A Project Report for

**Predictive Analysis (Program Elective 1)**

On

**Sentiment Analysis**

Submitted to Manipal University, Jaipur

Towards the partial fulfilment of for the Award of the Degree of

**BACHELOR OF TECHNOLOGY**

In Computer Science and Engineering

**2020-24**

**By**

**Akshat Charan – 209301286**

**Ansh Harjai - 209301140**

**Pranshu Raghuwanshi - 209301252**

\_\_\_\_\_\_\_

**Under the guidance of**

**Mr Lav Upadhyay**

**Assistant Professor**

\_\_\_\_\_\_\_\_\_\_

Department of Computer Science and Engineering

School of Computing and Information Technology

Manipal University Jaipur

Jaipur, Rajasthan

**Introduction**

The emotions in textual data is analysed and processed in Sentiment Analysis. In other words the Sentiment Analysis specifies that the given information is positive or negative about a specific topic or product. For processing the textual information Sentiment Analysis adapts the approaches of NLP (Natural language processing), AI (Artificial Intelligence) and ML (Machine Learning). Sentiment Analysis is competent in understanding the people’s opinion and to provide a basis of evidence and reasoning on particular product. Sentiment Analysis analyses abundant textual information presented in internet to provide future insight for the organization and aid the public to take decision on their purchase. For instance, in an e-commerce website, people purchase product and give their reviews with or without using the product. These reviews must help the customers who desire to purchase the product or item. But the issue is, as the number of reviews is more the customers may find it difficult to read all reviews. So, there is a need of automated process to give an appropriate conclusion for particular product or topic and this task is known as Sentiment Analysis.

**Literature Review**

**Natural Language Processing**

**Natural Language Processing or NLP** is a branch of Artificial Intelligence which deal with bridging the machines understanding humans in their Natural Language. Natural Language can be in form of text or sound, which are used for humans to communicate each other. NLP can enable humans to communicate to machines in a natural way.

**Text Classification** is a process involved in Sentiment Analysis. It is classification of peoples opinion or expressions into different sentiments. Sentiments include Positive, Neutral*,* andNegative*,*Review RatingsandHappy, Sad*.* Sentiment Analysis can be done on different consumer centered industries to analyse people's opinion on a particular product or subject.

**Data Preprocessing**

Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format.

**Steps Involved in Data Pre-processing:**

**1. Data Cleaning:**   
The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc. 

* **(a). Missing Data:**   
  This situation arises when some data is missing in the data. It can be handled in various ways.   
  Some of them are:
  1. **Ignore the tuples:**   
     This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
  2. **Fill the Missing values:**   
     There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.
* **(b). Noisy Data:**   
  Noisy data is a meaningless data that can’t be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
  1. **Binning Method:**   
     This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  2. **Regression:**   
     Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
  3. **Clustering:**   
     This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

**2. Data Transformation:**

This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**   
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
2. **Attribute Selection:**   
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
3. **Discretization:**   
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
4. **Concept Hierarchy Generation:**   
   Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

**3. Data Reduction:**

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

1. **Data Cube Aggregation:**   
   Aggregation operation is applied to data for the construction of the data cube.
2. **Attribute Subset Selection:**   
   The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute. The attribute having p-value greater than significance level can be discarded.
3. **Numerosity Reduction:**   
   This enable to store the model of data instead of whole data, for example: Regression Models.
4. **Dimensionality Reduction:**   
   This reduce the size of data by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are: Wavelet transforms and PCA (Principal Component Analysis).

**Text Preprocessing**

Tweet texts often consists of other user mentions, hyperlink texts, emoticons and punctuations. In order to use them for learning using a Language Model. We cannot permit those texts for training a model. So we have to clean the text data using various preprocessing and cleansing methods.

### **Stemming/ Lematization**

For grammatical reasons, documents are going to use different forms of a word, such as write, writing and writes. Additionally, there are families of derivationally related words with similar meanings. The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

Stemming usually refers to a process that chops off the ends of words in the hope of achieving goal correctly most of the time and often includes the removal of derivational affixes.

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base and dictionary form of a word.

### **Hyperlinks and Mentions**

Twitter is a social media platform where people can tag and mentions other people's ID and share videos and blogs from internet. So the tweets often contain lots of Hyperlinks and twitter mentions.

* Twitter User Mentions - Eg. @arunrk7, @andrewng
* Hyperlinks - Eg. [https://keras.io](https://keras.io/), [https://tensorflow.org](https://tensorflow.org/)

### **Stopwords**

Stopwords are commonly used words in English which have no contextual meaning in an sentence. So therefore we remove them before classification. Some stopwords are...

****

**Train and Test Split**

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem. Although simple to use and interpret, there are times when the procedure should not be used, such as when you have a small dataset and situations where additional configuration is required, such as when it is used for classification and the dataset is not balanced.

### **Tokenization**

Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens , perhaps at the same time throwing away certain characters, such as punctuation. The process is called **Tokenization.**

**Label Encoding**

We are building the model to predict class in encoded form (0 or 1 as this is a binary classification). We should encode our training labels to encodings.

# Word Embedding

In Language Model, words are represented in a way to intend more meaning and for learning the patterns and contextual meaning behind it.

**Word Embedding** is one of the popular representation of document vocabulary. It is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc.

Basically, it's a feature vector representation of words which are used for other natural language processing applications.

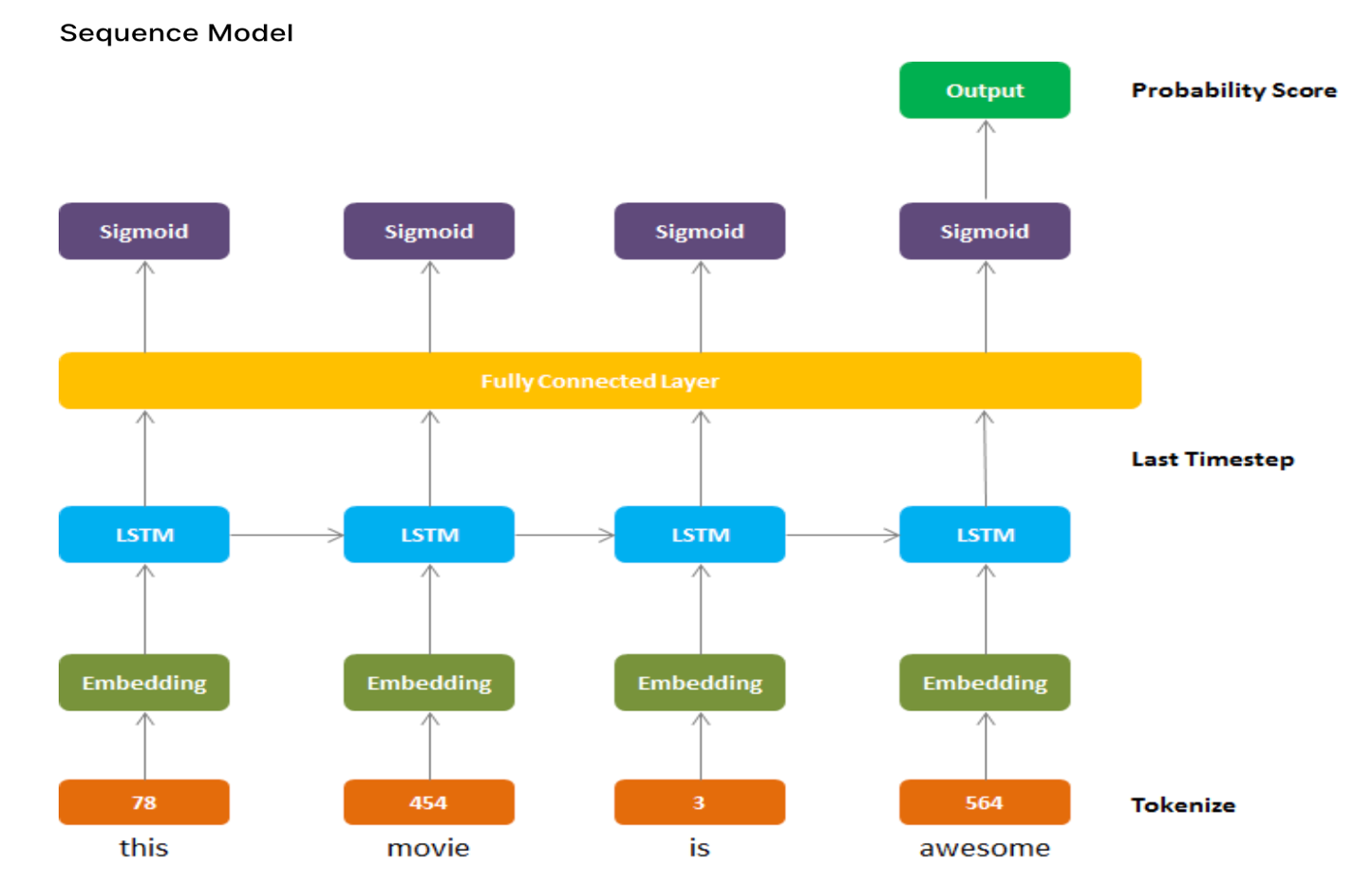
We could train the embedding ourselves but that would take a while to train and it wouldn't be effective. So going in the path of Computer Vision, here we use **Transfer Learning**. We download the pre-trained embedding and use it in our model.

The pretrained Word Embedding like **GloVe & Word2Vec** gives more insights for a word which can be used for classification.

**Model Training – LSTM**

Model training is the phase in the data science development lifecycle where practitioners try to fit the best combination of weights and bias to a machine learning algorithm to minimize a loss function over the prediction range. The purpose of model training is to build the best mathematical representation of the relationship between data features and a target label (in supervised learning) or among the features themselves (unsupervised learning). Loss functions are a critical aspect of model training since they define how to optimize the machine learning algorithms. Depending on the objective, type of data and algorithm, data science practitioner use different type of loss functions. One of the popular examples of loss functions is Mean Square Error (MSE).

In this analysis, some words are predominantly feature in both Positive and Negative tweets. This could be a problem if we are using a Machine Learning model like Naive Bayes, SVD, etc.. That's why we use **Sequence Models**.

****

Recurrent Neural Networks can handle a sequence of data and learn a pattern of input sequence to give either sequence or scalar value as output. In our case, the Neural Network outputs a scalar value prediction.

For model architecture, we use

1) **Embedding Layer** - Generates Embedding Vector for each input sequence.

2) **Conv1D Layer** - Its using to convolve data into smaller feature vectors.

3) **LSTM** - Long Short Term Memory, it’s a variant of RNN which has memory state cell to learn the context of words which are at further along the text to carry contextual meaning rather than just neighbouring words as in case of RNN.

4) **Dense** - Fully Connected Layers for classification

**Optimization Algorithm**

Our project uses Adam, optimization algorithm for Gradient Descent.

**Callbacks**

Callbacks are special functions which are called at the end of an epoch. We can use any functions to perform specific operation after each epoch. I used two callbacks here,

* **LRScheduler** - It changes a Learning Rate at specific epoch to achieve more improved result. In this notebook, the learning rate exponentially decreases after remaining same for first 10 Epoch.
* **ModelCheckPoint** - It saves best model while training based on some metrics. Here, it saves the model with minimum Validity Loss.

**Model Evaluation**

Model evaluation is the process of using different evaluation metrics to understand a machine learning model’s performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring.

**Classification**

The most popular metrics for measuring classification performance include accuracy, precision, confusion matrix, log-loss, and AUC (area under the ROC curve).

* **Accuracy** measures how often the classifier makes the correct predictions, as it is the ratio between the number of correct predictions and the total number of predictions.
* **Precision** measures the proportion of predicted Positives that are truly Positive. Precision is a good choice of evaluation metrics when you want to be very sure of your prediction. For example, if you are building a system to predict whether to decrease the credit limit on a particular account, you want to be very sure about the prediction or it may result in customer dissatisfaction.
* The **confusion matrix** (or confusion table) shows a more detailed breakdown of correct and incorrect classifications for each class. Using a confusion matrix is useful when you want to understand the distinction between classes, particularly when the cost of misclassification might differ for the two classes, or you have a lot more test data on one class than the other. For example, the consequences of making a false positive or false negative in a cancer diagnosis are very different.

# Confusion Matrix

# 

# Classification Scores

# Table Description automatically generated

**API USED**

1. **Tweepy  
   ­­­­**It is used to establish a connection with twitter website as it authenticates all the api credentials, and then it fetches ‘n’ number of tweets from a particular user (eg. ‘Elon Musk’)
2. **Textblob**

It is used to to determine the polarity and subjectivity of a given tweet.

It used functions: .polarity() and .subjectivity()

Each word in the lexicon has scores for:

**polarity**: negative vs. positive   
 (-1.0 => +1.0)   
The state of having two opposite or contradictory tendencies, opinions or aspects.  
  
**subjectivity**: objective vs. subjective

(+0.0 => +1.0)

Subjective sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information

The dataset of tweepy is:

https://github.com/sloria/TextBlob/blob/eb08c120d364e908646731d60b4e4c6c1712ff63/textblob/en/en-sentiment.xml

Negation multiplies the polarity by -0.5, and doesn't affect subjectivity.

Eg. In sentence: “Today is not a great day”, the polarity is multiplies by -0.5 because of ‘not’.

**Conclusion**

The growth of sentiment analysis as one of the most active research areas of the last 10 years is due to different reasons. First, sentiment analysis has a wide array of applications, in almost every domain. Second, it offers many challenging research problems that have never been studied before. Third, with the advent of the big data technologies, we now have a huge volume of opinionated data recorded and easily accessible in digital forms on the web. These reasons have motivated the recent advances in the state of the art presented in this chapter. Most of the work regarding polarity classification usually considers text as unique information to infer sentiment, disregarding that social networks are actually networked environments. To take advantage of both natural language and social networks relationships, a novel research branch is developing.

Using an API like textBlob, our accuracy is: 62.26%

While using machine learning model, our accuracy increases to: 78%