International Journal of Distributed Artificial Intelligence
Volume 11 • Issue 1 • January-June 2019

Review of Sentiment Detection:
Techniques and Challenges

Smiley Gupta, N.C. College of Engineering, Israna, India

Jagtar Singh, N.C. College of Engineering, Israna, India

ABSTRACT

A large volume of user-generated data is evolving on a day-to-day basis, especially on social media platforms like Twitter, where people express their opinions and emotions regarding certain individuals or entities. This user-generated content becomes very difficult to analyze manually and therefore requires a need for an intelligent automated system which mines the opinions and organizes them using polarity metrics, representing the process of sentiment analysis. The motive of this review is to study the concept of sentiment analysis and discuss the comparative analysis of its techniques along with the challenges in this field to be considered for future enhancement.

KEYWORDS

Challenges of Sentiment Analysis, Polarity Metrics, Sentiment Analysis Techniques, Sentiment Analysis, Social Media

INTRODUCTION

The opinions and thoughts of people especially on social networking platforms have a major influence on our daily decision making. This decision making includes buying a product, making investments, choosing a school, etc., all these decisions affect our daily life. People usually seek opinions from review sites, e-commerce sites and social media like Twitter, to get feedback on certain product or service. Similarly, organizations use opinion polls, questionnaires, surveys, and social media as a medium to get feedback on their products and services (Katrekar, 2005).

However, to analyze and summarize the opinions from the huge amounts of data gathered from the social networking sites, the process called sentiment analysis is required. Sentiment analysis is the computational study of mining the attitudes, opinions, and emotions of people from the networking sites through Natural Language Processing. These summarized opinions are then divided into categories like "positive", "negative" and "neutral" (Vohra & Teraiya, 2013). Riyadh (2017) implemented emotion detection using Naïve Bayes classifier for classification of emotions with unigram and unigram plus POS tagging for feature extraction in order to improve the results. Sentiment analysis can be classified into three levels namely: document level, sentence level, and aspect level. In document level, the complete document is considered as a source of information and then concluded as either positive or negative (Medhat, Walaa, Hassan, & Korashy, 2014). In sentence level, firstly a sentence is categorized as a subjective sentence or objective sentence. Objective sentences contain no opinions or judgments as they are completely based on facts. Whereas, subjective sentences involve opinions and therefore play a role in deciding polarity of the sentences and these polarities are further summed up to derive the final conclusion. In aspect level, the entities are identified and their features are extracted holding different opinions and sentiments. For example, Battery life is very long lasting. Here, "Battery" (noun) is the feature of the product and "very long lasting" (adjective) is the opinion word (Kolkur, Seema, Dantal, & Mahe, 2015).

DOI: 10.4018/IJDAI.2019010105

METHODOLOGY FOR SENTIMENT ANALYSIS

1. Data Collection: First and foremost, the user-generated data is collected from social networking sites, forums, and blogging sites. Twitter is one of the most frequently used data sources and the length of text in twitter is maximum 140 characters long. These data are unstructured, expressed in different ways by using the different context of writing along with slangs, acronyms, etc., due to which the manual analysis of text becomes really complex.
2. Data Preprocessing: Data preprocessing is nothing but cleaning and filtering out the unstructured data before analysis. In this, identification and elimination of non-textual content and the content that is irrelevant with respect to the following area of study occurs. Cleaning of data involves removal of URL's, removal of punctuations, case conversion and stemming.
3. Feature Selection: Several findings in feature selection specific to sentiment analysis are:
    - Term presence and frequency: Term presence is based individual word or n-grams and Term frequency is the number of repeated occurrences of the term in the text.
    - Parts Of Speech (POS): These features are selected to keep count of the number of verbs, adverbs, and nouns, etc., in the sentence or document.
    - Opinion words and Phrase: These include words and phrases which depict opinions such as 'good or bad,' 'like or hate,' etc.
    - Negation: The use of negation word in the text can reverse the whole polarity and meaning of opinion. For example: "not good" is the same as "bad."
4. Sentiment Classification Algorithm

At this stage, a particular sentiment analysis methodology is applied as per need and distributed levels of polarity are obtained which are added up and the total score is achieved and averaged (https://www.edureka.co/blog/sentiment-analysis-methodology/).

5. Experimental Results

Therefore, after the conversion of the unstructured text into useful information, the final results are displayed (Figure 1).

SENTIMENT CLASSIFICATION TECHNIQUES

The three most commonly used sentiment classification techniques represented by Figure 2 are machine learning, lexicon-based, and hybrid approach which are further subdivided into different types.

1. Machine Learning: This technique is used in training a machine on how to learn from data and produce output without being explicitly programmed. It includes both supervised and unsupervised methods. The supervised approach is used when there is a finite collection of classes that require the labeled data to train the classifiers and the test data to examine the performance (Jain, Anuja, & Dandannavar, 2016). The unsupervised learning approach is used when it is difficult to get the labeled data.

    a. Probabilistic Classifier: It is a generative and mixture model in which each class act as a component and then the sampling probability for the component is derived. These classifiers are further categorized into three types:
    i. Naïve Bayes Classifier: This classifier is the simplest and most frequently used supervised classification techniques. This probabilistic classifier is based on the Bay's Theorem

Related Content