Logo

Final Report

On

Sentiment Analysis Research

Submited to: …. Submited By: 1

2

3

4

Submitted Date:

**Table of Contents**

[**1.Finding a research topic** 1](#_Toc42343237)

[**1.1 Sentimental Analysis on Amazon Reviews** 1](#_Toc42343238)

[**1.2. Justification of chosen topic** 1](#_Toc42343239)

[**1.3 Aims and Objectives** 1](#_Toc42343240)

[**1.3.1. Aims** 1](#_Toc42343241)

[**1.3.2 Objectives** 1](#_Toc42343242)

[**1.4 Fact Finding Technique and Strategy** 2](#_Toc42343243)

[**1.5 Specific Library Search Tools** 2](#_Toc42343244)

[**1.6 Knowledge Gained from The Topic** 2](#_Toc42343245)

[**1.7 Outside from Research Library** 2](#_Toc42343246)

[**2. Professional Activities** 3](#_Toc42343247)

[**2.1 Gantt Chart** 3](#_Toc42343248)

[**2.3 Challenges in research** 4](#_Toc42343249)

[**3. Literature Review** 5](#_Toc42343250)

[**3.1 Mind Map** 5](#_Toc42343251)

[**3.2 Tools and resource to develop the Mind Map** 5](#_Toc42343252)

[**3.3 Literature Review** 5](#_Toc42343253)

[**3.3.1 Lexicon Based Sentiment Analysis** 5](#_Toc42343254)

[**3.3.2 Sentiment Analysis Using Naïve Bayes Classifier** 6](#_Toc42343255)

[**3.3.3 Sentiment Analysis of Tweets using Support Vector Machine** 9](#_Toc42343256)

[**3.4 Resources and techniques to write literature review** 11](#_Toc42343257)

[**3.5 Reason for not selecting specific resources** 11](#_Toc42343258)

[**3.6 Chosen research method** 11](#_Toc42343259)

[**3.7 Feasibility, challenges and Recommendation of content analysis** 12](#_Toc42343260)

[**4. Reflection of Research Project** 12](#_Toc42343261)

[**References** 14](#_Toc42343262)

**Table of Images**

[**Image 1 Gantt Chart 3**](#_Toc42347963)

[**Image 2 Mind Map 5**](#_Toc42347964)

[**Image 3 Model of document classification with the original Naïve Bayes classifier 7**](#_Toc42347965)

[**Image 4 Evaluation of Accuracy 8**](#_Toc42347966)

[**Image 5 The accuracy of 3 methods that applying to similar text table. 9**](#_Toc42347967)

[**Image 6 SVM 10**](#_Toc42347968)

[**Image 7 Accuracy comparison 10**](#_Toc42347969)

**PART 1- The Reflective Individual Report**

# **1.Finding a research topic**

## **1.1 Sentimental Analysis on Amazon Reviews**

Analysis of sentiment, also known as opinion mining, is clearly described as the process of discovering the opinions of people expressed in written language. Natural Language Processing (NPL) includes the application, text interpretation, detection and removal of intangible information in literary works. (IGI Global, 2019)

Customer reviews for a wide variety of goods and services are increasingly available online. While studies have shown the benefits of customer reviews to an online retailer, what makes customer reviews beneficial to a consumer in making a buying decision is a largely uninvestigated problem. An analysis of Amazon.coms’ 1,587 reviews across six products indicated that reviewing extremity, review depth, and product type affect the perceived usefulness of the review. (Susan M. Mudambi, 2010)

## **1.2. Justification of chosen topic**

When customers look for product details online to compare product options, they also have access to tens or hundreds of reviews of other consumers' goods. In addition to product information, expert reviews and personalized advice created by automated recommendation systems these customer reviews are given. Each of these solutions possesses the potential to add value to a prospective client. So, on amazon reviews, this subject is chosen to describe how sentimental analysis is working.

## **1.3 Aims and Objectives**

### **1.3.1. Aims**

The aim of this research is to learn and implement the Support Vector Machine (SVM) Sentimental Analysis Algorithm on Amazon reviews.

### **1.3.2 Objectives**

* To use the Google Scholar Library to find relevant information on sentimental analysis,
* To evaluate and analyze performance of various machine learning algorithms applied in sentimental research,
* To find the Amazon Reviews dataset, import the Jupyter Notebook dataset, preprocess data, and create Al model implementing Support Vector Machine (SVM),
* To learn how to model data in order to achieve greater accuracy,
* Understanding how sentimental analysis works

## **1.4 Fact Finding Technique and Strategy**

Good search practice could involve keeping a search diary or document detailing our search activities, so that effective search terms can be tracked.

The search strategies I used when I looked for different sources are:

* + Searching with proper keyword and/or subject headings,
  + combining my search terms with AND/OR, vs, - , etc.

## **1.5 Specific Library Search Tools**

**Google Scholar**

Google Scholar is one of the search tools that provides articles, theses, books, academic publishers abstracts, professional societies and websites across a wide range of disciplines and resources.

Google Scholar is chosen for the reason below:

* It has a function to scan through a broad variety of scholarly sources,
* To Retrieve novels, documents and other references with correct citations

## **1.6 Knowledge Gained from The Topic**

It was known that different journals, articles and books on sentiment analysis and different algorithms have already been used to extract emotions. Each paper had a unique and quality source of information on different techniques with varying accuracy depending on their datasets.

## **1.7 Outside from Research Library**

It was really getting difficult to implement the algorithm only with research papers and books. Research alone was not sufficient to find the information needed to implement the sentimental analysis algorithm in practice. As a result, different tutorial videos on sentiment analysis and Support Vector Machine (SVM) were watched. In addition, other search tools such as 'Catalogue' have been used to find more relevant research papers and book models for sentimental analysis.

# **2. Professional Activities**

## **2.1 Gantt Chart**

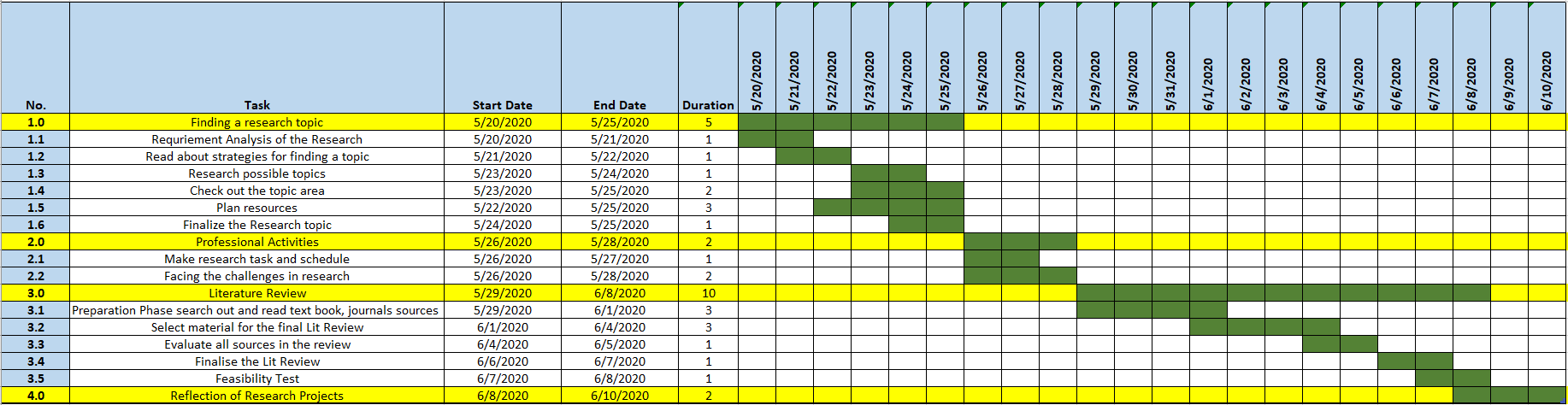


Image 1 Gantt Chart

**2.2 Justification for independent researcher role**

Initially, a researcher must pick a topic. And, with this task assigned to us, the Real Estate Consulting System was the particular region chosen to perform a joint survey. You have to know what you have chosen to look for and why. In hundreds of such systems the main reason why this recommended system is chosen is to exclude the illogical. The first of this study is to offer alternate image types to land / house websites that provide more highlights and enticing events, all in combination with a new look to current ones. Any priorities and targets were then set to kick-start the project. The role also includes reviewing the factors related to the real estate recommendation, assessing systems and algorithms to recommend the property, collecting and analyzing all requirements, content-based work and other forms of machine learning. For the property 's recommendation and accuracy, one must collect Web automation tool data sets and implement content-based algorithm and cosine similarity. To find the relevant data, the specific techniques and strategies that need to be applied to the research because it consists first of all of a survey of Google forms, so that information can be provided on how many people are involved and how many people are involved in immobilization.

Besides this, it is always necessary to investigate and analyze the secondary evidence. To run the computer program on jupyter's notebook on other research articles, records, journals, and archives, you'll need to explore the resources and data sets. The researcher must present a working flow and exact comparison of the recommendation system with other recommendation algorithms such as collaborative filtering and hybrid. The models focused mainly on quality, and each represented different facets of the selection behavior of the users. Customized models tailored to the particular actions of individual users and non-personalized models defined the common patterns of all apps. With the assistance of a software production tool, iteratively based on their results, various models were tested, evaluated and improved, as user experience slowly evolved.

## **2.3 Challenges in research**

At first, as this topic was new to me, I could not collect relevant information for the sentiment analysis. Finding and understanding the core of sentimental analysis had been very difficult for me. Then I started to search various books, journals, on sentimental analysis papers. I was able to understand the working mechanism of sentimental analysis, after reading several papers. Multiple papers have implemented different algorithms which left me very confused in choosing the algorithm to be implemented in my project. I came to learn Naïve Bayer 's research in binary classification effectively after learning different algorithms. Then I took an amazon feedback named dataset and attempted to apply Naïve Bayer 's Algorithm. I also had problems in the preprocessing phase in the implementation part but it was solved later after a colleague's help. And eventually, the software was developed and the algorithm's consistency tested.

# **3. Literature Review**

## **3.1 Mind Map**



Image 2 Mind Map

## **3.2 Tools and resource to develop the Mind Map**

Git-Mind was used as a tool for developing a mind map. It's a free online mind map builder for brainstorming, creation and other innovative activities.

**Resources**

* Git-Mind
* Related online Mind Map guideline tools
* Sentimental Analysis Research Papers
* Websites and YouTube videos

## **3.3 Literature Review**

### **3.3.1 Lexicon Based Sentiment Analysis**

Some of the most essential approaches to sentiment analysis is application of lexicon. When the semantic orientation of words or phrases occurs in a text, this approach involves in calculating the sentiments from it. In this approach, a dictionary is required in which can when termed has a positive and negative meaning. A combining feature, including sum or average, is implemented on the way making the final prediction of the overall message feeling. (Jurek et al., 2015)

The lexicon-based strategies ' simple steps are described as following:

1. Every document is pre-examined (i.e. eliminate loud variables, HTML tags,).

2. Instantiate the value of the complete text feeling: s ← 0.

3. Initiate tokenization; test if it is available in a vocabulary of feelings to each symbol:

(a) When the token in the vocabulary is existing,

i. s ← s + w; when positive

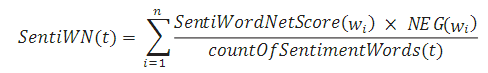
ii. s ← s – w; when negative

4. Overview of the overall text experience score's ',

(a) Text is considered as positive when s > threshold.

(b) Text is considered as negative when s < threshold

SentiWN technique is used in this method where the positive PosS and negative polarity scores of NegS are loosely assigned to synsets. That synset consists of different terms which, in a particular context, have the same meaning. Terms are ranked in this sense based on their prominence. Synset data is stored in the archive of SentiWordNet. Now the score i.e. + 1 and -1 is applied to the synsets with the aid of formula:



If a result is positive, it has a positive meaning for the subsequent word. Otherwise, the sense of it is negative. (Vu, 2017)

The program used both the binary and multi-class classification approach. In the binary classification, classification of documents as intrinsic and extrinsic, while in the multiclass classification, texts have been graded as intrinsic, extrinsic and neutral. Here, binary grouping and division in multiple classes achieved 92% and 87% accuracy respectively. (Kundi et al., 2014)

### **3.3.2 Sentiment Analysis Using Naïve Bayes Classifier**

In this approach, a document is presumed that the collection of words and the probability of a word in the document is irrespective of its position or the inclusion of another word in the document. Hence, extracting the Bayes law's definition of Support Vector Machine (SVM) (NB); p(c/d)= 𝑝(𝑐)𝑝(𝑑/𝑐) 𝑝(𝑑)

Where P(d) plays no role in selecting ‘c’. The text classification centered on Support Vector Machine (SVM) tends to play strongly even though the implicit principle of independence may not involve real-world scenarios. (Wawre & N Deshmukh, 2016)

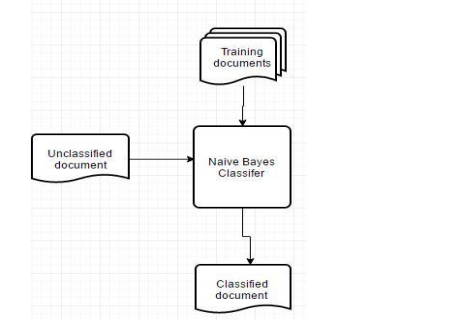
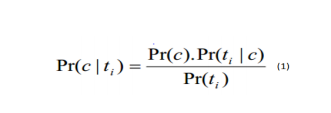


Image 3 Model of document classification with the original Naïve Bayes classifier

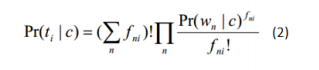
**(Le et al., 2019)**

As presumptuous as naïve bayes is for a specific amount of time, naïve bayes will be classified amongst the most basic of predictive classifiers. Such presumption vindicates every single word to be transparent, scattered and essentially irrelevant. ‘C’ is regarded as a collection of categories (negative and positive) in this algorithm and fresh documentation of ‘ti’ with the necessity to be categorized; the likelihood that ‘ti’ belongs to ‘ci’ class is:



As **‘ci ∈ C’**;

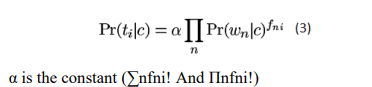
For the Real class ‘C’ document, Pr(c) can be determined segregated through the Real documents of all grades. Whilst, calculating the highest likelihood of Pr(c|ti), the measurement of Pr(ti) may be skipped because it is not necessary to equate Probability Pr(c|ti) by:



Where,

'fni' is the frequency of the word ‘n’ on the test data ‘ti’

Pr(wn) is the probability of the word ‘n’ given class ‘c’ due to the fact that the term ∑nfni! And Пnfni! And probability Pr(ti|C) can be calculated with no after outcome adjustment due to class ‘c’ independence.



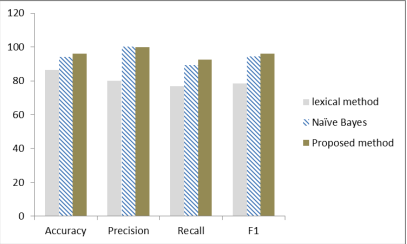


Image 4 Evaluation of Accuracy

**(Le et al., 2019)**

The text has more than 10 terms for practical uses, although may amount up to millions, leading in a long vector. One phrase and 1 of 1000 pages can be shown via vectors of 100,000 or 1 million dimensions. Accuracy, recall, precision, and F-measurement are the most popular methods to test the effects of the machine learning experiment. Envisioning in depth is the degree to which Real Positive cases are anticipated correctly and positive, while consistency means the extent to which estimated positive cases are actually ' True/Real Positives ' in one's system. (Le et al., 2019)

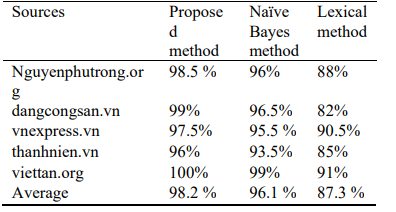


Image 5 The accuracy of 3 methods that applying to similar text table.

**(Le et al., 2019)**

### **3.3.3 Sentiment Analysis of Tweets using Support Vector Machine**

Support Vector Machine (SVM) works to find out the best hyperplane which separates the data points of two different classes. As a sentiment tool is to be created, these classes are simply called ‘positive’ and ‘negative’. Both sides of the plane represent a different class. The plane can be seen as a decision boundary for any new point as the new point can be easily defined on the basis of which side of the plain it lies. (NEKST-ONLINE, 2014)

SVM is used for tasks of text classification such as assignment of categories, identification of spam and examination of feelings.It is also widely used for challenges ofimage recognition, performing particularly well in aspect based recognition and classification based on color. (Noel Bambrick, 2019)

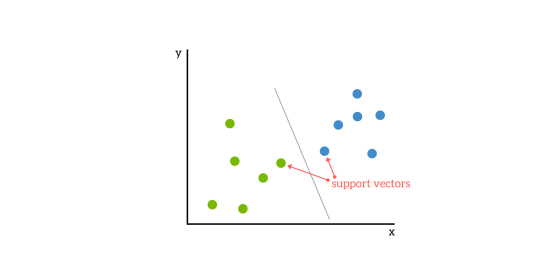


Image 6 SVM

Based on confusion matrix, four effective measures are used in Support Vector Machine. They are: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN).

Precision = TP/ (TP + FP)

Recall = TP/(TP + FN)

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

(Rani & Singh, 2017)

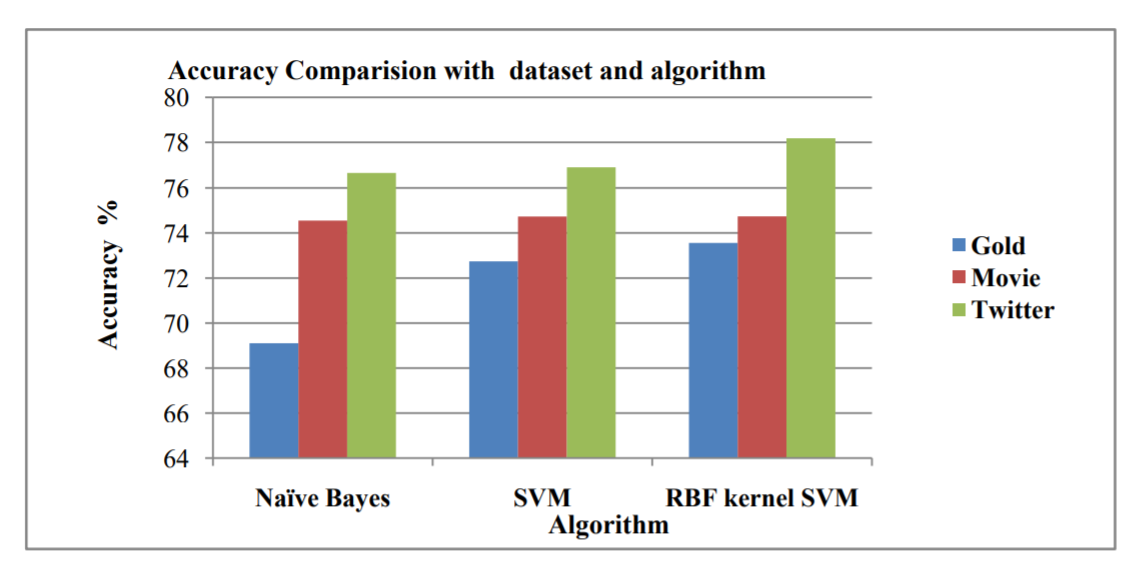


Image 7 Accuracy comparison

**(Rani & Singh, 2017)**

Precision and recall evaluate an effective measure. This system has gained around 80% accuracy. (B. Vaghela & M. Jadav, 2016)

## **3.4 Resources and techniques to write literature review**

**Resources**

Different journal papers were taken as resources to write review of literature about sentiment analysis. All the papers had implemented various algorithms for an analysis of sentiments. Three separate articles were thoroughly researched and a thorough study was conducted, and the following literature review was eventually completed with the instructor recommendations and reading several samples of literature reviews from reputable sources.

**Techniques**

* Search for relevant literature
* Evaluate sources
* Outline the structure
* Writing literature
* Introduction
* Body
* Summarize, Analyze, Critical evaluate
* Conclusion
* Analysis of the findings

## **3.5 Reason for not selecting specific resources**

While searching research papers, different journal papers were appeared promising that was about sentiment analysis. Those papers had in-depth research and analysis on different techniques for sentiment analysis. Those papers had implemented different algorithms and achieve around 85% accuracy. Those papers were based on rating of the product, visualization of the image, automated system which were not suitable for my analysis. My topic is based in text from which sentiment is extracted using machine learning technique.

## **3.6 Chosen research method**

**Content analysis:**

Content analysis is a method of research that identifies patterns from the content. Then various journals, websites, and articles were investigated and all the papers collected were read. The papers conducted a thorough research involving the implementation of various techniques and analysis. A complete analysis was carried out in all the papers collected, and was reflected in the section on the review of report literature.

Information for the content review is obtained from different sources such as:

* Books
* Websites
* Articles
* Journal papers

**3.7 Feasibility, challenges and Recommendation of content analysis**

Content investigation is an inexorably valuable strategy, especially as subjective methodologies proceed to enlarge and incorporate a scope of substance designs for example test. It tends to be a plausible decision for specialist as in this strategy information can be produce either by essential or optional technique. However, with this technique doesn't give measurable portrayal. It just gives look into information from points of view as it were. It absolutely depends on scientist which may impact predisposition and make misdirecting ends. This technique assists with finding the knowledge of a specific theme as it depends on realities rather than feelings. Likewise, it is adaptable in view of that prepared analyst are allowed to catch up on any answer they wish to create more profundity and unpredictability to the information being gathered. Along these lines, this strategy can be better for the individuals who need to examine any information inside and out and reach inference all alone as this technique assists with discovering realities and knowledge of any exploration point.

# **4. Reflection of Research Project**

The overall learning of how natural language processing works has been learnt throughout the research process. It was that machine learning project NLP is one of the hot topics. Significant volumes of data were in the form of a document on the web network. Creating such text is one of the demanding tasks for extracting sentiments. From this study it was learned that different techniques were applied with respective accuracy for the examination of sentiments. After deep work for sentiment analysis, it was discovered that measures such as preprocessing, function, abstraction and classification take place to introduce the particular technique of Natural Language Processing (NLP). Although different NLP technique has already been implemented for analysis of sentiments, the accuracy is still relatively low.

This research has helped me much in the research field. Now I am well known for specific topics, the valid sources. I also know the sentiment analysis for a text project very well. I have done a detailed work on NLP topics and I have a strong knowledge of pre-processing text, visualizing data and applying numerous algorithms for sentiment analysis. I have already introduced numerous algorithms for sentiment analysis in Neural Network, such as SVM, Naïve Bayes and Support Vector Machine (SVM) algorithm, so it may be a little challenging for me to implement Neural Network algorithm for sentiment analysis.

Still, the implemented model does not accurately predict the words grammatically error text for spelling errors. There are also plenty of terminology to know to develop this style of vocabulary. This model is also limited to binary classification, there are also other features like happy, sad, angry, neutral that is considered multi classification. So, for sentiment analysis, these pieces are very important that I really need to know and apply in the future.

If I had another chance than I would love to apply the same dataset as the Support Vector Machine (SVM) algorithm in Neural Network techniques because it gives higher. In addition, I will select a dataset comprising more than two functions. And will incorporate Impartial Network for purposes of multi-classification. Still, these existing sentiment analysis models have a huge scope for improvement. It is therefore necessary to do a lot of work in this field in the future to identify spam blogs, consider spelling errors and other challenges. Since Neural network can deal with the problem of big data. So, that Neural network can be the best technique for sentiment analysis among different technique of Natural Language Processing (NLP).

# **References**

B. Vaghela, V. & M. Jadav, B., 2016. Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis. *International Journal of Computer Applications,* 146(13), pp. 975-8887.

IGI Global, 2019. *IGI Global.* [Online]   
Available at: https://www.igi-global.com/dictionary/using-the-flipped-classroom-to-improve-knowledge-creation-of-masters-level-students-in-engineering/21327  
[Accessed 1 6 2020].

Jurek, A., Mulvenna, M. & Bi, Y., 2015. Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics,* 4(1), p. 9.

Kundi, F. M., Khan, A., Ahmad, S. & Asghar, M. Z., 2014. Lexicon-based sentiment analysis in the social web. *Journal of Basic and Applied Scientific Research,* 4(6), pp. 48-238.

Le, C.-C.et al., 2019. Text Classification: Naïve Bayes Classifier with. *IAENG International Journal of Computer Science,* 46(2), pp. 141-148.

NEKST-ONLINE, 2014. *NEKST-ONLINE.* [Online]   
Available at: https://nekst-online.nl/sentiment-analysis-using-support-vector-machines/  
[Accessed 23 December 2019].

Noel Bambrick, A., 2019. *KDnuggets.* [Online]   
Available at: https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html  
[Accessed 25 December 2019].

Rani, S. & Singh, J., 2017. SENTIMENT ANALYSIS OF TWEETS USING SUPPORT VECTOR MACHINE. *International Journal of Computer Science and Mobile Applications,* 5(10), pp. 83-91.

Susan M. Mudambi, D. S., 2010. *WHAT MAKES A HELPFUL ONLINE REVIEW? A STUDY OF CUSTOMER REVIEWS ON AMAZON.COM,* USA: MIS Quarterly.

Vu, L., 2017. *A lexicon-based method for Sentiment Analysis using social network data.* [Online]   
Available at: https://www.researchgate.net/  
[Accessed 29 11 2019].

Wawre, S. V. & N Deshmukh, ,. S., 2016. Sentiment Classification using Machine Learning. *International Journal of Science and Research (IJSR),* 5(4), pp. 819-821.