International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

SENTIMENT ANALYSIS ON NEWS ARTICLES

J. Reshma[1] , G. Sai Sruthi[2], T. Yeshaswi[3]
D. Radhika[4]
Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Telangana, India

ABSTRACT

Sentiment analysis or opinion mining is the computational study of people's opinion, sentiments, attitudes and emotions expressed in written language. It is one of the most active research areas in natural language processing and text mining in recent years.

It is a way to analyze the subjective information in the text and then mine the opinion. Sentiment analysis is the procedure by which information is extracted from the opinion appraisal and emotions of people in regards to entities, events and their attributes. In decision making, the opinion of others have a significant effect on customer ease, making choices with regards have a significant effect, product entity. The approach of text sentiment analysis typically works at a particular level like phrase, sentence or document level. This project aims at analyzing a solution for the sentiment identification at a fine-grained level,namely the sentence level in which polarity of the sentence can be given by three categories as positive, negative and neutral. The data set is gathered from inshorts.com and the project is restricted to 3 news article domains namely- Sports, World and Politics. Lexicon based approach is used for Sentiment Analysis. VADER gives the polarity of negativity, neutrality, positivity and also the consolidated compound score for the given text. The Data Labeling, Data Processing and Finalization is done using Compound score from VADER.

Key Words :

Machine Learning, Lexicons, Web Scraping, Data Cleaning, Tokenization, Data Labeling,

Data Finalization.

1. Introduction

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis, and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service. The objective of Sentiment Analysis is evaluating the sentiments and opinions of a writer respectively, one topic domain or multi-topic domain. It calculates the aggregate sentiment polarity of a text or online reviews for one topic based on sentiment classification levels, such as positive or negative. Existing analysis approaches to sentiment reviews can be grouped into four main categories: word level, sentence level, document level, and aspect/ entity level.

Sentiment analysis refers to the use of natural language processing text analysis and computational linguistics to identify and extract subjective information in source materials. Generally speaking sentiment analysis aims to determine the attribute of the speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attribute may be his or her judgment or evaluation affective state, or the intended emotional communication. Sentiment analysis is the process of detecting a piece of writing for positive, negative or neutral feeling bound to it. Human have the innate ability to determine sentiment; however, this process is time consuming,inconsistent, and costly in a business context.It's just not realistic to have people individually read tens of thousand of user customer reviews and score them for sentiment

For Example if we consider Semantria's cloud based sentiment analysis software. Semantria's cloud-based sentiment analysis software extracts the sentiment of a document and its components through the following steps:

- A document is broken in its basic parts of speech called POS tags which identify the structural elements of a document, paragraph, or sentence(i.e; Nouns, adjectives, verbs, and adverbs)
- Sentiment bearing phrases, such as "terrible services",are identified through the use of specially designed algorithms.
- Each sentiment-bearing phrase in a document is given a score based on a logarithmic scale that ranges between 1-10.
- Finally, the scores are combined to determined the overall sentiment of the document or sentence Document score ranges between -2 and 2

Semania's cloud-based sentiment analysis software is based on Natural Language

Processing and delivers you more consistent results than two humans.Using automated sentiment analysis, Semantic analyzes each document and its components based on sophisticated algorithm developed to extract sentiment from you content in a similar manner as a human - only 60,000 times faster

Existing approaches to sentiment analysis can be grouped into three main categories:

- Keyword spotting
- Lexical affinity
- Statistical methods

Keyword spotting is the most naive approach and probably also the most popular because of its accessibility and economy. Text is classified into effect categories based on the presence of fairly unambiguous affect words like ‘happy’, ‘sad’, ‘afraid’, and ‘bored’. The weakness of this approach lie inn two ares:
poor recognition of affect when negation is involved and reliance on surface feature.About its first weakness, while the approach can correctly classify the sentence “today was a happy day” as being happy, it is likely to fail on a sentence like “today wasn’t happy day at all”
About its second weakness, approach relies on the presence of obvious effect words that are only surface features of the prose.

In practice , a lot of sentences combine through underlying meaning rather than first affect adjectives. For example,the text “My husband just filed for divorse and he wants to take custody of my children away from me “ certainly evokes strong emotion, but uses no effective keywords, and therefore, cannot be classified using a keyword spotting approach.

Lexical affinity is slightly more sophisticated than keyword spotting as, rather than simply detecting obvious affect words, it assign arbitrary words a probabilistic ‘affinity’ for a particular emotion For example ‘accident’ might be assigned a 75% probability of being indicating a negative affect, as in ‘car accident’ or hurt by accident’ These probabilities are usually trained from linguist corpora.

1.1 About Project

Sentiment Analysis is a way to analyze the subject information in the text and then mine the opinion. Sentiment analysis is the procedure by which information is extracted from the opinion, appraisal and emotion of people in regards to entities, events and their attributes.
Sentiment Analysis is used to discover people’s opinions, emotions and feelings about a

product or service. It is a computational study of opinions and views expressed in text. This text can be in a variety of formats like Reviews, Blogs, News or Comments.

The ability to extract insights from this type of data is a practice that is widely adopted by many organizations across the world. Its applications are broad and powerful.

1.2 Objectives of Project

The objectives of the project are as follows

- To collect the Data set from inshorts.com and selected 3 domains of news articles namely- Sports, World and Politics.
- To Classify polarity of the text at the sentence level for all the 3 domains— whether the expressed opinion in the sentence level is “positive”, “negative”, and “neutral”.
- To visualize data by plotting Pie charts.

1.3 Scope of the Project

The main scope of this project is to understand the emotions of the writer of a particular text and perform sentiment analysis on the dataset containing different news domains from inshorts.com . Namely- Sports, Politics and World. Tools like VADER give the polarity of positivity, negativity, neutrality and also the consolidated compound score for the given text in the data set.

1.4 Advantages

The advantages of Sentiment Analysis are as follows:

1. Brand Monitoring
2. Customer Service
3. Finance and Stock Monitoring
4. Business Intelligence Buildup
5. Market Research and Analysis

The detailed explanation for the advantages is as follows:

1. Brand Monitoring

Sentiment analysis enables you to quantify the perception of potential customers. Analyzing social media and surveys, you can get key insights about how your business is doing right or wrong for your customers.

Companies tend to use sentiment analysis as a powerful weapon to measure the impact of their products and campaigns on their customers and stakeholders. Brand monitoring allows you to have a wealth of insights from the conversions about your brand in the market. Sentiment analysis enables you to automatically categorize the urgency of all brand mentions and further route them to the designated team.

Keeping the feedback of the customer in knowledge, you can develop more appealing branding techniques and marketing strategies that can help make quick transitions.

2. Customer Service

Customer service companies often use sentiment analysis to automatically classify their user incoming calls into "urgent" and "not urgent" classes. The classification is based on the sentiments of the emails or proactively identifying the calls of frustrated customers.

The customer expects their experience with the companies to be intuitive, personal, and immediate. Therefore, the service providers focus more on the urgent calls to resolve users' issues and thereby maintain their brand value. Therefore, analyze customer support interactions to make sure that your employees are following the appropriate process. Moreover, increase the efficiency of your services so that customers aren't left waiting for support for longer periods.

3. Finance and Stock Monitoring

Making investments, especially in the business world, is quite tricky. The stocks and market are always on the edge of risks, but they can be condensed if you do correct research before investing.

4. Business Intelligence Buildup

Digital marketing plays a prominent role in business. Social media often displays the reactions and reviews of the product. When you are available with the sentiment data of your company and new products, it is a lot easier to estimate your customer retention rate.

Sentiment analysis enables you to determine how your product performs in the market and what else is needed to improve your sales. Business intelligence is all about staying dynamic. Therefore, sentiment analysis gives you the liberty to run your business effectively.

5. Enhancing the Customer Experience

A satisfying customer experience means a higher chance of returning the customers. A

successful business knows that it is important to take care of how they deliver compared to what they deliver.

Brand Monitoring offers us unfiltered and invaluable information on customer sentiment. However, you can also put this analysis on customer support interactions and surveys.

This classification will help you properly implement the product changes, customer support, services, etc.

6. Market Research and Analysis

Business intelligence uses sentiment analysis to understand the subjective reasons why customers are or are not responding to something, whether the product, user experience, or customer support.

Sentiment analysis will enable you to have all kinds of market research and competitive analysis. It can make a huge difference whether you are exploring a new market or seeking an edge on the competition.

1.5 Disadvantages

Some of the challenges faced for this approach are follows:

- Misspellings and grammatical mistakes may cause the analysis to overlook important words or usage.
- Sarcasm and irony may be misinterpreted.
- Analysis is language-specific.
- Discriminating jargon, nomenclature, memes, or turns of phrase may not be recognized.

1.6 Applications

Sentiment Analysis has a wide range of applications. They are as follows:

- Social Media: For instance the comments on social media sites such as Instagram, over here all the reviews are analyzed and categorized as positive, negative, and neutral.
- Customer Service: In the play store, all the comments in the form of 1 to 5 are done with the help of sentiment analysis approaches.
- Marketing Sector: In the marketing area where a particular product needs to be reviewed as good or bad.
- Reviewer side: All the reviewers will have a look at the comments and will check and give the overall review of the product.

1.7 Hardware Requirements & Software Requirements:

The Hardware and Software requirements for the project are as follows:

Processor : Pentium IV(minimum)

Hard Disk : 40GB to 80GB

RAM : 256MB (minimum)

Operating System : Windows or Linux or Mac

Technology : PYTHON

IDE : Google colab

Explanation:

PYTHON:

It is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance.

Google Colab:

Google Colab is a powerful platform for learning and quickly developing machine learning models in Python. It is based on the Jupyter notebook and supports collaborative development. The team members can share and concurrently edit the notebooks, even remotely. The notebooks can also be published on GitHub and shared with the general public. Colab supports many popular ML libraries such as PyTorch, TensorFlow, Keras and OpenCV.

2. Literature Survey

This section summarizes some of the scholarly and research works in the field of Machine Learning and data mining to analyze sentiments on the Article and preparing

prediction model for various applications. As the available social platforms are shooting up, the information is becoming vast and can be extracted to turn into business objectives, social campaigns, marketing and other promotional strategies . The benefit of social media to know public opinions and extract their emotions are considered by authors in [2] and explained how twitter gives advantage politically during elections. Further, the concept of the lexicon is used for text classification as it conveys emotion in a few words. They suggested how previous research work suffered from lack of training set and misses some features of target data. They opted for a two stage approach for their framework- first preparing training data from news article using mining conveying relevant features and then propounding the Supervised Learning Model. After collecting and preprocessing the data, the training data set was created first by manual labeling of lexicon and forming clusters, next by using online Sentimental Analyzer VADER which outputs the polarity in percentage. This approach reduced the number of training set .The metric they used to determine the winner was the "PvT ratio" which is Positive number of data to the total count of data for each domain.

Sentiment Analysis by researchers Imran et al. [1] exploited the technology 'Apache Spark' for fast streaming of data and presented the approach StreamSensing to handle real time data in unstructured and noisy form. They conducted the approach on lexicon data to find some useful and interesting trends which further can be generalized to any real-time text stream. Unsupervised learning approach is used to locate interesting patterns and trends from lexicon processed on Apache Spark. Inspired by the approach described by Zhu et al. [7] and Li et al. [8] for mining data by selecting a time window, authors [1] opted for a sliding window method for capturing the live streams of tweets.

The common approach found in almost all relevant research works constitutes data collection using Twitter API, preprocessing of data, filtering of data then approaches in feature extraction, classification and pattern analysis makes the distinction. Authors used a sliding window of 5 minutes during data collection and further created Term Document Matrix(TDM) for feature extraction. The pattern analysis was carried out by using the score of TF-IDF for finding most important keywords as explained in [9] by Wu et al. The trending topic or hashtag is fed and tweets relevant to it are filtered to form TDM and

computing the weights of TF-IDF to find the most important words is the key idea of this sentiment analysis.

Parallel computation of TDM, TF-IDF score and determining top 5 keywords generated from TDM in each minute as the sliding window moves are one of the highlighting features of this research work. Thus, it leverages the fast computation power of Apache Spark.

In another work [5] of Sentiment Analysis and Influence Tracking on News Articles, authors also predicted the polarity – positive, negative or neutral of data by creating a classifier. In addition, they used multiple algorithms and methods to determine the influence of active entities on the data patterns of users exhibiting certain emotions. They mined data only at the entity level i.e. brand, product, celebrity elements present in tweets rather than the whole sentence in the data posted by users. The approach they followed using algorithms to extract features and track the impact and influence made their work different from rest of literature. The feature extraction process after preprocessing included constructing n grams along with POS taggers taking care of the negation part and improving accuracy of classification. For further analysis and measuring influence, they opted for two algorithms – People Rank Algorithm inspired by Page Rank Algorithm [6] used by Google. The main idea behind this algorithm is the more the value of People Rank, the more central the node in the graph means its importance on data in terms of followers, data and mentions.

2.1 EXISTING SYSTEM

We are truly living in the information age where the data is generated by both humans and machines at an unprecedented rate, therefore it's nearly impossible to gain insights into data for making intelligent decisions, manually. One such insight is assessing/calculating the sentiment of a big (rather huge) dataset.

One way of assessing, rather calculating the emotion on top of such a huge dataset is by way of employing sentiment analysis techniques.

2.2 Proposed System

To overcome the drawbacks of the methods we have reviewed above, we propose a

new model for sentiment analysis. In this model we combine many techniques to reach our final goal of emotion extraction. The steps for the process are documented below.

1. Retrieval of Data: Public Paper data is mined using the existing inshorts.com for data extraction. Articles would be selected based on a few chosen keywords pertaining to the domain of our concern, i.e; sports, world and politics news article domains. We have elected to use the inshorts.com website due to ease of data extraction.

2. Preprocessing: In this stage, the data is put through a preprocessing stage in which we remove identifying information such as Fullstops, commas and embedded links and images. Such information is largely irrelevant and may cause false results to be given by our system.

3. Data Cleaning: Removed Special characters by using regular expression operations library of python, stopwords(words that don't add much meaning to the sentence), HTML Tags by using HTML parser and also expanded contractions.

4. Polarity detection: In this step we begin the second phase of our proposed system, in which we try to identify the polarity of the sentence in question. We aim to find sentences where the polarity detection is not very clear or where the expressed sentiment may be low. We also try to isolate the opinion words in the sentence in relation to a given concept in the sentence.

a. We train the system to understand the relation between words in various contexts. Pre-existing dictionaries like SenticNet can be used in this phase to segregate the emotion from the context it is in.

b. Once the opinion words are identified with context, we can find the polarities of the words using NLTK-SentiWordNet.

c. To help with detection of the concepts associated, we train our system on a large dataset that expresses a wide variety of complex and ambiguous emotions. The system is given this data in an unsupervised fashion and will proceed by clustering.

5. Emotion Extraction: Text is represented as a bag of words. Each word is based on the intensity of emotions in the text on the scale of -1 to +1.

The Compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1 (most extreme positive).

- positive sentiment : (compound score >= 0.05)
- neutral sentiment : (compound score > -0.05) and (compound score < 0.05)
- negative sentiment : (compound score <= -0.05)

a. Mapping:

Once the word is found using VADER(Valence aware Dictionary and Sentiment Reasoner), it is mapped to the corresponding compound score.

b. Once the system calculates the summation of all the words in the news article, the membership of the emotion or emotions expressed in the statements is used to determine the most significant emotions. This value is decided after comparing all the compound scores of membership given by the opinion words in the statement.

3. Proposed Architecture

The below figure describes the overall architecture of proposed system

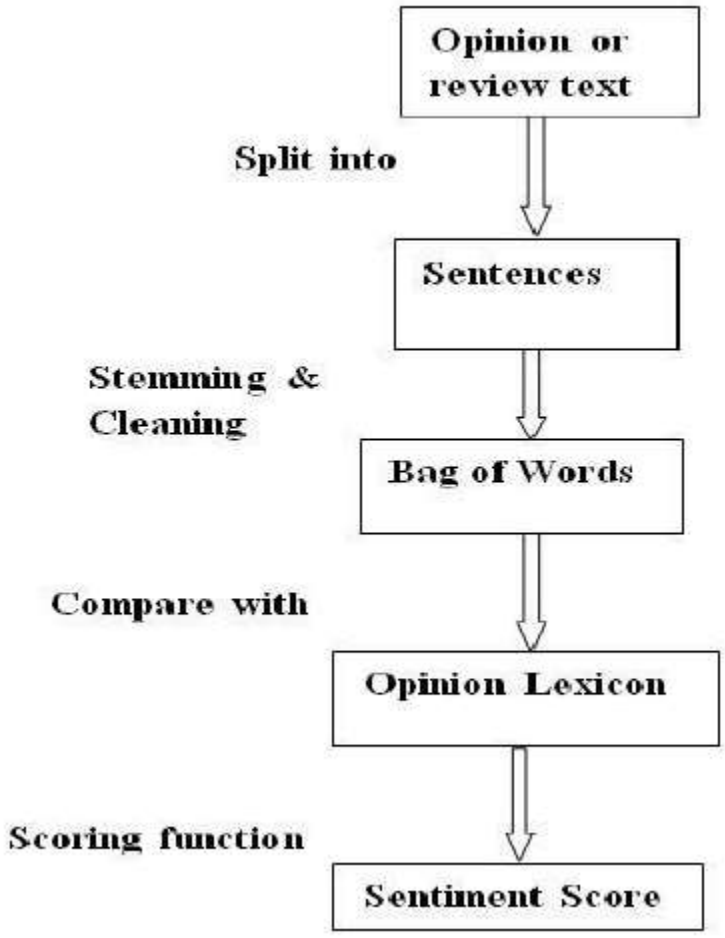


Fig 3.1. The model of sentiment analysis

Figure shows the model for sentiment analysis of news article comments using lexicon based approach. Initially, we gathered news from the inshorts. We used these comments as an input. The next step is preprocessing that removes the unwanted and noisy data. This step includes case conversion, split sentences, replacing “n't” with “not”, punctuation removal, and tokenization.

The next step is opinion or review text . This system compares each tokenized word in the comment with blind negation word or negation word or positive opinion word or negative opinion word or intensifier word by using sentiment word database. And then the polarity scores are assigned to each word by using a sentiment word database. The presence of the blind negation word indicates negative sentiment value (-2).

A heuristic technique is used to calculate the semantic orientation score of combining words for automated students' feedback comments analysis. In the following equations, ‘Ws’ is the semantic orientation score of combining words. ‘Sinf’ is the intensifier value of a word based on 100%. ‘Os' is the score of an opinion word from a sentiment word database.

$$Ws = Os \tag{1}$$

$$Ws = (100\% + Sinf) * Os \tag{2}$$

$$Ws = (100\% + Sinf) * (100\% + Sinf) * Os \tag{3}$$

$$Ws = Ws^{*}(-1) \tag{4}$$

If there is only one opinion word in a sentence, the corresponding positive scores or negative scores are assigned using (1) . If one intensifier word and one opinion word are found together, i.e. the location of intensifier word is adjacent with the location of the opinion word, (2) is used to get the semantic orientation score of combining words. If two intensifier words and one opinion word are found in a sentence, moreover the index of the first intensifier word must be the index by reducing two of the index of opinion word, (3) is used. If a negation word in front of the opinion word is found in a sentence, reversed polarity scores are given by (4).

The semantic orientation score of combining words in all sentences is summed up to get the total polarity scores by (5) . In (5), 'PTs' is the total polarity score of all words in all sentences from one comment. 'Wsi' is the semantic orientation score of combining words for each term in one comment. 'i' is the order of combining opinion words from 1 to n. 'n' is the total number of combining opinion words in all sentences from one comment. 'T' is a set of teaching sentiment terms from a sentiment word database. 'PTsi' is the total polarity score of the term for all comments. 'N' is the total number of opinion words in all comments. 'P' is the average polarity score of all comments.

$$\mathrm{PTs} = \Sigma_{\iota=1}^{n} \mathrm{Wsi}, (\mathrm{Wsi} \in \mathrm{T}) \quad (5)$$

$$\mathrm{P} = \sum_{i=1}^{N} \mathrm{PTsi}/\mathrm{N} \quad (6)$$

The average polarity score of all comments can be calculated by (6).

Methodology :

Sentiment analysis interprets the subjective information from online reviews to implicit assessment based on score. The scientific domain has hundreds of thousands people who care about news in the world . It takes a long time to select suitable papers. Online reviews on papers are the essential source to help them. The reviews save reading time and paper costs. In this chapter, we present a new technique to analyze online reviews which is called: "Sentiment Analysis Of Online Articles"

Article reviews are analyzed for their sentiment to predict the article response as positive or negative.

The overall methodology follows four steps

1) Data Collection
2) Preprocessing
3) Sentiment Extraction
4) Classify sentiment as positive or negative.

1.Data Collection: Data collection is defined as collecting and analyzing data to validate and research using some techniques. Public Paper data is mined using the existing inshorts.com for data extraction. Articles would be selected based on a few chosen keywords pertaining to the domain of our concern, i.e. sports, world and politics news article domains. We have

elected to use the inshorts.com website due to ease of data extraction.

2. Preprocessing: In this stage, the data is put through a preprocessing stage in which we remove identifying information such as Fullstops, commas and embedded links and images. Such information is largely irrelevant and may cause false results to be given by our system.

3. Sentiment Extraction: In this process the sentiment that is expressed in the text is extracted using VADER , in which we try to identify the polarity of the sentence in question.

4. Classify Sentiment as positive negative or neutral: In the process the text is classified as positive negative or neutral using compound score

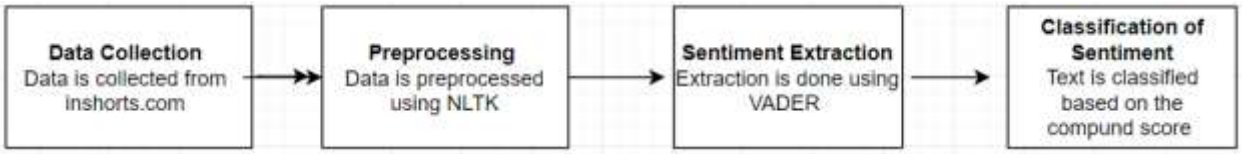


Fig 3.2 Block Diagram of the proposed System

Tools used: The tools used as part of the project are as follows:

A. NLTK(Natural Language Toolkit)

The following figure shows how NLTK classifies a review into either positive or negative. It first tokenizes the reviews into words and then removes the stop words such as a, an, the, for, is, etc. After removing stop words, the words are stemmed to get their root words. For example, “disappointed” is reduced to “disappoint”. This helps in reducing the time while searching the word in the SentiWordNet. All special symbols and numbers are also removed from the reviews. Now it performs the POS (Parts of Speech) tagging on the purified reviews. It involves stringent grammar rules while performing the tagging. Thus, the data is ready for classification by extracting the positive and negative words from the given review and match them with the respected sentiment score given in the SentiWordNet. Finally, by counting the positive and negative terms which are found in the review, and using sentiment polarity, the class receives the highest score.

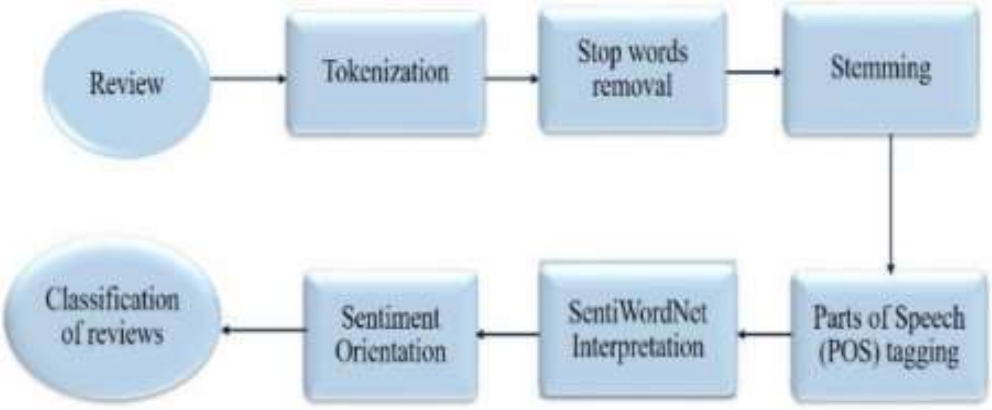


Fig. 3.3 Overall process of NLTK to classify a review

For each Synset the score in SentiWordNet lexicon, can be calculated by

SynsetScore = PosScore — NegScore (2)

For a term with specific POS tag, if k synsets contain it, then the sentiment score of the term can be calculated by following expression

$$TermScore = \frac{\sum_{n=1}^{k} SynsetScore(r)/r}{\sum_{n=1}^{k} 1/r} \quad (3)$$

Where n is the sense number.If a term is not in the SentiWordNet, we assume that its sentiment score is 0. If a negation word appears in front of a term, we simply reverse the sentiment value of the respected term. The sentiment score of the target review can be calculated by adding up all the term sentiment scores as shown in below:

$$PosScore(p) = \sum_{i=1}^{m} TermScore(T_i) (4)$$
$$PosScore(p) = \sum_{i=1}^{m} TermScore(T_i) \quad (5)$$
$$SentiScore(p) = PosScore(p) + NegScore(p) (6)$$

Where p is a review which contains m positive terms and n negative terms.PosScore(p) and NegScore(p) represents the positivity and negativity of the corresponding review p. SentiScore of p represents the final sentiment score of the review p.

B. TextBlob

Textblob is a python library that provides text mining, text analysis and text processing modules for python developers. Textblob reuses NLTK corpora, and if NLTK has been installed beforeTextblob, then the Textblob will be installed with a great ease. Textblob supports the python versions 2.6 and the latest. Installation: $ pip install -U textblob $ python -m textblob.download_corpora The above commands install Textblob and download necessary NLTK corpora, and if NLTK is installed before Textblob, there is no need to download corpora. Textblob is a sentence level analysis. First, it takes a dataset as the input then it splits the review into sentences. A common way of determining polarity for an entire dataset is to count the number of positive and negative sentences/reviews and decide whether the response is positive and negative based on the total number of positive and negative reviews. Polarity and subjectivity of a given review can be known using sentiment() function. It returns a named tuple with two parameters called polarity and subjectivity. The polarity score is ranging from -1 to 1 and subjectivity ranges are from 0 to 1 where 0 is most objective and 1 is most subjective. Example: review=Textblob ("the movie was interesting.") review.sentiment # Sentiment(polarity=0.5, subjectivity=0.5)

C. VADER

As mentioned earlier, VADER is a lexicon and rule-based sentiment analysis tool. It uses a combination of a sentiment lexicon, a list of lexical features which are generally labeled according to their semantic orientation as either positive or negative.VADER has been quite successful when dealing with social media texts, movie reviews, and product reviews. This is because VADER not only tells about the positivity and negativity score but also tells about how positive or negative a sentiment is. The developers of VADER have used Amazon's Mechanical Turk to get most of their ratings.

1. Advantages of VADER

a. Works perfectly on social media type text.

b. It does not require any training data but is constructed from generalizable, valence-based, human-curated gold standard lexicon.

c. VADER supports emoji for sentiment classification.

d. It is fast enough to be used online.

e. It does not severely suffer from a speed-performance tradeoff.

Installation: C:\Users\Admin>pip installvaderSentiment

VADER analyzes a piece of text to see if any of the words from the text are present in the VADER lexicon. It can find the polarity indices using polarity_scores() function. This will return the metric values of the negative, neutral, positive, and compound for a given sentence. The compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1 and +1 where -1 indicates most extreme negative and +1 indicates most extreme positive. It is useful to set the standardized thresholds for classifying sentences as positive, neutral or negative. The typical threshold values are given below

Positive Sentiment: compound score >= 0.05

Neutral Sentiment: compound score > -0.05 and < 0.05

Negative Sentiment: compound score <= -0.05

These are the most useful metrics for multidimensional measures of sentiment for a given textual review.The below figure shows the VADER lexicon containing words along with their sentiment ratings.

VADER analyses sentiments primarily based on certain key points such as Punctuation, Capitalization, Degree modifiers, Conjunctions, Preceding Tri-gram [10]. There are more than 7,500 lexical features with validated valence scores that indicate both the sentiment polarity, and sentiment intensity ranging from -4 to +4.

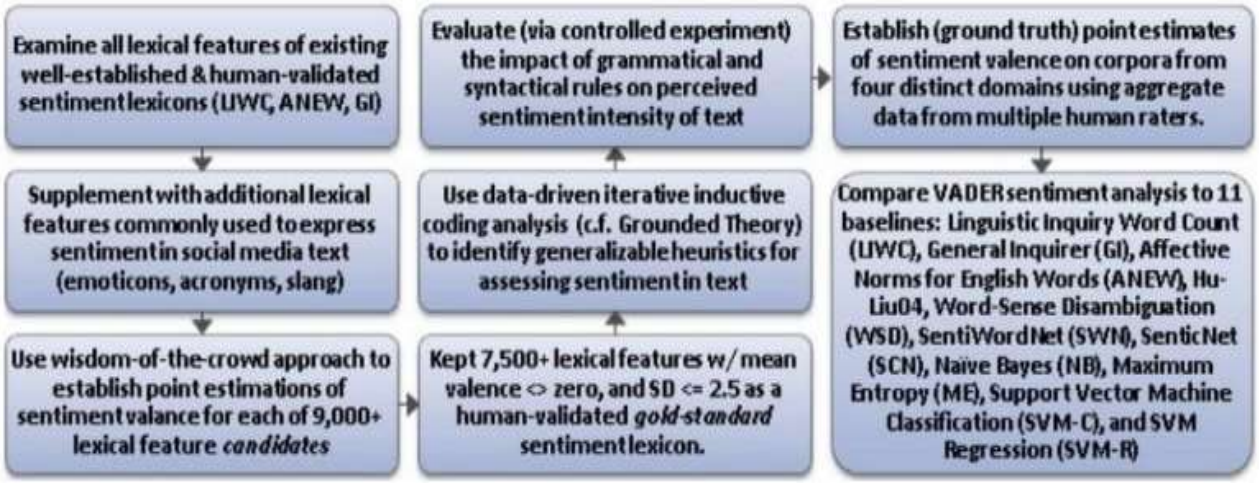


Fig 3.4 Methods and process approach overview of VADER

4. Implementation

The Algorithm of the Project is as follows:

4.1 Algorithm

Step	Description
1	The input.
2	Web scraping/ Web extracting.
3	Text analysis process. o Split into sentences.
4	Check on class for each sentence in each review. o Classify each sentence in each review. o Assume SA score=0.
5	Text analysis process (continue) o For each w in sentence (tokenization) o Remove stop list. o Remove punctuation list. o Convert all words into Upper case.
6	Create a reduced lexicon. o Assume each word has two values (positive and negative). o Convert word to infinitive o Construct the reduced lexicon.
7	Check on words. o Check on the word list. o Check on the prefixes list. o Else, Check on Nouns (Features, keywords, noun).
8	Check on Future words.
9	Assign sentiment classification for each review.
10	Get the total sentiment score for each word. (Real total score and Average total score).

Steps for proposed system

4.2 Code

```
!pip install vaderSentiment
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
vs=SentimentIntensityAnalyzer()
#web scraping
import requests
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
import os
url='https://www.inshorts.com/en/read/sports'
news_data=[]
new_category=url.split('/')[-1]
data=requests.get(url)
soup=BeautifulSoup(data.content)
print(soup)
url=['https://www.inshorts.com/en/read/sports']

def build_dataset(url):
  news_data=[]
  for u in url:
    soup=BeautifulSoup(requests.get(u).content)
    category=u.split('/')[-1]
    news_article =[{'news_headline': headline.find('span',
attrs={"itemprop":"headline"}).string,
                'news_article': article.find('div', attrs={"itemprop":"articleBody"}).string,
                'news_category': category}
                for headline,article in zip(soup.find_all('div',class_=["news-card-title news-
right-box"]),
```

```
                soup.find_all('div',class_=["news-card-content news-right-box"]))
            ]
    news_article = news_article[0:20]
    news_data.extend(news_article)
df=pd.DataFrame(news_data)
df=df[['news_headline','news_article', 'news_category']]
return df
df=build_dataset(url)
df.to_csv('news.csv',index=False)
!pip install nltk
import nltk
nltk.download('stopwords')
stopword_list = nltk.corpus.stopwords.words('english')
len(stopword_list)
# function to remove HTML tag
def html_tag(text):
  soup=BeautifulSoup(text,"html.parser")
  new_text = soup.get_text()
  return new_text
# Expand Contraction
!pip install contractions
import contractions
def con(text):
  expand=contractions.fix(text)
  return expand
#removal of special characters
import re
def remove_sp(text):
  pattern= r'[^A-Za-z0-9\s]'
  text= re.sub(pattern,'',text)
  return text
from nltk.tokenize.toktok import ToktokTokenizer
```

```
tokenizer=ToktokTokenizer
#removal of stopwords
tokenizer = ToktokTokenizer()
def remove_stopwords(text):
  tokens = tokenizer.tokenize(text)
  tokens = [token.strip() for token in tokens]
  filtered_tokens = [token for token in tokens if token not in stopword_list]
  filtered_text= ' '.join(filtered_tokens)
  return filtered_text
#1. Lower case
#2. HTML tags
#3. Contractions
#4. Stopwords
df.news_headline=df.news_headline.apply(lambda x:x.lower())
df.news_article=df.news_article.apply(lambda x:x.lower())
df.news_headline=df.news_headline.apply(html_tag)
df.news_article=df.news_article.apply(html_tag)
df.news_headline=df.news_headline.apply(remove_sp)
df.news_article=df.news_article.apply(remove_sp)
df.news_headline=df.news_headline.apply(remove_stopwords)
df.news_article=df.news_article.apply(remove_stopwords)
# dataset labeling and processing
df['compound'] = df['news_headline'].apply(lambda x: vs.polarity_scores(x)['compound'])
df.head()
# data finalization
def predict(comp):
  comp=float(comp)
  if (comp>0):
    return 'positive'
  elif (comp==0):
    return 'neutral'
  else:
    return 'negative'
```

```
df['type_pred'] = df['compound'].apply(predict)
df.head()
pip install matplotlib
Polarity = df.compound
print(Polarity)
df.plot(y="compound")
plt.title("Plot For the News Articles")
plt.xlabel("Sports World Politics")
plt.ylabel("Compound Score")
total_score = df['news_category'].groupby(df['type_pred'])

# printing the means value
print(total_score.count())

total_Polarity=df.groupby(df['compound'])
print(total_polarity.mean())
Emotion = ['NEGATIVE','NEUTRAL','POSITIVE']

data = [26,26,8]
# Creating plot
fig = plt.figure(figsize =(10, 7))
plt.pie(data, labels =Emotion)

# show plot
plt.show()
```

5. Results

The output screens are as follows:

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

	news_headline	news_article	news_category	compound	type_pred
0	i will eat icecream with you after the tokyo o...	during an online interaction with indias tokyo...	sports	0.0000	neutral
1	icc do not want me to use the universe boss st...	talking about using just the boss sticker inst...	sports	-0.0572	negative
2	fans 2013 tweet predicting italys win over eng...	a football fans eightyearold tweet correctly p...	sports	0.5859	positive
3	f1 driver landos 40000 watch stolen outside we...	britishbelgian formula one driver lando norris...	sports	-0.4939	negative
4	crawl back under your rock uk pm to those raci...	condemning people who are racially abusing eng...	sports	-0.4588	negative

Fig 5.1 Output screen for with the predicted compound score and type prediction.

The above figure shows the first five records of the dataset represented in the form of a table. The attributes or column names are as follows: Namely- news_headline, news_article, news_category, compound and type_pred.

Predicted attributes are compound and type_pred. Based on the compound score, the emotion expressed is classified as positive or negative or neutral emotion.

Table containing the values to be plotted for the graph.

Index (X-axis)	**Compound Score (Y-axis)**	**Index (X-axis)**	**Compound Score (Y-axis)**
0	0.0000	31	0.0000
1	-0.4939	32	-0.5106
2	0.5859	33	0.0000
3	-0.4939	34	-0.6249
4	-0.4588	35	-0.5423
5	0.0000	36	-0.1027
6	-0.3612	37	-0.8020
7	-0.4997	38	0.0000
8	0.0000	39	-0.7650
9	0.0000	40	-0.5106

10	0.5859	41	0.0000
11	0.0000	42	0.0000
12	-0.3818	43	0.0000
13	0.2023	44	0.0000
14	0.0000	45	-0.3400
15	0.1531	46	0.0000
16	-0.5994	47	0.2263
17	0.0000	48	0.0000
18	0.0000	49	0.0000
19	-0.4404	50	-0.8176
20	-0.4939	51	0.0000
21	0.3818	52	0.0000
22	-0.4588	53	-0.2500
23	0.5106	54	0.0000
24	0.0000	55	0.0000
25	-0.2263	56	-0.8271
26	-0.8910	57	0.0000
27	0.0000	58	-0.4023
28	-0.3400	59	-0.5267
29	0.0000		
30	0.3612		

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

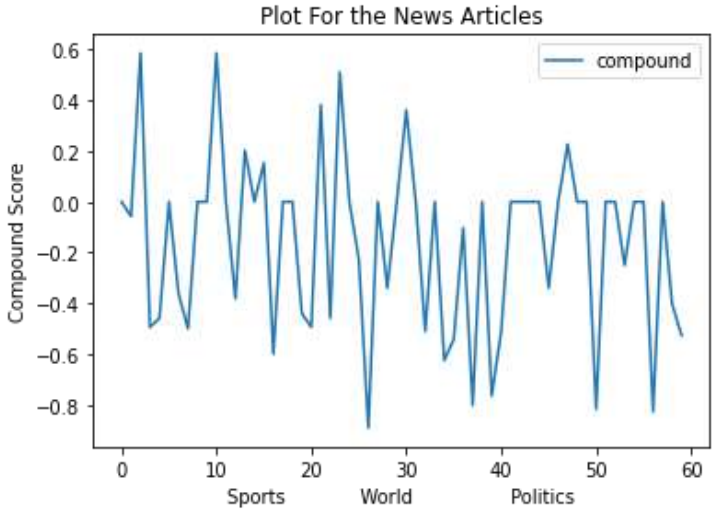
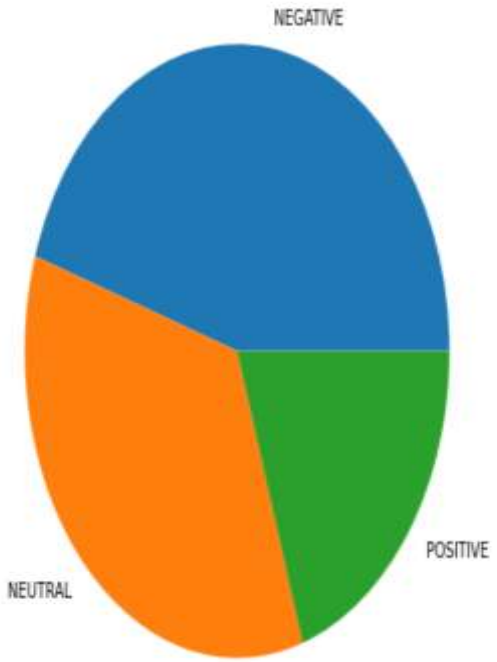


Fig 5.2 Plot for the News Articles of different domain with respect to their compound scores

The figure contains the plot for all the data items with respect to their index and corresponding compound scores across various Domains.

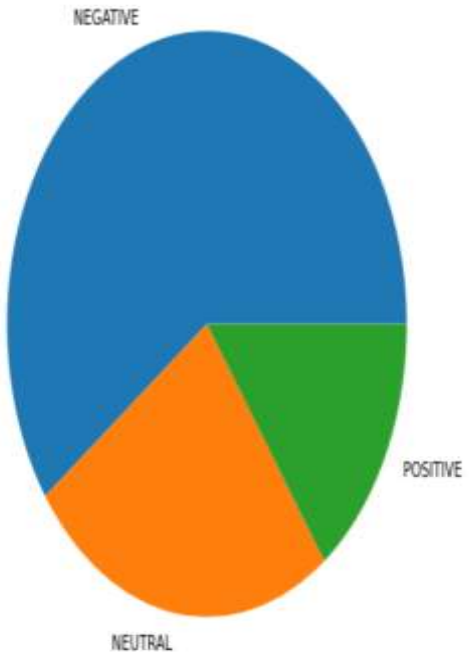
i) Index 0 - 19 : contains plot for Sports Domain

0	0.0000	10	0.5859
1	-0.4939	11	0.0000
2	0.5859	12	-0.3818
3	-0.4939	13	0.2023
4	-0.4588	14	0.0000
5	0.0000	15	0.1531
6	-0.3612	16	-0.5994
7	-0.4997	17	0.0000
8	0.0000	18	0.0000
9	0.0000	19	-0.4404



ii) Index 20-39 : contains plot for World Domain

20	-0.4939	30	0.3612
21	0.3818	31	0.0000
22	-0.4588	32	-0.5106
23	0.5106	33	0.0000
24	0.0000	34	-0.6249
25	-0.2263	35	-0.5423
26	-0.8910	36	-0.1027
27	0.0000	37	-0.8020
28	-0.3400	38	0.0000
29	0.0000	39	-0.7650



iii) Index 40-59: contains plot for Politics Domain

Fig 5.3 Pie chart representing the prediction of the dataset

0	0.0000	31	0.0000
1	-0.4939	32	-0.5106
2	0.5859	33	0.0000
3	-0.4939	34	-0.6249
4	-0.4588	35	-0.5423
5	0.0000	36	-0.1027
6	-0.3612	37	-0.8020
7	-0.4997	38	0.0000
8	0.0000	39	-0.7650
9	0.0000	40	-0.5106

10	0.5859	41	0.0000
11	0.0000	42	0.0000
12	-0.3818	43	0.0000
13	0.2023	44	0.0000
14	0.0000	45	-0.3400
15	0.1531	46	0.0000
16	-0.5994	47	0.2263
17	0.0000	48	0.0000
18	0.0000	49	0.0000
19	-0.4404	50	-0.8176
20	-0.4939	51	0.0000
21	0.3818	52	0.0000
22	-0.4588	53	-0.2500
23	0.5106	54	0.0000
24	0.0000	55	0.0000
25	-0.2263	56	-0.8271
26	-0.8910	57	0.0000
27	0.0000	58	-0.4023
28	-0.3400	59	-0.5267
29	0.0000		
30	0.3612		

In the Data set, there are 60 records.

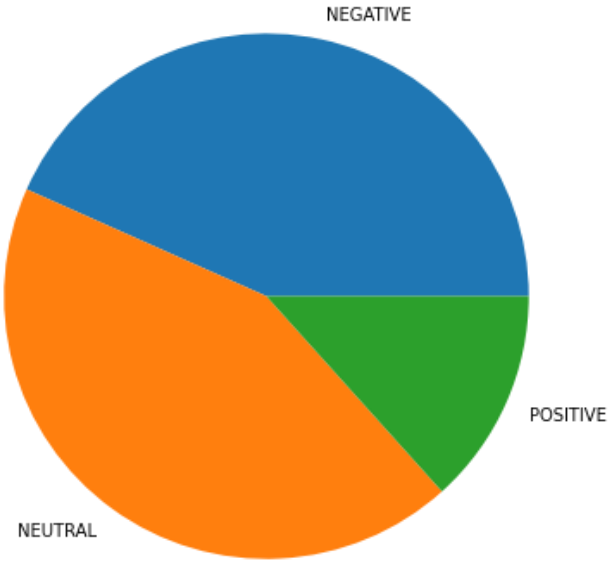
Out of those 60 records, there are -

negative 26

neutral 26

positive 8

For this data, we have plotted the pie chart.




```
In [ ]: from sklearn.metrics import accuracy_score,confusion_matrix
        confusion_matrix(y_pred,y_test)

Out[ ]: array([[ 7,  0,  1],
               [ 0, 11,  0],
               [ 1,  0,  7]])
        time: 10.7 ms (started: 2021-06-16 10:30:17 +00:00)

In [ ]: accuracy_score(y_pred,y_test)

Out[ ]: 0.9259259259259259

        time: 14.7 ms (started: 2021-06-16 10:30:49 +00:00)
```

Fig 5.4 The output screen for Accuracy and Confusion Matrix.

The Accuracy of the project after implementation is found to be 92.59%.

Accuracy:

The percentage of accurate predictions for the test results is known as accuracy in Machine

Learning.

The accuracy_score method is used to calculate the accuracy of either the fraction or count of correct prediction in Python Scikit learn.Mathematically it represents the ratio of the sum of true positives and true negatives out of all the predictions.

```
In [ ]: accuracy_score(y_pred,y_test)

Out[ ]: 0.9259259259259259

        time: 14.7 ms (started: 2021-06-16 10:30:49 +00:00)
```

The Accuracy of the project after implementation is found to be 92.59%.

Accuracy Score = (TP+TN)/ (TP+FN+TN+FP)

Confusion Matrix : A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. A much better way to evaluate the performance of a classifier is to look at the confusion matrix.

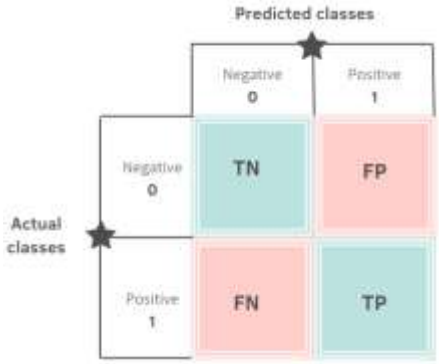


Fig 5.5 Confusion Matrix

```
In [ ]: from sklearn.metrics import accuracy_score,confusion_matrix
        confusion_matrix(y_pred,y_test)

Out[ ]: array([[ 7,  0,  1],
               [ 0, 11,  0],
               [ 1,  0,  7]])
        time: 10.7 ms (started: 2021-06-16 10:30:17 +00:00)
```

6. Conclusion

It has been confirmed by this investigation that it is possible to collect, pre-process, analyze and visualize article data using Python open source tool packages. It is viable to apply text mining tasks and sentiment analysis for the data to analyze writer contributed reviews of the sports world and political domain . It will provide a competitive advantage for researchers and service providers to analyze views regarding news articles using social media data. This will help them improve their business value and better manage their customer relationship. The described approach is applicable on other social media data sources such as Facebook. It can be generalized that, Businesses can utilize their consumer opinions generated from social media tracking and analysis by adapting their marketing plans, products and business intelligence respectively. An important perspective for future work could be building social media tracking and monitoring systems as opinions are changing over time. Moreover, it is also valuable to use un-supervised techniques in sentiment analysis and opinion mining for improving the business competitive value and the customer relationship management. In addition to comparing various sentiment classification techniques utilized for opinion mining. What is more, social media can be used as a tool for sales prediction using opinion mining. The accuracy of the model is found to be 92.59%.

7. Future Scope

At this stage, the project is limited to only three domains in news articles . In the future, the system can also be extended to analyze sentiments about entertainment , finance and other affairs. Complete removal of ambiguity is an uphill task indeed. Therefore, interpretation and classification of sarcastic sentences are not a part of the current scope. However, in the future, the scope can be extended to accommodate the same. Finally, the project can be extended to work for natural languages other than English

.

8. References

[1] Vikas Malik and Amit Kumar. "Sentiment Analysis of Twitter Data Using Naive Bayes Algorithm", International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 6, No. 4, 2018.

[2] J.Ge, M.Alonso Vazquez, and U.Gretzel,"Sentiment analysis: a review", In Sigala, M. &Gretzel, U. (Eds.), Advances in Social Media for Travel, Tourism and Hospitality: New Perspectives, Practice and Cases,pp. 243-261. New York: Routledge, 2018.

[3] Z. Nanli, Z. Ping, L. Weiguo, and C. Meng, "Sentiment analysis: A literature review", Proceedings of the International Symposium on Management of Technology (ISMOT), Hangzhou, IEEE, pp. 572-576, 2012.

[4] J.W. Pennebaker, R.L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015", Austin, TX: University of Texas at Austin, 2015

[5] S. Baccianella, A. Esuli, and F. Sebastiani, "SENTIWORDNET 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", pp. 2200–2204, 2008

[6] M. Hu and B. Liu, "Opinion Extraction and Summarization on the Web", pp. 1621–1624.

[7] B. Pang, L. Lee, H. Rd, and S. Jose, "Thumbs up ? Sentiment Classification using Machine Learning Techniques", pp. 79–86, 2002

[8] Wordnet.com, "WordNet, a Lexical database for English", [online] Available: http://wordnet.princeton.edu/

[9] Textbolb.com, "Textblob Tutorial, Quickstart", [online] Available at: https://textblob.readthedocs.io/en/latest/quickstart.html#quickstart

[10] C. J. Hutto and E. Gilbert, "VADER : A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text", Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, , pp. 216–225, 2014

[11] Cornell.edu, "Movie Review data", [online] Available: http://www.cs.cornell.edu/people/pabo/movie-review-data/

[12] Steven Bird and Edward Loper. "NLTK: The Natural Language Toolkit", 2006

[13] Bing Liu,"Sentiment Analysis andOpinionMining",Morgan&ClaypoolPublishers, May 2012.

[14] Adamo and David. "A Text Similarity Approach to Sentiment Classification (of Movie Reviews) using SentiWordNet" .10.13140/ RG.2.1.3271.1120, 2015

[15] H. Han, Y. Zhang, J. Zhang, J. Yang, and X. Zou, "Improving the performance of lexicon-based review sentiment analysis method by reducing additional introduced sentiment bias",, pp. 1–11, 2018

[16] Steven Loria. "Textblob Documentation",Release 0.15.2,2018.