Sentiment Analysis on Twitter Data (Final Report)

Durva Chobe

(Comp)

D.Y Patil College of Engineering, Akurdi

Pragati Narote

(Elec)

College of Engineering, Pune

1. Abstract

Social media helps engaging business with consumers. It helps identify and extract the social sentiment of their brand, product or service while monitoring the online conversations. Twitter is one of the most used social media platforms worldwide with almost 330 million active users. Twitter offers organizations a fast and effective way to analyse customers' perspectives toward the critical success in the market place. According to a source 67% of B2B businesses are using twitter as a marketing tool. Twitter sentiment analysis therefore refers to using advanced techniques to analyse the sentiment of the tweet as positive, negative or neutral. A sentiment analysis system for text analysis combines natural language processing (NLP) and machine learning techniques to assign weighted sentiment scores to the entities, topics, themes and categories within a sentence or phrase.

2. Introduction

In the past decade forms of communication have evolved. A large portion of the human population are using social media to express their views. *People usually depend on user generated content on any product to a great extent when it comes to perform any desired action. When people want to buy a product through online, they will first look up its reviews in that particular product website through online, before making up a decision. Some analysis is to be done on all these reviews so that the final outcome says whether the product is good to buy or not*. While there is no limit to the range of information conveyed by tweets and texts, often these short messages are used to share opinions and sentiments that people have about what is going on in the world around them. As humans we are able to classify a text as positive, negative or neutral subconsciously, but this opinion is relative and will not always be the same for different people. Twitter is an online networking site driven by tweets which are 140-character limited messages. Tweets are short texts: a sentence or a headline rather than a document. The language used is very informal, with creative spelling and punctuation, misspellings, slang, new words, URLs, and genre-specific terminology and abbreviations, such as, RT for "retweet" and "#" hashtags, which are a type of tagging for Twitter messages. How to handle such challenges so as to automatically understand the opinions and sentiments of people is the main challenge here. The project would heavily rely on techniques of "Natural Language Processing" in extracting significant patterns and features from the large data set of tweets and on "Machine Learning" techniques for accurately classifying individual unlabelled data samples(tweets) according to whichever pattern model best describes them.

3. Problem Description

Classify whether the text messages(tweets) are of positive, negative or neutral sentiment.

For messages conveying combination of positive, neutral or negative sentiment partially, whichever is the stronger sentiment should be chosen.

4. Motivation

Motivation for sentiment analysis is two-fold. Both consumers and producers highly value “customer’s opinion” about products and services. The active user count of Twitter increased from approx. 30 million in 2010 to approx. 321 million in mid-2020; also, large number of companies all over the world have chosen Twitter as a marketing platform, since the last decade. This narrows the communication gap between the consumers and producers. Rightly using sentiment analysis would be a competitive advantage for a company. Thus, Sentiment Analysis has seen a considerable effort from industry as well as academia.

Consumer’s Perspective: For a consumer it is very important to know the reviews or opinion about certain product from the people who have already used the product. Earlier, the reviews from only friend and family were considered. But, as the Internet and Social Networking Platforms are experiencing hick in usage, we see a large number of people expressing their opinion in blogs, forums and even tweets. Manually reading every review is not optimistic. Hence, the need of classifying the opinions as good/bad or positive/neutral/negative. Further, labelling the opinions would give a sentimental summary to the readers.

The Producer’s Perspective: With the increase in use of blogs, forums and twitter, the consumer voices have made a way to reach a larger crowd. According to Pang and Lee (2008), these consumer voices can wield enormous influence in shaping the opinions of other consumers and, ultimately, their brand loyalties, their purchase decisions, and their own brand advocacy. The consumers have started using review sites, blogs and other social networking platforms to discuss, admire or criticize various features of different products. These opinions thus shape the future of the product or the service. The producers need a system that can identify trends in customer reviews and use them to improve their product or service and also identify the requirements of the future.

The Societies’ Perspective: Recently, certain events, which affected Government, have been triggered using the Internet. The social networks are being used to bring together people so as to organize mass gatherings and oppose oppression. On the darker side, the social networks are being used to insinuate people against an ethnic group or class of people, which has resulted in a serious loss of life. Thus, there is a need for Sentiment Analysis systems that can identify such phenomena and curtail them if needed

5. Sentiment Analysis

Sentimental analysis is the process of computationally determining the opinion or attitude of the writers as positive, negative or neutral. Data mining is another name for sentimental analysis. In many fields like business, politics and public actions, determining the sentimental analysis is very important. Considering business, it is very useful to understand the customer's feelings in order to develop their company. Next in politics: It can be used to predict the election results.

There are two ways of classifications and they are:
(1) Machine Learning (2) Lexicon-based Approach.

In this paper machine learning classifiers are implemented in sentimental analysis and is done in twitter because most of the product consumers like politicians, famous personalities and even general people regularly update their moods as well as opinions on currents affairs, technical advancement and recently launched products in the form of tweets.

4. Related Work

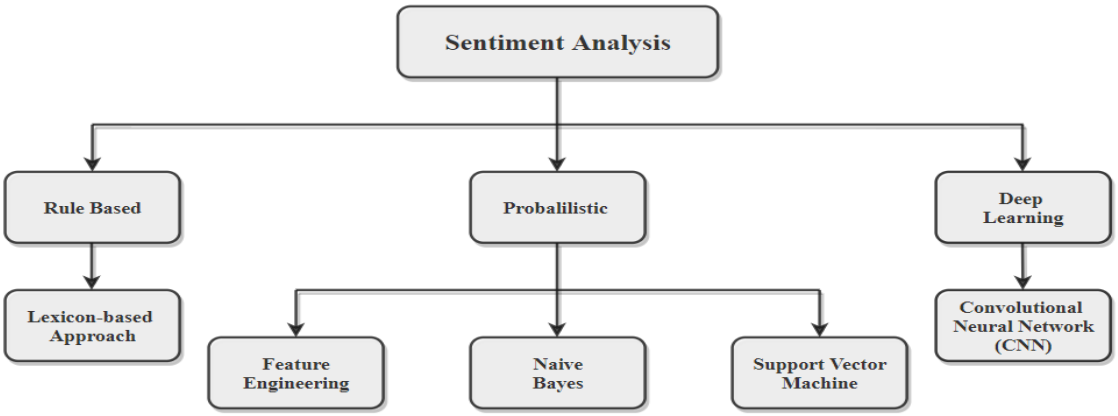
The best results reached in sentiment classification use supervised learning techniques such as Naive Bayes and Support Vector Machines, but the manual labelling required for the supervised approach is very expensive.

Naïve Bayes Classifier is a 'bag-of-words' approach for subjective analysis of content.

The 'bag-of-words' (BOW) model is one of the most widely used feature models for almost all text classification tasks due to its simplicity coupled with good performance. In this model, a text (such as a sentence or a document) is represented as bag of its words, disregarding grammar and even word order but keeping multiplicity. The bag-of-words model has also been used for computer vision. The bag-of-words model is commonly used in methods of document classification where the frequency of each word is used as a feature for training a classifier.

Bidirectional Encoder Representations from Transformers (BERT) is a technique for NLP (Natural Language Processing) pre-training developed by Google. BERT was created and published in 2018 by Jacob Devlin and his colleagues from Google. Google is leveraging BERT to better understand user searches. The original English-language BERT model used two corpora in pre-training: Book Corpus and English Wikipedia.

5. General Approaches in NLP

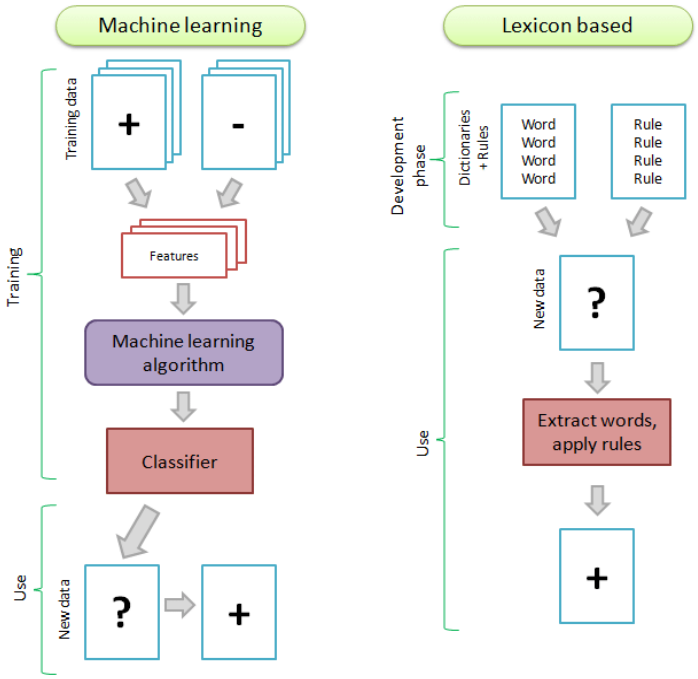


Different approaches for Sentiment Analysis

Lexicons and Machine Learning:

Lexicon-based techniques were the first to be used for sentiment analysis. They are divided into two approaches: dictionary-based and corpus-based [8]. In the former type, sentiment classification is performed by using a dictionary of terms, such as those found in SentiWordNet and WordNet. Nevertheless, corpus-based sentiment analysis does not rely on a predefined dictionary but on statistical analysis of the contents of a collection of documents, using techniques based on k-nearest neighbours (k-NN) [9], conditional random field (CRF) [10], and hidden Markov models (HMM) [11], among others

Neural Networks for information retrieval or document searches use lexicons as their sample data. Each lexeme can be assigned an associated vector, where the substance of the lexeme is defined as coordinates, and its frequency within a database defines its magnitude (length). Using techniques like cosine similarity, machine learning algorithms can quickly distinguish and compare documents to each other based upon their lexical similarities and overlap of subject matter. *These techniques employ dictionaries of words annotated with their semantic polarity and sentiment strength. This is then used to calculate a score for the polarity and/or sentiment of the document. Usually this method gives high precision but low recall.*



Naive Bayes:

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is referred as Naive Bayes because:

- Naive: It is called Naive because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Hence each feature individually contributes to identify without depending on each other.
- Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

Bayes' Theorem: Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability. The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Here;

A,B : Events

P(*A*|*B*) : Probability of A given B is true.

P(*B*|*A*) : Probability of B given A is true.

P(*A*), P(*B*): Individual Probabilities.

The Naive Bayes classifiers provide a simple, yet effective way to train a neural network to classify and identify data. Sometimes represented simply as a Bayesian network. Naive Bayes is essentially a technique for assigning classifiers to a finite set. It can be used for Binary as well as Multi-class Classifications. It performs well in Multi-class predictions as compared to the other Algorithms.

Pros and Cons of Naïve Bayes classifiers

Pros:

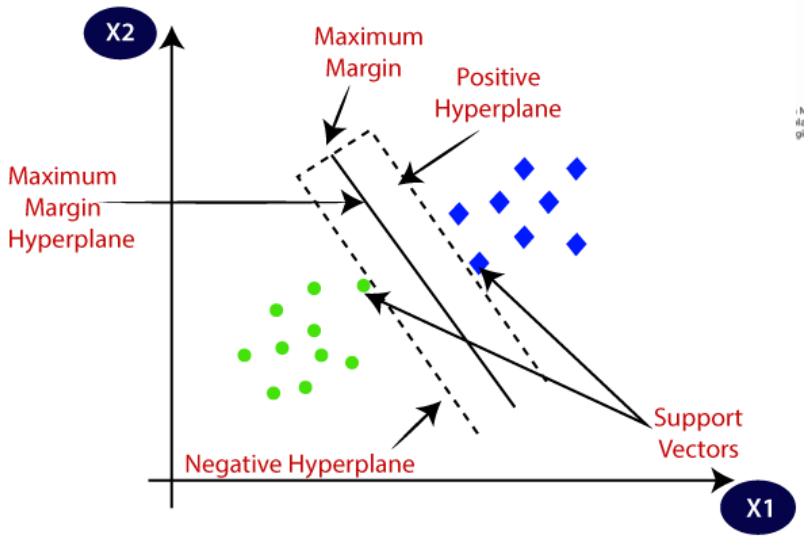
- It is easy and fast to predict class of test data set. It also performs well in multi class prediction
- When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
- It performs well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

Cons:

- *If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as "Zero Frequency".* To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
- On the other side naive Bayes is also known as a bad estimator, so the probability outputs from predict_proba are not to be taken too seriously.
- Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

Support Vector Machines:

A non-probabilistic model which uses a representation of text examples as points in a multidimensional space. Examples of different categories (sentiments) are mapped to distinct regions within that space. Then, new texts are assigned a category based on similarities with existing texts and the regions they're mapped to. SVM is a supervised machine learning algorithm that can be used for both classification or regression challenges. Classification is predicting a label/group and Regression is predicting a continuous value. SVM performs classification by finding the hyper-plane that differentiate the classes we plotted in n-dimensional space. Support vector machine is highly preferred by many as it produces significant accuracy with less computation power. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data point to separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e. the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line).



Types of SVM

SVM can be of two types:

- ***Linear SVM:*** *Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.*
- ***Non-linear SVM:*** *Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.*

Hyperplane and Support Vectors in the SVM algorithm*:*

- ***Hyperplane:***

  There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.

We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

- ***Support Vectors:***

  The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

Every classification algorithm has its own advantages and disadvantages that are come into play according to the dataset being analysed.

Some of the advantages of SVMs are as follows:

- The very nature of the Convex Optimization method ensures guaranteed optimality. The solution is guaranteed to be a global minimum and not a local minimum.
- SVM is an algorithm which is suitable for both linearly and nonlinearly separable data (using kernel trick). The only thing to do is to come up with the regularization term, C.
- SVMs work well on small as well as high dimensional data spaces. It works effectively for high-dimensional datasets because of the fact that the complexity of the training dataset in SVM is generally characterized by the number of support vectors rather than the dimensionality. Even if all other training examples are removed and the training is repeated, we will get the same optimal separating hyperplane.
- SVMs can work effectively on smaller training datasets as they don't rely on the entire data.

Disadvantages of SVMs are as follows:

- They are not suitable for larger datasets because the training time with SVMs can be high and much more computationally intensive.
- They are less effective on noisier datasets that have overlapping classes.

Deep Learning:

It is a diverse set of algorithms that attempt to mimic the human brain, by employing artificial neural networks to process data. One algorithm from deep learning can be used for sentiment analysis is Recurrent Neural Network (RNN)

Recurrent neural network (RNN)

Recurrent neural networks are a class of neural networks whose connections between neurons form a directed cycle, which creates feedback loops within the RNN. The main function of RNN is the processing of sequential information on the basis of the internal memory captured by the directed cycles. Unlike traditional neural networks, RNN can remember the previous computation of information and can reuse it by applying it to the next element in the sequence of inputs. A special type of RNN is long short-term memory (LSTM), which is capable of using long memory as the input of activation functions in the hidden layer. This was introduced by Hochreiter and Schmidhuber (1997). Figure 4 illustrates an example of the LSTM architecture. The input data is preprocessed to reshape data for the embedding matrix (the process is similar to the one described for the CNN). The next layer is the LSTM, which includes 200 cells. The final layer is a

fully connected layer, which includes 128 cells for text classification. The last layer uses the sigmoid activation function to reduce the vector of height 128 to an output vector of one, given that there are two classes to be predicted.

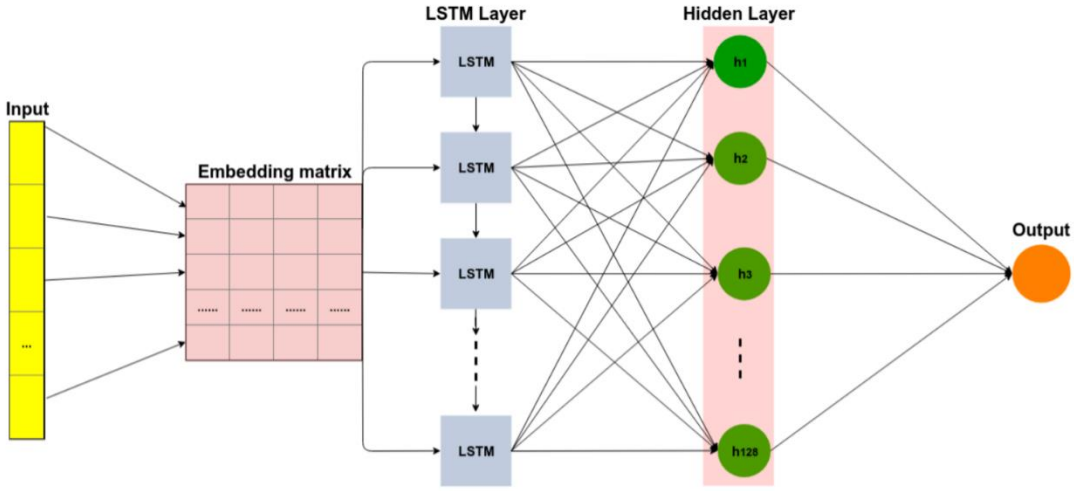
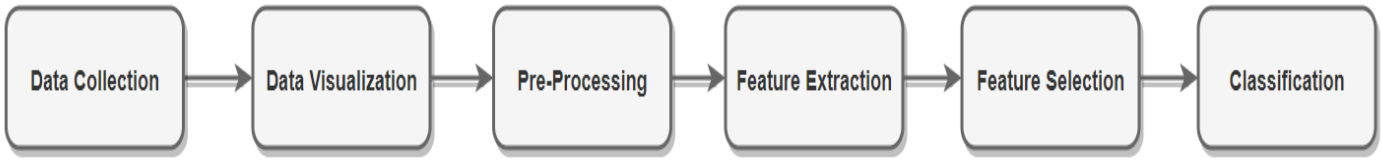


Figure 4. A long short-term memory network. LSTM, long short-term memory.

6. Proposed Methodology

Methodology of sentimental analysis in twitter mainly involves 6 steps.



1. Data Collection:

Data collection is the process of gathering the data. There are various ways to collect the data. In our case, the dataset is already available on Kaggle.

Other ways to gather data are:

Twitter API — A Python wrapper for performing API requests. For fetching the twitter data from the twitter API includes the following steps 1] Installation of the needed software 2] authentication of twitters data. The main installation software's include tweepy, text blob, nltk etc, Authentication involves different steps

step1: visit the twitter website and click the button 'create new app'. Step2:fill the details in the form provided and submit. Step3:It will be redirected to the app page where the "'consumer keys', 'consumer access', 'access token' and 'access token secret' "that is needed to access the twitter data will be present. Step4:implement in python.

There are different sources for storing the data taken from the twitter. They are like MongoDB , open source document storage database and is the go-to "No SQL" database. It makes working with a database feel like working with JavaScript.

PyMongo, a Python wrapper for interfacing with a MongoDB instance. This library lets you connect your Python scripts with your database and read/insert records.

2. Data Visualization:

Data visualization is a very important step in sentiment analysis. Since, the data in this case is unstructured and has a lot of variations, understanding and visualizing the dataset becomes a crucial step.

```
# We will now take a look at random tweets
# to gain more insights
rand_indexs = np.random.randint(1,len(train_data),50).tolist()
train_data["tweet_text"][rand_indexs]
```

```
18507                                    @WWERetweeting Ric Flair on the 20th anniversary of #Raw
12774    The Google+ for Business Workshop 18 August is SOLD OUT! Don`t worry you can book into the 10 Se...
5048     Kainis! Bukas na pala first day ng Secret Love (Sungkyunkwan Scandal) sa ABS-CBN. Eh may pasok k...
8329     Maniaci on Casimir Pulaski Day (March 1\u002c 2004): \""Some people call it a holiday.\"" Judge:...
7770     Today was an awesome day but it is quite sad for me and my teammates :) . Qian \u002c Yue \u002c...
14185    But what's that off in the distant? Is it a bird? Is it a plane? No. It's Joe Biden! https://t.c...
4629     Mosconi Cup from Monday 10th to Thursday 13th December at the York Hall\u002c Bethnal Green\u002...
18596    @RBPundit I really like Rick Perry,but Rubio is my strong 2nd choice,where can I read about his ...
10069    Happy 20th Birthday Amazon! Amazon Prime day is bigger than black friday! Support QCHF, save mon...
9545     Show em what it isRT @Trujohnson2: Love it when scouts be at our Practice..  I counted 4 today.....
12838         Listening to some Grateful Dead on the YouTubes. It's giving my Friday afternoon a great vibe.
16353    @jah_alpha and todd!! Need to see it! Might watch it tomorrow night! Firework central round here...
1971     Just watched the What Makes You Beautiful video and I may have teared up. I shouldn\u2019t be al...
12685    @tuna_lucy Sorry am rubbish at posting links. Google 'Spike's disease' +border terrier and it wi...
9803     It's Friday night, at home sipping on some Crown Royal Whisky, listening to AC/DC, talking and t...
9999     Did I miss something?  Just placed an order from Amazon Prime with free 2 day delivery and got a...
7213     RT @Nick136 Astonishing! Police raid Milan offices of Standard & Poor's http://t.co/QKXH4vb || D...
12819    Getting the most from Google+ for your Business? Join our next Workshop in Derby 10 September ht...
```

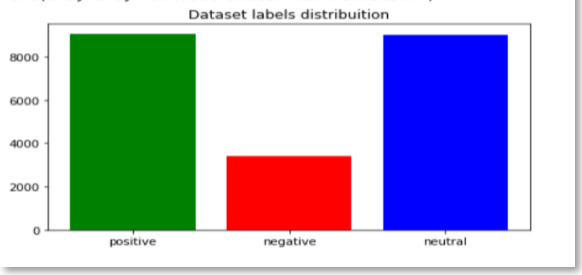
Take a look at some random tweets to gain more insights of the dataset.

Using a plot from matplotlib.pyplot visualization the distribution of tweet's label as positive, negative or neutral.

From the graph it the clear that the count of positive and neutral tweets is nearly equal, but the count of negative tweets is too less in comparison.

```
# Matplot
import matplotlib.pyplot as plt
%matplotlib inline
from collections import Counter
sentiment_cnt = Counter(train_data.sentiment)
plt.figure(figsize=(7,4))
plt.bar(sentiment_cnt.keys(),
        sentiment_cnt.values(),
        color=('g','r','b'))
plt.title("Dataset labels distribuition")
```

```
Text(0.5, 1.0, 'Dataset labels distribuition')
```



3. Pre-Processing:

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Twitter Data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. Data pre-processing is a proven method of resolving such issues.

After retrieval of tweets, Sentiment analysis tool is applied on raw tweets but in most of cases results to very poor performance. Therefore, pre-processing techniques are necessary for obtaining better results. We extract tweets i.e. short messages from twitter which are used as raw data. This raw data needs to be pre-processed. So, pre-processing involves following steps which constructs n-grams:

- Filtering

```
import re as regex
def clean_tweets(tweet):
    tweet = str(tweet)
    # remove URL
    tweet = regex.sub(r"http\s+", '', tweet)
    # Remove usernames
    tweet = regex.sub(r"@[^\s]+[\s]?", '', tweet)
    # remove special characters
    tweet = regex.sub('[^ a-zA-Z0-9]','', tweet)
    # remove Numbers
    tweet = regex.sub('[0-9]','', tweet)
    return tweet
#Apply function to Tweet column
train_data['tweet_text'] = train_data['tweet_text'].map(clean_tweets)
```

Filtering is nothing but cleaning of raw data. In this step, URL links (E.g. http://twitter.com), special words in twitter (e.g. "RT" which means Retweet), usernames in twitter (e.g. @Ron - @ symbol indicating a username), numbers, special characters are removed.

- Tokenization

Tokenization is nothing but Segmentation of sentences. In this step, we will tokenize or segment text with the help of splitting text by spaces and punctuation marks to form container of words. Tokenization generally done by installing the NLP package, nltk. After filtering and tokenizing, it was found that there were 55700 different words in the dataset.

```
import nltk
from nltk.tokenize import word_tokenize
nltk.download('punkt')
def most_used_words(text):
    tokens = word_tokenize(text)
    frequency_dist = nltk.FreqDist(tokens)
    print("There is %d different words" % len(set(tokens)))
    return sorted(frequency_dist,key=frequency_dist.__getitem__, reverse=True)
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

- Removal of stopwords

Words such as articles and some verbs are usually considered stop words because they don't help us to find the context or the true meaning of a sentence. These are words that can be removed without any negative consequences to the final model that you are training. There's no universal stop words list because a word can be empty of meaning depending on the corpus you are using or on the problem you are analysing. This means that any word can be a stop word depending on what you are trying to do. Stopwords are generally removed by installing the NLP package. Articles such as "a", "an", "the", etc and other stopwords such as "to", "of", "is", "are", "this", "for", etc are removed. After the removal of stopwords the number of unique words in the dataset reduced to 55562 from 55700.

```
from nltk.corpus import stopwords
nltk.download("stopwords")
mw = most_used_words(train_data.tweet_text.str.cat())
most_words = []
for w in mw:
  if w in stopwords.words("english"):
    continue
  else:
    most_words.append(w.lower())
print("There are ",len(most_words),"after removing stopwords")
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
There is 55700 different words
There are  55562 after removing stopwords
```

- Replacing emoticons

Emoticons are considered to be handy and reliable indicators of sentiment, and hence could be used either to automatically generate a training corpus or to act as evidence feature to enhance sentiment classification. Emoticons are introduced as expressive, non-verbal components into the written language, mirroring the role played by facial expressions in speech. Their role is mainly pragmatic: emoticons give a positive or negative sense to written sentences by a visual expression. According to this consideration, there is a relationship between the sentiment orientation of emoticons and messages. Emoticons have been distinguished in two main categories, i.e. positive and negative. Instances of positive emoticons are :-), :), =), :D, while examples of negative ones are :-(, :(, =(, ;(. These emoticons surely are an important source of information for polarity classification. In fact, on social media positive and negative messages have a high percentage of emoticons.

```
HAPPY_EMO = r" ([xX;:]-?[dD)]|:-?[\)]|[;:][pP]) "
SAD_EMO = r" (:'?[/|\(]) "
print("Happy emoticons:", set(re.findall(HAPPY_EMO, tweets_text)))
print("Sad emoticons:", set(re.findall(SAD_EMO, tweets_text)))
```

```
Happy emoticons: {'XD', 'x)', 'xD', ';P', ':P', ';D', ';)', ':p', ';-)', ':)', ':-)', ':D'}
Sad emoticons: {':/', ":'(", ':('}
```

4. Feature Extraction:

Selection of useful words from the tweet is called as feature extraction. In the feature extraction method, we extract the aspects from the pre-processed twitter dataset. Feature extraction was done using stemming and lemmatization functions.

- Stemming
  Stemming is a method for collapsing distinct word forms. This could help reduce the vocabulary size, thereby sharpening one's results, especially for small data sets. The Porter_stemmer_is one of the earliest and best-known stemming algorithms. It works by heuristically identifying word suffixes (endings) and stripping them off, with some regularization of the endings. The Porter stemmer often collapses sentiment distinctions, by mapping two words with different sentiment into the same stemmed form.

```
from nltk.stem.snowball import SnowballStemmer
from nltk.stem import WordNetLemmatizer
nltk.download('wordnet')
def stem_tokenize(text):
    stemmer = SnowballStemmer("english")
    stemmer = WordNetLemmatizer()
    return [stemmer.lemmatize(token) for token in word_tokenize(text)]

def lemmatize_tokenize(text):
    lemmatizer = WordNetLemmatizer()
    return [lemmatizer.lemmatize(token) for token in word_tokenize(text)]
```

```
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Unzipping corpora/wordnet.zip.
```

- Lemmatisation
  Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form.
  In computational linguistics, lemmatisation is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatisation depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighbouring sentences or even an entire document.

5. Feature Selection:

Correct feature selection techniques are used in sentiment analysis that has got a significant role for identifying relevant attributes and increasing classification (machine learning) accuracy.

6. Classification:

We used Recurrent Neural Networks, Naïve Bayes Classifier and Support Vector Machines to classify the tweets as positive, negative or neutral.

Classified Tweets: We labelled the tweets in three classes according to sentiments expressed/observed in the tweets: positive, negative and neutral. We gave the following guidelines to our labellers to help them in the labelling process:

- Positive: If the entire tweet has a positive/happy/excited/joyful attitude or if something is mentioned with positive connotations. Also, if more than one sentiment is expressed in the tweet but the positive sentiment is more dominant. Example: “4 more years of being in shithole Australia then I move to the USA! :D”.
- Negative: If the entire tweet has a negative/sad/displeased attitude or if something is mentioned with negative connotations. Also, if more than one sentiment is expressed in the tweet but the negative sentiment is more dominant. Example: “I want an android now this iPhone is boring :S”.
- Neutral: If the creator of tweet expresses no personal sentiment/opinion in the tweet and merely transmits information. Advertisements of different products would be labelled under this category. Example: “US House Speaker vows to stop Obama contraceptive rule.

RNN lead to an accuracy of 42%.

Naïve Bayes classifiers like Bernoulli Naïve Bayes and Multinomial Naïve Bayes from sklearn library were also used on the dataset. Bernoulli Naïve Bayes gave an accuracy of 61% and Bernoulli Naïve Bayes gave an accuracy of 56%.

Later, Support Vector Machines was used to classify the tweets. Linear kernel and 0.98 regularization term gave the highest accuracy – 67.230% for the test samples when removal of stopwords and lemmatization were not done while pre-processing the text data, and the entire training dataset was used for training the model.

Moreover, as it is observed during data visualization that there are only 15% negative tweets in the training dataset, i.e. the dataset is moderately skewed. SMOTE undersampling and oversampling was performed before training the model. But, both didn't contribute positively towards the accuracy of the model. So, the skewness won't affect the predictions in a positive sense.

Since, SVM gave the highest accuracy, it is the final classifier used in the code.

7. Conclusion

The task of sentiment analysis, especially in the domain of micro-blogging, is still in the developing stage and far from complete. In our project we have experimented sentiment analyser model using a Naïve Bayes classifier, Recurrent Neural Network and Support Vector Machine (SVM). We have presented the classification of tweets based on its polarity as positive, negative or neutral. We found that SVM model gave a better accuracy then RNN and Naïve Bayes classifiers. So, the best model is SVM classifier. The performance of classifier highly depends on the training data. Thus, we can say that SVM classifier is the most suitable one for our dataset. Moreover, we also found that pre-processing and feature extraction also affects the accuracy of the model, and depends on the classifier used. In case of SVM classifier, removal of stopwords and lemmatization affects the accuracy of the model in a negative way. Retaining stopwords and skipping lemmatization increased the accuracy by 0.3%. In case of SVM, the parameters like kernel and regularization term highly affect the accuracy of the model. For our dataset, linear kernel and 0.98 regularization parameter suits the best. Kernels like poly, sigmoid, rbg gave accuracy approximately 20% less than that by linear kernel. The size of dataset used for training also affect the accuracy. The larger the training dataset the better the accuracy of the model.

Note: Model of this project is embedding in a real-world application, where you can enter the username of any twitter handle and the tweets on that twitter handle would be labelled as positive, negative or neutral and then displayed.

Use the URL given below to experience the real-world application:

18.225.11.191:8000

REFERENCES

1. S. Mukherjee and P. Bhattacharyya. Sentiment analysis in Twitter with lightweight discourse analysis, December 2012.
2. P. Nakov. Developing a Successful SemEval Task in Sentiment Analysis of Twitter and Other Social Media Texts, January 2018
3. Nurulhuda Zainuddin and Ali Selamat.Sentiment Analysis Using Support Vector Machine, September 2014.
4. Kavya Suppala, Narasinga Rao. Sentiment Analysis Using Naïve Bayes Classifier, June 2019
5. Kai Sheng Tai. Sentiment Analysis of Tweets: Baselines and Neural Network Models, December 2013
6. Muhammad Zubair Asghar , Aurangzeb Khan , Shakeel Ahmad, Fazal Masud Kundi. A Review of Feature Extraction in Sentiment Analysis, February 2014.
7. Salas-Zárate, M.P.; Medina-Moreira, J.; Lagos-Ortiz, K.; Luna-Aveiga, H.; Rodriguez-Garcia, M.A.; Valencia-García, R.J.C. Sentiment analysis on tweets about diabetes: An aspect-level approach. Comput. Math. Methods Med. 2017, 2017. [CrossRef] [PubMed]
8. Huq, M.R.; Ali, A.; Rahman, A. Sentiment analysis on Twitter data using KNN and SVM. IJACSA Int. J. Adv. Comput. Sci. Appl. 2017, 8, 19–25.
9. Pinto, D.; McCallum, A.; Wei, X.; Croft, W.B. Table extraction using conditional random fields. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, Toronto, ON, Canada, 28 July–1 August 2003; pp. 235–242. Electronics 2020, 9, 483 28 of 29
10. Soni, S.; Sharaff, A. Sentiment analysis of customer reviews based on hidden markov model. In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015), Unnao, India, 6 March 2015; pp. 1–5.