الجامعة الإسلامية العالمية ماليزيا
INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA
يونيورسيتي إسلام انتارابغسا ملايسيا

Course:

NATURAL LANGUAGE PROCESSING

CSC 4309

i-Rate Movies
Sentiment analysis of movie reviews

Name	Matric No
Abdirahman Abdullahi	1432401
Asem Hamood Al-Abdali	1513599
Mohamed Alrefaei	1617111
MHD Khaled Maen	1523591

1. Introduction

Sentiment analysis is language processing task that uses a computational approach to identify uncompromising content and categorize it as positive, negative or neutral. Sentiment analysis is also a well-known task in the domain of natural language processing the unstructured textual data on movie reviews often carries the expression of opinions of users. Sentiment analysis tries to differentiate the expressions of opinion and mood of writers. A simple sentiment analysis algorithm attempts to distinguish a document as 'positive' or 'negative', based on the opinion expressed in it.

2. Related Work

A method of automatic sentiment analysis of movie reviews was presented by (Thet, Na, & Khoo, 2010). The method focuses on a linguistic approach of computing the opinion of a clause from the prior sentiment scores assigned to individual words. Sentences will be divided into individual portion and then check if there is any grammatical mistakes. The approach found to be effective for short documents such as message posts on discussion boards. This method can be used to sentiment summarization multiple review aspects, unlike other methods that only focuses on the positive and negative opinions.

On a recent survey of dimensional and categorical emotion recognition system that was provided by (Z., M., G.I., & T.S., 2009). A new line was seen in a video retrieval field of research addressing the multimodal fusion of language. Some features such as visual gestures and acoustic features used subtitle project that uses all three modalities to perform video retrieval

In (Bahrainian & Dengel, 2013) work, a novel solution was introduced to solve short informal texts focusing on Twitter posts (Tweets). Two methods were compared which are the state-of-the-art SA methods against a novel hybrid method. The hybrid method uses SL with Machine learning for opinionated texts in

the domain of consumer-products. Results showed that the proposed method was more accurate and outperforms the unigram baseline.

In the simplest setting of sentiment analysis a discrimination only between positive and negative sentiment. (Mesnil, Mikolov, Ranzato, & Bengio, 2014) compared different machine learning algorithm to this problem. The algorithms were a generative approach (language models), continuous representations of sentences and a clever re-weighing of tf-idf bag-of-word representation of the document. The three algorithms were combined, each model contributed to achieve a strong score on IMDB dataset reviews.

3. Problem Statement

movies review is becoming a more useful and important information resource for people, and it is very important to get people's reviews on a particular topic of interest. However, due to the lack of strong grammatical structures in movie reviews, which makes it hard for language processing to classify the review (positive or negative). In this project, a method to solve this problem is proposed.

4. Motivation

The motivation behind this approach is that the people who review the movies, their reviews are used to rate the movies. Therefore, if the users want to watch a movie, they can look directly to the rate and the reviews, and they can decide whether to watch the movie or not.

5. Technical background

this phase identifies what used in this system, such programming language, libraries, and tools, we will introduce all possible tools that used in this system.

Python

Python is a programming language that has efficient high-level data structures and approach to object-oriented programming. Python has great packages to use for machine learning and deep learning algorithms.

Machine Learning

Machine learning is the core part of artificial intelligence (AI) and provides applications the ability to learn from past experience, the learning ability started from observation, in order to look for insights or patterns that could be analysed.

Natural Language Processing

Natural language is a subfield of artificial intelligence that gives the machine ability to understand and process human language. NLP has so many techniques such as information extraction which helps in extracting entity names, fact extraction, relationship extraction, and also NLP can do sentiment analysis to discover whether the sentence is positive or negative or uses in any other intention or purpose.

Regular Expression:

Regular expression is a function used in programming to find pattern matched. It provides string matching of a text. For example, a regular expression can search a word through large text and also replace particular word to different word.

One Hot Encoding

One hot encoding is a numerical representation, it converts a categorical value into binary vectors.

Logistic Regression

Logistic Regression is a predictive model that used to describe the relationships between binary variables, means its a method to solve binary classification.

6. Experimental setup & Results

Data set

The data set used for the experiment is taken from IMDb[1], which is 25000 reviews for movies 50% of the data is positive reviews and the other half is negative reviews.

Data preprocessing

Since users' reviews are not controlled the data might be messy and it needs to be cleaned to have a good performance model, in our case regular expressions being used to delete some unrelated text such as spaces and Punctuation.

```
<br />All in all, it's worth a watch, though it's definitely not
Friday/Saturday night fare.<br /><br />It rates a 7.7/10 from...<br
/><br />the Fiend :."
```

```
of the story all in all its worth a watch though its definitely not
friday saturday night fare it rates a   from the fiend"
```

Data before and after preprocessing

Vectorization

Vectorization is the method of converting the words in the reviews into numerical values so machine learning algorithm can understand it, after that store

[1] (n.d.). IMDb: Ratings and Reviews for New Movies .... Retrieved May 8, 2019, from https://www.imdb.com/

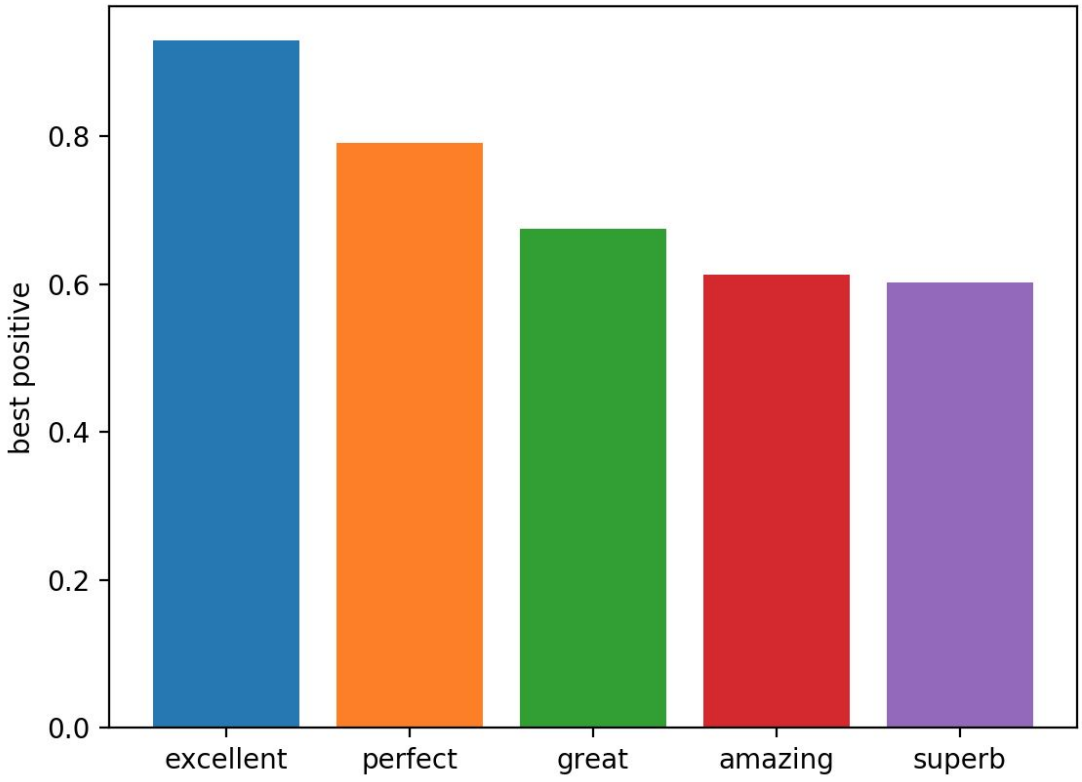
the data into a matrix with one column for each unique word, then each review's word is stored in one row as 0 or 1, where 0 means that the word is not in the corpus and 1 means that it's there.

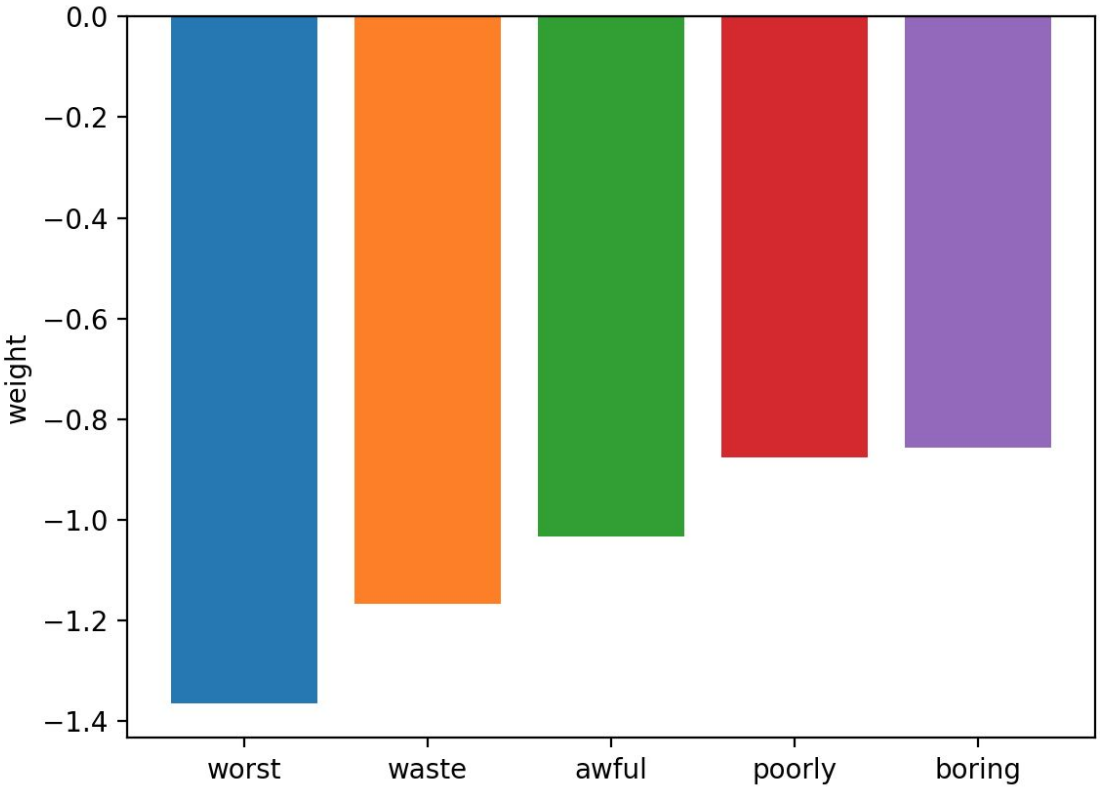
Algorithm

The algorithm being used in this experiment is Logistic Regression, Because linear models tend to work well sparse dataset and the learning is very rate is very fast comparing to the other algorithms.

Results

After training data with different C (Inverse of regularization strength), turned out that is the best value for regularization is C=0.05 with accuracy 0.88152.





The most effective positive and negative words

REFERENCES

Bahrainian, S. A., & Dengel, A. (2013). Sentiment Analysis using sentiment features. *Proceedings - 2013 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops,WI-IATW2013*,*3*,26–29. https://doi.org/10.1109/WI-IAT.2013.145

Mesnil, G., Mikolov, T., Ranzato, M., & Bengio, Y. (2014). Ensemble of Generative and Discriminative Techniques for Sentiment Analysis of Movie Reviews, 1–5. Retrieved from http://arxiv.org/abs/1412.5335

Thet, T. T., Na, J. C., & Khoo, C. S. G. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, *36*(6), 823–848. https://doi.org/10.1177/0165551510388123

Z., Z., M., P., G.I., R., & T.S., H. (2009). A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 39–58.