Information Processing and Management 53 (2017) 764–779

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Twitter sentiment analysis using hybrid cuckoo search method

Avinash Chandra Pandey*, Dharmveer Singh Rajpoot, Mukesh Saraswat

Jaypee Institute of Information Technology, Noida, India

ARTICLE INFO

Article history:
Received 25 June 2016
Revised 25 January 2017
Accepted 1 February 2017

Keywords:
Sentiment analysis
Cuckoo search
Twitter
Data preprocessing
K-means

ABSTRACT

Sentiment analysis is one of the prominent fields of data mining that deals with the identification and analysis of sentimental contents generally available at social media. Twitter is one of such social medias used by many users about some topics in the form of tweets. These tweets can be analyzed to find the viewpoints and sentiments of the users by using clustering-based methods. However, due to the subjective nature of the Twitter datasets, metaheuristic-based clustering methods outperforms the traditional methods for sentiment analysis. Therefore, this paper proposes a novel metaheuristic method (CSK) which is based on K-means and cuckoo search. The proposed method has been used to find the optimum cluster-heads from the sentimental contents of Twitter dataset. The efficacy of proposed method has been tested on different Twitter datasets and compared with particle swarm optimization, differential evolution, cuckoo search, improved cuckoo search, gauss-based cuckoo search, and two n-grams methods. Experimental results and statistical analysis validate that the proposed method outperforms the existing methods. The proposed method has theoretical implications for the future research to analyze the data generated through social networks/medias. This method has also very generalized practical implications for designing a system that can provide conclusive reviews on any social issues.

1. Introduction

The unrivalled increase in the acceptance as well as penetration of social media platforms, such as Facebook, Twitter, Google plus, etc., in a day to day life, have changed the pattern of online communication of people. Formally, user's online access was highly restricted to professional contents such as news agencies or corporations. However, these days they can seamlessly interact with each other in a more concurrent way by creating their own content within a network of peers. According to Howard (2011), "We use Facebook to schedule the protest, Twitter to coordinate, and YouTube to tell the word". Social media has emerged as a vital platform of representing people's sentiment, boosting the requirements of data mining in the field of the sentiment analysis.

In the sentiment analysis, the raw data is the online text that is exchanged by users through social media (Tang, Tan, & Cheng, 2009). Twitter, which is one of such social medias, has become the prominent source to exchange the online text, providing a vast platform of sentiment analysis. Twitter is a very popular social networking website that allows registered users to post short messages, also called tweets, up to 140 characters. Twitter database is one of the largest database having

* Corresponding author.
E-mail addresses: avish.nsit@gmail.com (A. Chandra Pandey), dharamveer.rajpoot@jiit.ac.in (D. Singh Rajpoot), saraswatmukesh@gmail.com (M. Saraswat).

http://dx.doi.org/10.1016/j.ipm.2017.02.004
0306-4573/

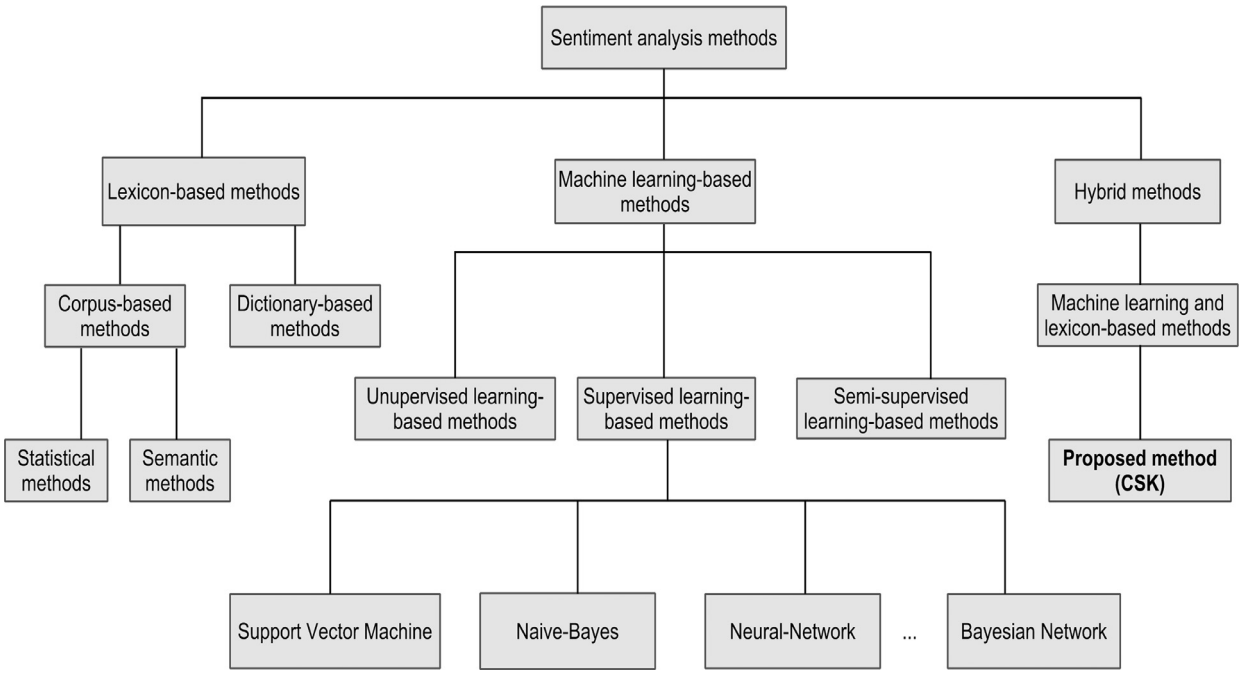

Fig. 1. Sentiment classification methods.

200 million users who post 400 million messages/tweets in a day (Ritter, Clark, Etzioni et al., 2011). At Twitter, users often share their personal opinion on different subjects such as acceptance or rejection of politicians and viewpoint about products, talk about current issues and share their personal life events. However, users post their tweets with fewer characters by using a short form of words and symbols such as emoji. Therefore, analysis of these tweets can be used to find strong viewpoints and sentiments for any topic. Twitter data has already been used by different people to predict stock market prediction (Bollen, Mao, & Zeng, 2011), box office revenues for movies (Asur & Huberman, 2010), identify the clients with negative sentiments (Thet, Na, & Khoo, 2010), etc. The main aim of sentiment analysis is to determine the attitude of users on a particular topic. Therefore, this paper proposes a novel clustering method for sentiment analysis on Twitter dataset.

Sentiment analysis methods can be broadly categorized into lexicon-based methods, machine learning-based methods, and hybrid methods (Medhat, Hassan, & Korashy, 2014) which can be further classified into sub-category as depicted in Fig. 1. Lexicon-based methods require predefined sentiment lexicon to determine the polarity of any document. However, the accuracy of lexicon-based method is reduced drastically in the presence of emoticons and short hand texts, as they are not the part of predefined sentiment lexicon (Khan, Atique, & Thakare, 2015). Emoticons are the visual emotional symbols used by the users at social medias (Hu, Tang, Gao, & Liu, 2013a). Hu, Tang, Tang, and Liu (2013b) proposed a novel method of sentiment analysis that considers the short texts like “gud nite” and emotional symbols such as “:)”, in a unified framework. The performance of this method does not show stability on some of the emotional signals, such as emoticons, when used on datasets from different domains (Hu et al., 2013a). This problem can be resolved by examining the contributions of other emotion indication information existing in social media, like product ratings, restaurant reviews, and other emotion correlation information (Hu et al., 2013a; Yusof, Mohamed, & Abdul-Rahman, 2015) such as correlation between two words in a post. Emotion indication represents the sentiment polarity of a post and further, it is classified into post level emotion indication (emoticons) and world level emotion indication (publicly available sentiment lexicons) (Hu et al., 2013a). Moreover, emotion correlation for posts are usually represented by a graph in which nodes represent the data points and edge represent correlation between the words. Further, Canuto, Gonçalves, and Benevenuto (2016) proposed a new sentiment-based meta-level features for effective sentiment analysis. This method has a capability to utilize the information from the neighborhood effectively and efficiently to capture important information from highly noise data.

Bravo-Marquez, Mendoza, and Poblete (2013) introduced a novel supervised method to combine strengths, emotions, and polarities for improving the Twitter sentiment analysis process. Kontopoulos, Berberidis, Dergiades, and Bassiliades (2013) proposed ontology-based sentiment analysis of tweets. In this method, a sentiment grade has been assigned for every distinct notion in the tweets. Further, Mohammad, Zhu, Kiritchenko, and Martin (2015) analyzed US presidential electoral tweets by using supervised automatic classifiers and identified the emotional state, emotion stimulus, and intent of these tweets. Coletta, da Silva, Hruschka, and Hruschka (2014) combined the strength of SVM classifier with a cluster ensemble for refining the tweet classification. SVM classifier is executed first to classify tweets, thereafter C3E-SL algorithm has been used to enhance the classification of tweets.