International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

Consumer Behaviour Analysis In Social Network Using Big Data

M.Arumugam[1], Dr.C.Jayanthi[2], N. Magesh kumar[3]

[1]Assistant Professor, Department of Computer Technology-PG, Kongu Engineering College,Perundurai
[2]Assistant Professor, Department of Computer Science, Government Arts College,Karur
[3]Student, Department of Computer Technology-PG, Kongu Engineering College,Perundurai

Abstract. In present, the social network plays a vital role in every person of the society irrespective of age, community, gender, business background etc… At present, 3.4 billion to 3.6 billion people were connected through social networks, and also there will be an increase in the number of around 4.50 billion by 2025. This massive usage of social media platforms paves the way to many E-Commerce companies to promote their growth by targeting the customers according to their needs. But it is not easy for the companies to reach their respective customers as they want to analyze huge data that were generated in a social network of different age groups and backgrounds. To analyze these huge volumes of data, Big Data comes into the role which handles both the structured and unstructured data. An algorithm is needed to implement on these data that analyses and classifies the consumer based on their behavior that satisfies their needs. To provide an optimal solution to this problem two mechanisms are applied in this paper. One is applying Word2vec on data that gives the numerical representation (vectors) of the words and the second is applying K – Means Clustering on those numerical values which form the cluster of similar words. By using those clusters companies can identify which product is more needed or liked by the people. Therefore it is believed that this method provides a reliable solution to businessenterprise.

Keywords: Big Data, Word2vec, K – Means Clustering

1. INTRODUCTION

Today every business depends on the consumer data. Even though it is an offline business, customer reviews or feedback are collected in the form of hard copies, later that can be manually analyzed[1]. So analyzing consumer data plays a vital role in bridging the gap between consumer needs and companies grow. Especially in E-Commerce business analysis of the consumer data determines the company's revenue and exponential growth by predicting and suggesting the right product to the right customer[2]. Consumer data can be collected in various platforms like Facebook, Twitter, Instagram etc…which comes under the category of social networks and another type is collecting through the online shopping websites like Amazon, Flipkart, Snapdeal, etc…When business enterprises collect data only in their respective websites it will not be sufficient to predict the customer needs[3]. On the other hand, when the data are collected through social networks it is more sufficient and efficient to predict the consumer needs because every E-Commerce or other business has its page in social networks that reveals the upcoming products and ideas[4]. So people in the social network will like, share and comment on those products according to their view. For example, on Facebook and Instagram, the product will get likes, comments, and share among the groups and Twitter people will tweet about the product[5]. These likes, comments, shares, and tweets are considered to be the data generated by the consumer for a product that is to be analyzed to determine whether the product will reach the maximum people that satisfy their needs or end up in vain[6]. In this paper, data were collected from Twitter (i.e tweets) using Twitter Stream API and those data were analyzed with the help of Apache Spark framework as it is faster than Hadoop in handling real-timedata[7].

5. LITERATURE SURVEY

This paper **[1]**proposed the concept of distributed implementation of the C4.5 decision tree algorithm with the help of the Hadoop MapReduce Framework. Data visualization was done using D3.js.

In this paper[2], the author revolves around the concept of Big Data Analysis process / analyze a huge volume of data to extract some meaningful information from that data in real-time. Twitter Stream API and Apache Spark are the research instruments used. Top ten words, languages, and number of times a particular "word" used in twits collected were the following scenarios executed for experimental purpose.

The paper**[3]**describes efforts to analyze datasets from social media platforms using community networks with the help of big data. Many Social media platforms provide public APIs that are used to retrieve data. (Eg: Twitter Stream API, Facebook API). Apache Spark is used for the analysis and extraction of information in a very efficient timeframe. Oracle NoSQL database with Avro Schemas was used for this research. Java Universal Network/Graph Framework (JUNG) library in Java or Python's NetworkX library was used to built a network library. Java's Abstract Window Toolkit (AWT) was used for graphvisualization.

In this paper[4], Apache Spark and Shark were used to provide Real-Time Healthcare Analytics. Spark, a real-time analytics software access data directly from HDFS that overcomes data migration problems. Apache Shark4 is a large-scale data warehouse system that allows SQL queries to run up to 10 times faster than Hive (which uses MapReduce).

The aim of this paper [5] is to use Big Data Analysis (BDA) for commercial purpose by using Elevate Performance algorithms .PCA (Principal Component Analysis) and SVM (Support Vector Machine) algorithms were used to extract and classify the consumer opinions.

6. PROPOSED METHODOLOGY

In this proposed system, the data (i.e. tweets) were collected from the Twitter platform by using Twitter Stream API with the help of a python package called tweepy. Then those data were preprocessed (i.e.) cleaning of data because real- world data contains many unwanted information that is needed to be remove. After preprocessing, data were converted into tokenized forms. These tokenized data were converted into a structured Data frame using spark. This data frame was the final dataset that was passed as an input parameter to the Word2vec algorithm and then output data were collected. Those output data were feed as an input to the K – Means Clustering algorithm which forms the cluster of most tweeted product as we took the screen name of Amazon shopping twitterpage.

1.1 WORD2VEC:

Word2vec is a NLP (Natural Language Processing) algorithm. Word2vec uses a neural network model that learns word associations from a large collection of text or document. It gives the numerical representation of words called Vectors with several dimensions.Word2vec consists of models that comprises of input , hidden and output layer and it is made up of two architecture : CBOW and Skip Gram.

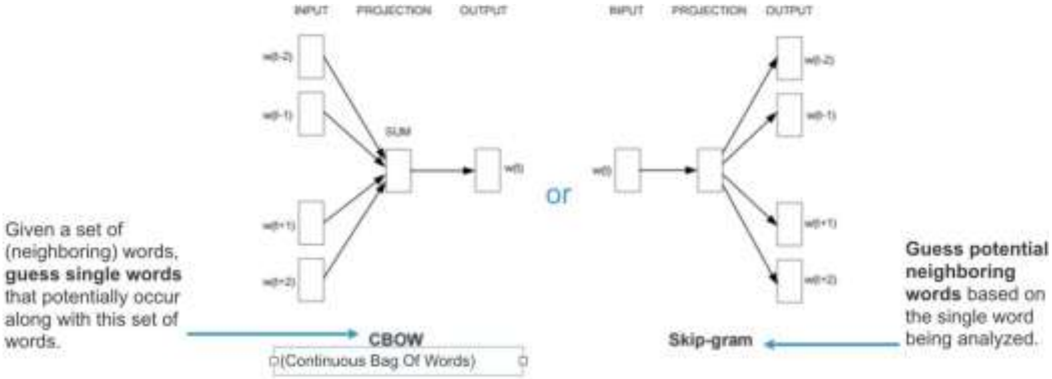
INPUT
PROJECTION
OUTPUT
w(t-2)
w(t-1)
SUM
w(t)
w(t+1)
w(t+2)
or
INPUT
PROJECTION
OUTPUT
w(t)
w(t-2)
w(t-1)
w(t+1)
w(t+2)
Given a set of (neighboring) words, guess single words that potentially occur along with this set of words.
CBOW
(Continuous Bag Of Words)
Skip-gram
Guess potential neighboring words based on the single word being analyzed.

Fig.3.1.1 Working of Word2vec

CBOW (Continuous Bag Of Words) predicts the dimension of the current word and Skip Gram predicts the dimension of context word. The main concept of using this algorithm is identifying words that appear in resemblance context will be closer to each other in linear space.

```
+---+--------------------+
|ID|                    words|
+---+--------------------+
|    0|[cook, some, of,...|
|    1|[jab, poora,desh...|
|    2|[this, prime,day...|
|    3|[when, balan,wan...|
|    4|[today, on,natio...|
|    5|[the, past, few,...|
|    6|[prime, members,...|
|    7|[did, you, know,...|
|    8|[the, smarter,vi...|
|    9|[it, s,official,...|
| 10|[from, products,...|
| 11|[watch,cricketer...|
| 12|[hundreds, of,th...|
| 13|[you, can, now,p...|
| 14|[it, s, so,heart...|
| 15|[great, news,for...|
| 16|[serving, the,co...|
| 17|[it, is,humbling...|
| 18|[smbs, power,and...|
| 19|[do, not,believe...|
+---+--------------------+
```

Fig.3.1.2 Input

```
+---+--------------------+--------------------+
|ID|                 words|            features|
+---+--------------------+--------------------+
|  0|[cook, some, of,...|[0.00881423667493...|
|  1|[jab, poora,desh...|[-1.9295284829030...|
|  2|[this, prime,day...|[0.00468003259064...|
|  3|[when, balan,wan...|[0.00806664588691...|
|  4|[today, on,natio...|[0.00795175624422...|
|  5|[the, past, few,...|[0.00700200653768...|
|  6|[prime, members,...|[0.00921624225156...|
|  7|[did, you, know,...|[0.00491017071255...|
|  8|[the, smarter,vi...|[0.00918910880996...|
|  9|[it, s,official,...|[0.00459522105908...|
| 10|[from, products,...|[0.00900642693656...|
| 11|[watch,cricketer...|[0.00865458792108...|
| 12|[hundreds, of,th...|[0.00653734833168...|
| 13|[you, can, now,p...|[0.00908771805475...|
| 14|[it, s, so,heart...|[0.00633239251267...|
| 15|[great, news,for...|[0.00493897967040...|
| 16|[serving, the,co...|[0.00721900272765...|
| 17|[it, is,humbling...|[0.01070436039541...|
| 18|[smbs, power,and...|[0.00870783687073...|
| 19|[do, not,believe...|[0.00858345756811...|
+---+--------------------+--------------------+
```

Fig.3.1.3 Output

As earlier said, the data frame that contains tokenized words in a column will be passed as input data (Fig.3.1.2) and output (Fig.3.1.3) will be the numerical representation of words. The output column name must be defined as “features” because it is a default parameter that should be passed as an input to the K – Means clustering algorithm. To implement Word2vec using spark, import Word2Vec from pyspark.ml.features.

1.2 K – MEANSCLUSTERING:

K-Means Clustering is a vector quantization algorithm that divides n observation into k clusters. Each observation belongs to the cluster with the nearest mean. This algorithm comes under the category of unsupervised learning algorithm since there is no target to predict. It is implemented to form the cluster of the similar words identified in the tweet that is related to particular product so that company can be easily identify whether the product meet the success or failure at earlystage.

```
                              Cluster Centers:
 [5.92594733e-03                    2.90937053e-03 -4.02627539e-03 - 3.50278111e-
                                    03
     -4.40793399e-03-3.72795788e-03                         4.66709873e-03 -
                              1.84110163e-03
     1.20994161e-03                       5.56380436e-03     6.15264394e-03
                              9.35506677e-03
     6.01331284e-03-3.22324763e-03                          5.11941672e-03
                              7.54321937e-03
              8.48886519e-03 -4.78706774e-03 -8.19317072e-04 -
```

```
                                 5.27250051e-03
          4.25416173e-03-3.84213887e-03                             9.37958682e-06
                                 9.05732939e-03
          6.85188016e-03-4.06282211e-03                             4.34808329e-03-
                                 3.03055366e-03
                   6.21949993e-03 -4.87426332e-04 -6.84660842e-03
                                 2.04869114e-03
     -1.90982772e-03                        4.76067708e-04   3.32735240e-03 6.37278722e-
                                       03
          -1.27865407e-02-2.46277629e-03                            9.28424416e-04 -
                                 2.45228319e-04
          6.92796019e-03                         9.18220943e-03     4.06354801e-03
                                 2.86113977e-03
          -7.13339541e-04                        1.70515731e-04     1.15023969e-02 -
                                 8.54837992e-04
           2.16410935e-03                          4.70359126e-03-1.62468682e-03
                                 3.02553501e-03
          -2.24953861e-03-8.30969651e-03                            7.00294680e-04 -
                                 1.28430570e-02
          2.32122281e-03                         6.10855826e-03     1.14470675e-02-
                                 1.10821973e-03]
```

Fig.3.2.1 ClusterFormation

```
+----------+-----+
|prediction|count|
+----------+-----+
|         0|   47|
|         1|   26|
|         2|   53|
|         3|   11|
|         4|    9|
+----------+-----+
```

Fig.3.2.2 No.of tweets formed per cluster

```
                            +--------------------+----------+
|                      words|prediction|
+--------------------+----------+
                      |[cook, some, of, ...|         2|
                      |[jab, poora, desh...|         4|
                      |[this, prime, day...|         0|
                      |[when, balan, wan...|         2|
                      |[today, on, natio...|         2|
                      |[the, past, few, ...|         2|
                      |[prime, members, ...|         2|
                      |[did, you, know, ...|         0|
                      |[the, smarter, vi...|         1|
                      |[it, s, official,...|         0|
+--------------------+----------+
```

Fig.3.2.3 Cluster identification for each tweet

As earlier said, the output parameter "features" of Word2vec was passed as an input parameter to K – Means clustering with k value 5. Here k value was chosen randomly. According to K value number of clusters were formed. It is recommended to choose k value as minimum as possible so that a better result of the cluster will be formed. Fig.3.2.1 shows the result of implementing this algorithm that reveals the cluster centers (centroid). Then using this cluster centers produce predicted cluster center that counts the number of tweets in each cluster (Fig.3.2.2) and also able to identify each tweet belongs to which cluster (Fig.3.2.3). The output column name of the predicted cluster center must be defined as "prediction" because it is a default parameter that is used to calculate the silhouette score. Silhouette with squared Euclidean distance measure of how exactly the data were clustered for the overall dataset with greater consistency. Silhouette score value ranges from -1 to +1. To implement this algorithm using spark import KMeans and ClusteringEvaluator from pyspark.ml.clustering, evaluation.

7. RESULTS ANDDISCUSSION:

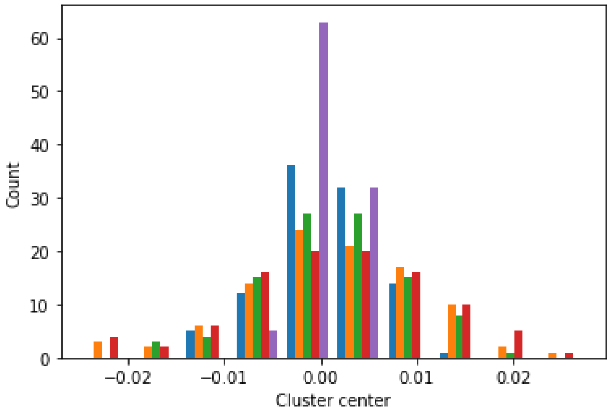

Fig.4.1 Visualization of Clusters and Count of tweet used

In the above figure (Fig.4.1), it clearly shows that cluster center value 0.00 counts more number of tweets when compared to other cluster center values. By using this, companies can easily predict which product was reached maximum to the consumer. For example, when we see the cluster 0 in Fig.6, it is easy to understand that more people tweet about the amazon prime day where prime members can buy the products at a cheap rate. So the company can predict that sales will be high on amazon primeday.

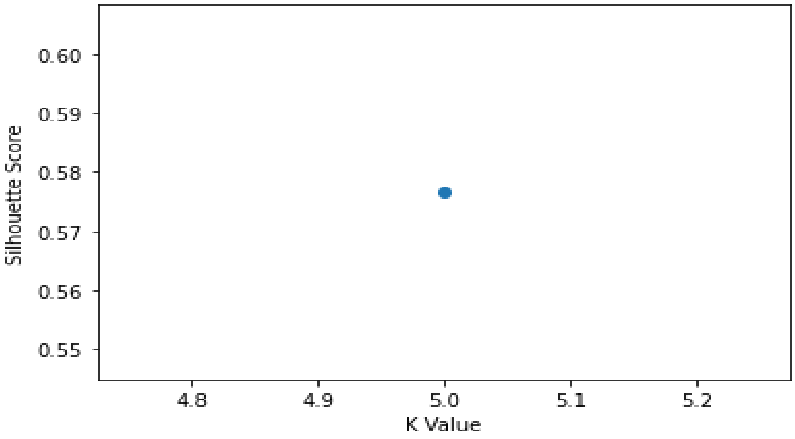

Fig.4.2 Silhouette Score based on K Value

As mentioned earlier Silhouette Score determines the quality of the cluster formed. If the score is -1 the clusters are wrongly formed and when it is +1 clusters are exactly formed. In our experiment silhouette score with squared Euclidean distance is 0.576596…. which is approximately equal to 0.6. So the clusters are best formed according to the K value. If k value is changed the silhouette score will also bechanged.

8. CONCLUSION

This paper concludes that it is mainly focused on the concept of business growth especially E-Commerce business by analyzing consumer behavior efficiently. This proposed method can be further enhanced by applying K – Nearest Neighbor algorithm which will provide suggestions to consumers based on their past likes and purchase. For example, if a consumer buys smartphones then they will also by the appropriate back case for protection. This enhancement further leads to the company'srevenue.

9. REFERENCES

[1] Anindita A Khade "Performing Customer Behavior Analysis using Big Data Analytics", 7th International Conference on Communication, Computing and Virtualization, 2016.

[2] Abdul Ghaffar Shoro, Tariq Rahim Soomro (2015) "Big Data Analysis: Ap Spark Perspective", Global Journal of Computer Science and Technology: C Software & Data Engineering.

[3] Humam K.Majeed Al-Chalabi, Ufuoma Chima Apoki, Hiba Akram Ali Abu- Alsaad "Big Data Analysis Using Social Networks", 1st International Conference on Recent Trends of Engineering Sciences and Sustainability, 2017.

[4] Abhi Basu, Terry Toy (2014) "Real-Time Healthcare Analytics on Apache Hadoop using Spark and Shark ", Intel® Distribution for Apache Hadoop Software [5]. M.Arumugam , S.Deepa , Dr.C.Jayanthi " BDA For Commercial Furtherance By Using Elevate Performance Algorithms " , International Journal of Scientific & Technology Research , 2020

[5] Classification and prediction of social attributes By K-Nearest Neighbor Algorithm with Socially-aware wireless networking-A study To cite this article: Sujatha Krishanmoorthy et al 2020 IOP Conf. Ser.: Mater. Sci. Eng. 937 01205

[6] V.R. Balaji, Maheswaran S, M. Rajesh Babu, M. Kowsigan, Prabhu E., Venkatachalam K,Combining statistical models using modified spectral subtraction method for embedded system,Microprocessors and Microsystems, Volume 73,2020.