Chapter 33
Classification of Code-Mixed Bilingual Phonetic Text Using Sentiment Analysis

Shailendra Kumar Singh
https://orcid.org/0000-0001-9658-1441
Computer Science and Engineering, Sant Longowal Institute of Engineering and Technology, India

Manoj Kumar Sachan
Computer Science and Engineering, Sant Longowal Institute of Engineering and Technology, India

ABSTRACT

The rapid growth of internet facilities has increased the comments, posts, blogs, feedback, etc., on a large scale on social networking sites. These social media data are available in an unstructured form, which includes images, text, and videos. The processing of these data is difficult, but some sentiment analysis, information retrieval, and recommender systems are used to process these unstructured data. To extract the opinion and sentiment of internet users from their written social media text, a sentiment analysis system is required to develop, which can work on both monolingual and bilingual phonetic text. Therefore, a sentiment analysis (SA) system is developed, which performs well on different domain datasets. The system performance is tested on four different datasets and achieved better accuracy of 3% on social media datasets, 1.5% on movie reviews, 1.35% on Amazon product reviews, and 4.56% on large Amazon product reviews than the state-of-art techniques. Also, the stemmer (StemVerb) for verbs of the English language is proposed, which improves the SA system's performance.

1. INTRODUCTION

With the enhancement of internet services and facilities, social networking sites such as YouTube, Google Plus, LinkedIn, Twitter, and Facebook have increased rapidly (Press Trust of India, 2013). These social networking sites provide facilities to share the users' feelings, emotions, comments, feedbacks, and reviews over the internet. Thus, the size of such content over social media (SM) increases exponentially day

DOI: 10.4018/978-1-6684-6303-1.ch033

by day. Most of the SM text contents are written using more than one language and called code-mixed language. The text of languages other than English is written using Roman script's alphabets called phonetic text. The phonetic text mixed with English language text, but there is no fixed format for these SM texts (Dutta et al., 2015). These contents are used as input text to extract information, opinion, text summarization, etc., using various linguistic computations, natural language processing, text mining, and information retrieval systems (S. K. Singh & Sachan, 2019b).

Opinion mining or sentiment analysis (SA) is a sub-field of text mining and is one of the most recent research topics of interest (Pang & Lee, 2008). The SA is related to predicting and analyzing hidden information, emotion, and feelings from the written text. The SA is widely used to analyze feedbacks on government regulation and policy proposed, to analyze the customers' likes/dislikes, to know the product demand, brand reputation, real-world event monitoring and analyzing of political party demand, competitors products' merit or demerit analyzes, and subtask component of recommender system (Bonadiman et al., 2017; D'Andrea et al., 2015). In April 2013, 90% of consumers decided to purchase things or services based on online reviews (Peng et al., 2014).

The user-written texts are classified using SA into two or three classes based on different formats such as positive/negative/neutral, like/dislike, and good/bad (Hopken et al., 2017; S. K. Singh & Sachan, 2019b). Mostly two approaches are used to classify the text using SA, such as (i) feature-based and (ii) bag-of-words. Machine learning techniques are based on features, while lexicon-based techniques use bag-of-words approaches. In SA of products and services, machine learning is widely used, but the bag-of-words are used for social issues (Karamibekr & Ghorbani, 2012). Machine learning systems are trained on the labeled dataset(s) and classify the testing dataset(s) based on the trained system. The lexicon-based techniques are categorized as two approaches such as corpus-based and dictionary-based (S. K. Singh & Sachan, 2019b). The text classification using the dictionary-based approach depends upon the opinion word's sentiment score. This dictionary consists of opinion words and their sentiment scores (Alharbi & Alhalabi, 2020). The opinion words are nouns, verbs, adverbs, and adjectives, which act as features in dictionary-based approach as discussed in the articles (Hopken et al., 2017; Shamsudin et al., 2016; P. K. Singh et al., 2015; R. K. Singh et al., 2020; S. K. Singh & Sachan, 2019b). These opinion words are used to develop opinion dictionaries such as SentiWordNet 3.0 (Baccianella et al., 2010), the latest SenticNet 4 (Cambria et al., 2016), etc.

The SM text is available in monolingual, bilingual, and multilingual. The processing of multilingual texts is a difficult task as compared to bilingual and monolingual text. Monolingual text can be easily processed, but bilingual text up to some extent. Taboada et al. (2011) developed "Semantic Orientation CALculator (SO-CAL)" using a dictionary, which includes opinion words (nouns, adverb, verbs, and adjectives) along with their polarity and strength value (Taboada et al., 2011) and achieved the best performance in term of the accuracy 70.10% on movie dataset among other datasets. In 2012, Karamibekr & Ghorbani developed a sentiment classification system in which verbs are considered the core element of the system and created an opinion verb dictionary of 440 verbs and 1726 terms for the term opinion dictionary. The value of polarity is assigned to each word in the dictionary ranging from +2 to -2, and their system was tested on a dataset related to social issues with an accuracy of 65% (Karamibekr & Ghorbani, 2012). P. K. Singh et al., (2015) used negation handling rules and a dictionary-based approach to classify the social issues related dataset into positive or negative class and achieved an accuracy of 79.16% for negative sentences. Iqbal et al., (2015) proposed a Bias-aware thresholding (BAT) method with the combination of AFINN and SentiStrength to reduce the bias in the lexicon-based method for SA and obtained 69% of accuracy using Naïve Bayes classifier.