International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 04 | April - 2021 ISSN: 2582-3930

Sentiment Analysis Project in R

Aditi Hemant Patil[1], Arya Ramesh Patil[1], Sayali Satish Sangare[1], Sachi Pratosh Singh[1]

[1]*Electronics and Telecommunication Engineering, Pillai College of Engineering, Panvel, Maharashtra, India*

Abstract - Sentiment Analysis or text analysis is all about finding polarity (positive or negative) in narrative of script. The scripts or text can be as short as a sentence and as large as a paragraph or even a book for that matter. With growing age of technology, it is evident that data is being generated at almost every click and is studied to improve the quality of certain product or service. An extensive research work is being carried out in these areas by using different technologies. The two main methodologies used for opinion mining are lexicon-based approach and machine learning approach. A hybrid approach is a combination of both lexicon-based and machine learning approach for the optimum results. In this paper, sentiment analysis is carried out on janeaustenr dataset using lexicon-based approach in R language. The tidytext package is used which consists of all three lexicons used to retrieve sentiments from the dataset. The sentiment polarity would be visualized using wordcloud.

***Key Words*:** Sentiment analysis, lexicons, reviews, extraction, polarity, visualization**.**

1. INTRODUCTION

Sentiment Analysis is a process of extracting opinions that have different polarities. By polarities, we mean positive, negative or neutral. It is also known as opinion mining and polarity detection. With the help of sentiment analysis, we can find out the nature of opinion that is reflected in documents, websites, social media feed, etc. Sentiment Analysis is a type of classification where the data is classified into different classes. These classes can be binary in nature (positive or negative) or, they can have multiple classes (happy, sad, angry, etc.).

The Fig 1 shows the positive and negative sentiments. Since customers express their thoughts and feelings more openly than ever before, sentiment analysis is becoming an essential tool to monitor and understand that sentiment. Automatically analyzing customer feedback, such as opinions in survey responses and social media conversations, allows brands to learn what makes customers happy or frustrated, so that they can tailor products and services to meet their customer's needs.

Fig -1: Classification of sentiments

The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organizations across the world. Shifts in sentiment on social media have been shown to correlate with shifts in the stock market.

The Obama administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages ahead of 2012 presidential election. Being able to quickly see the sentiment behind everything from forum posts to news articles means being better able to strategize and plan for the future.

Traditionally, individuals usually ask for opinions from friends and family members, while businesses rely on surveys, focus groups, opinion polls and consultants. In the modern age of Big Data, when millions of consumer reviews and discussions flood the internet every day, while individuals feel overwhelmed with information, it is as well impossible for businesses to keep that up manually. Thus, there is a clear need of computational methods for automatically analyzing sentiment using unstructured text from social media to aid people on information indigestion.

Every sentiment analysis project has a basic framework of collecting data, cleaning it, analyzing it and then visualizing it. Sentiment Analysis approach is classified in the following manner:

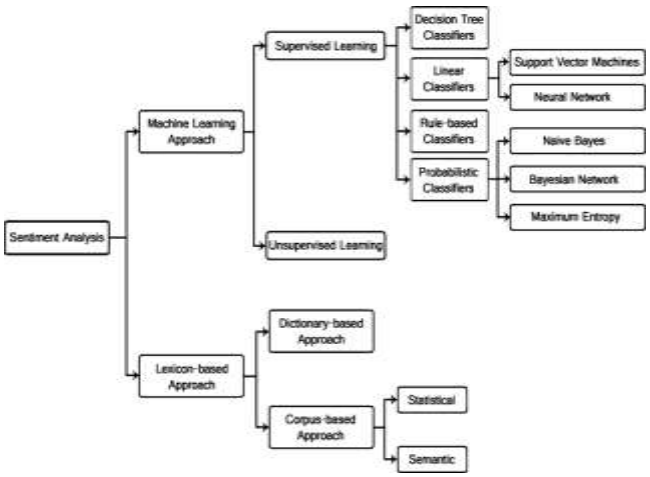

Fig -2: Classification of domain techniques

The present study has adopted lexicon-based approach which makes use of lexicons ("Afinn", "bing", "nrc") to analyze the sentiments of the books written by Jane Austen. These lexicons are based on unigrams that is on single words. All these lexicons contain many English words and each word is pre-tagged with some sentiment values.

Section 2 gives the methodology that has been adopted for the analysis of six novels written by Jane Austen. Section 3 presents the application of experimental implementation of the study. Section 4 describes the result of the main framework and the application. Section 5 concludes the study and Section 6 gives the future scope of the present study.

2. METHODOLOGY

To be able to apply the analytical tools to practical applications, the foundational understanding of sentiment analysis is of utmost importance. The aim is to analyze the emotional quotient of the books written by Jane Austen. In our approach, we are making use of 'janeaustenR' package in R. This package consists of the six novels (namely: "Pride and Prejudice", "Sense and Sensibility", "Mansfield Park", "Emma", "Northanger Abbey", "Persuasion") written by Jane Austen. The proposed methodology is illustrated in form of flowchart below.

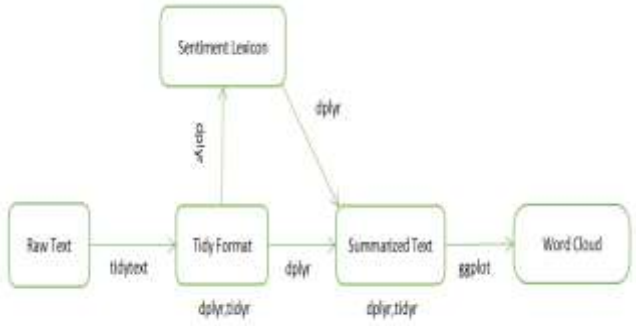

Fig -3: Proposed system architecture

The methodology consists of five steps:

1. Cleaning of data

The books are in their original format which is raw in the context of the project. The text written in the books need to be converted into structured format. Cleaning involves removal of stop words (Antijoin). Stop words consists of articles, nouns, pronouns and conjunctions. This process of converting unstructured data to structured format is called tidying.

2. Tokenization

Tokenization is the process of splitting of text into words with each word per row. Tokenization even strips off the punctuation marks and converts the text into lowercase.
For example: The movie is great!
Output: movie
great

3. Applying Sentiment Lexicons

Lexicons are very similar to dictionaries; however, it consists of meaningful units of words pertaining to certain application like in this case of Natural Language Processing. These lexicons consist of sentiment words, its synonyms and antonyms. Applying these lexicons to the tidy format data, we will be able to gauge the emotional quotient behind the text.

4. Using innerjoin

By definition, it returns all rows from x where there are matching values in y and all columns from x and y. If there are multiple matches between x and y, all combination of the matches is returned. The output obtained after analyzing the sentiments is then summarized using functions from dplyr and tidyr packages.

5. Visualization – Wordcloud

Tidy text mining approach works well with ggplot2 that is the output can be visualized using plots. However, we make use of Wordcloud for visualization. It displays the words in different sizes depending upon the frequency of each word occurring in the text and divides the positive and negative words in upper and lower half of the window.

3. EXPERIMENTAL IMPLEMENTATION

The present research is carried out in R language. R is a functional programming language developed by Hadley Wickham. R is mainly used for statistical inference, data analysis, data visualization and many more. It offers multiple packages and has a huge community.

Here, we made use of a referenced dataset from a repository which consists of 1000 amazon reviews. Alternatively, the reviews can be scrapped using a selector gadget and then creating a dataset in the desired file such as .txt file or a .csv file. Methodology adopted is identical as the main research of Jane Austen novels. Elaboratively, procedure starts

by importing the dataset, tokenizing it, cleaning the imported dataset, and applying the three lexicons ("Afinn", "bing", "nrc") and analyzing the opinions of user reviews.

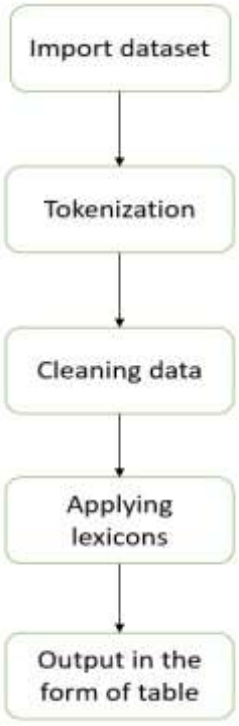

Fig -4: Amazon reviews methodology

R consists of numerous libraries, packages and functions for natural language processing. Here, we present some of them which have been used to carry out research around opinion mining of Jane Austen novels and the amazon reviews.

Table -1: Table of libraries and packages

Libraries and Packages	Usage
library(janeaustenr)	Provides the six novels written by Jane Austen.
library(tidytext)	Provides conversion of text to and from tidy formats.
library(stringr)	Provides functions to work with strings.
library(dplyr)	Provides ease for data manipulation.
library(ggplot2)	It is dedicated for data visualization.
library(wordcloud)	Helps to analyze text and visualize keywords.
library(tm)	Provides functions for text mining.

Table -2: Table of some of the functions

Functions	Usage
unnest_tokens()	Split a column into tokens with each token per row.
mutate()	Adds new variables and preserves the existing ones.
group_by()	Groups by one or more variables.
ungroup()	Removes grouping.
anti_join()	Return all row from x where there are not matching values in y.
inner_join()	Joins tables.
filter()	Used to subset dataframe.

4. RESULT

This section presents the result of comprehensive study of the Jane Austen novels and its experimental implementation on amazon reviews. The study is adopted using lexicon-based approach. Describing the lexicons namely: afinn lexicon ranges the words in the range of -5 to 5 where negative range indicates the negative words and the positive range indicates positive words; bing lexicon gives binary output that is, it categorizes the words into positive and negative sentiment; nrc lexicon is an elaborative lexicon since it presents the words into various sentiments such as happy, anger, joy, surprise and many more. It has been observed out of all the three lexicons, bing seems to be more accurate and reliable. The final result is visualized using wordcloud.

Charts

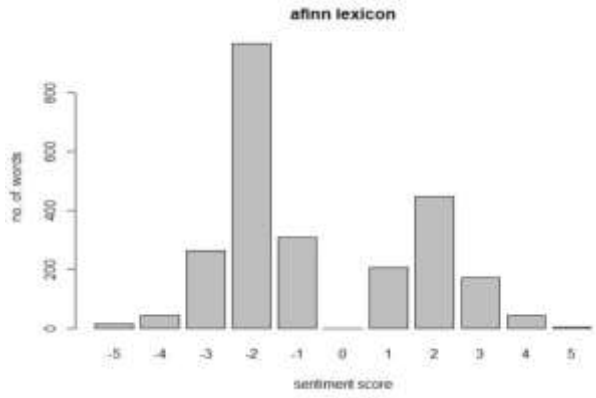

Chart -1: Sentiment score of afinn lexicon

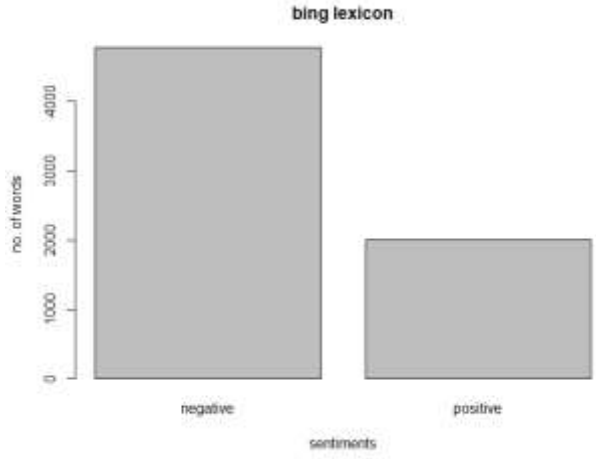

Chart -2: Sentiment presentation of bing lexicon

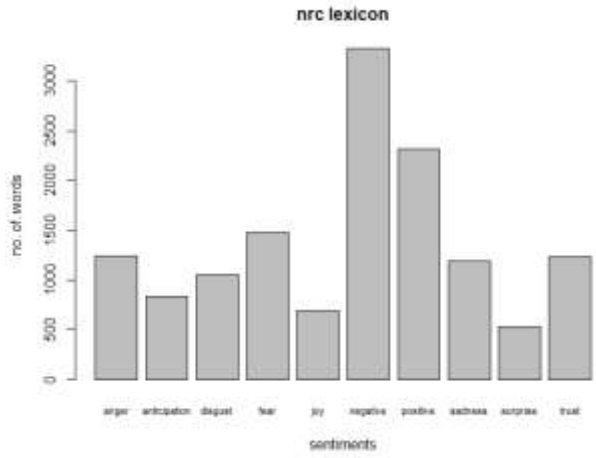

Chart -3: Sentiment presentation of nrc lexicon

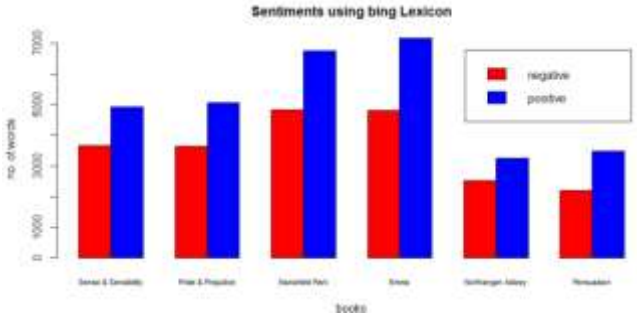

Chart -4: Bing lexicon over all six novels

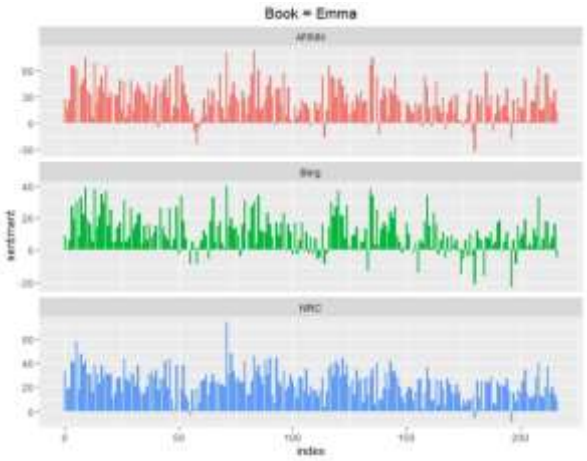

Chart -5: Sentiment narrative of book Emma

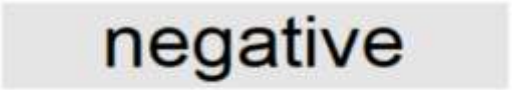

Fig -5: Wordcloud

```
> table(amazon_sent$bing_judgement,amazon_sent$actual_sentiment)

          negative positive
  negative      251       14
  neutral       150       65
  positive       99      421
> table(amazon_sent$nrc_judgement,amazon_sent$actual_sentiment)

          negative positive
  negative      183       53
  neutral       217      229
  positive      100      218
> table(amazon_sent$afinn_judgement,amazon_sent$actual_sentiment)

          negative positive
  negative      212       18
  neutral       187      105
  positive      101      377
```

Fig -6: Tabled output of Amazon Reviews

5. CONCLUSION

In this paper, the study of sentiments existing in the janeaustenr dataset is presented using sentiment analysis. The sentiment analysis of the six novels of the Jane Austen has been performed using tidytext package in R language. The study is carried out using lexicon-based approach. The sentiment lexicons such as afinn, bing, nrc are described and implemented. All the three lexicons have different sentiment intensities and methods of giving value to the words. The bing lexicon gives the result in positive and negative category; afinn lexicon gives score-based values to the words. Similarly, nrc lexicon gives result in different categories of sentiments like happiness, joy, anger, fear, anticipation etc. It is evident that the choice of lexicons is subjective to the user and should be implemented optimally. The sentiment analysis approach is proposed with innerjoin modification. Innerjoin is used to match every token with the lexicons to access the sentiments in the dataset.

6. FUTURE SCOPE

Sentiment Analysis is a very active area of research and development in the field of Data Science. Natural Language Processing has been and will be an exquisite field of opinion mining. Sentiment Analysis can be carried out using different techniques including machine learning and updated package in R. The future applications include social media monitoring, businesses, public actions, recommendation system etc.

ACKNOWLEDGEMENT

We would like to extend our appreciation to our Principal Dr. Sandeep Joshi and our H.O.D Dr. Avinash Vaidya for always encouraging and providing the opportunity to implement and demonstrate our project in the best possible way.

We would also like to express our gratitude to our project guides, Prof. Uma K.S. and Prof. Florence Simon for always helping and guiding us throughout the project.

REFRENCES

[1] Priyavrat and Nonita Sharma, Research related to sentiment analysis: - "Sentiment Analysis using tidytext package in R" Contribution: Sentiment Analysis of first seven novels of Harry Potter using tidytext package in R, 2018.

[2] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede, "Lexicon-based methods for Sentiment Analysis", Contribution: Presented word-based method for extracting sentiments from texts by using dictionaries and incorporates intensification and negation, 2011.

[3] Nitika Nigamand, Divakar Yadav, "Lexicon-based approach to Sentiment Analysis of tweets using R language", MTech. Department of CSE, M.M.M. University of Technology Gorakhpur-273010, U.P., India.

[4] I. Hemalatha, G. P Saradhi Varma and A. Govardhan, "Sentiment Analysis Tool using Machine Learning Algorithms", JNT University Kakinada, Kakinada, A.P., Department of Information Technology, S.R.K.R. Engineering College, Bhimavaram, A.P., JNT University, Hyderabad, A.P.,India. 2013.

[5] Thakare Ketan Lalji, Sachin N. Deshmukh, "Twitter Sentiment Analysis Using Hybrid Approach", Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, e-ISSN: 2395 -0056, p-ISSN: 2395-0072, June-2016.

[6] W. Medhat, A. Hassan, and H. Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams Engineering Journal 5, no. 4, pp.1093-1113, 2014.

[7] Rahul Rajput, Arun Kumar Solanki, Review of Sentimental Analysis Methods using Lexicon Based Approach, School of ICT, Gautam Buddha University, India.

[8] Amit Agarwal, Durga Toshniwal, "Application of Lexicon Based Approach in Sentiment Analysis for short Tweets", Computer Science and Engineering Indian Institute of Technology Roorkee Roorkee, India, 2018.

[9] Anna Jurek*, Maurice D. Mulvenna and Yaxin Bi, "Improved lexicon-based sentiment analysis for social media analytics", DOI 10.1186/s13388-015-0024-x, 2015.

[10] Practical Text Analytics, Anandranjan, 2019. (Book)