Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 166-175
Received 15 April 2021; Accepted 05 May 2021.

Sentiment Analysis for E-Commerce Products Using Natural Language Processing

Bineet Kumar Jha [1], Sivasankari G.G [2], Venugopal K.R[3]

[1] Research Scholar (AMC Engineering College, VTU), CMR Institute of Technology, Bengaluru, India. Email Id: yoursbineetjha@gmail.com
[2] Professor, Computer Science and Engineering, AMC Engineering College, Bengaluru, India. Email Id: shivashankarigg@gmail.com
[3] Vice –Chancellor, Bangalore University, Bengaluru, India, Email Id: venugopalkr@gmail.com

ABSTRACT

Sentiment analysis is one of the ways to evaluate the attitude of consumers towards products and services. E-commerce businesses have grown to a larger level in recent years. Customers' opinions and preferences are collected to analyze them further to boost online businesses. Collecting real-time structured and unstructured data and performing sentiment analysis on them are challenging and need to be addressed. We have used PySpark, and resilient distributed dataset (RDD) based sentiment analysis using Spark NLP to address scalability and availability issues in sentiment analysis on the e-commerce platform. We have also used FLASK-based Restful APIs and Scrapy for web scrapping to collect useful data from an e-commerce site. Our findings indicate that the proposed method of Natural Language Processing (NLP) for e-commerce products in real-time has enhanced efficiency in terms of scalability, availability, and faster data collection

INTRODUCTION

Online Shopping has become significantly important for people nowadays as they can save time and effort to buy a product. E-commerce has grown greatly and customer feedback has become crucial to determine their interest and activities. Sentiment analysis is used to determine to know what consumers think about the product. This helps other customers in making buying decisions about the product.

A recommender system built on this can provide suggestions to other customers or show them related products while shopping. In recent years, sentiment analysis has drawn great interest, as it does the text classification based on consumer reviews.

The reviews are given in form of textual reviews, star ratings, and emojis. Sentiment analysis is used to analyze the huge amount of data that helps the retailers or service provider to achieve their targets. The opinion and features based on the given feature-wise details of the product. A myriad of information available in social media about a specific subject. People share their opinion on Twitter, Facebook, and other social platforms. Customer reviews their product and utilities after buying or using it. They put a tremendous amount of information on various platforms. Interpreting these reviews gives a significant business advantage, as many decisions related to the quality of services or products can be taken by the vendors. Also, these reviews help to build the recommender system. Based on their review and purchase behavior we also provide the frequently purchased or together purchased information.

Problem statement

The major challenge in sentiment analysis is performing real-time sentiment analysis in a distributed environment for structured and unstructured sentiment data. In addition to the conventional machine learning-based NLP approach, an RDD-based SparkNLP, Spark MLlib and pySpark with Restful APIs, and a Scarpy (or Sparkler) web scraper are applied to resolve the issue. It provides improved results in terms of scalability, availability, and faster data collection. The Spark OCR (optical character recognition) is another extension of Spark NLP which support to extract textual data written on images or PDF for sentiment analysis.

RELATED WORK

In the research article suggested by P. D. Turney *et al.*[1], in which word relationship is not considered, the Bag of words algorithm is clarified. The sentiment analysis of every single word is calculated individually and aggregated to determine the sentiment analysis of the entire sentence. The summarization techniques of all opinions are used. Relevant characteristics and attributes are obtained and the general characteristics of each product class are obtained from them. Each function is then assigned a Support Vector Machine and Sequential Minimum Optimization polarity. The polarity categorization polarity based on review-level and sentence-level has been reviewed by Xing Fang *et al.*[2]. Better results are given by the experimental output obtained for the categorization. D.M. E.-D. M. Hussein *et al.*[3] identified a technique for finding the significant limit. J.Khairnar *et al*[4] proposed Support Vector Machine and supervised Machine Learning for opinion mining. P. V. Rajeev *et al.*[5], defined a framework that uses machine learning and python to extract customer reviews.

C. Rain *et al.*[6], suggested a technique, based on Natural Language Processing (NLP) and an AdaBoost classifier to improve the performance of customer review processing on the e-commerce platform. M. Trupthi *et al.*[7] proposed a Hadoop based interactive system to predict the polarities of sentiment to improve marketing policies.

NhanCach Dang *et al.* [9] described the issues related to Natural Language Processing. Deep Learning provides solutions to the challenges faced by NLP. The frequency-inverse document frequency (TF-IDF) and word embedding are added to the datasets.

A tremendous amount of research is done in the area of Sentiment analysis, such as sentiment classifications[14, 11, 15, 10, 18], analysis of effect[16, 19], analysis of related survey[12, 15, 21], opinion extraction[13, 20] and the recommendation system[17]. Such strategies are used to separate emotions into positive, negative and neutral feelings. The literature concludes that sentiment analysis plays a key role in the e-commerce platform and several methods are used for sentimental analysis. Sentimental analysis and opinion mining are emerging fields of research that investigate the mindset of the consumer, thoughts, feelings, or emotions. There are three categorical types of sentiment: positive, negative, and neutral.

A machine learning and bigdata based approach can be used to evaluate customer experience[22]. The proposed system for opinion mining has shown 96% accuracy. A similar approach is described in another research work[23]. Nowadays we get many reviews about product on social media platform such as twitter, Facebook etc. We can analyse twitter sentiment data for the performace of the products and customers' opinion mining 24, 25, and 26.

In this paper, we have included the sentiment analysis of electronics products in the Amazon e-commerce website. We have predicted sentiments or satisfaction of purchased electronics products based on features and review text.

PROPOSED SYSTEM

In our approach we have developed the application in two phases in the first phase we have collected the sentiment data from the e-commerce website and separated it into words using Parts of Speech Tagging (POST). The second phase of our proposed system model gives the real-time analytics of the data gathered from an e-commerce site. The collected data is fed to the trained model which determines the sentiment in real-time. For this purpose, we have developed a Restful web service approach which is implemented using FLASK and python[8]. Scalability, fault tolerance, and availability are the challenging aspects of sentiment analysis. To overcome these challenges the Spark NLP and SparkML based PySpark is used. It has an in-built Natural Language Toolkit (NLTK) library to perform sentiment analysis.

Many reviews are collected in the form of structured and unstructured forms. We need in-memory distributed data processing for doing sentiment analysis. The Apache Spark framework gain importance due to in-memory distributed processing. It has in-built SparkML library to accomplish certain natural language processing tasks but doesn't provide fully fledged solution for natural language processing. To overcome this issue the John Snow Labs contributed in the development of Spark NLP. It is built to does NLP tasks completely. The data collection and data modeling process of sentiment analysis is shown in the proposed system architecture as hown in Figure 1.

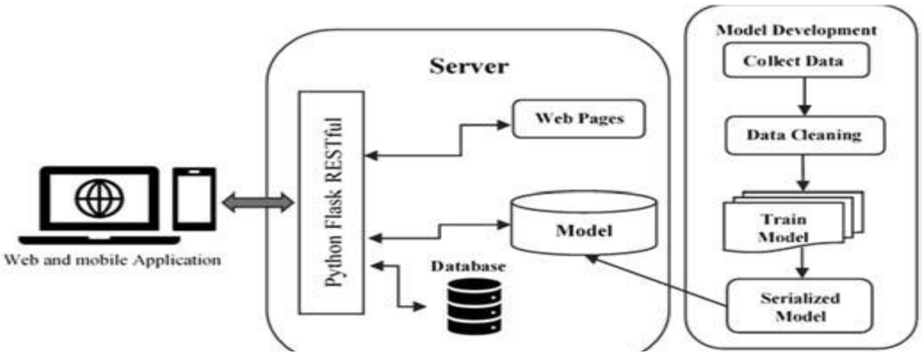


Figure 1 Proposed system architecture

METHODOLOGY

Data Collection

The related data are collected from the Amazon E-commerce site. These data include customer, product, seller, and payment-related information used for sentiment analysis. Data are read from the CSV files and kept in the data frame. To collect the real-time data using Scrapy application framework is used. This framework crawl the Amzon shopping website to collect structure or unstructured data. It is one of the fastest web crawling techniques.

Exploratory Data Analysis

E-commerce has growing trends worldwide. In this section, we will examine the different datasets available in the context of the Amazon e-commerce site for an electronics product. The total orders
on the e-commerce platform is obtained based on the dataset is shown in Figure 2.

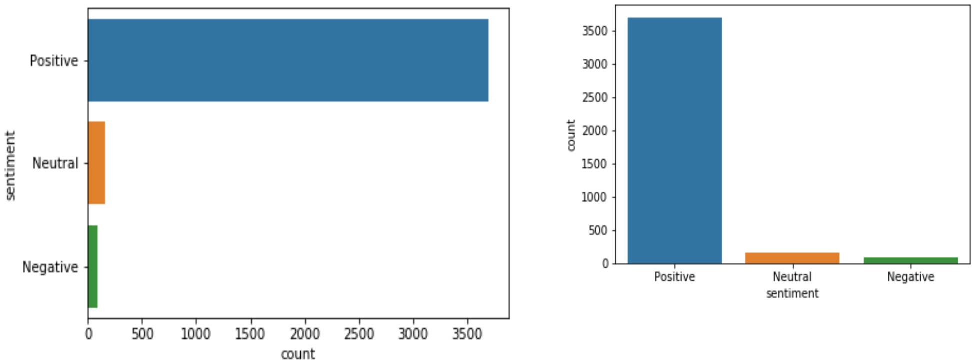


Figure 2 Sentiments details

Data Preprocessing

Post visualization removal of mentions, hashtag, stopwords, and links are needed from the training dataset. Stop words do not have any significance in search queries and can be removed as they contain a huge amount of unnecessary information. Stopwords are removed using Natural Language Toolkit (NLTK).

Sentiment Sentence Extraction

The sentiment sentence is extracted from the review. In PySpark, TextBlob library is used to compute sentiment polarity to identify sentiment sentences. After identifying sentiment sentences many other Spark NLP steps are performed to extract sentiments as described in section F. The overall sentiments from the extracted sentiment sentence is computer is obtained using sentiment analysis as shwn in Figure 3.

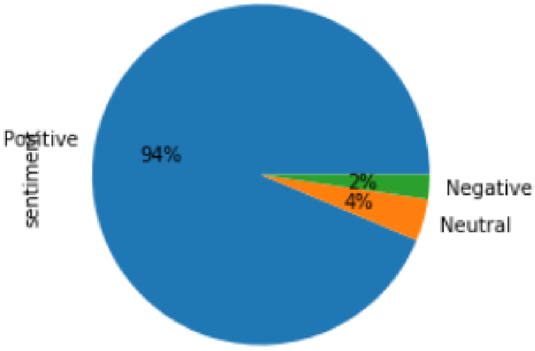


Figure 3 Overall sentiments statistics

In our proposed study for given dataset most of the reviews are classified into positive and few into negative and neutral sentiments.

Training Dataset

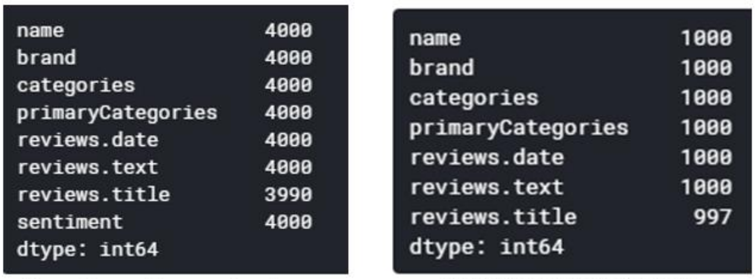
Table 1 shows the training dataset for Amazon electronic store that is collected from kaggle.com and real-time data updated from the Amazon website using a web crawler. It has a name, brand, categories, primary categories, review_date, review_text, review_title, and sentiment as the main attributes taken for training the proposed model.

Table 1 Sample Training dataset

S. no	Name	Brand	Categories	Primary categories	Reviews date	Reviews text	Reviews title	Sentiment
1	Fire Tablet, 7 Display, Wi-Fi, 16 GB - Includes Special Offers, Black	Amazon	Fire Tablets, Computers/Tablets & Networking, Tablets, All Tablets, Amazon Tablets, Frys, Computers & Tablets, Tablets & eBook Readers	Electronics	2017-06-23T00:00:00.000Z	It's a good device for children because they don't know any better	Good for kids	Negative

2	Amazon - Echo Plus w/ Built-In Hub - Silver	Amazon	Amazon Echo, Smart Home, Networking, Home & Tools, Home Improvement, Smart Home Automation, Voice Assistants, Amazon Home, Amazon, Smart Hub & Kits, Digital Device 3	Electronics, Hardware	2017-12-29 T00:00:00.000Z	Great sound quality. Easy to set up. Easy to use. Looks good too.	Great product.	Positive
3	All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi, 16 GB - Includes Special Offers, Magenta	Amazon	Electronics, iPad & Tablets, All Tablets, Fire Tablets, Tablets, Computers & Tablets	Electronics	2017-01-31T00:00:00.000Z	This was a great purchase! Love the fact I can download TV or movies and watch off-line. Easy set up.	Great tablet!	Positive
4	Fire Kids Edition Tablet, 7 Display, Wi-Fi, 16 GB, Green Kid-Proof Case	Amazon	Fire Tablets,Tablets,All Tablets,Amazon Tablets,Computers & Tablets	Electronics	2017-04-22T00:00:00.000Z	There is nothing spectacular about this item but also nothing majorly wrong with it	Does what it says, missing one key feature	Neutral
5	Brand New Amazon Kindle Fire 16gb 7 Ips Display Tablet Wifi 16 Gb Blue	Amazon	Computers/Tablets & Networking, Tablets & eBook Readers, Computers & Tablets, Tablets, All Tablets	Electronics	2017-01-01 T00:00:00.000Z	Too bad Amazon turned this tablet into a big advertising tool. Many apps dont work and the camera is not good.	Amazon Fire 7 Tablet	Negative

The number of categories and reviews used in our research work for training is obtained using sentiment analysis as shown in Figure 4.



```
name                4000
brand               4000
categories          4000
primaryCategories   4000
reviews.date        4000
reviews.text        4000
reviews.title       3990
sentiment           4000
dtype: int64
```

(a) Training attributes and reviews

```
name                1000
brand               1000
categories          1000
primaryCategories   1000
reviews.date        1000
reviews.text        1000
reviews.title        997
dtype: int64
```

(b) Test attributes and reviews

Figure 4 Training and test attributes

Spark NLP and Feature Selection Algorithm

The Spark NLP is an open-source natural language processing library. It has annotators which are either transformers or estimators, which utilize machine learning, deep learning, and rule-based algorithm. The estimators are basically the learning algorithm whereas the transformers convert one dataframe from one format to another. The result of Spark NLP is an annotation. Some of the important annotators are Tokenizer, Normalizer, Stemmer, Lemmatizer etc. which does the natural language processing. It has wide ranges of pre-trained model also known as Annotator model such as NerDLModel, Deep Sentence Detector, Lemmatizer Model etc. This model helps in transforming on DataFrame into another. The transformers help to convert one annotator type into another. Some of the important transformers are-

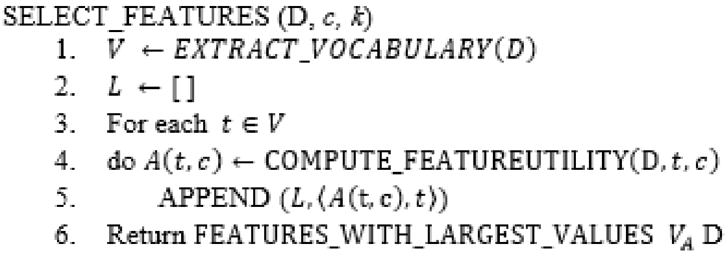
- DocumentAssembler: It annotates the raw data for further processing.
- TokenAssembler: It reconstruct document annotations from tokens
- Doc2Chunk: It convert document type to chunk type.
- Chunk2Doc: It convert chunk type data into document for the processing
- Finisher: It gives the annotation values as string.

Finally Spark NLP has set of APIs which are integrated with Spark ML. To process a simple text document a set of steps are done such as-

- Splitting text document into sentances and documents (Converting into words)
- Normalizing words through text processing such as cleaning, lemmatizing, stemming etc.
- Transforming token into feature vector (word embeddings, TF-IDF).

These steps are done by Spark NLP pipelined with sklearn library. Feature selection is a process of selecting useful words in a given document or web page. The algorithm eliminates unwanted stopwords or punctuations. The Spark NLP has also set of pre trained model to accomplish complex NLP tasks.

Algorithm 1: Feature Selection Algorithm



```
SELECT_FEATURES (D, $c$, $k$)
    $V \leftarrow EXTRACT\_VOCABULARY(D)$
    $L \leftarrow [\,]$
    For each $t \in V$
    do $A(t,c) \leftarrow$ COMPUTE_FEATUREUTILITY(D, $t$, $c$)
        APPEND ($L$, ⟨$A$(t, c), $t$⟩)
    Return FEATURES_WITH_LARGEST_VALUES $V_A$ D
```

The proposed algorithm extracts the vocabulary from the given dataset D and appends it to another list L based on its utility and importance as shown in Algorithm 1. After doing computations at the last the algorithm returns the features which have more impact on the sentiment analysis. In steps 4, 5, and 6 of the feature selection algorithm, for each vocabulary word, a utility measure *A(t,c)* is computed for a given class *c* and selects the *k* terms with the highest value of *A(t,c)*. The frequency-inverse document frequency (TF-IDF) is a feature vectorization method used for the text mining supported by RDD based Spark MLlib whereas Spark NLP does the text segmentation. In the text mining inputs are generally a corpus that includes many documents. For the feature extraction, we need to find the importance of terms in a document. Following calculation find the same.

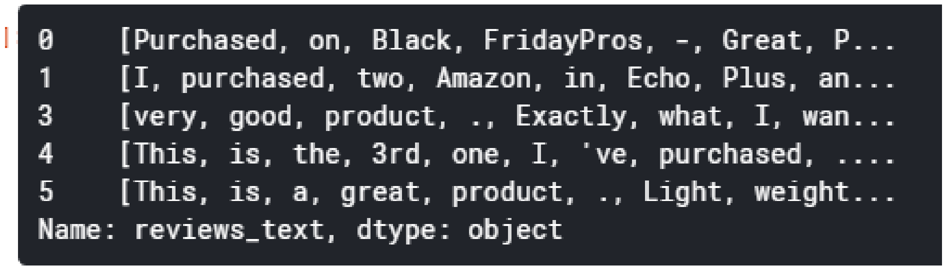
$$IDF(t,D) = log\frac{|D|+1}{DF(t,D)+1} \qquad (1)$$

$$TFIDF(t,d,D) = TF(t,d) \cdot IDF(t,D) \qquad (2)$$

The term frequency (TF) defines the number of times a term appears in a document whereas the document frequency (DF) represents the number of documents in corpus to contain the term. The Inverse of DF (IDF) can project the actual importance of terms *t* in document *D* as calculated in equation (1). The base of the *log* is any number greater than 1. The TFIDF is computed in equation (2) which is the product of $TF(t,d)$ and $IDF(t,D)$. It provides the actual importance of the term. To calculate the TFIDF in Spark first we need to collect various documents having sentences stored in them, later they are tokenized and pass into TF Transformer. The transformed data is provided as input to IDF Estimator. Word2Vec computes the distributed vector representation of words.

RESULTS AND DISCUSSION

This section we have shown the results obtained from our proposed system. The first step of the proposed system tokenizes the phrases whereas the second step it passes through preprocessing functions (Figure 5).



```
0    [Purchased, on, Black, FridayPros, -, Great, P...
1    [I, purchased, two, Amazon, in, Echo, Plus, an...
3    [very, good, product, ., Exactly, what, I, wan...
4    [This, is, the, 3rd, one, I, 've, purchased, ....
5    [This, is, a, great, product, ., Light, weight...
Name: reviews_text, dtype: object
```

Figure 5 Tokenizing and preprocessing steps

The most common keywords are visualized afterward. The top 20 common keywords extracted from the customers reviews and their counts. Most of the keywords have positive sentiments (Figure 6).

Received 15 April 2021; Accepted 05 May 2021.

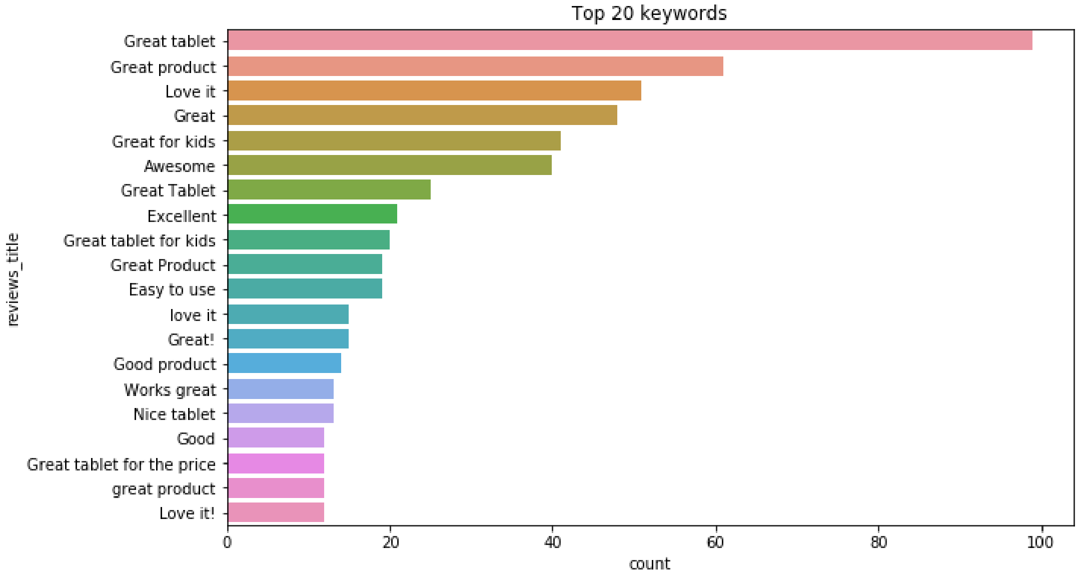
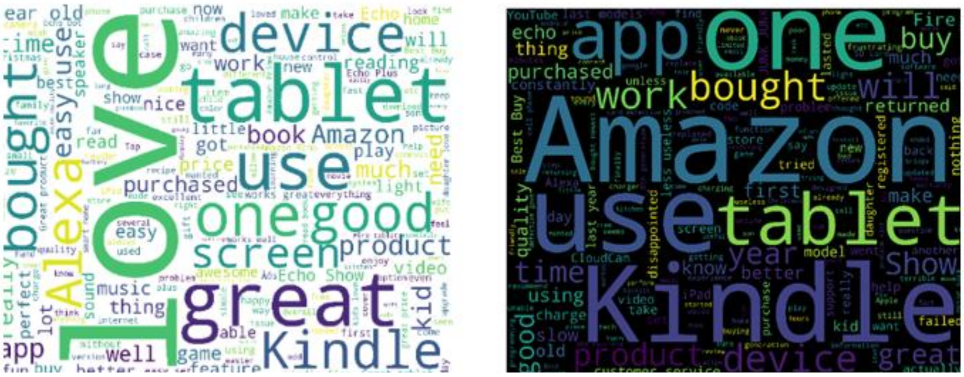


Figure 6 Most common keywords

The positive and negative sentiments are displayed using WordCloud visualization as shown in Figure 7. It is a collection of words, in that most specific words such as frequently occurring words appear bigger and bolder. These words are collected from sentiment sentences.



(a) Positive words (b) Negative Words

Figure 7 Sentiment classification

The density plots for the frequency distribution of customer age and positive feedback is obtained as shown in Figure 8.

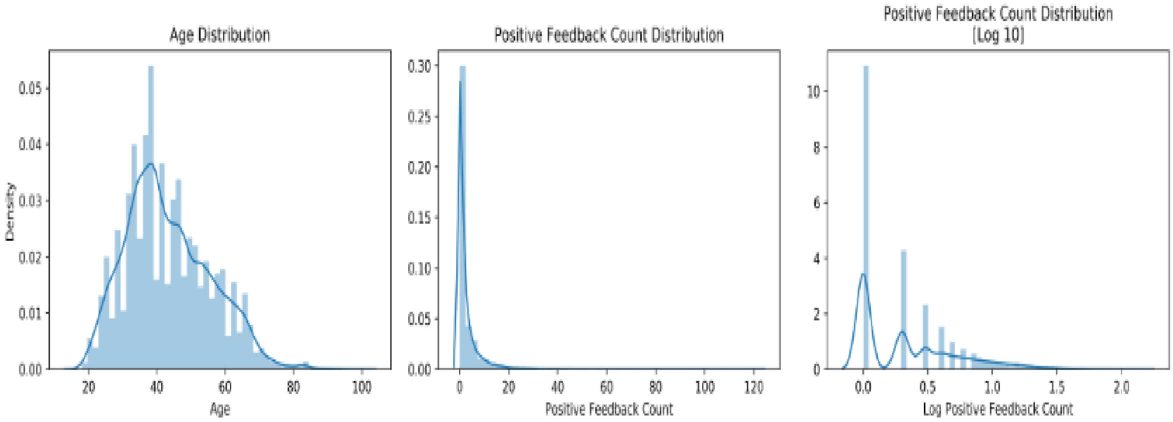


Figure 8 Frequency distribution of reviewer age and positive opinions

The sentiment score is computed based on the feedback given by different age group customers. The result indicates that most of the positive reviews are given by customers of age between 30-50 years. It also shows the logarithmic distribution of the same. It also represents the log-transformed distribution of the same to fit in the standard model.

CONCLUSION

We have applied the Spark NLP technique for sentiment analysis. The proposed technique uses a set of steps along with a features extraction algorithm. The accuracy level of the algorithm is 90% to 93% whereas the precision is 85% to 90%. The proposed technique performs well in terms of scalability and distributed processing as the sentiment analysis is performed in the Spark environment which has in-memory computation for large real-time sentiment analysis. The proposed Spark NLP has a rich set of libraries and it uses the pipeline approach to connect with standard MLlib library through APIs. The proposed technique is based on the RDD approach which ensures the availability and fault tolerance. However, a better methodology can be suggested for sentiment analysis with an improved deep learning approach in Spark NLP to relationship analysis.

ACKNOWLEDGEMENTS

The author is thankful to all authors who have participated to conduct the experiment and research in the proposed research work. We are thankful to AMC Engineering college for providing all resources and computing facilities to conduct this research.

REFERENCES

1. P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in Proceedings of the 40th annual meeting of the association for computational linguistics, Dec. 2002, pp. 417-424. https://arxiv.org/abs/cs/0212032.
2. X. Fang and J. Zhan, "Sentiment analysis using product review data," J. Big Data, vol. 2, p. 5, June 16 2015, https://doi.org/10.1186/s40537-015-0015-2.
3. D. M. E.-D. M. Hussein, "A survey on sentiment analysis challenges," J. King Saud Univ. - Eng. Sci., vol. 30, pp. 330-338, Oct. 2018, https://doi.org/10.1016/j.jksues.2016.04.002
4. J. Khairnar and M. Kinikar, "Machine learning algorithms for opinion mining and sentiment classification," Int. J. Sci. Res. Publ., vol. 3, pp. 1-6, 2013, doi: 10.1.1.414.8110.

5. P. V. Rajeev and V. S. Rekha, "Recommending products to customers using opinion mining of online product reviews and features," in 2015 International Conference on Circuits, Power and Computing Technologies (ICCPCT-2015), 2015, pp. 1-5, doi: 10.1109/ICCPCT.2015.7159433.
6. C. Rain, "Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning," Swarthmore College, 2013.
7. M. Trupthi, S. Pabboju, and G. Narasimha, "Sentiment analysis on twitter using streaming API," in 2017 IEEE 7th International Advance Computing Conference (IACC), 2017, pp. 915-919, doi: 10.1109/IACC.2017.0186.
8. “Quickstart–Flask” 2019 [Online]. Available: http://flask.pocoo.org/docs/1.0/quickstart/#routing.
9. NhanCach Dang, María N. Moreno-García, and Fernando De la Prieta, “Sentiment Analysis Based on Deep Learning:A Comparative Study” Electronics 2020, 9, 483; doi: 10.3390/electronics9030483.
10. S. Das and M. Chen, “Yahoo! for anazon: Extracting market sentiment from stock message boards”, In Proc. of the 8th APFA, 2001.
11. M. Hearst, “Direction-based text interpretation as an information access refinement”, Text-Based Intelligent Systems, 1992
12. H. Li and K. Yamanishi, “Mining from open answers in questionnaire data”, In Proc. of the 7th ACM SIGKDD Conf., 2001.
13. S. Morinaga, K. Yamanishi, K. Teteishi, and T. Fukushima, “Mining product reputations on the web”, In Proc. of the 8th ACM SIGKDD Conf., 2002.
14. B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques”, In Proc. of the 2002 ACL EMNLP Conf., pages 79–86, 2002.
15. W. Sack, “On the computation of point of view”, In Proc. Ofthe 12th AAAI Conf., 1994.
16. P. Subasic and A. Huettner, “Affect analysis of text usingfuzzy semantic typing”, IEEE Trans. on Fuzzy Systems, Special Issue, Aug., 2001.
17. L. Terveen, W. Hill, B. Amento, D. McDonald, and J. Creter “PHOAKS: A system for sharing recommendations”, CACM, 40(3):59–62, 1997.
18. R. M. Tong, “An operational system for detecting and tracking opinions in on-line discussion”, In SIGIR Workshop on Operational Text Classification, 2001.
19. C. Whissell, “The dictionary of affect in language”, Emotion: Theory, Research, and Experience, pages 113–131.
20. Savitha Mathapati, S H Manjula, and Venugopal K R, “Sentiment Analysis and Opinion Mining from Social Media: A Review”, Global Journal of Computer Science and Technology: C Software & Data Engineering. Volume 16. Issue 5 Version 1.0. 2016
21. Prasad B S, Akhilaa (2019), “Supervised Machine Learning Algorithms for Early Diagnosis of Alzheimer’s Disease”, 10.35940/ijrte.C6646.098319.
22. J. Jabbar, I. Urooj, W. JunSheng and N. Azeem, "Real-time Sentiment Analysis On E-Commerce Application," 2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC), Banff, AB, Canada, 2019, pp. 391-396, doi: 10.1109/ICNSC.2019.8743331.
23. Yi, S., Liu, X, “Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers’ review”, Complex Intell. Syst. 6, 621–634 (2020). https://doi.org/10.1007/s40747-020-00155-2
24. Jianqiang Z, Xiaolin G, Xuejun Z, “Deep convolution neural networks for twitter sentiment analysis”, IEEE Access 6:23253– 23260. https://doi.org/10.1109/ACCESS.2017.2776930
25. Jianqiang Z, Xiaolin G, “Comparison research on text preprocessing methods on twitter sentiment analysis”, IEEE Access 5:2870–2879. https://doi.org/10.1109/ACCESS.2017.2672677
26. McAuley J (2020) Recommender systems datasets. https://csewe b.ucsd.edu/~jmcauley/datasets.html. Retrieved 6 June 2020.