter: patterns are extracted in pragmatic way from the training set and we define a set of parameters to optimize the collection of patterns.

2) In addition to patterns, we propose an approach that collects, also in a pragmatic way, words and expressions showing hate and offense, and use them with

Patterns, along with other sentiment-based features to detect hate speech.

3) The proposed sets of unigrams and patterns can be used as already-built dictionaries for future works related to hate speech detection.

4) The system classifies tweets into three different classes (instead of only two) where we make distinction between tweets showing hate, and those being just offensive.

**Advantages**

**The system has covered following Approaches**

* **Fast** technique which consists of tweets that are neutral, non-offensive and present no hate speech.
* **An Affective techniques which** contains tweets that are offensive, but do not present any hate or segregative / racist speeches.
* **The technique that** includes tweets which are offensive and present hate, racist and segregative words and expressions.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV

➢ RAM - 4 GB (min)

➢ Hard Disk - 20 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL