Sentiment Analysis: A Comparative Study

Hawas Alsadeed

Abstract

The study of public opinion can provide us with valuable information. Analyzing the large textual information manually is tough and time-consuming. Sentiment Analysis is an automated process that uses computing (AI) to spot positive and negative opinions from text. Sentiment Analysis is used to extract the subjective information in the source material by applying various techniques such as Natural Language Processing (NLP), Computational Linguistics and text analysis to classify the polarity of the opinion. Both Machine learning and Deep learning techniques can be used for Sentiment Analysis but deep learning techniques give more promising results. This project is more focused on Recurrent Neural Networks (LSTM), because in recent years they have shown the most promising results. I found an article in *Analytics Vidhya*, a famous Data Science magazine, about Sentiment Analysis using LSTM [1]. Using this article as a starting point for our study, I will build a comparative study around it.

1. Introduction

1.1 Motivation

The advent of the internet and smartphones has led to an explosion of blogs, forums, reviews, opinions, recommendations, ratings and online social media that enables the user to discuss topics and express their opinions online. They might, for

example, express their political views or express an opinion about the movie they recently watched. Deriving insights from such data is the crux of many applications such as recommendation systems, analyzing public opinion, organizing political campaigns, getting insights on products or businesses and analyzing reviews on products online.

These writer generated sentiment content can be about movies, hotels, events and restaurants. Sentiment Analysis becomes an extremely useful tool for business and social media. Opinion mining or Sentiment Analysis is the use of NLP for identification and classification of opinions expressed in the text. The sources of data for Sentiment Analysis (SA) are online social media, the users of which generate an ever-increasing amount of information. We create 1.5 quintillion bytes of information daily which can't be possibly processed manually and automation comes in handy. Some of the major uses of Sentiment Analysis are social media monitoring, brand monitoring and reputation management, product analysis, listening to the voice of the customer and competitive research. For example, in recommender systems or personalization, the understanding of the user opinion/sentiment can be very useful if explicit user feedback is not available on the said service/product. Another example is labeling a product review positive or negative would provide a succinct summary to the reader.

Sentiment Analysis is an active research area where collective efforts are being made to investigate the problems encountered in Sentiment Analysis (SA) and in increasing the accuracy.

1.2 Approaches for Sentiment Analysis

The combination of Natural Language Processing, Statistics and Computer Sciences is used for Sentiment Analysis (SA). Natural Language Processing is the branch of

Artificial Intelligence where computers are trained to understand human languages and communicate with humans.

Both Machine learning and Deep Learning techniques can be used for Sentiment Analysis. And in this project I have explored both the approaches conducting a comparative study of various techniques - Starting with a simple Machine Learning model like Logistic Regression and progressively moving towards more complex techniques like Recurrent Neural Network.

2. Background

2.1 Various methods for Sentiment Analysis

The dataset used for this project is IMDB movie review dataset, which is a public benchmark dataset widely used for SA. The dataset contains 50K highly polar movie reviews.In traditional Machine Learning techniques predictor variables are extracted manually from the reviews by using NLP libraries.Then these are used for binary classification using traditional ML models such as SVM, Random Forest, Logistic Regression, Naive Bayes. Accuracy of ML models mainly depends on the predictor variables chosen. In Deep learning techniques, namely Artificial Neural Networks predictor variables are automatically defined and extracted which can lead to better accuracy.

The data cleaning and data preprocessing remains somewhat the same for both the techniques.

. Fig1: source[2]

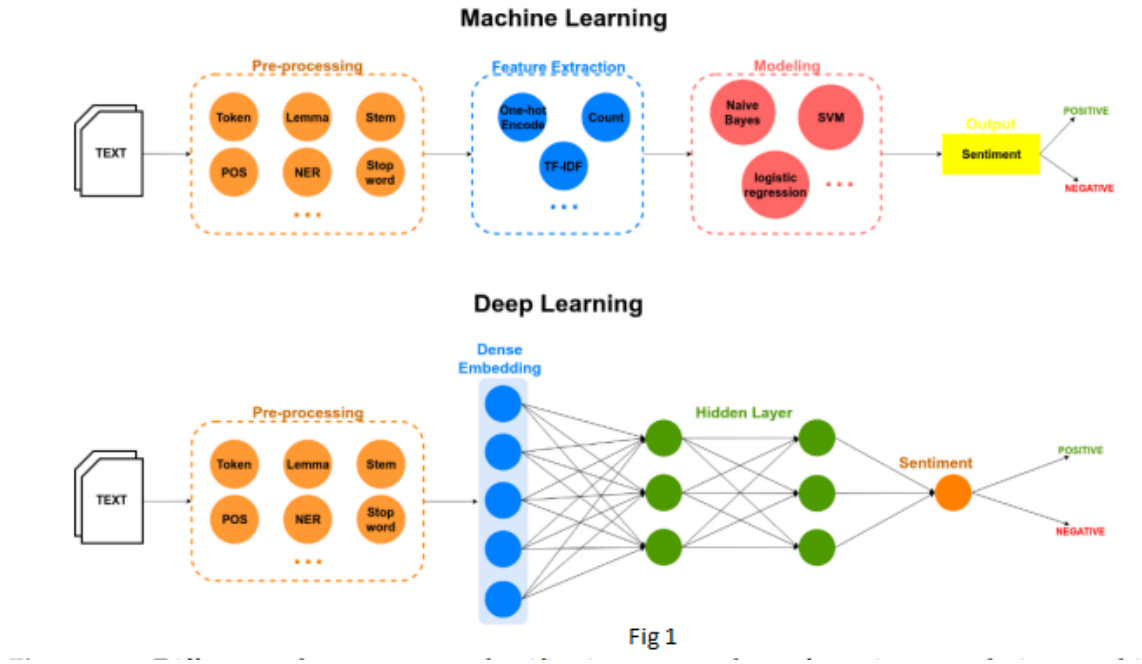


Fig 1

For Sentiment Analysis majorly two types of Deep learning techniques are used, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). A Simple Neural network basically maps the inputs to outputs using a mathematical function. CNN is very popular in computer vision as it has multiple sub classification layers to filter different aspects of an image helping in image classification.

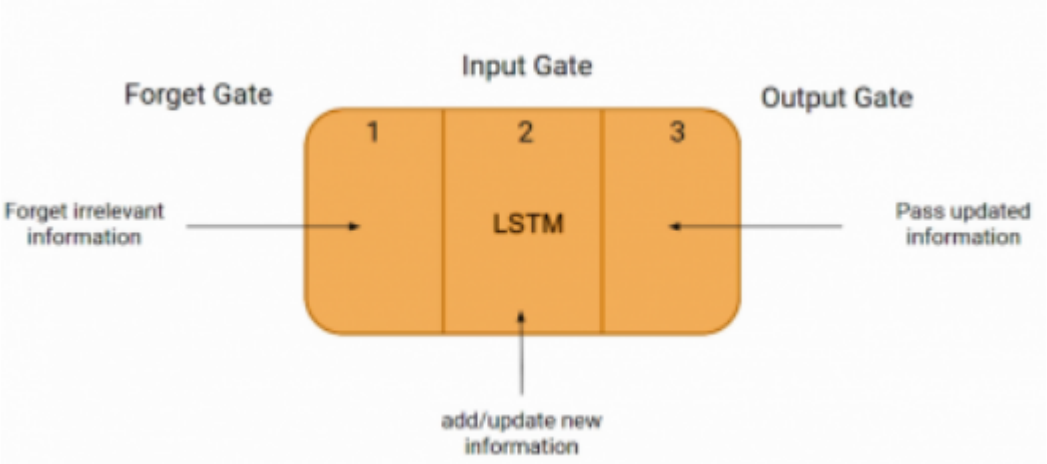
Recent research shows that RNN has better accuracy with text data and also for SA because of its ability to work with sequential data, it does so by storing information in the internal memory. A sentence is an example of sequential data where the order of the words in the sentence is most important. RNN can remember the previous words in a sentence and then predict the next word based on it. A classic example would be the auto complete feature in Gmail.

A special and popular type of RNN is long short term memory (LSTM). It's built to selectively remember patterns for a long duration of time. Hence it is one of the best choices for sequential data.

2.2. LSTM in a nutshell

The LSTM architecture has long term memory called cell state. Important words from a sentence are stored here. Forget gate erases those words to input another important word using input gate, and finally output gate determines which output should be passed on.

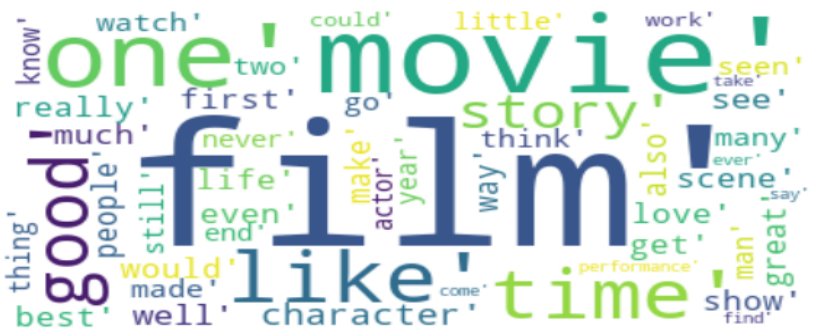
Fig 2: source[5]



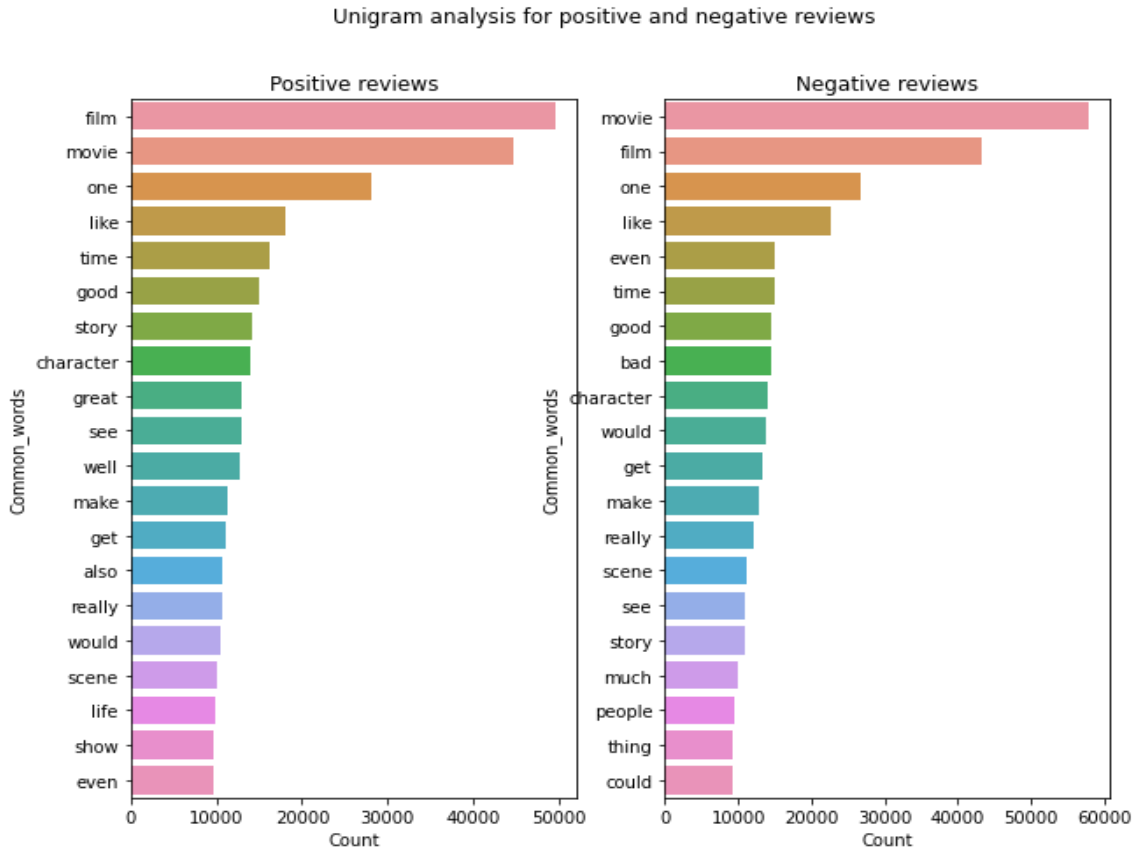
3. Sentiment Analysis on IMDB dataset

3.1 Data cleaning and EDA

The IMDB movie review dataset has 25000 positive reviews and 25000 negative reviews. Data cleaning was done by removing special characters, null values and duplicate reviews. Lemmatization and removing stop words were done as a part of data preprocessing. EDA showed the more frequent words in the positive reviews as well as the negative reviews. An image of those words was then created using WordCloud.



Most frequent words in the positive reviews appear larger in the WordCloud image. Above figure was generated during EDA using WordCloud for positive reviews. In EDA the distribution of most frequent words was studied. Unigram, Bigram and Trigram analysis showed which words were used frequently and were used together.



3.2 Machine learning

For traditional Machine learning models we need to define the predictor variables / features. Turning text into features can be done by two popular methods, Bag of words (BOW) and TF-IDF (Term Frequency- Inverse Document Frequency). One major drawback of these techniques is that they don't retain the contextual information about the words or sentences. As the TF-IDF technique assigns weights to words with higher frequency it was preferred over BOW. Following Machine Learning models were used: Support Vector Classifier (SVC), Logistic Regression

(LR), Random Forest (RF), Multinomial Naive Bayes, XgBoost. As it's a balanced dataset, accuracy was used as a model performance metric.

ML Model	Accuracy (%)
Support Vector Classifier (SVC)	89.80
Logistic Regression(LR)	89.35
Random Forest (RF)	84.48
Multinomial Naive Bayes	86.75
XgBoost	79.45

3.3 Deep learning

TF-IDF and BOW sequence information is discarded and it can affect the accuracy. For Sentiment Analysis sequence, information is very important. Word2vec is a word embedding technique where each word is mapped to a vector space and similar meaning words are mapped close by in that vector space.

After the data is cleaned, it is lemmatized to its root form then converted into numbers by word embedding techniques. Word2vec is a popular word embedding technique for SA. After word2vec the embedding layer was created by following steps

- Determine vocabulary size
- Use tokenizer to create index for words
- Define embedding layer dimensions = 100
- Use sequential model to create embedding layer

After the data preprocessing this can be used for simple RNN, GRU, LSTM and CNN Following are the results after running RNN, GRU and LSTM for 5 epoch and CNN for 10 epochs

Model	No. of epochs	Accuracy %
RNN	5	84.80
GRU	5	86.53
LSTM	5	86.12
CNN	10	78.99

The accuracy of the GRU model matches the LSTM model but LSTM took a longer time to train.The simple RNN model has accuracy lesser than all the other deep learning models as it can't retain sequential information. The CNN model took less time to train hence it was possible to train it for 10 epochs as compared to other RNN models. But its accuracy does not match LSTM's. Getting a higher accuracy for the LSTM model comes with a higher computational time.[2] And if we deal with big data we will have to keep the computational cost in mind.

4. Comparative Study and Conclusion

TF-IDF yielded good results in Traditional Machine learning models. SVC and logistic regression had 89% accuracy. The ML models took less computational resources and time compared to deep learning models and yielded a good accuracy. Further study needs to be conducted to measure the ML models performance against big data or a different dataset.

RNN, GRU and LSTM took a longer duration of time to train but in 5 epochs yielded similar accuracy to traditional machine learning models. With more computational power higher accuracy can be achieved with RNN models by hyper parameter tuning and running for many epochs.

In recent research RNN and CNN are the most popular techniques for Sentiment Analysis but little light has been shed on computational time needed for such accuracy.

5. Future Work

BERT (Bidirectional Encoder Representations from Transformers) is the state of the art NLP Model which was published in 2018 by Jacob Devlin. It has achieved state of art performance on many NLP tasks. Proposed future work is to use the BERT model on the IMDB dataset. Also expand this study to include Hybrid Models which combine CNN with LSTM.

Many popular studies on Sentiment Analysis include Twitter dataset and Amazon review dataset. In the future, I would like to compare Deep learning Model performances for different benchmark datasets to create an extensive study of performance of different deep learning techniques on different datasets.

6. Acknowledgments and References

[1].Sentiment Analysis with LSTM - *Analytics Vidhya* by Koushiki Dasgupta Chaudhuri

[2].Sentiment Analysis based on deep learning: A comparative study published in MDPI journal in March 2020.

[3].Understanding LSTM Networks -- colah's blog : Christopher Colah former Researcher at Google.

[4].Chapter 14 of Grokking Deep Learning by Andrew W. Trask (Manning Publications)

[5]. LSTM | Introduction to LSTM | Long Short Term Memor (analyticsvidhya.com) by shipra_saxena

Data Documentation

IMDB review dataset is a public benchmark dataset, which was widely used for Sentiment Analysis and text analytics. This is a dataset for binary sentiment classification, It includes 50k highly polar movie reviews.

Dataset contains 2 columns, 1st column being the review and the 2nd column being the sentiment associated with the review. Sentiment column is classified as Positive or Negative.

Following is the link to the dataset: Sentiment Analysis (stanford.edu)

Notable Publications using this dataset : Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word

Vectors for Sentiment Analysis. *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).*