International Journal of Decision Support System Technology
Volume 14 • Issue 1

Hate Speech Detection Using Text Mining and Machine Learning

Safae Sossi Alaoui, Faculty of Sciences and Techniques, Moulay Ismail University of Meknes, Morocco*
Yousef Farhaoui, Faculty of Sciences and Techniques, Moulay Ismail University of Meknes, Morocco
https://orcid.org/0000-0003-0870-6262
Brahim Aksasse, Faculty of Sciences and Techniques, Moulay Ismail University of Meknes, Morocco

ABSTRACT

Automatic hate speech detection on social media is becoming an outstanding concern in modern countries. Indeed, hate speech towards people brings about violent acts and social chaos; hence, law prohibits it, and it engenders moral and legal implications. It is crucial that we can precisely categorize hate speech and not hate speech automatically. This allows us to identify easily real people who represent a threat for our society. In this paper, the authors applied a complete text mining process and naïve bayes machine learning classification algorithm to two different data sets (tweets_Num1 and tweets_Num2) taken from Twitter to better classify tweets. The results obtained demonstrate that the model performed well regarding different metrics based on the confusion matrix including the accuracy metric, which achieved 87. 23% on the first dataset and 93. 06% on the second.

KEYWORDS

Hate Speech, Machine Learning, Naïve Bayes, Sentiment Analysis, Text Mining

1 INTRODUCTION

Recently, people communicate and discuss their opinions in digital form more and more by taking advantage of online social networks like Twitter, Facebook, and Instagram and so on. These social media have many benefits to humanity in enhancing culture diversity; otherwise, the dark side of social media causes hazardous consequences when it comes to attack others by harassing, bullying and threatening them using hateful expressions known as hate speech. Hate speech (Chetty & Alathur, 2018) can be defined as a threating and abusive language which expresses hatred against a particular group especially on the basis of race, color, religion, ethnicity and even gender.

Generally speaking, media have a significant impact on individuals' beliefs and perceptions (Mastrorocco & Minale, 2018). Indeed, when social media have been exploited as a tool to convey hate, racist and terroristic contents, it can engender crimes and violent acts (Jendryke & McClure, 2019). For example, Chetty and Alathur (2018) emphasized the strong correlation between hate speech and terrorist activities. As a result, collaborative efforts between government, Internet service providers and online social networks will effectively define policies to combat both hate speech and terrorism.

In order to fight Cyber hate, many organizations have enforced their policies towards law, technology and education so as to prevent and reduce its negative influences (Blaya, 2019).

To handle hate speech automatically, it can be seen as a part from sentiment analysis or opinion mining (Hussein, 2018) which utilizes the natural language processing (NLP), text mining and

DOI: 10.4018/IJDSST.286680 *Corresponding Author

computational algorithms to automate the identification and extraction of subjective information from text. The hate speech is a behavior built from education, TV and many other factors. It is hard to design a hate speech detector since it depends on the language of the hater. There exist three sentiment analysis techniques (Medhat et al., 2014) lexicon based method, machine learning approach, and hybrid approach. Effectively, the mechanisms of hate speech detection are part of the mentioned approaches; the **lexicon-based methods** tend to calculate semantic orientation of words or phrases in a text by means of a dictionary which provides words with a positive or negative sentiment value assigned to each of the words. The **machine learning approaches** are used to get a discriminative function that can separate hate speech from normal speech. Machine learning algorithms are programs; that considered as an evolution of the regular algorithms, which can automatically learn from data and improve from experience, without human intervention. In our case, the hate speech detection can be seen as a supervised learning problem where both the inputs and outputs are already known, which means that the data used to train the algorithm is already labeled with correct answers in order to generate reasonable predictions for the response to new data. Since the outputs are discrete, the classification algorithms (Sossi Alaoui et al., 2018, 2017) are used to categorize the data into specific groups or classes. Finally, **the hybrid approach** that combines machine learning methods with lexical-based approaches.

In this paper, we focus on machine learning approach because lexical based method tend to confuse between terms used in hate speech and offensive language and therefore it gives low precision (Davidson et al., 2017).

The main objective of this paper is to propose both a framework and a model to detect automatically hate speech in social media for both binary and multiclass problems by using two public datasets taken from Twitter. The framework will describe a full implementation of a text mining process and the model will be based on Naïve Bayes a supervised machine learning algorithm. The reason behind choosing Naïve Bayes and not another machine learning classification algorithm (Sossi Alaoui et al., 2018, 2017) is according to numerous research works that will be discussed in the next section, which were conducted a comparative study of several machine learning algorithms; Naïve Bayes was the best method in terms of different performance measures and which proved its efficiency and simplicity in dealing with almost all problems related to sentiment analysis (Alam & Yao, 2019). Precisely, the accuracy of Naïve Bayes algorithm as S. Alam and N. Yao (2019) has been considerably improved after the application of preprocessing steps compared to maximum entropy (MaxE), and support vector machines (SVM) for sentiment analysis.

The motivation behind this work is to overcome the difficulties of generalizing the resulting models to detect hateful text-based content, which are present in the literature, and to propose a framework for the construction of a precise model capable of detecting hate speech automatically by performing a complete text mining process based on Naïve Bayes; a probabilistic classification machine learning algorithm and applied to two different datasets in terms of data size, type of classification task (binary or multiclass) and data sources (data.world and Kaggle).

The remainder of the paper is organized as follows. Previous related works are discussed in section II. The methodology of this work is explained in detail by designing and describing the process of text mining analytics in section III. Section IV presents the results obtained. Finally, section V concludes the paper.

2 RELATED WORKS

In order to offer an overview of the main works related to this area of research, this section has focused on numerous papers on sentiment analysis and hate speech detection. First, Gonçalves et al. (2013) made a comparison between eight popular sentiment methods expressly SentiWordNet, SASA, PANAS-t, Emoticons, SentiStrength, LIWC, SenticNet, and the Happiness Index using a web service named iFeel. They additionally developed a competitive approach in terms of coverage and agreement.