Sentiment Analysis

Deva Narayana Babu

May 2023

This project was undertaken as a course requirement for Bachelor of Science (Hons.) in Mathematics. It was done under the guidance of Dr.Darshan Gera.

Contents

1 Introduction

1.1 Motivation

When buying a product the opinion of people who have used it, i.e. reviews have always been a crucial part of the decision making process, whether it be the act of buying a cricket bat or purchase of real estate, an informed choice is made possible by inputs from relatives, domain experts, etc. With the invent of the world wide web and the evolution of the online review communities, we can harness them to make and recommend better products. The primary motivation behind this project has been the idea that web scraping combined with sentiment analysis has the potential to do a mass harvesting of information that could potentially lead to better decision making in business application and for government operations. A company keeping track of customer feedback about product launches, using it to drive growth and innovation. This would be the description of an ideal company, a mouthwatering ideal for the booming startup ecosystem. Natural Language Processing takes us one step closer towards it. Financial markets are very susceptible to the market sentiments i.e. the view of people. So keeping a track of it means having an upper hand.[1] Detection of "flames"[6] directed at VIPS like the President, Prime Minister is also possible. It can be used for poll prediction on the basis of social media activity.

1.2 Objective

- The prime objective of this project is to compare the performance of the machine learning techniques on different data sets.
- To Learn the functioning and Mathematics behind these models.
- To Understand how the problem of sentiment classification is solved in the literature and the different ways to do it.

1.3 Problem Statement

"Sentiment Polarity Classification Using Machine Learning Techniques"

2 Approach

Literature survey revealed that the following are the most commonly used approaches to Text Representation in Natural Language Processing.

- Parts of speech tagging
- Bag of word model
  - Term presence
  - Term frequency
- Word embeddings

The earliest works in sentiment classification began with part of speech tagging to determine the polarity of the text. Adjectives were predominantly used. For instance, if a movie review data sample had the words ‘dazzling’, ‘awesome’, ‘breathtaking’ then the review was most likely to be positive in nature and if the review had words like ‘horrible’, ‘disgusting’, then the review is negative. Subtler and more efficient forms of POS tagging based approaches have emerged that perform good.

2.1 BOW model

The process behind the Bag of words model is creating a ‘bag’ which stores all the words that exist in a data frame or any other data storage entity, and the all the samples become vectors, whose components can have binary values signifying absence or presence of the particular word. For instance, If I want to represent the word “I like ice cream” in a the corpus “ice” ,“monkeys”,”jumps”,”am”,”I”,”cream”,”hi”,”like”,”hello”,”more”,”than” then the vector would look like [1,0,0,0,1,1,0,1,0,0,0]. A more formal definition given in [5] is as follows:

Let f_1, . . . , f_m be a predefined set of m features that can appear in a document; examples include the word “still” or the bigram “really stinks”. Let n_i(d) be the number of times f_i occurs in document d.
Then, each document d is represented by the document vector

$$d := (n_1(d), n_2(d), ..., n_m(d)).$$

2.2 Word Embedding

The idea behind word embeddings is to map each word to a point in a high-dimensional vector space, where words with similar meanings are closer to each other than words with different meanings. This representation allows NLP models to work with words as continuous, numerical values rather than discrete symbols, making it easier for them to learn from text data and generalize to new words and phrases.

3 Challenges

- Presence of no ostensibly negative words
  "If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut." This forementioned review of a perfume does not contain any "negative word" like "not happy","not user-friendly","not cheap", yet it expresses a very strong negative opinion about the described product. These kinds of reviews are hard to classify purely on the basis of the presence of negative words, a similar example follows: "Presence of Jane Austen's books madden me so that I can't conceal my frenzy from the reader. Everytime I read 'Pride and Prejudice' I want to dig her up and beat her over the skull with her own shin-bone."

- Frequency Effects can often be misleading
  "This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up."

  This review is dominated by positive words, however the last sentence completely flips the table, showing how the ordering of the statement plays an important role and how these frequency effects are not so reliable.

In more subtler forms of analysis, this becomes clearer. The sentence "Hanuman is the most devoted being" gets processed fine, but "Hanuman is more devoted than Vibhashana" is tricky since this statement is equivalent to "Vibhishana is more devoted than Hanuman" in the Bag of Words representation.

4 Methodology

4.1 Datasets Used

Datasets Used The Datasets used in this project are as follows:

- Amazon Fine Foods Reviews [3] This was an imbalanced dataset where positive reviews were greater than negative. It has the following data characteristics:
    - Reviews from Oct 1999 - Oct 2012
    - 568,454 reviews
    - 256,059 users
    - 74,258 products
    - 260 users with ¿ 50 reviews
- Sentiment140 [2]
    - Balanced Dataset with 1.6M Data Samples
    - Extracted using the Twitter API.
- Movie Review [4]
    - Balanced Dataset with 1000 Reviews.
    - To make the corpus more representative the contribution of some prolific reviewers was limited.
    - More than 200+ reviewers were represented.

4.2 EDA

The Dataset "Movie Reviews"[4] available in raw form had to be preprocessed owing to some characters not present in utf-8 encoding, since the only 'illegal' character was the copyright symbol, all non-ascii characters were stripped; no useful information was lost.

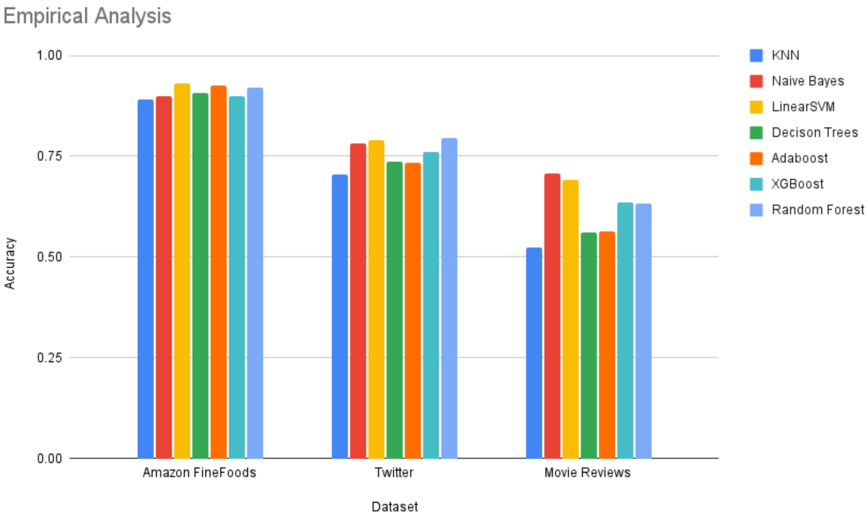
Followed by which a dataframe was created out of this raw data, and then correlation between the target variable and the other features was computed using a heatmap. [insert all those images from sys40]

Then after the most relevant features were determined for future work, the work to determine the best model for each respective dataset began.

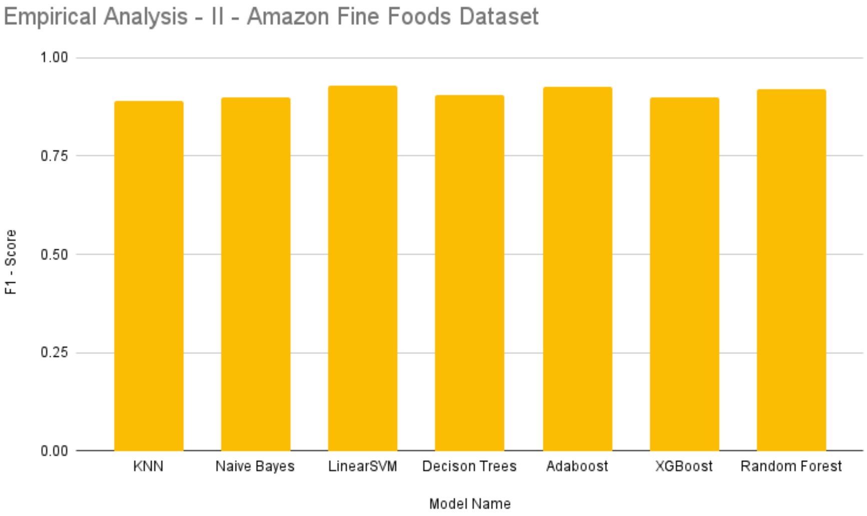
The models used in this project were as follow:

1. K-Nearest Neighbors
2. Linear SVM
3. Decision Trees
4. XGBoost
5. Random Forest

5 Results



(a) Empirical Analysis - I



(b) Empirical Analysis - II - Amazon Fine Foods Dataset

6 Conclusions

- The “Movie Reviews” dataset had a lot of noise, so the complex models like XGBoost and Random Forest overfitted due to which it had relatively poor accuracy, while the simpler Naive Bayes model gave the best accuracy.
- All the models performed substantially better in the Amazon Fine Foods dataset, possibly due to the presence of the “review summary” feature’s presence. To validate this hypothesis, the Sentiment140 dataset - which does not contain such a feature - was used.
- Models performed better in this dataset, compared to Movie Reviews but nowhere as good as in the AFF dataset.
- Linear SVM has been the best performing model in AFF and second best in both Movie Reviews, and Sentiment140.

References

[1] Ann Devitt and Khurshid Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 984–991, 2007.

[2] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

[3] Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908, 2013.

[4] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*, 2004.

[5] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*, 2002.

[6] Ellen Spertus. Smokey: Automatic recognition of hostile messages. In *Aaai/iaai*, pages 1058–1065, 1997.