Abstract

Sentiment analysis is a branch of natural language processing that identifies context and meaning in a text to understand the sentiment behind it. It allows a business to understand the social sentiment of their products and brand based on how people describe it online and also tell them the public opinion about their competitors.

It can be used to classify the emotion in a text into three main categories that are positive, negative or neutral. Depending on the business problem, it can be tuned to make n-number of classifications such as very good, good, average, bad, etc. It can also be used to identify the intent behind a certain text. A major area where sentiment analysis is used is in customer reviews of various products and services. There are a number of websites that provide automatic summaries of product reviews and their distinct features. Sites automatically filter their top reviews so that new customers can see them first. Candidates running for various posts can benefit greatly from sentiment analysis. It allows campaign managers to track how people feel about various subjects and how they connect to the contestants' words and behaviour.

There are different approaches to sentiment classification techniques. One of them is the lexicon based approach where the number of positive and negative words are found in a text and the sentiment is determined based on the one having a higher count. In a machine learning based approach, the document is split into two and one is used for training and the other is used for testing. Commonly used algorithms include support vector machines, naive bayes, random forest and maximum entropy. A hybrid approach is suggested to improve the classification performance.

Some of the major challenges faced are dealing with noisy text. This refers to data that includes spelling mistakes, grammatical errors or improper use of punctuation. It also has a hard time recognising slang words that are not part of the mainstream language. With social media, new words are coming into use and the language itself is evolving. Similarly, there are not many systems that can properly identify the use of sarcasm in a text. With new technology, text analytics is moving to be able to identify more sophisticated human emotions such as hope, anxiety or excitement.

The main objectives of this project are:

- To identify the sentiment of tweets posted by a user or based on a particular hashtag.
- To train and build a model that can correctly identify the sentiment in a text.

Keywords: sentiment analysis, machine learning, natural language processing, twitter