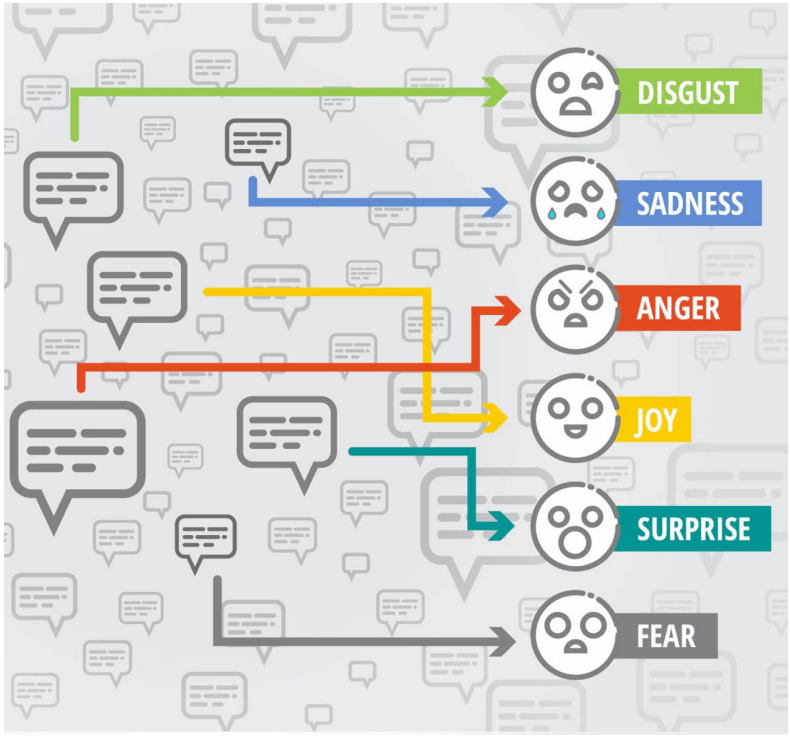
International Burch University
71000 Sarajevo

Research paper

Sentiment Analysis on IMDB reviews



Student: Biševac Nežla
Mentor: Samed Jukić, assist.prof.dr

Sarajevo, 05 January, 2021.

Abstract:

Social networks sites have, in a short period of time, become an extremely popular place for people to express their emotions, thoughts and the way they think, through short texts. The emotions they express, such as excitement, joy, fear, sadness, anger, allow us to analyse better whether they liked something or not.
Since the 1990's, when IMDB was released, the urge of rating and commenting movies has expanded. Whenever someone sees a film/tv series/tv show, one of the first thing they do is to visit the IMDB site to share their opinion with the others and to give a certain rating.

Sentiment Analysis on IMDb movie reviews identifies the overall sentiment or opinion expressed by a reviewer towards a movie. Many researchers are working on pruning the sentiment analysis model that clearly identifies and distinguishes between a positive review and a negative review. In the proposed work, we used Bag of Word for feature representation and the ANN for model training.

2 INTRODUCTION

SOCIAL media has become an integral part of human living in recent days. People want to share each and every happening of their life on social media. Nowadays, social media is used for showcasing one's pride or esteem by posting photos, text, video clips, etc. The text plays a vital aspect in information shared, where users share their opinions on trending topics, politics, movie reviews, etc. These opinions which people share on social networking sites are generally known as Short Texts (ST) because of its length [1].

ST have gained its importance over traditional blogging because of their simplicity and effectiveness in influencing the crowd. They are even used by search engines in the form of queries. Apart from their popularity, ST has certain challenges like identification of sarcasm, sentiment, use of slang words, etc. Therefore it becomes important to understand short texts and derive meaningful insights from them, which is generally known as Sentiment Analysis (SA) [2].
SA played an important role in the elections in the USA and Canada and is also known as a great tracker of customer feedback at many companies.
According to Cambridge English Dictionary, review is a report in a newspaper, magazine, or programme that gives an opinion about a new book, film, etc. Review can also be a short text that generally expresses an opinion about the topic. These reviews play an important role in the success of a movie/tv series/show because people generally look into blogs or

review sites like IMDB to know more about movie cast, crew, review and ratings. Therefore, Sentiment Analysis on movie reviews makes the task of Opinion Summarization easier by extracting the sentiment expressed by the reviewer.

The task of Sentiment Analysis on movie reviews includes Preprocessing, Feature Extraction, Training the model and finally the analysis of results. Preprocessing involves removal of stop words, replacing slangs, abbreviating short forms etf. Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones (and then discarding the original features). These new reduced set of features should then be able to summarize most of the information contained in the original set of features. Here as a feature extraction model we chose Bag-of-Words. The Bag-of-Words is a simplifying representation used in natural language processing (NLP) and information retrieval (IR). In this model, a text is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity.

The polarity of a review depends on the intensity of each word present in the review and the context used by the reviewers to express their opinion. Therefore, identifying the features that extract the intensity of words based on context that inclines the polarity either towards positive or negative polarity is a challenging task.
The major contribution of this paper includes:

- Use of nltk's stopword which reduces memory overhead, reduces noise and false positives and can potentially improve power of prediction
- Already having splitted dataset into train, valid and test csv's
- Using ANN model that gave us great accuracy.

2 METHODOLOGY

Previously, sentimental analysis of text or more precisely negative/positive classifcation depends on using a classifer and a dataset that applies the classifer into sets of two such as negative –ve and positive +ve.
Generally, sentiment detection using machine learning techniques is divided into two types: supervised method and unsupervised methods. Supervised methods focus on creating sentiment classifcation models by labeling the dataset or the document and thus help in making proper decisions, whereas in unsupervised methods, the dataset or the document is not labeled; instead, they rely on the statistical features of the documents.

Feature Extraction identifies the features that have a positive effect towards classification. In this work, extraction is carried in a Machine learning based feature extraction. Machine learning based feature extraction method is used to extract the features using popularly known technique Bag of Words, wherein the column corresponds to words and row corresponds to value of weighing measures such as Term Frequency (TF) and Term FrequencyInverse Document Frequency (TF-IDF).

2.1 Preprocessing

IMDB reviews data was collected, processed and analysed using Python programming language and its packages.

For each review in the dataset, a series of standard text preprocessing operations were executed before applying a sentiment analysis approach. For the sake of dimensionality reduction and to remove noisy and inconsistent data, the following text pre-processing steps were performed on the dataset:

- Elimination of punctuations. The removal of special characters such as “ ’ ? ; : # $ % & ( ) * + − ⁄ < > = [ ] \ ˆ _ | ~ aims to avoid unwanted concatenations that render the words unavailable to a dictionary and resolve unwarranted discrepancies during polarity assignment phase.

```
dataframe['text'] = dataframe['text'].str.replace('[^\w\s]','')
```

Elimination of punctuations

- Elimination of numbers and stop words. Both numbers and stop words(e.g., “the”, “a”, “an”, “ang”, “ito”, 146 “mga”, etc.) available in English do not contribute to polarity. These words were removed to ensure complexity reduction.

```
sw = stopwords.words('english')
dataframe['text'] = dataframe['text'].apply(lambda x: " ".join(x for x in x.split() if x not in sw))
```

Elimination of stop words

- Elimination of rare characters

```
#rare characters deleting
sil = pd.Series(' '.join(dataframe['text']).split()).value_counts()[-1000:]
dataframe['text'] = dataframe['text'].apply(lambda x: " ".join(x for x in x.split() if x not in sil))
```

Elimination of rare characters

Also, many words that are present in the reviews will not be in their root forms. For example, words like ‘studying’, ‘studied’ belong to same root word ‘study’.

2.2 Bag-of-Words

The bag-of-words model is a way of representing text data when modeling text with machine learning algorithms. The bag-of-words model is simple to understand and

implement and has seen great success in problems such as language modeling and document classification.
A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:

1. A vocabulary of known words.
2. A measure of the presence of known words.

It is called a “bag” of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document.

“A very common feature extraction procedures for sentences and documents is the bag-of-words approach (BOW). In this approach, we look at the histogram of the words within the text, i.e. considering each word count as a feature.” [3]

The intuition is that documents are similar if they have similar content. Further, that from the content alone we can learn something about the meaning of the document.

2.3 Limitations of sentiment analysis

One of the best descriptions of sentiment analysis limitations is given by Matthew Russell in one of his interviews:
Like all opinions, sentiment is inherently subjective from person to person, and can even be outright irrational. It’s critical to mine a large — and relevant — sample of data when attempting to measure sentiment. No particular data point is necessarily relevant. It’s the aggregate that matters. An individual’s sentiment toward a brand or product may be influenced by one or more indirect causes; someone might have a bad day and tweet a negative remark about something they otherwise had a pretty neutral opinion about. With a large enough sample, outliers are diluted in the aggregate. Also, since sentiment very likely changes over time according to a person’s mood, world events, and so forth, it’s usually important to look at data from the standpoint of time. As to sarcasm, like any other type of natural language processing (NLP) analysis, *context matters*. Analyzing natural language data is, in my opinion, the problem of the next 2-3 decades. It’s an incredibly difficult issue, and sarcasm and other types of ironic language are inherently problematic for machines to detect when looked at in isolation. It’s imperative to have a sufficiently sophisticated and rigorous enough approach that relevant context can be taken into account. For example, that would require knowing that a particular user is generally sarcastic, ironic, or hyperbolic, or having a larger sample of the natural language data that provides clues to determine whether or not a phrase is ironic. [4]

Therefore, while sentiment analysis is useful, it is not a complete replacement for reading survey responses. Often, there are useful nuances in the comments themselves. Where sentiment analysis can help you further is by identifying which of these comments you should read.

3. State-of-the art implementation

The ANN (artifcial neural network) is a “tool for processing information that is inspired by the way biological nervous systems try to replicate the way the brain processes information. The key feature of this method is the structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working together to solve specifc problems.” Neural networks, like humans, learn by examples and replication of those examples. A neural network (NN) could be skilled for “specifc application, such as pattern recognition or data classifcation, through a learning process.” Learning in nature suggests changes in the synaptic connections existing between the neurons. That also applies to neural networks.

3.1 Neural networks (NNs)

The artifcial neural network is a developed tool that is used to process information. The motivation behind developing ANN was to imitate “how the human brain processes information,” which was stimulated by the biological nervous system. “Information processing system” is the main factor behind the structure of artifcial neural networks. Neural networks are a set of huge numbers of strongly connected processing elements called neurons, operating simultaneously to solve a specifc problem.

In human beings, neural networks can be studied by the behavior and its repetition. Nowadays, because of machine learning, there are wide research areas that are using neural network application to train for a “specifc purpose, i.e., pattern recognition or text classifcation, using a learning process.” “The artifcial neural network is a powerful data-driven, self-adaptive, fexible computational tool possessing the capability of capturing nonlinear and complex underlying characteristics of any physical process at a high degree” [34]. Because of this very mature nature, it is often known as a “Black Box.” These ANNs perhaps seem to work most of the time independently but still the involvement of humans is required at a certain point. To defne the complication or learning rate, it is required to set some “higher level” properties before the model knows about the real parameters. “Hyperparameters” are known as “knobs” of the big machines. Selecting hyperparameter standards corresponds to the prosecution of model selection. Hyperparameters could be “distinct (as in model selection) or continuous.” The values of the parameter can be

selected manually or adjusted by an algorithm, but the most important thing is to select them carefully.

3.2 A Simple neuron

An artifcial neuron is basically having a structure that consists of many inputs and only one output. . Fundamentally, “the training mode” and “the operation mode” are the two functioning mechanisms that are set in the artifcial neuron. The neuron could be trained, i.e., it executes or not in the training mode. And when it approaches the later one (operating mode) for a model in which neuron has been trained through input which is defned, it identifes the pattern of the network, and related output is fred. The fring rules defne if it should fre or not if the trained model is not renowned by the input pattern. Below is the fgure of a simple neural network. X is defned as an input function, and W is the weight of the inputs.

3.3 Multilayer perceptron

Multilayer perceptron is also referred to ANN and which nowadays is the most widely used as a type of neural network. An MLP is made up of a minimum of three layers making it nonlinearly for deep neural networks, which consist of input, hidden, and output layer.A perceptron is a particular neuron model that placed the basis of deeper and more complicated networks. A supervised learning approach can be used for MLP to train and learn about the correlations in between the given input and outputs. In order to reduce the error, we can modify the weights, parameters, and biases during the training process. MLP is also known as feedforward networks; it is just like a real neural network in human being’s information which is continuously moving between input layers to output layers. The basic purpose of it is to create ingenuous replicas of real neural networks so that people may take beneft of solving problems while creating strong data pattern techniques.

3.4 Experimental work

Network architecture (NA) used in the experiments is explained below. With one output layer and two hidden layers for binary classifcation, the network employs a “stochastic” gradient lineage for learning and a sigmoid activation function (Fig 1)

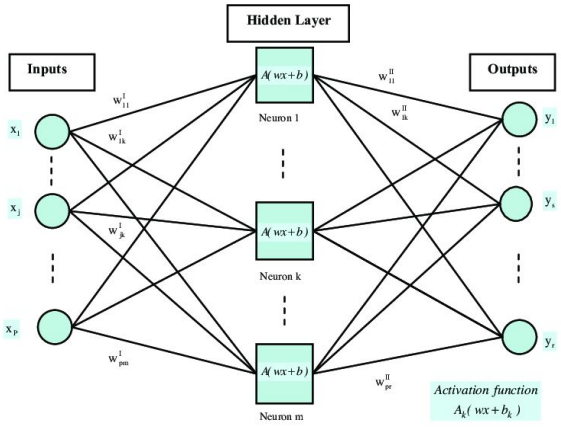


Fig 1: Neural network

3.5 Evaluation

Based on this approach, the proposed sentiment polarity relies on word weight instead of term frequency for each word. Every term has two values and polarity, as proposed in an assumption equation:

$$[Vw = Wp + Wn = 1]$$

where Vw is the “word value,” Wp goes for “positive word value,” and Wn stands for “negative word value.” The selection between negative polarities and positive polarities is afected by the meanings of a word and each of its polarity. The classifcation scale of sentiments is between 0 and 1. In other words, all probable values of each instance cannot exceed one. However, the sentence contains “negative that difers in the word value. If the word is positive, convert to negative polarity and the negative score will be as in the equation”: In the case of a negative word, the score will be calculated by the given equation: where 0.4 refers to the assumption level of one to fve (1–5) sentiments classifcation.

```
import tensorflow as tf
model = tf.keras.models.Sequential()
model.add(tf.keras.layers.Dense(50,input_dim=x_train_count.shape[1], kernel_initializer="uniform", activation="relu"))
#model.add(Dense(6, #ernel_initializer="uniform", activation="relu"))
model.add(tf.keras.layers.Dense(1, kernel_initializer="uniform", activation="sigmoid"))
model.compile(loss="binary_crossentropy", optimizer="adam", metrics=["accuracy"])
# fit the model
history = model.fit(x_train_count, train_y.values.reshape(-1,1), batch_size=128,epochs=2)
```

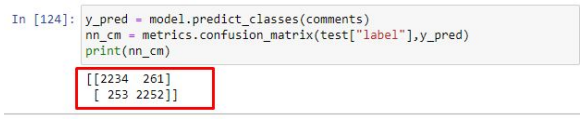
```
Epoch 1/2
313/313 [==============================] - 22s 66ms/step - loss: 0.4562 - accuracy: 0.7786
Epoch 2/2
313/313 [==============================] - 21s 67ms/step - loss: 0.1326 - accuracy: 0.9532
```

ANN *model*

4 CONCLUSION

Unarguably, sentimental analysis techniques are among the utmost signifcant bases in the decision-making process. A lot of people depend on sentimental analysis for achieving efficient results of services or products. We started with a model that was decent in

producing IMBD movie reviews. So, the idea of applying a pre-trained language model to actually outperformed the cutting-edge research in academia as well. It is an undeniable fact that human languages are relatively complex to be understood by the machine, which leads to conditions where a negatively said word has a positive association and vice versa. So, a sentimental analysis of movie reviews was a challenging task. In this study, we used neural networks to see whether the review is negative or positive.
Here is our confusion matrix:



From the dataset with 5000 rows, and from confusion matrix, here are the results:

	Predicted: NO	**Predicted: YES**	
Actual: NO	TN = 2234	FP = 261	2495
Actual: YES	FN = 253	TP = 2252	2505
	2487	2513	

From the information above, let's compute the accuracy:

$$\frac{TP + TN}{TP + TN + FP + FN} = 89.72\%$$

The accuracy is 89.72%, which means that it has almost reached 90%. So, we can say that the trained system accomplished to accomplish an ultimately exceptional final precision.

[1] C. D. Santos and M. Gatti. “Deep convolutional neural networks for sentiment analysis of short texts.” In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 69-78.

[2] A. Ortigosa, J. M. Martín, and R. M. Carro. “Sentiment analysis in Facebook and its application to e-learning.” Computers in Human Behavior Vol. 31, pp.527-541. 2014.

[3] Yoav Goldberg, Neural Network Methods in Natural Language Processing, 2017

[4] Matthew Russel, co-founder and Principal of Zaffra, interview by A.R. Guess, April 4,2011