Final Project Literature Review

Kateryna Solonenko

CIND820: Big Data Analytics Project

Tamer Abdou, Ph.D

February 14, 2022

Abstract

With the rise of online review platforms and social media, it is becoming increasingly important for businesses to analyze public opinion about a product or service through customer feedback to have a leading edge over their competitors. Manually sorting through thousands of reviews and social media posts is very time consuming and inefficient, thus it is more beneficial for various industries to invest in machine learning research. Insights from this exploration can be used to monitor the strengths and weaknesses of a product, while comparing it to how other products are received online. This project will focus on text mining and sentiment analysis, which are commonly used methods for categorizing and analyzing bodies of text. The dataset that I will be using is the Large Movie Review Dataset[1], which is a publicly available dataset. This dataset has a total of 50,000 movie reviews labeled as having a negative or positive sentiment. The research topic is to build a model that can predict sentiment classification based on the content of the reviews, and to analyze which common factors distinguish positive and negative reviews. To achieve this goal, the NLTK Python library will be used to process the text, TF-IDF scores will determine word weight across reviews, and a Naive Bayes classifier model will predict outcomes.

Introduction

Sentiment analysis, or opinion mining, is the field of study that analyzes the opinions and attitudes that people have towards products, services, topics, etc. and the attributes involved (Liu, 2012).

Sentiment analysis is currently one of the fastest growing areas of research in computer science The rapid growth of this field coincides with and is reliant on the continual rise of social media. (Liu, 2012). In 2019, there were 2.95 billion people active on social media globally, with users expected to rise to nearly 3.43 billion by 2023 (Dwivedi et al, 2021).

People are drawn to the interactivity of the platforms and the immediacy of information (Samuels & Mcgonical, 2020). Thus, the decline of traditional media necessitates that businesses optimize their digital and social media strategies to survive in the current market. It is no coincidence that over 88% of businesses use Twitter to market and build their brand (Dwivedi et al, 2021).

Social networks provide consumers the ability to easily and visibly express their opinions about products and services online. Peer recommendations notably influence consumer trust and brand loyalty (Dwivedi et al, 2021).

Social media channels such as Facebook and Snapchat are monoliths of consumer sentiment data, which can provide powerful insights to marketing and advertising (Lexalytics, 2019). For example, a company would be able to pivot their marketing strategy in real time by monitoring market trends and the retention rate of customers of a new product, potentially reducing marketing costs (Gupta, 2018).

However, the enormous volume of these posts is a problem, as it would be impossible for humans to manually parse all of that information in real-time. Machine learning is a great tool to more efficiently digest this data, but there is also a barrier to accessing useful information from these sources, as social media posts are full of emoticons, acronyms, and other features which make it difficult for machine learning models to properly assess information (Lexalytics, 2019).

Sentiment analysis allows businesses to understand customer sentiment on a large scale and assess their standing relative to their competition, making it an important strategy for success in the highly competitive market.

Related Work

In Sentiment Analysis for IMDB Movie Review (Li, 2019), Ang Li looks into the IMBD Movie Review dataset from Stanford Artificial Intelligence Laboratory to find which features may be indicative of whether a review is positive or negative. To do this, Li introduces a sentiment analysis classification model which uses embedded context information.

Li aims to increase classification performance through parameter tuning and extending the feature space of the movie review dataset. LightSIDE and WEKA were the tools used in this study. LightSIDE is an open-source data analysis software used for text mining, and WEKA is an open source software with a collection of tools for data analysis and predictive modeling. Python was used for data preparation, through numpy and pandas.

Decision Tree, Support Vector Machine (SVM), Naive Bayes and Logistic Regression models were compared. The Logistic Regression model was found to be most accurate, and was further fine-tuned for better accuracy, achieving significant improvement over the basic model.

Though the same dataset is used and there is a similar goal of creating a model to predict whether a movie review is positive or negative, the tools and means are different to those that I would like to use for this project. Instead of WEKA and LightSIDE softwares, Python and its available libraries will be used for data preparation, model building and analysis.

News Sentiment Analysis, a paper by Samuels and Mcgonical, provides a lexicon-based approach for sentiment analysis of news articles.

There are two popular approaches to sentiment analysis, one is lexicon based and the other is based on machine learning. The lexicon-based approach uses a dictionary of weighted words to determine the polarity of a text, whereas the machine learning method is a classification based approach which is trained to predict polarity based on previous examples (Samuels & Mcgonical, 2020).

The general methodology and techniques used in the Samuels and Mcgonical paper are similar to those that will be employed in this project. Text is preprocessed by cutting out irrelevant information, remaining words are weighted with IT-IDF scores, and bodies of text are classified based on the total weight of all combined weighted words. However, this project will focus on a machine learning based approach rather than a lexicon based one.

Data Set

The dataset that will be used is the Large Movie Review Dataset[1], which is a publicly available dataset from the Stanford Artificial Intelligence Laboratory. This dataset has a total of 50,000 movie reviews split into a 25,000 review training set and a 25,000 review testing set, with every review labeled as having a negative or positive sentiment.

Only polarized reviews are used in the training and test set. Negative reviews have a score of <= 4 out of 10, and positive reviews have a score of >= 7 out of 10. The data set is split evenly between positive and negative reviews.

A random sample of 1000 reviews will be taken from the overall data set, as 50,000 reviews would require a lot of computation.

Approach

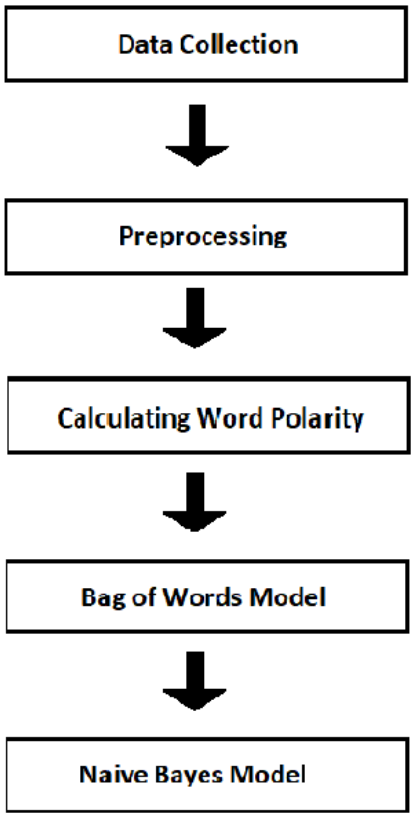
Sentiment analysis can be carried out using supervised or unsupervised approaches. A supervised approach uses labeled training data to build a classification model to later be used to predict outcome of unlabeled data. Unsupervised, or lexicon-based, approaches do not require training data, and use word polarity to determine sentiment (Samuels & Mcgonical, 2020). .

A supervised approach will be used to determine review sentiment, as the data set being used for this project is labeled. The Natural Language Toolkit (NLTK) library in Python, TF-IDF scores and Naive Bayes classifier will be used to achieve this goal.

The major constraint of TF-IDF is that the algorithm cannot identify or differentiate words that are similar in meaning but differ grammatically. For example, “playing” and “plays” use the same base verb, but because the overall form is changed, they are treated as unrelated words (Qaiser & Ali, 2018).

To overcome this issue, stemming will be included in the preprocessing step. Stemming is the process of removing grammatical forms from words, leaving only the roots. For example, “playing” and “plays” would become “play” after removing the -ing and -s stems.

Methodology



The methodology consists of 5 steps: Data collection, Preprocessing, Calculating Word Polarity, Bag of Word Model, and Naive Bayes Model.

1. Data Collection
   IMDB Movie Review dataset will be imported as a dataframe for processing.

2. Data Preprocessing
   The dataset will be processed for redundant information for more effective analysis. Punctuation, numbers, spaces between words, and infrequent words will be removed.

   Stop words provided in the NLTK stopword corpus will be removed.
   Stop words are words that do not add semantic information to text, but are syntactic fillers to complete sentences. For example, words such as “the”, “a”, “are”.

   Word stems will be removed by a process of stemming, leaving only the roots of words. This will lessen redundancy by grouping words with similar meaning.

3. Calculate Word Polarity
   Term Frequency - Inverse Document Frequency, or TF-IDF, will be used to assign weight to words based on their frequencies relative to other word frequencies in the data.
   TF-IDF is a numerical measure of how essential a word is to a particular text within a corpus.

4. Bag of Words Model
   The Bag of Words Model, or BOW, is used to create a vector of all words and their frequency. This step will be necessary for predictive modeling and analysis, as these processes cannot be applied to strings.

5. Naive Bayes Model
   The dataset will be split up into 70% training data, and 30% testing data.
   A Naive Bayes model will be created using the scikit-learn package, and evaluated using the metrics module.

Limitations and Future Considerations

One of the most notable reasons leading to errors in sentence-level Sentiment Analysis is the difficulty in parsing sentence negation, particularly in how negation words influence other words in a sentence. Classification performance is improved by negation identification, as it improves accuracy in identifying sentence polarity (Mukherjee et al, 2021).

According to research by Mukherjeea et al, the difference in performance is not as relevant to neural networks, but is more notable for Naive Bayes and Support vector machines. Deep learning algorithms outperform traditional machine learning models in both cases.

Though it is a big leap to delve into deep learning algorithms, a sentence negation parser is something to be considered in future approaches to Sentiment Analysis models.

References

Cardie, C. (2014). Sentiment Analysis and Opinion Mining Bing Liu (University of Illinois at Chicago) Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, 5(1)), 2012, 167 pp; paperbound, ISBN 978-1-60845-884-4. *Computational Linguistics, 40*, 511-513.

Dwivedi, Y.K., Ismagilova, E., Hughes, D.L., Carlson, J., Filieri, R., Jacobson, J., Jain, V., Karjaluoto, H., Kefi, H., Krishen, A.S., Kumar, V., Rahman, M.M., Raman, R., Rauschnabel, P.A., Rowley, J.E., Salo, J.T., Tran, G.A., & Wang, Y. (2021). Setting the future of digital and social media marketing research: Perspectives and research propositions. *Int. J. Inf. Manag., 59*, 102168.

Gupta, S. (2018, March). Applications of sentiment analysis in business. Towards Data Science. Retrieved from https://towardsdatascience.com/applications-of-sentiment-analysis-in-business-b7e660e3de69

Lexalytics. (2019, December). Top applications of sentiment analysis & text analytics. Retrieved from https://www.lexalytics.com/applications

Li, A. (2019). Sentiment Analysis for IMDb Movie Review (Carnegie Mellon University) https://www.andrew.cmu.edu/user/angli2/li2019sentiment.pdf

Mukherjee, P., Badr, Y., Doppalapudi, S., Srinivasan, S.M., Sangwan, R.S., & Sharma, R. (2021). Effect of Negation in Sentences on Sentiment Analysis and Polarity Detection. *Procedia Computer Science, 185*, 370-379.

Samuels, A., & Mcgonical, J. (2020). News Sentiment Analysis. *ArXiv, abs/2007.02238*.

Qaiser, S., & Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*.