MIT-WPU
॥ विश्वशान्तिर्ध्रुवं ध्रुवा ॥
Dr. Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

Project Report

on

Sentiment Analysis of IMDB Reviews Using Natural Language Processing

Submitted by:

Shwetha Iyer (1032211195)

Aarushi Gupta (1032211933)

Ayushi Sachan (1032211768)

in

Data Science

SY B.Tech ECE AI-ML

Under the Guidance of

Dr. Arti Khaparde

MIT WORLD PEACE UNIVERSITY

School of Electronics & Communication Engineering

2022 - 2023

Table of Contents

1. Introduction to Natural Language Processing:

Natural language processing refers to the field of artificial intelligence that deals with the creation of computational models that give computers the ability to process and understand natural language in the same way humans do. Natural language processing is a combination of computational linguistics, statistical models, machine learning and deep learning. These technologies allow computers to process text or voice data and understand its meaning, along with its intent and sentiment. The main goal of NLP is to make computers understand complex human language so that they can help humans with various language-related tasks. NLP techniques are used in many applications like machine translation, search engines, text summarization, sentiment analysis, question answering, etc.

NLP is a large part of everyday life and is utilized in voice-operated GPS systems, language translators, speech-to-text converters, digital assistants like Amazon's Alexa and Apple's Siri, customer service chatbots, and in several other diverse fields like retailing, medicine and business. One such famous computer program, GPT-3, can produce sophisticated prose on a wide range of topics and is even capable of holding coherent conversations. NLP is also employed by Google to improve search engine results, and by Facebook and other such social networks to identify and filter hate speech.

The working of NLP is divided into three parts:

1. Speech Recognition, which is the translation of spoken language into text.
2. Natural Language Understanding (NLU), which is the ability of a computer to understand the text.
3. Natural Language Generation (NLG), which is the ability of a computer to produce natural language.

2. Sentiment Analysis:

Sentiment analysis, also known as *opinion mining*, is the field of study that involves interpreting and classifying people's opinions, sentiments, and emotions from text using text analysis techniques. It is implemented by using a combination of natural language processing and statistics. Sentiment analysis mainly deals with the evaluation of opinions into positive, negative or neutral categories. This output is based on dependencies that are captured by word n-grams, TF-IDF features, or deep learning models.

Sentiment analysis allows organizations to identify public sentiment towards entities such as products, services, issues, events or topics. It is most commonly used in social media analysis to discover consumer insights or to identify signs of mental illness. Companies and organizations can then use this information to develop better products and campaigns. In this project, sentiment analysis was performed on IMDB reviews to understand the overall opinion of movies and television shows.

In order to build a language understanding model for sentiment analysis, three pre-requisites are needed:

- a particular problem statement
- a relevant data set, and
- an appropriate machine learning algorithm

3. Understanding the Dataset:

The CSV file used contains 50,000 rows of movie reviews from IMDB. There are two columns: ‘review’, which contains the review in the form of text, and ‘sentiment’, which categorizes the review as having a ‘positive’ or ‘negative’ sentiment. On inspecting the dataset, it was found that positive and negative reviews are evenly distributed. This sentiment must be converted to a binary classification for our convenience.

Both positive and negative reviews have ‘one’ as a recurrent word, which might be due to reviews like *“One of the best characters/movies/films/scripts I’ve ever seen”* or *“One of the worst characters/movies/films/scripts I’ve ever seen”*. There are some other words that are common to both positive and negative reviews, such as ‘movie’, ‘film’, ‘see’, ‘make’, ‘character’, ‘good’, etc.

4. Text Pre-processing:

Before processing occurs, the text should be pre-processed into a format that the model can understand so that the model’s performance is improved.

The steps taken in text pre-processing are:

1. ***Lower Casing:*** All text should be converted to lowercase. Words like great and Great have the same meaning, so they need not be represented as two different words.
2. ***Tokenization:*** It is the process of breaking text into smaller tokens, like words, word fragments, or sentences. Tokens help to understand the context of the text and are needed

to develop the NLP model. For example, “It is raining” can be tokenized into ‘It’, ‘is’, ‘raining’.

3. ***Noise Removal:*** This includes the removal of punctuation and special characters, html formatting, domain-specific keywords (like ‘RT’ for retweet), source codes, URLs, etc.
4. ***Stopword Removal:*** Stopwords are commonly used words (like ‘a’, ‘an’, ‘the’, etc.) that do not add much meaning to the sentence, and can be safely removed without affecting the text. This helps to reduce noise and to decrease the size of the dataset.
5. ***Stemming:*** It is an informal process of converting a word into its root form. Stemming works by slicing the end of the word, using a list of common prefixes and suffixes like (-ed, -ing, -s). One limitation of stemming is that it can result in meaningless words. For example, ‘university’ and ‘universe’ might all be mapped to the base ‘univers’, even though they have no close semantic relationship.
6. ***Lemmatizing:*** It is the process of converting a word to its base form by analyzing the word’s linguistic morphology. Unlike stemming, lemmatization always reduces the word to a meaningful word, for example, ‘better’ to ‘good’. Hence, lemmatization is preferred over stemming.

5. Feature Extraction:

After cleaning the initial text, its features need to be extracted to be used for modeling. Text data needs to be converted into the form of numerical data such as a vector space model. The process of transforming text data into numbers is called *Feature Extraction* or *text vectorization*. This data can then be fed into the machine learning algorithm.

The numbers that describe the corpus are created in the following ways:

a. **Bag-of-Words:** It is one of the most common text vectorization techniques. It represents a text by the number of occurrences of each word or n-gram (combination of n words). The text is converted into a sparse matrix, where each column represents a unique word (or n-gram), and each row represents the count. The advantage of considering n-grams instead of individual words is that the local ordering of words can be preserved. Bag-of-Words can be implemented directly by using CountVectorizer class in scikit-learn.

b. **TF-IDF:** TF-IDF is a statistical measure that weighs each word in a document by its significance. Two parameters are considered in TF-IDF:

$$Term\ Frequency\ (TF) = \frac{\text{No. of occurences of word in document}}{\text{No. of words in document}}$$

$$Inverse\ Document\ Frequency\ (IDF) = \log\left(\frac{\text{No. of documents in the corpus}}{\text{No. of documents where word appears}}\right)$$

c. **Word2Vec:** Introduced in 2013, Word2Vec is a deep learning-based technique that uses a neural network to convert a given word to a collection of numbers. It has two types: *Skip-Gram*, in which surrounding words are predicted from a target word, and *Continuous Bag-of-Words (CBOW)*, in which a target word is predicted from surrounding words.

These models take a word as input and generate a word embedding that can be used for NLP tasks. Word2Vec embeddings capture context and semantic meaning. Two words will have similar embeddings if they appear in similar contexts.

d. **GloVe:** Like Word2Vec, GloVe also produces word embeddings, but it does so by using matrix factorization techniques instead of neural learning. The advantage of GloVe is that it does not just rely on local contexts, but also incorporates global statistics (like counts of word-to-word co-occurrence) to generate vectors. The resulting vector representations are linear substructures of the word vector space.

6. NLP Techniques:

Traditional NLP is done using Machine Learning techniques:

- **Logistic Regression:** It is a supervised classification algorithm that aims to classify the data into categories or to predict the probability of an event based on some input. Logistic regression models can be applied to solve NLP problems like spam detection, sentiment analysis and toxicity classification.
- **Naive Bayes:** It is a supervised learning algorithm that works on the principle of conditional probability, as given by Bayes theorem. The naive word implies the assumption that individual words are independent. Naive Bayes is mainly used in text classification for spam detection or finding bugs in software code.
- **Decision Trees:** This is a hierarchical classification model that splits the dataset using a tree-like structure to predict the possible outcomes.

It was observed that logistic regression and Naïve Bayes model perform much better compared to other techniques in terms of speed and accuracy (~80%). Linear SVM also yields high accuracy (~85%) but processing time is much longer.

Newer NLP models employ Deep Learning techniques. Deep Learning is a subset of

Machine Learning that uses neural networks to learn unsupervised or unstructured data. Deep Learning applications of NLP include machine translation, language modelling, caption generation and question answering.

- **Convolutional Neural Network (CNN)**: These are neural networks that help to process data having grid-like topology, such as images. CNN assigns importance (learnable weights and biases) to the pixels in the image so that one object can be differentiated from another. CNN can be used in NLP by looking at a text document as an image. Instead of pixels, the input to the model is sentences or a matrix of words.
- **Recurrent Neural Network (RNN):** Most deep learning techniques like CNN do not learn the sequential structure of the data. RNNs are ideal for sentences where every word is dependent on the previous word or a word in the previous sentence. An RNN is a fully connected neural network that refactors some of its layers into a loop. It remembers previous information by using hidden states and connecting it to the current task. They have been used to translate human thoughts into words and generate mathematical proofs.

7. Challenges:

An important point to consider is that sentiment analysis is not a foolproof method and the outputs may be affected by other factors such as cultural differences, sarcasm and personal preferences or biases. Moreover, these reviews are submitted by individual users whose opinions may or may not be representative of the general population. Hence, when performing sentiment analysis of IMDB reviews, it is important to take into account other factors such as critical reviews and target demographics.

8. Conclusion:

With the size and sophistication of unstructured information rapidly growing, NLP is becoming more and more relevant. Engineers continue to find opportunities to apply NLP to a variety of diverse fields. Newer researches aim to further narrow the gaps between human language and computer-generated language.

References:

1. Ren, P., Xiao, Y., Chang, X., Huang, P. Y., Li, Z., Gupta, B. B., Chen, X., & Wang, X. (2020, August 30). *A Survey of Deep Active Learning*. arXiv.org. https://arxiv.org/abs/2009.00236v2

2. Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022, July 14). *Natural language processing: state of the art, current trends and challenges* - Multimedia Tools and Applications. SpringerLink. https://doi.org/10.1007/s11042-022-13428-4

3. Kang, Joon Yoo, & Han. (2012). *Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews*, Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2011.11.107.

4. Natural Language Processing (NLP) - A Complete Guide. (n.d.). Natural Language Processing (NLP) [a Complete Guide]. https://www.deeplearning.ai/resources/natural-language-processing/

5. Top Applications of NLP in 2023 - Intellipaat. (2020, April 30). Intellipaat Blog. https://intellipaat.com/blog/applications-of-nlp/