Chapter 101
Sentiment Analysis of Twitter Data: A Hybrid Approach

Ankit Srivastava
The NorthCap University, Gurgaon, India

Vijendra Singh
The NorthCap University, Gurgaon, India

Gurdeep Singh Drall
The NorthCap University, Gurgaon, India

ABSTRACT

Over the past few years, the novel appeal and increasing popularity of social networks as a medium for users to express their opinions and views have created an accumulation of a massive amount of data. This evolving mountain of data is commonly termed Big Data. Accordingly, one area in which the application of new techniques in data mining research has significant potential to achieve more precise classification of hidden knowledge in Big Data is sentiment analysis (aka optimal mining). A hybrid approach using Naïve Bayes and Random Forest on mining Twitter datasets is presented here as an extension of previous work. Briefly, relevant data sets are collected from Twitter using Twitter API; then, use of the hybrid methodology is illustrated and evaluated against one with only Naïve Bayes classifier. Results show better accuracy and efficiency in the sentiment classification for the hybrid approach.

1. INTRODUCTION

Nowadays, one way to aid individuals and/or organizations in making intelligent decisions such as choosing among available options wisely is to draw upon the opinion of the crowd. Traditionally, many of us have depended on other people's opinions, particularly those of family members, friends and relatives, when making decisions on critical issues (Pang & Lee, 2008; Saif, He, & Alani, 2012; Kharde &

DOI: 10.4018/978-1-6684-6303-1.ch101

Sonawane, 2016; Xia, Zong, & Li, 2011; Cambria, Schuller, Xia, & Havasi, 2013). However, with rapid technological advances and the increasing ubiquity of the Internet in all corners of the world, many of us are now showing interests in social platforms, as these have made it relatively easy for us to know the thinking of not only family members and friends, but also of strangers around us (including willing experts who do not mind providing their educated advice) (Godbole, Srinivasaiah, & Skiena, 2007; Tan, Lee, Tang, Jiang, Zhou, & Li, 2011).

Accordingly, around 6,000 tweets are generally disseminated on Twitter every second; on average, this amounts to 500 million tweets daily or, 200 billion tweets annually. Platforms such as Facebook, Yelp and Amazon have accumulated a huge traffic of texts and opinions being generated daily. Such huge numbers means a lot of texts and data from all around the globe. Consequently, it has become crucial for individuals and/or organizations to be able to analyze these data meaningfully so as to be able to profit from, and/or capitalize on, these opinions to enhance one's reputation (Balahur & Jacquet, 2015; Kumar, Morstatter, & Liu, 2014; Isah, Trundle, & Neagu, 2014; Jiang & Kotzias, 2016).

Sentiment analysis (SA), a process by which sentiment over the accumulated tweets can be automatically detected, is an increasingly popular means of analyzing "big data" such as "tweets" arising from the use of Twitter. Furthermore, such analysis allows the text polarity (whether it is neutral, positive or good, negative or bad), to be aggregated. Briefly, in order to classify the polarity of the accumulated text via sentiment classification (West, Paskov, Leskovec, & Potts, 2014; Cogburn & Espinoza-Vasquez, 2011; Gamallo & Garcia, n.d.), SA entails five fundamental steps: (1) collecting the data to be analyzed; (2) preprocessing the data; (3) extracting feature(s) linked to the data; (4) performing sentiment classification on the data; and (5) presenting result(s).

In essence, SA can be conducted at four different levels: Word, Sentence, Document and/or the Feature/Aspect level (Karlgren & Ericsson, 2013; Recupero & Cambria, 2014; Irsov & Cardie, 2014). At the Document level, the aim will be to aggregate the single sentiment polarity of the entire document by seeking out the sentiment polarities of all sentences combined in the document and then summarizing them. At the Sentence level, sentiment polarity of a sentence is first computed by identifying the sentiment polarity of each and every word in the sentence. These are then aggregated (Tan et al., 2011; Vijendra & Laxman, 2013; Vijendra, Sahoo, & Ashwini, 2010). At the Word level, sentiment polarity of each and every word is determined. At the Aspect/Feature level, the main concern will be to identify and extract product features from the source data. In this approach, the entities for which the sentiment may be directed will have to be identified, for example, if the sentiment analysis encompasses that of phone reviews, the differing aspects/features may include the camera, the screen, and the phone speaker.

In Twitter data, tweets often contain noises, incomplete data, slangs, unstructured sentences and many irregular expressions. Before preforming classification analysis, a preprocessing of the text is needed such as the removal of URLs (Uniform Resource Locators), numbers, stop words and the like (Feldman, 2013; Vinodhini & Chandrasekaran, 2012; Vijendra & Laxman, 2015; Quadri, Prashanth, Pongpaichet, Esmin, & Jain, 2017). In an earlier work, six different pre-processing methods were used to filter out the necessary and useful data from the complete dataset with a series of experiments using four classifiers conducted to verify the effectiveness of several pre-processing methods on Twitter datasets. Results indicate that in N-grams model, removing URLs and stop words reduce the vocabulary size while no change in the performance of all of the classifier approaches was observed. In this follow up work, we aim to improve the performance by incorporating a hybrid model.

The rest of this paper is organized as follows. Section 2 describes the background or related work. Section 3 presents the hybrid methodology for the suggested evaluation approach. Section 4 details the