Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 9
September 2023

Lexicon Based Sentimental Analysis Using R

Electa Alice Jayarani A[1*], Bhagyalakshmi V[2], Reegan J[1]
[1]Department of Electronics and Communication Engineering,
St. Mother Theresa Engineering College
Vagaikulam, THoothkudi, Tamil Nadu, India.
[2]Department of Electronics and Communication Engineering,
KC College of Engineering and Management Studies and Research
Thane, India.

Abstract-In recent years, the use of internet among the people has exponentially increased because of the growth of technology. All share their view of any product, political, or movie on social media. The sentiment of the shared views is analyzed by machine without any human intervention. In this paper, the sentimental analysis of the novel is proposed. Afinn lexicon-based sentimental analysis model is designed and experimented with the R programming language. The results show the polarity of the novel. The same proposed algorithm can be tested against various other novels also.
Keywords: social media, Novel, Sentiment analysis, Lexicon, R programming language

1. Introduction

Sentimental Analysis

The sentimental analysis tool is used to analyze the sentiment of text, such as positive and negative. In recent days, the popularity of social media among people are very high for sharing their opinion. The most commonly used social media are Facebook, Twitter, Instagram, Snapchat, etc. People can share their opinion on any shared media, starting from a political view, movie review, or product review. The sentimental analysis is a machine learning tool. The tool contains a set of train data, and using the train data, it will analyze the polarity of the text whether positive or negative.

Application Of Sentimental Analysis

1. Decision making

In the olden days, customers ask their friends or relatives to find a review about any product. But in recent days, all review is available on social media. Thus, it helps them to identify the best products. Even to choose the school, hospital, or restaurant, people can view the review and decide the best choice. The sentimental analysis also helps to identify the best scheme to invest the money.

2. Feedback for organization

In the olden days, the organization gets written feedback or suggestions about their manufactured products, schemes, etc. The process is long and sometimes the feedback or suggestions doesn't reach the higher officials. Therefore, the development of the organization is withheld. Whereas in recent days, using social media, the feedback and suggestions from the customers can reach everyone inside and outside the company. The polarity of the feedback and suggestion can be analyzed using the sentimental analysis tool.

Levels Of Sentimental Analysis

The different levels of sentimental analysis are discussed below.

1. Document level

At this level, the whole review document is analyzed for the polarity of the review Ashish Katrekar(2014). The polarity can be positive, negative, and neutral. Efficient output can be obtained, if the review is from the same person. For example, consider that customer A brought a Mercedes Benz. The customer is writing the review of the car, " I brought a Mercedes Benz a few days ago. The car looks very stylish and spacious. Although the mileage is less, the other features are cool, and I love it.

2. Sentence level

The sentimental analysis of the sentence level is of two classes; objective and subjective classes Ashish Katrekar (2014). In Objective, the analysis is done using facts, not by emotions or feelings. In Subjective, the analysis is done using the emotions or feelings of the concerned person.

3. Feature level

The document and sentence level of sentimental analysis identifies whether the text is positive,

negative, or neutral Ashish Katrekar(2014). But it doesn't give the likes or dislikes of the customer. Therefore, the intermediate step is needed to extract the targeted opinion in the text. The feature level analyzes the extracted target for polarity.

Sentimental Analysis Classifiers

1. Naïve Bayes classifier

Naïve Bayes classifiers are grouping of algorithms based on Bayes theorem. Bayes theorem states that the conditional probability of an event A, given the occurrence of another event B, is equal to the product of the likelihood of B, given A and the probability of A. It is given in equation (1).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

Here, P(A) = Probability of A

P(B) = Probability of B.

P(A/B) = Probability of A occurring given evidence B has already occurred.

P(B/A) = Probability of B occurring given evidence A has already occurred.

The basic condition for the Naïve Bayes classifier is the events or features are independent and equal contribution to the outcome. If the condition is satisfied, then Bayes theorem is applied to the prescribed dataset.

2. Support Vector Machine

Support Vector Machine (SVM) classifier is used for classification and regression. It is widely used for classification in machine learning. The aim of SVM is to identify the best boundary decision. The best boundary region is called hyperplane.

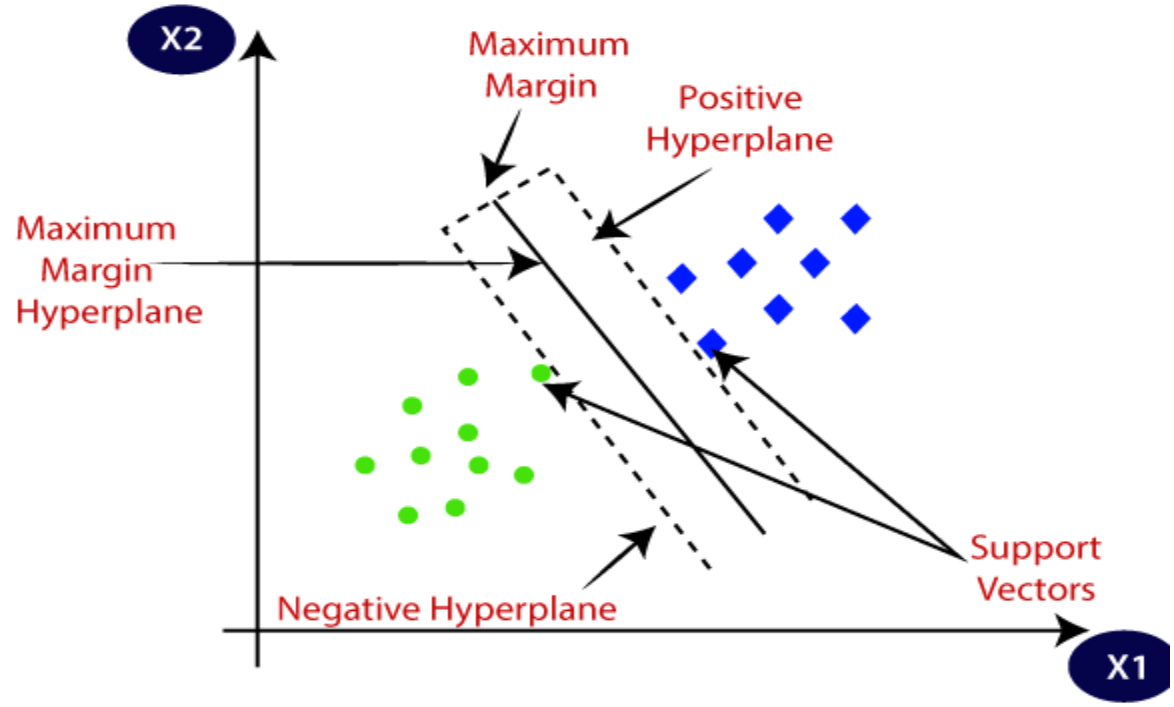

Figure 1. SVM classifier

The figure 1 shows the maximum margin hyperplane and the classified positive and negative hyperplane. For example, in the forest, if the camera captures a tiger but has a feature of a lion also. The animal can be classified using SVM. To do that, train our model using different tiger and lion images. It will extract the different features of tiger and lion. Now using the SVM, the best boundary decision called hyperplane is derived. Using the If the margin extracted from the dataset is not a straight line, then it is called Non-Linear support vector machine classification.

hyperplane, the newly captured image is analyzed and can be identified as a tiger. The Support vector machine process is shown in figure 2. The SVM is of two types. They are Linear and non-linear SVM.

Linear SVM

If the margin extracted from the dataset is a straight line, then it is called Linear support vector machine classification.

Non–linear SVM

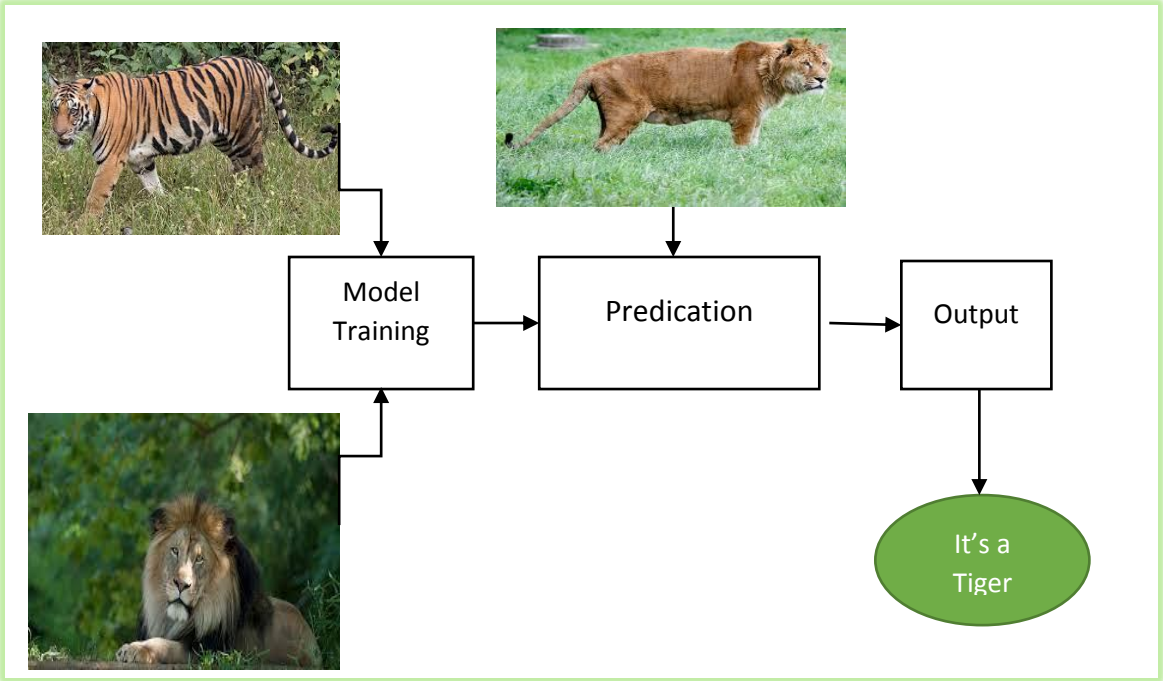

Figure 2. SVM process

2. Literature Review

The author Hassan Raza et al. (2019) proposes an algorithm to sentiment analysis of the citation sentence in research article. The proposed method is experimented on 8736 citation sentences. The algorithm uses six different machine learning algorithm including Naïve-Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbor (KNN) and Random Forest (RF).

The author Kawade et al. (2017) proposes an algorithm for sentimental analysis of the tweets about terrorist attack, The tweets considered are mainly focussed on Uri attack. Nearly 5000 tweets are processed and the analysis shows 94.3% is disturbed by the attack.

The author Singh J et al. (2017) proposes an algorithm to sentiment analysis of product review and movie review. Nowadays peoples use a social media for the review of any product. The paper focuses on the product review from amazon and the movie review from IMDB.

The author Hadley Wickham et al. (2017) introduces the R programming language. The author demonstrates how the raw data can be analysed using the R and R studio. The book also explains various library packages such as tidyverse, tidytext, dplyr, ggraph, ggplot, etc. Many data analyst beginners are using R Studio for their research work.

3. Sentimental Analysis Of Novel

The sentimental analysis of the novel is performed in this paper. The analysis is performed in R programming languages. R Studio is open source and is widely used by data analysts. Figure 3 shows the process of the analysis. Once the novel of interest is loaded, the whole document is tokenized. Before the word count and cloud word, the stop words were removed. The snowball stop words are considered in this paper. Once the stop words are removed, it is ready for analysis. In this paper, lexicon-based sentimental analysis is performed.

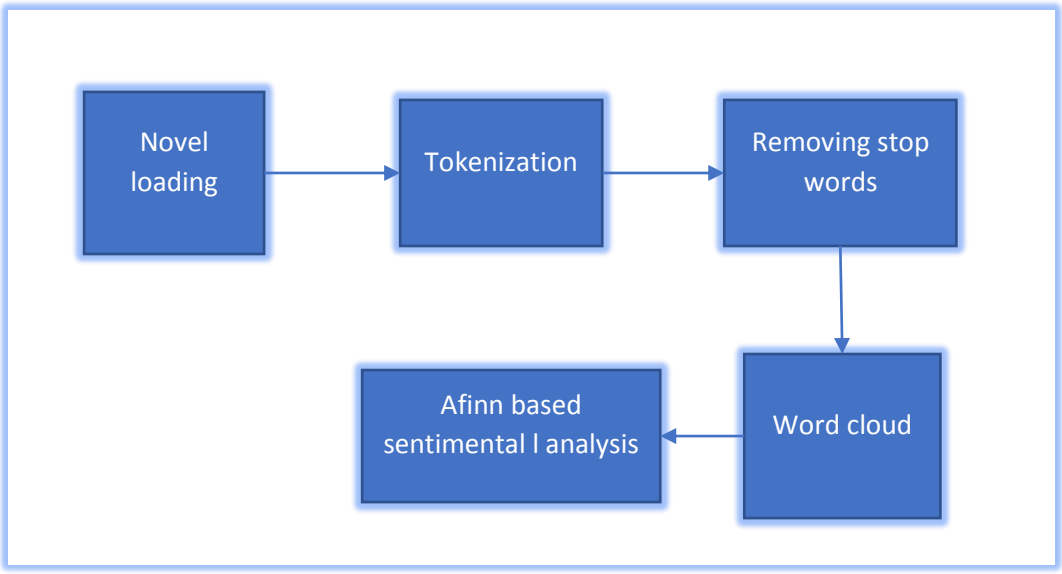

Figure 3. Sentimental analysis process

The steps to analyse the novel is

1. Importing the library packages

The packages used are tidyverse and tidy text. Tidyverse package was invented by R studio chief scientist, Hadly Wickham. Tidyverse is used for data modeling, wrangling, and visualization tasks. As the output is better efficient and fast and well-documented overflow, this package is widely used by many data scientists. To support the datasheet to convert from text to tidyformat, the tidy test package is used.

2. Load the novel for the analysis.

In this step, load the novel of interested to analysis. The input file can be of any format txt, csv, SPSS, XLSX, SAS, Stata, etc. Identify the line number by extracting the text from the input novel and outputs separate line as a character string.

3. Tokenization

Tokenization is defined as allocating the tokens to the data which is having more value. The main purpose of tokenization is to provide security to the data. Only the allotted token is released for the further process. By doing this, the security of the data is maintained. The tokenization is easily understood by remembering how the casino token is used for playing in Casinos. The casino token is exchanged for the currency, and the players use their token for playing and betting. The token has to be exchanged at the end to get the money. Similar way, the token provided to each data in the novel is called tokenization. The reverse process is called detokenization.

4. Remove the stop words from the text

The most commonly used words in any language are called stop words. The different lists of stop words are determiner (the, a, an), Coordinating conjunction (for, nor, but, or, yet, so), and prepositions (in, under, towards, before). The data scientist commonly uses the published stop words. The published stop words include snowball stop words, Terrier stop words and minimal stop words. In this project, the snowball stop words are used to remove all the commonly used words. It is always better to remove the common words before starting the analysis. For example, if the user is searching, “how to write a journal” in google. If the Google search machine searches for “how”, “to”, “write”, “a”, and “journal”, it will end with a lot of pages. However, if the commonly used words such as “how”, “to”, and “a” are removed and the search is made, it can produce more relevant data. Therefore, it is very much necessary to clean the text before starting the analysis.

5. Count word frequency

Using the command *count*, calculate the word frequency. The graphical representation of the word frequency is done using the command, *ggplot*. The graphical representation of text data is called a word cloud Celine (2015). The data analyst prefers graphical representation rather than a tabular representation. Because, in the word cloud, the most widely used data is

highlighted. Hence, it is easy to understand. It also gives clarity to the text data.

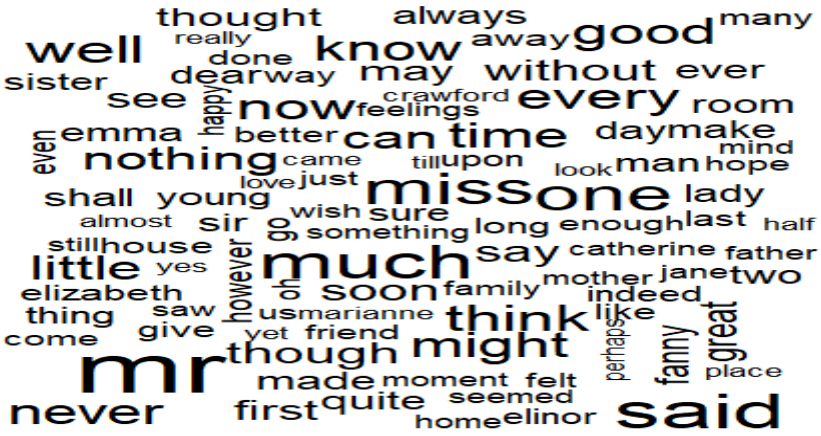

Figure 4. Word cloud

6. Sentimental Analysis

In this paper, the sentimental analysis of the novel "Emma" is performed using R programming. The novel Emma is written by Jane Austen and it is the last novel published by her during her lifetime. It was published in the year Dec 1815. The novel is about the girl Emma woodhouse and describes her marriage and social status. R is the programming language used for statistical computing. The data analyst and text miners widely use R studio for their analysis.

The set of rules followed for the sentimental analysis is called lexicons. The most commonly used lexicons are AFINN, bing, and Loughran. In the AFINN Finn Ärup Nielsen (2011), the text is valued between – 5 and +5. The Negative polarity is indicated by -5 and the positive is indicated by +5. This lexicon is developed by Finn Årup Nielsen. In the Bing lexicon, the text is valued in the binary form Minqing Hu et al. (2004). For the financial document sentiment analysis, the most widely used lexicon is Loughran. This lexicon will value the text in six different forms, they are negative, positive, litigious, uncertain, constraining, and superfluous Loughran, T et al. (2011). In this paper, the AFINN lexicon based sentimental analysis is performed to evaluate the sentiment of the novel. The number of positive and negative words are calculated and by comparing it, the sentiment of the novel is analysed. For better understanding and sharing, the graphical representation of the analysis is processed.

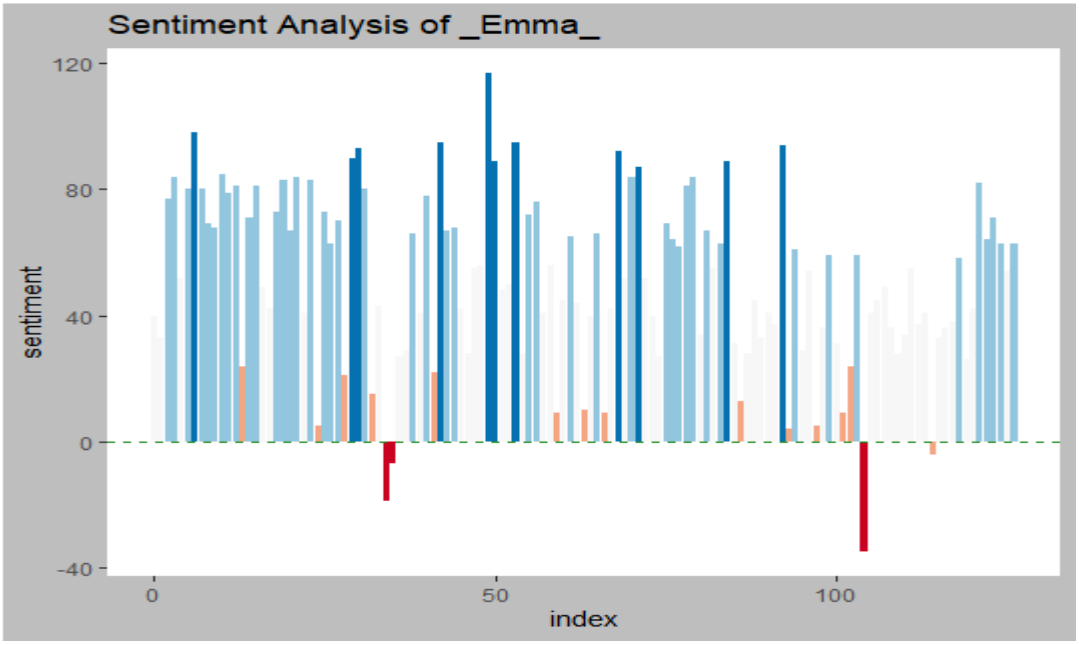

Figure 5. Sentimental analysis of the novel Emma

The figure 5 shows the sentimental analysis of the novel based on Afinn lexicon.

4. Conclusion

The process of the sentimental analysis of the novel using R programming shown in the figure is simulated using the R programming language. The novel considered for analysis is "Emma". The figure shows the analysis of the novel. From the figure, the sentiment of the novel is positive. The blue lines indicate the positivity of the novel, whereas the orange and red indicate negativity in the novel. Thus, it is concluded the novel "Emma" has a polarity of positive. In the future, the algorithm can be extended to other novels, reviews, and tweets.

References

[1] Ashish Katrekar, AVP, Big Data Analytics, An Introduction to Sentiment Analysis, Global logic Inc., 6 pages, October 2014.

[2] Celine Van Den Rul, "How to generate word cloud in R", Towards Data science, October 2015.

[3] Finn Ärup Nielsen (2011), "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs", Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages (2011) 93-98.

[4] Hadley Wickham, Garrett Grolemund, "R for Data Science", O'Reilly Media, Inc., 2017/1/7.

[5] Hassan Raza, M. Faizan, Ahsan Hamza, Ahmed Mushtaq and Naeem Akhtar, "Scientific Text Sentiment Analysis using Machine Learning Techniques" International Journal of Advanced Computer Science and Applications (IJACSA), 10(12), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0101222

[6] Kawade, Dipak & Oza, Kavita. (2017). Sentiment Analysis: Machine Learning Approach. International Journal of Engineering and Technology. 9. 2183-2186. 10.21817/ijet/2017/v9i3/1709030151.

[7] Loughran, T. and McDonald, B. (2011), "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." The Journal of Finance, 66: 35-65.

[8] Singh, J., Singh, G. & Singh, R. Optimization of sentiment analysis using machine learning classifiers. Hum. Cent. Comput. Inf. Sci. 7, 32 (2017). https://doi.org/10.1186/s13673-017-0116-3

[9] Minqing Hu and Bing Liu, "Mining and summarizing customer reviews.", Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004), 2004.