Project Title:	Exploring different approaches for sentiment analysis.
Student ID:	20210378 20210149
Student name:	Rohit Nair Ashwin Bharadwaj
Student email	rohit.nair2@mail.dcu.ie ashwin.sridhar3@mail.dcu.ie
Chosen major:	Data Analytics
Supervisor	Dr. Jennifer Foster

Exploring different approaches for sentiment analysis.

Introduction

The digital era has given users access to the internet and this generates a huge amount of data each second. Every click on the internet generates data which can be captured by the website or the application. In our digital world, data is the new oil. Twitter, Facebook, LinkedIn and other such social media platforms give users a stage to voice their opinion. Humans are social beings and our decisions are influenced by our emotions and the judgement of others. Millions of users share their views on a wide range of topics, which in turn influences more users. Marketing firms try to predict the market trend by conducting extensive market research and surveys. Data produced on social media platforms can play a vital role in decision making with respect to market trend and customer reviews.

Sentiment analysis involves analysing the data and extracting the sentiment attached to it. Natural Language Processing (NLP) enables us to extract emotions from human language. It involves tokenisation of sentences, i.e. breaking down sentences into words and deriving meaning from it based on the sentiment value associated with each word. Each word has a sentiment value attached to it based on its intensity and level of emotions that it conveys. This project explores lexicon based and machine learning based algorithms to compares their accuracy and performance on different data sources.

Literature Review

There are systems that have currently implemented both, lexicon and machine learning based algorithms. R. K. Bakshi et al. (2016) implemented a lexicon-based approach to extract sentiment from micro blogging websites. They created a JDBC-ODBC connection for storing tweets. The sentiment score was calculated by calculating the difference in positive terms and negative terms and dividing it with the total number of terms. Xing Fang & Justin Zhan (2015) compare 3 classification models in their study. They compared Naïve Bayesian, Random Forest and SVM classification. They also implemented parts of speech (POS) tagging to remove parts of speech that do not contribute to the sentiment. The dataset consists of 5.1 million reviews from amazon belonging to 4 major categories: beauty, books, electronics and home.

Yige Wang (2017) combined POS tagging with Naïve Bayes classifier to train the model which is used to predict the sentiment of movie reviews. With POS tagging, the complexity was reduced by a huge margin as the number of words that had to be trained decreased, but the overall accuracy of the model decreased by 7%. A. M. Rahat et al. (2019) compare Naïve Bayes approach and SVM to predict sentiment using review dataset. The dataset they considered of airline reviews which was collected from twitter. They found the accuracy of SVM to be higher than Naïve Bayes, the accuracy of SVM was found to be 83% and Naïve Bayes provided an accuracy of 77%. H. Parveen and S. Pandey (2017) use HDFS and map reduce architecture with Naïve Bayes algorithm to reduce the overhead. They also used emoticons to extract sentiment from the data. The whole emoticon was converted into its equivalent word and the sentiment associated with the word was used. U. Kumari et al. (2017) use SVM classification technique to find the sentiment from a product review dataset. The accuracy of the model was tested with datasets of 4 different products. The highest and the lowest accuracy the model achieved were 90.99% and 88.03% respectively. Though the accuracy of the model is high, the datasets used for testing were considerably small. M. Wongkar & A. Angdresey (2019) analysed twitter user sentiment towards political figures using Naïve Bayes algorithm. They implemented data crawlers to gather relevant data from twitter. The model they implemented had an accuracy of 80.9%. They also noticed that the precision for positive class was higher than the negative class.

Proposed Work

This project would explore how lexicon based and supervised machine learning approaches perform for different data sets. Naïve Bayes algorithm will be used for the machine learning based approach. For the lexicon-based approach, a dictionary will be used to identify positive and negative words. The sentiment score will be calculated in the following manner (Bakshi, et al., 2016):

$$\frac{Positve\ Words - Negative\ Words}{Total\ Number\ Of\ Words}$$

Sentiments from emoticons will also be extracted by the model (Parveen & Pandey, 2017). Not all words contribute towards sentiments, the data would have to be cleaned before the model begins analysing it. Graphical representation of the analysed data would help in conveying the message in a well ordered and organised manner.

In supervised learning the dataset used to train the model plays a vital role in determining the accuracy of the model. How different datasets affect the accuracy of the system and which approach performs better for each type of data set will be analysed. The accuracy of the model and how it is affected when data sets of different class get introduced will be examined.

3 standard data sets will be used to train and test the models. Stanford Artificial Intelligence Laboratory contains labelled data sets of IMDB movie reviews and amazon user reviews. There are 12000 positive and negative movie reviews for training as well as testing. Amazon data set has approximately 120000 records and the ratings are on a scale of 5 which was acquired from Stanford Network Analysis Platform. This will have to be converted to positive or negative to train the model. The twitter data set consists of 1.5 million tweets labelled with 1 and 0. 1 being positive and 0 being negative. This data set was acquired from Thinknook.com.

Spyder and Jupyter Notebook will be used for building the model. Additional libraries such as NLTK and Numpy will also be used for analytics

and Matplotlib or ggplot can be used for visualisation. Different models will be trained using various algorithms on Python. Implementation of Naïve Bayes and one lexicon-based algorithm will be done.

Conclusion

Data is being generated by everyone these days. Users post their views and form opinions based on what others share online. The goal of this project is capitalising on this huge source of data, find which approach is better for each dataset and explore the reason behind an approach outperforming the other. A decision, if not backed by data, will not yield good results. The results will be evaluated by calculating the accuracy percentage of the model based on the difference in predicted sentiment and actual sentiment. This will help us determine which approach would be better for extracting user sentiment for each subdomain.

Bibliography

Bakshi, R. K., Kaur, N., Kaur, R. & Kaur, G., 2016. *Opinion mining and sentiment analysis.* New Delhi, IEEE.

Fang, X. & Zhan, J., 2015. Sentiment analysis using product review data. *Journal of Big Data,* Volume 2.

Kumari, U., Sharma, A. & Soni, D., 2017. *Sentiment analysis of smart phone product review using SVM classification technique.* Chennai, IEEE.

Parveen, H. & Pandey, S., 2017. *Sentiment analysis on Twitter Data-set using Naive Bayes algorithm.* Bangalore, IEEE.

Rahat, A. M., Kahir, A. & Masum, A. K. M., 2019. *Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset.* Moradabad, IEEE.

Wang, Y., 2017. *Advanced Naïve Bayes Algorithm Design with Part-of-Speech Tagger on Sentiment Analysis.* Dalian, IEEE.

Wongkar, M. & Angdresey, A., 2019. *Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter.* Semarang, IEEE.