**Hate speech Detection Using Machine learning**

Objective

This aims to classify textual content into non-hate or hate speech, in which case the method may also identify the targeting characteristics (i.e., types of hate, such as race, and religion) in the hate speech.

Abstract

In this work, we argue for a focus on the latter problem for practical reasons. We show that it is a much more challenging task, as our analysis of the language in the typical datasets shows that hate speech lacks unique, discriminative features and therefore is found in the ‘long tail’ in a dataset that is difficult to discover. We then propose Deep Neural Network structures serving as feature extractors that are particularly effective for capturing the semantics of hate speech. Our methods are evaluated on the largest collection of hate speech datasets based on Twitter, and are shown to be able to outperform state of the art by up to 6 percentage points in macro-average F1, or 9 percentage points in the more challenging case of identifying hateful content.

Introduction

The exponential growth of social media such as Twitter and community forums has revolutionised communication and content publishing, but is also increasingly exploited for the propagation of hate speech and the organisation of hate-based activities [1, 3]. The anonymity and mobility afforded by such media has made the breeding and spread of hate speech – eventually leading to hate crime – effortless in a virtual land scape beyond the realms of traditional law enforcement. The term ‘hate speech’ was formally defined as ‘any communication that disparages a person or a group on the basis of some characteristics (to be referred to as types of hate or hate classes) such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics’. In the UK, there has been significant increase of hate speech towards the migrant and Muslim communities following recent events including leaving the EU, the Manchester and the London attacks. In the EU, surveys and reports focusing on young people in the EEA (European Economic Area) region show rising hate speech. And related crimes based on religious beliefs, ethnicity, sexual orientation or gender, as 80% of respondents have encountered hate speech online and 40% felt attacked or threatened. Statistics also show that in the US, hate speech and crime is on the rise since the Trump election. The urgency of this matter has been increasingly recognised, as a range of international initiatives have been launched towards the qualification of the problems and the development of counter-measures.

Existing System

Existing methods primarily cast the problem as a supervised document classification task [33]. These can be divided into two categories: one relies on manual feature engineering that are then consumed by algorithms such as SVM, Naive Bayes, and Logistic Regression [3, 9, 11, 15, 19, 23, 35–39] (classic methods); the other represents the more recent deep learning paradigm that employs neural networks to automatically learn multi-layers of abstract features from raw data [13, 26, 30, 34] (deep learning methods).

Disadvantage

* Existing studies on hate speech detection have primarily reported their results using micro-average Precision, Recall and F1 [1, 13, 30, 36, 37, 40].
* The problem with this is that in an unbalanced dataset where instances of one class (to be called the ‘dominant class’) significantly out-number others (to be called ‘minority classes’), micro-averaging can mask the real performance on minority classes.

Proposed System

All datasets are significantly biased towards non-hate, as hate Tweets account between only 5.8% (DT) and 31.6% (WZ). When we inspect specific types of hate, some can be even scarcer, such as ‘racism’ and as mentioned before, the extreme case of ‘both’. This has two implications. First, an evaluation measure such as the micro F1 that looks at a system’s performance on the entire dataset regardless of class difference can be biased to the system’s ability of detecting ‘non-hate’. In other words, a hypothetical system that achieves almost perfect F1 in identifying ‘racism’ Tweets can still be overshadowed by its poor F1 in identifying ‘non-hate’, and vice versa. Second, compared to non-hate, the training data for hate Tweets are very scarce. This may not be an issue that is easy to address as it seems, since the datasets are collected from Twitter and reflect the real nature of data imbalance in this domain. Thus to annotate more training data for hateful content we will almost certainly have to spend significantly more effort annotating non-hate.

Advantage

* Also, as we shall show in the following, this problem may not be easily mitigated by conventional methods of over- or under-sampling.
* Because the real challenge is the lack of unique, discriminative linguistic characteristics in hate Tweets compared to non-hate.
* As a proxy to quantify and compare the linguistic characteristics of hate and non-hate Tweets, we propose to study the ‘uniqueness’ of the vocabulary for each class.

System Architecture

Classification CNN

Accuracy Score

Feature Selection

Natural Language Processing

Pre-processing

Dataset