ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 3, March 2018

Building Competitive Advantage in Business by using Opinion Word Mining

Rohini A [1], Pooja M [2], Rakshana R [3], Hemalatha B [4]

U.G. Student, Department of Computer Science and Engineering, Velammal Engineering College, Chennai, India[1]

U.G. Student, Department of Computer Science and Engineering, Velammal Engineering College, Chennai, India[2]

U.G. Student, Department of Computer Science and Engineering, Velammal Engineering College, Chennai, India[3]

Assistant Professor, Department of Computer Science and Engineering, Velammal Engineering College, Chennai, India[4]

ABSTRACT: In this highly digitalized world, businesses are carving their competitive advantage by decision making using data analytics. One of the sources to harness data through product reviews, as it is a clear depiction of the customer's interest in a particular product or service. Opinion Mining techniques are employed to identify the perception of the product or service by the textual analysis of public reviews obtained by computational extraction. In the proposed system, opinion mining is utilized to recognize opinion words from the subjective information shared by customers' experiences of a product. In this work, the algorithm is applied on a public dataset of reviews extracted from the Amazon website to discover customer requirements and preferences. A sentiment analysis is further performed using the Bing lexicon to identify the sentiment of the reviews. These analyses allow businesses to refine their products by adapting to the user demand, thereby improving the competitive advantage of the business.

KEYWORDS: Data Mining**,** Opinion Word Mining, Sentiment Analysis.

I. INTRODUCTION

With the growth of digitalization, the way businesses are thought of and are conducted have become radically different. With the digital transformation of businesses, it has become more and more pertinent to use this data as a source of competitive advantage. Likewise, the notion of competitive advantage for businesses, the leverage a company has over its competitors, has undergone a massive change. Today, data is a source of competitive advantage.

Data Mining is an upcoming field of research which extracts patterns from a given data. It has been used so far in a variety of applications such as climate modelling, or predicting the results of elections. Data Mining is carried out in the following stages: first, the problem is identified. Then the relevant data is gathered. After this, the model is built and evaluated. Only after this the relevant knowledge is obtained.

Opinion Word Mining is an advancement of Data Mining. Here, the Opinion Words are extracted from a given text and the data is further analysed. Opinion Words are the descriptive words in a text. They are identified along with the Opinion Targets which are the focal points which the Opinion Words describe. Sentiment Analysis is generally used to identify the sentiment of the given text.

In this system, the reviews of products hosted on Amazon, extracted from public datasets are mined to produce the most frequently used Opinion Words to obtain insights into the customer perception of the given product.

This organization of the Paper is as follows. Section II contains the Literature Survey. The algorithm description and system architecture diagram are presented in Section III. Section IV contains the results of the system and also presents the screenshots of the same. Finally, Section V proposes the future work and also presents the conclusion.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 3, March 2018

II. LITERATURE SURVEY

This section presents the analysis and discussions on the previous work completed in this area of Opinion Word Mining.

2.1 DOUBLE PROPOGATION

This model proposed in [1] is a semi-supervised model. An existing set of words or opinion lexicons are used to mine the text for other Opinion Words. The mining process takes place after the relationship and dependencies between themselves. Research conducted as in [1] showed that this process demonstrated a higher accuracy and a higher precision. As per the research in [2], it was identified that the possibility of introduction of noise into large databases was high.

2.2 WORD-BASED TRANSLATIONAL MODEL

This unsupervised approach is presented in [3]. In this approach, all the nouns and adjectives in a given sentence are identified. Their associations are further calculated Using factors such as word positions and word co-occurences, the candidate confidence is measured. The words with the highest confidence are considered as the Opinion Targets.

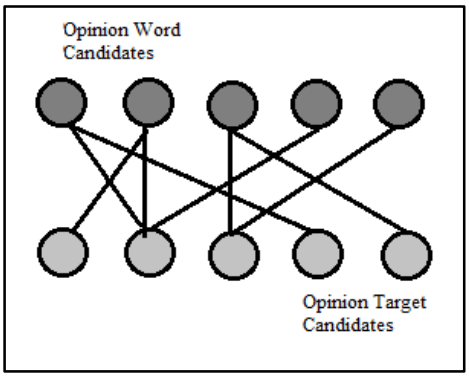


Fig. 1. Opinion Relation Graph

In small sets of data, precision is lost to due to the inability of the algorithm to frame associations due to the sparseness of data. A drastic improvement in performance is seen here as is depicted in [3].

2.3 PARTIALLY SUPERVISED WORD ALIGNMENT MODEL

The algorithm proposed by [3] has shown a high rate of performance, yet there is a high possibility of forming the wrong associations. This model proposed by [4] suggests the use of certain links to train the data. To identify alignment in the sentences, a constrained EM algorithm is used. Using a graph based algorithm, the Opinion Targets can be extracted. Further, a bipartite graph is used to identify opinion target candidates. To identify the Opinion Targets, high confidence has to exist. This can be estimated by using a random walking algorithm. This method is comparatively more precise. The only disadvantage is that the use of semantic relations isn't suitable for the nature of online reviews.

2.4 TWO-STAGE FRAMEWORK

This framework, proposed by [5] performs mining in two stages. A Sentiment Graph Walking Algorithm is employed in the first stage to avoid the identification of false opinion relations. The ones with low confidence are thus discarded. A semi-classifier is used in the next stage to avoid the use of stop words in the given text. This method outperforms

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 3, March 2018

all the other methods in terms of performance[5], but the opinion words aren't limited to Opinion Words. So extracting them will prove to be a difficult task.

2.5 WORD ALIGNMENT MODEL

The paper [6] proposes this model. It suggests that the Opinion Words are often aligned with the Opinion Targets. It doesn't parse text or use near-neighbour rules. It considers word co-occurrence frequencies. Therefore, it is quite effective. Since the use of Supervised or Unsupervised models could produce unsatisfactory results, a Partially Supervised approach is taken. A co-ranking algorithm based on random walks is utilised to estimate candidate confidence and in turn identify the Opinion Words and Targets. This algorithm greatly reduces error and noise.

III. PROPOSED METHODOLOGY

The proposed system utilises a public dataset comprising of reviews from Amazon provided by [7]. This data is prepared and processed. The algorithms are then applied on the processed data. The algorithm is applied and the output obtained is represented in the form of a Data Table. Further data analysis is performed and it is represented in the form of a word cloud. Sentiment Analysis using Bing lexicon is conducted to understand the sentiments of the reviews.

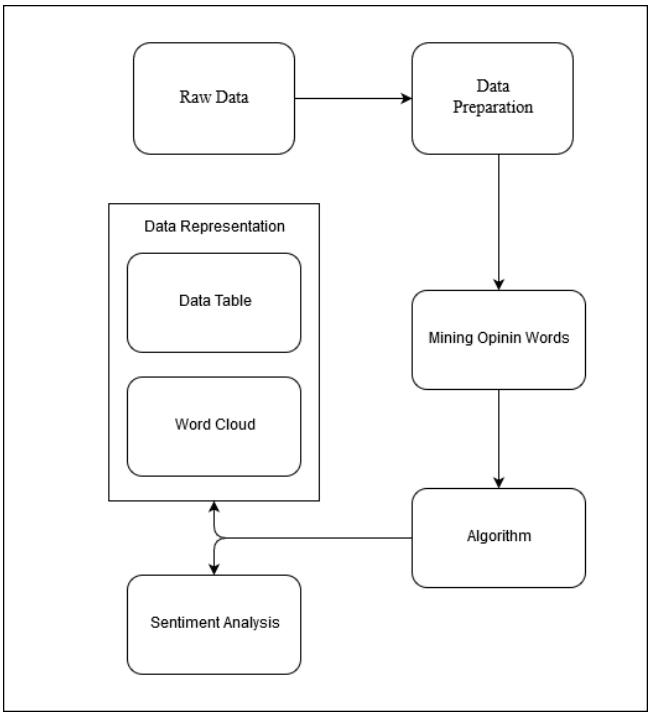


Fig. 2. System Architecture

2.1 DATA PREPARATION

A public dataset comprising of reviews of Office Products listed in Amazon is obtained from [7]. This extracted data is structured and in the JSON format. This JSON data is first read from a '.csv' file in RStudio. Then, they are split into columns using the delimiter ';' .It is ensured that characters excluding the relevant data are deleted. This is possible by using regex or regular expressions. Appropriate and useable column names are given. Two columns from this data are

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 3, March 2018

extracted converted into a 'dataframe' structure for it to be operated on. They are 'ProductID' and 'reviewText' columns. The stop-words, the words that add no value when mined are removed in advance.

2.2 APPLICATION OF ALGORITHM

The processed data is split into one-word tokens. Each token is present in a different row along with the product ID that it represents. Next, the list of adjectives and adverbs are obtained from the package 'RCorpora'. This list of verbs is each joined with the original token list. Finally, all the three joined dataframes are further joined into one 'key' dataframe to obtain the Opinion Words. The opinion words obtained in this key frame are further arranged in the order of increasing frequency, giving more priority to the words that repeatedly appear in one sentence.

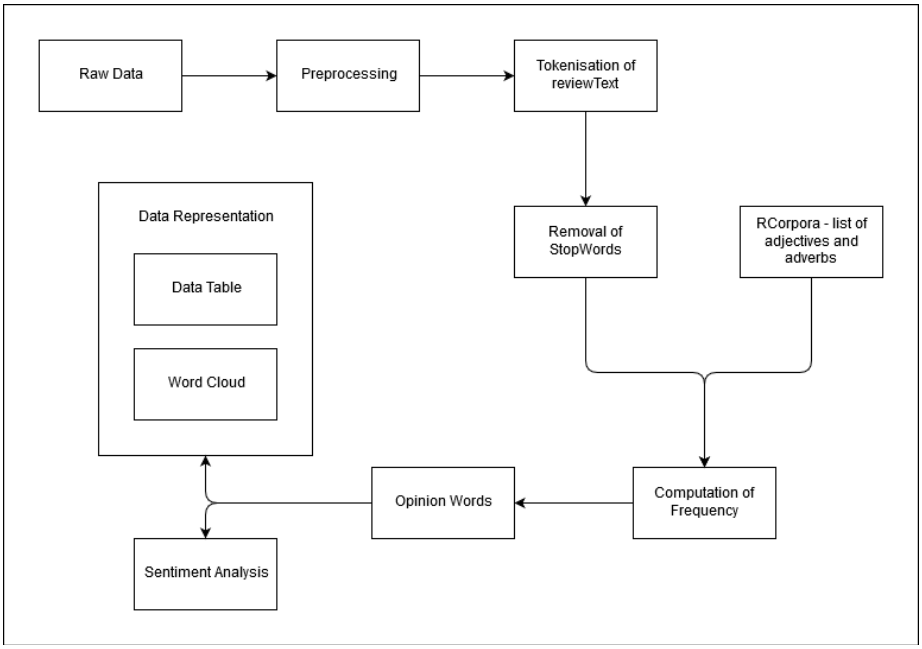


Fig. 3. Detailed Flowchart of the Proposed System

2.2.1 ALGORITHM

The algorithm for the proposed system is presented here:

Let sample be the dataframe with columns ProdID and reviewText
Let x be the required ProdID
tokens <- sample %>%
filter (x in ProdID) %>% //filters ProdID where it is equal to x
tokenizer(split reviewText by 'word') //tokenizes the text
//loads the package 'rcorpora'
library(rcorpora)
//After reading from rcorpora, converts it into a dataframe
adjs <- dataframe(read(list of adjs from rcorpora))
advs <- dataframe(read(list of advs from rcorpora))
//full_join(a,b) completely includes rows from both dataframes 'a' and 'b'
temp <- full_join(adjs, advs)

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 3, March 2018

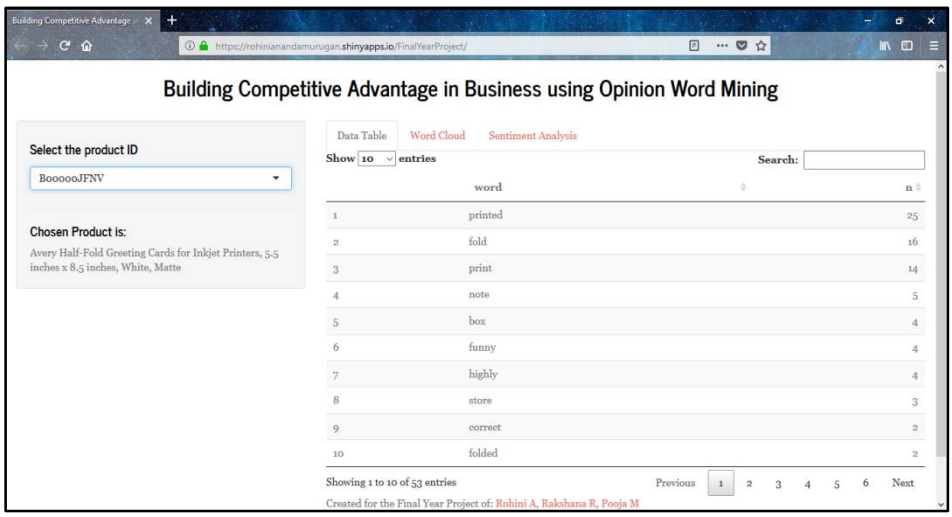
```
key <- full_join(tokens, temp)
//prioritize word co-occurences
Let d be dist between two words
for all words in key
        if d is less
                assign high priority
sort(key based on words) //in ascending order
count(the words) //to obtain frequency
```

2.3 DATA ANALYSIS AND REPRESENTATION

Shiny Web App is used to represent the data obtained from applying the algorithm to processed data. The data obtained is directly visualized in the form of a Data Table. Another representation of data is also obtained. The resulting data from Module 2 is presented in the form of a word cloud. A word cloud is an image obtained from the words used in the review and the size of the words depends on the frequency of the words used. Sentiment Analysis to identify the emotions of the reviewers, expressed via the reviews. Using the lexicon Bing, the sentiments can be classified as positive or negative. So the reviews are measured against these parameters and are plotted in the form of a bar graph.

IV. RESULTS AND DISCUSSIONS

This section presents the output of the various modules in screenshots. The results are also discussed in this section.
Fig. 4(a) represents the Opinion Words arranged in the order of decreasing frequency of appearance in the reviews of the product “B00000JFNV” in the form of a Data Table.
The Fig. 4(b) represents the Word Cloud obtained from the data table for the product “B00000JFNV” seen in Fig. 4(a). As one can infer from the given Word Cloud in Fig. 4(b), the most commonly used words to represent the product “B00006HO37” are “fold”, “printed” and “print”. Since the product is a greeting card, those words are inevitable. So they can be eliminated. If a closer look is taken at the other words, words such as “fancy”, “correct”, “love”, “perfectly”, “trouble” and “miss” may be found among other words. This is a clear depiction of the presence of positive reviews over the negative reviews.



(a)

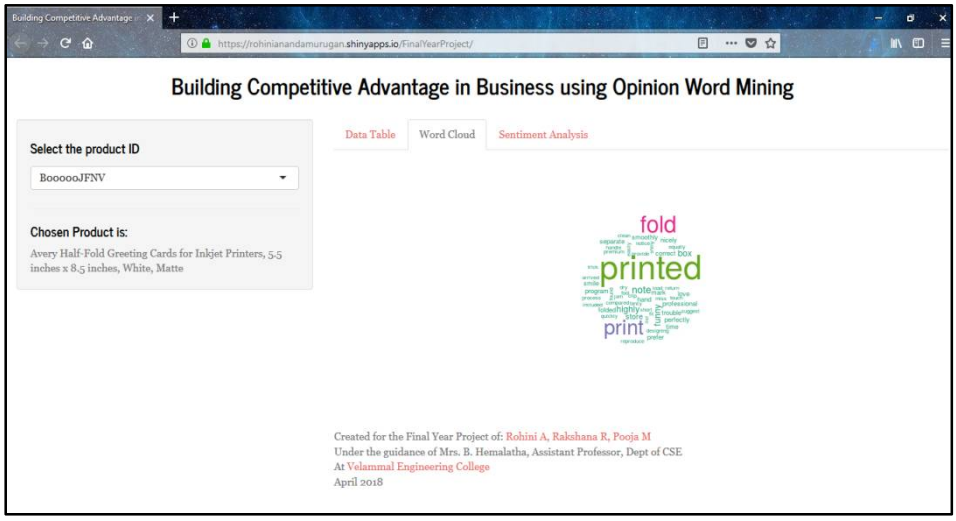
ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

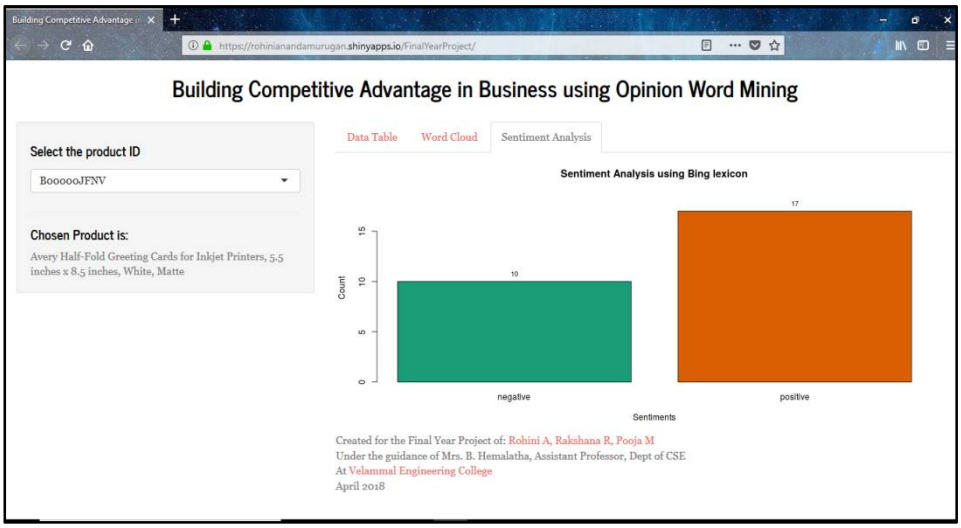
(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 3, March 2018



(b)



(c)

Fig. 4. Results of Performing Opinion Word Mining on product "B00000JFNV" [Avery Half-fold Greeting Cards]. (a) The Data Table (b) Word Cloud (c) Sentiment Analysis

The Word Cloud in Fig. 4(b) paints a picture of the commonly used words. To measure the general emotions represented in the reviews, sentiment analysis is required. As inferred previously, the analysis, represented in Fig. 4(c) also shows a high positive sentiment and compared to the negative sentiments. So it can be concluded that the productive is faring pretty well currently. There, there is a high chance of this product to be purchased by the customers.

V. CONCLUSION

Data provides competitive advantage in this fast-paced, growth-oriented world. Businesses rely on this data to analyze their own processes. The growth of e-commerce platforms like Amazon have revolutionized the way shopping is done. Customers leave reviews for the products they buy online and these reviews provide a window into the minds of the

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 3, March 2018

target customers to identify their perception of the product in question. These reviews were mined using the algorithm mentioned in 2.2.2 and the resultant Opinion Words were represented in the form of a Data Table. The Opinion Words in the Data Table were represented as a Word Cloud for visualisation. Sentiment Analysis using Bing lexicon was undertaken to identify user sentiment of the product. These analyses can be utilized to identify the customer perception of the given product. This valuable data can be used to rectify or improve upon certain aspects in the system. Thus, they help improve the competitive advantage of businesses. Future work can comprise of the extraction of Opinion Targets along with the Opinion Words so that a clear Opinion Relation can be established. This will help the business peruse the reviews in detail.

REFERENCES

[1] Qiu et al., Opinion word expansion and target extraction through double propagation, in Computational Linguistics archive, Volume 37 Issue 1, March 2011, Pages 9-27, MIT Press Cambridge, MA, USA

[2] Zhang et al., Extracting and ranking product features in opinion documents, COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Pages 1462-1470, Association for Computational Linguistics Stroudsburg, PA, USA ©2010

[3] Kang Liu, Liheng Xu, Jun Zhao, Opinion Target Extraction Using Word-Based Translation Model, Published in EMNLP-CoNLL '12 Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Pages 1346-1356, Association for Computational Linguistics Stroudsburg, PA, USA ©2012.

[4] Kang Liu, Liheng Xu, Jun Zhao, Opinion Target Extraction Using Partially-Supervised Word Alignment Model, published in IJCAI '13 Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, Pages 2134-2140, AAAI Press ©2013.

[5] Robert. C. Moore, A discriminative framework for bilingual word alignment, HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Pages 81-88, Association for Computational Linguistics Stroudsburg, PA, USA ©2005

[6] Kang Liu, Liheng Xu, and Jun Zhao, Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model, in IEEE Transactions on Knowledge And Data Engineering, Vol. 27, NO. 3, March 2015.

[7] J. Mcauley, C. Targett, Q. Shi, and A. V. D. Hengel, "Image-Based Recommendations on Styles and Substitutes," Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 15, 2015.