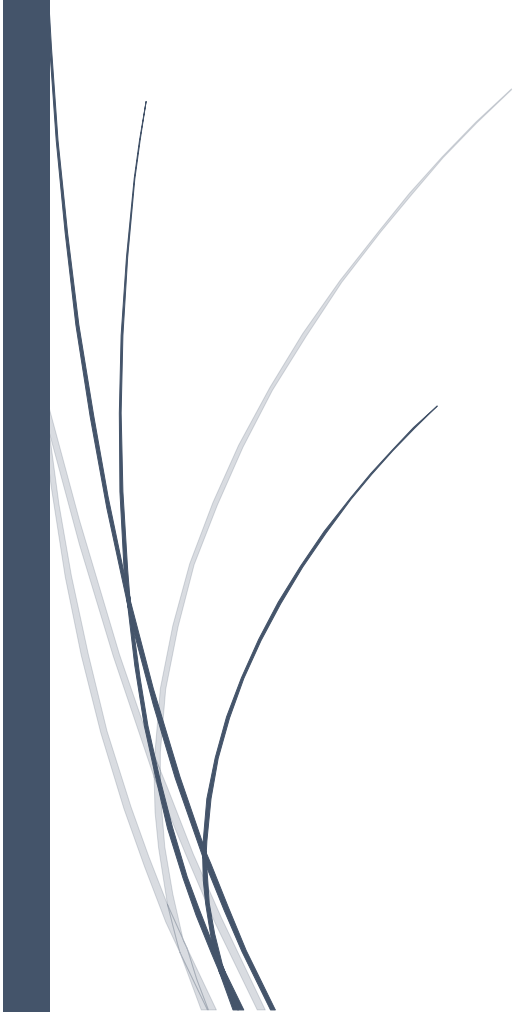
1/23/2023

Twitter Sentiment Analysis

CIND820-XJH



Parth Dhananjay Desai
parth.desai@torontomu.ca

Theme & Introduction

Twitter Sentiment analysis is a sub machine learning task where we want to determine which is the general sentiment of a given tweets or hashtag. Using machine learning techniques and natural language processing we can extract the subjective information of a data and try to classify it according to its polarity such as positive, neutral or negative. It is a really useful analysis since we could possibly determine the overall opinion about a selling object, or predict stock markets for a given company like, if most people think positive about it, possibly its stock markets will increase, and so on. Sentiment analysis is far from to be solved since the language is very complex (objectivity/subjectivity, negation, vocabulary, grammar) but it is also why it is very interesting to working on. In this project I choose to try to classify tweets from Twitter into “positive” or “negative” sentiment by building a model based on probabilities. Twitter is a microblogging website where people can share their feelings quickly and spontaneously by sending a tweet limited by 140 characters. You can directly address a tweet to someone by adding the target sign “@” or participate to a topic by adding a hashtag “#” to your tweet. Because of the usage of Twitter, it is a perfect source of data to determine the current overall opinion about anything.

Problem Definition.

Sentiment analysis in the field of micro-blogging is a relatively new area of research, with a lot of potential for further study. There has been a significant amount of prior work on sentiment analysis in other areas such as user reviews, documents, web blogs, and general phrase-level sentiment analysis. However, sentiment analysis of tweets is unique due to the 140-character limit, which requires users to express their opinions in a very condensed form. The most successful sentiment classification methods to date have used supervised learning techniques such as Naive Bayes and Support Vector Machines, but these methods require a lot of manual labelling, which can be expensive. Some researchers have also explored unsupervised and semi-supervised approaches, but there is still room for improvement. Researchers testing new features and classification methods often compare their results to baseline performance, but there is a need for formal comparisons between different methods to determine the best features and most efficient classification techniques for specific applications.

Data Set

Twitter sentiment analysis will typically use tweets as the data source. Tweets are short text messages of up to 280 characters that users post on the Twitter platform. These tweets can include text, hashtags, mentions, and links. Sentiment analysis of tweets aims to determine the sentiment or emotion expressed in the text of the tweet, which can be classified as positive, neutral, or negative. The tweets that are used for analysis can be collected in real-time using the Twitter API or from pre-existing datasets of tweets. The dataset can be filtered based on various criteria such as keywords, hashtags, or user accounts to focus on specific topics or communities.

Techniques and Tools

- Python
- libraries: textblob, tweepy

Summary