Enhanced Bootstrapping Algorithm for Automatic Annotation of Tweets

Mudasir Mohd, University of Kashmir, Srinagar, India
https://orcid.org/0000-0003-1597-146X

Rafiya Jan, Central University of Kashmir, Srinagar, India

Nida Hakak, Mahareshi Dayanand University, Haryana, India

ABSTRACT

Annotations are critical in various text mining tasks such as opinion mining, sentiment analysis, word sense disambiguation. Supervised learning algorithms start with the training of the classifier and require manually annotated datasets. However, manual annotations are often subjective, biased, onerous, and burdensome to develop; therefore, there is a need for automatic annotation. Automatic annotators automatically annotate the data for creating the training set for the supervised classifier, but lack subjectivity and ignore semantics of underlying textual structures. The objective of this research is to develop scalable and semantically rich automatic annotation system while incorporating domain dependent characteristics of the annotation process. The authors devised an enhanced bootstrapping algorithm for the automatic annotation of Tweets and employed distributional semantic models (LSA and Word2Vec) to augment the novel Bootstrapping algorithm and tested the proposed algorithm on the 12,000 crowd-sourced annotated Tweets and achieved a 68.56% accuracy which is higher than the baseline accuracy.

KEYWORDS

Bootstrapping, Emotion Classification, Emotions, Semantic Similarity

INTRODUCTION

Twitter is leading microblog service used by over 974 million users with 500 million tweets/day, thus is playing an active role in the new form of media. Twitter posts are called tweets and are limited to 280 characters. Users also upload photos and short videos for broadcasting their experience and feelings about daily life (McFedries, 2007). Twitter is acting as an essential communication channel for governments and heads of state to highlight their governance initiatives and interact with their citizens directly. The evolution of Internet and mobile based communications, led to increase in social interaction among multiple users ("social networking sites"), and thus huge data ("Big Data") is equipped depicting the public attitude and acknowlegments related to different events like world events, consumer product events, political and movies events (Salton, 1991). According to the Twitter blog, recently, something remarkable happened on Twitter: #NuggsForCarter was the most retweeted

DOI: 10.4018/IJCINI.2020040103

tweet of the year 2017. A high scholar's call for free nuggets to Wendys became the highest retweeted tweet of all time with 3.24 million retweets[1]. In general, Twitter users now share excessive tweets near about 500 million tweets per day that is about 5,700 Tweets per second, according to mean based mentioned on a later report in Twitter blog[2]. This shows the considerable popularity Twitter is gaining and the role it's playing in changing people's lives. People use Twitter for various reasons. (Java, Song, Finin, & Tseng, 2007) in their study categorize user intentions as: (1) source of information; (2) being social; and (3) retrieving information. (Hakak, Mohd, Kirmani, & Mohd, 2017) have given an excellent summary of the state of work done so far in the area.

Twitter is becoming a reliable media to search for timely information then the web and this information is mined extensively for opinion mining, emotion detection and sentiment polarity by different business and researchers. Automatic affect detection on Twitter is attracting much research since users continuously express their opinions' regarding anything that they are interested in. These opinions include reviews of products, general feelings, etc. Affect detection finds its applications in various applications like (Rodriguez, Ortigosa, & Carro, 2012) monitored how affect and emotional factors determine the outcome of the e-learning environment; (Desmet & Hoste, 2013) showed how affect monitoring on social media can help suicide prevention; (Cherry, Mohammad, & De Bruijn, 2012) used emotion classification to detect depression on social media; (Dadvar, Trieschnigg, Ordelman, & de Jong, 2013) showed how to improve detection of cyberbullying from user content.

Opinion analyzers and emotion detection tools for social media text streams use supervised learning classifiers which rely heavily on the manually annotated corpus. The manually annotated corpus for use in supervised learning is difficult to create and human annotators, who associate different sentences with different categories, traditionally produce annotated corpus. However, this process is arduous and time-consuming and also obtaining an inter-annotator agreement is difficult in such tasks as human judgment is subjective. This research aims to create an auto-annotation tool capable of annotating twitter corpus by analyzing tweets, i.e., to create a bootstrapping algorithm for automatic annotation of the Twitter corpus. Bootstrapping processes lack subjectivity and overlook the inherent semantics of underlying text. Thus, there is a greater need for extending bootstrapping algorithms for achieving better accuracy in the automatic annotation of tweets. For this reason, we propose an extended bootstrapping algorithm for the automatic annotation of tweets.

The proposed enhanced bootstrapping algorithm takes semantics of text as a feature and annotates the corpus using distributional semantic hypothesis. We exploited distributional semantic models for enhancing the bootstrapping algorithm and achieved comparable results. The existing bootstrapping algorithms overlook the semantics of the text and work on the presence of either critical terms in the text or some other statistical features to annotate the sentences and thus are not scalable. The key idea of our proposed enhanced bootstrapping algorithm is thus to extend the bootstrapping process by using semantic models to create any domain annotations and thus have a scalable bootstrapping algorithm.

The proposed algorithm constitutes five important steps: 1) Preprocessing of tweets; 2) Lexicon generation; 3) Enhancement of lexicon of seed words using word2Vec model; 4) Seed extension using and another dictionary-based approaches 5) Using LSA to compute semantic similarity; 5) Using big vectors created using Word2Vec to calculate semantic coherence. The proposed system was evaluated on Kashmir 2016 unrest dataset collected from Twitter. Around 12,000 tweets were manually annotated using crowd-sourcing to check the efficiency of the proposed approach. The results are above the traditional baseline approaches, and thus confirm that the competitive performance of our proposed approach.

Rest of the paper is organized as follows: background, enhanced bootstrapping algorithm in detail, experiments, evaluation and results, discussion and comparative analysis and then conclusions and future work.