ISSN 2278-3091

Volume 10, No.2, March - April 2021

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse1371022021.pdf

https://doi.org/10.30534/ijatcse/2021/1371022021

VADER Sentiment Analysis without and with English Punctuation Marks

Ammar Oad[1*], Imtiaz Hussain Koondhar[2], Pinial Khan Butt[3], Huang lei[1], Aneel Oad[4], Mansoor Ahmed Khuhro[5] and Sajida Raz Bhutto[4]

[1] Faculty of Information Engineering, Shaoyang University, Shaoyang 422000, China,
[*]ammar_2k309@yahoo.com, 87431539@qq.com

[2]Isrra University Hyderabad Sindh Pakistan, imtiaz@yahoo.com

[3] Information Technology Center, Sindh Agriculture University Tando Jam 70060, Pakistan, pinial@sau.edu.pk,
khushk.ghulammujtaba@gmail.com

[4] Information Engineering, Central South University, China, oad_aneel@yahoo.com,
sajida.raz@muetkhp.edu.pk

[5] Department of Computer Science, Sindh Madressatul-Islam University (SMIU), Karachi, Sindh, Pakistan,
74000, makhuhro@smiu.edu.pk

ABSTRACT

Sentiment analysis is the foremost task in Natural Language Processing to understand the user's attitude (positive, neutral, or negative) by capturing their thoughts, opinions, and feeling about a particular product. This helps companies to fulfill customer satisfaction and make better future decisions about the product. Various techniques have been used in the literature for sentiment analysis, such as polarity scores, classifications, and automated sentiment analysis. In this paper, Valence Aware Dictionary and sEntiment Reasoner (VADER) sentiment analysis tool has been employed on a Twitter dataset (downloaded from https://www.kaggle.com). The study aims to measure the performance of VADER sentiment while concatenating fourteen English language punctuations marks, including Exclamation (!), Comma (,), Full Stop (.), Question Mark (?), Round Brackets (), Curly Brackets {}, Square Brackets [], Colon (:), Apostrophe ('), Dash (-), Hyphen (--), Semi-Colon (;), Slash (/), Quotation Mark (" ") and to observe whether the polarity (positive, neutral and negative) of a sentence changes or remains the same. After the analysis, the study found that Exclamation (!) maximizes the average positive polarity and average negative polarity and lowers the average neutral polarity. The Hyphen (--) and Comma (,) increase the average positive and neutral polarity and decrease the aver-age negative polarity. For Round Brackets (), Curly Brackets {}, Square Brackets [], Colon (:), Apostrophe ('), Dash (-), Semi-Colon (;), Slash (/) and Full Stop (.) the average positive and average neutral polarity decreases and average negative polarity increases.

Key words: Sentiment Analysis, Valence Aware Dictionary and sEntiment Reasoner (VADER), Natural Language Toolkit (NLTK), Punctuation Marks
.

1. INTRODUCTION

Sentiment analysis alludes to the distinguishing proof of feeling and assessment in the info messages that are generally client-created remarks. In practice, sentiment analysis includes a wide range of explicit assignments, such as opinion extraction sentence and aspect-level sentiment classifications. Traditional methods frequently study these assignments independently and plan exact models for each assignment in the light of manually designed features [1]. In the market, various brands are available, and choosing the right one is an intense job for a buyer. Also, the progression of E-Commerce influences the purchasing routine of clients. Thus, nowadays, buyers usually make their decision based on the review present in E-commerce (for example, the ratings and summary of relevant text about the items) [2]. Consequently, it becomes one of the most active areas in the research that tries to classify a piece of text containing opinions based on its polarity and determine whether an expressed opinion about a particular topic or event about the product is positive or negative [3]. In this regard, sentiment analysis has numerous applications in different areas, for instance, in organizations to get criticisms for items by which organizations can become familiar with clients' input and surveys on social media [4]. Meanwhile, Valence Aware Dictionary and sEntiment Reasoner (VADER) has the advantages of customary notion dictionaries alongside improved ones, which can be effortlessly utilized and broadened. VADER sentiment lexicons are considered of better quality since people have approved them. Utilizing punctuation in a sentence assists the reader with obviously understanding the message that is being passed on. Punctuation basically assists with demonstrating the pauses and the accentuation on specific thoughts that are discussed about in the content. Specifically, in academic writing, it is fundamental to precisely utilize punctuation as it assists with reinforcing contentions that are made in the text. The main aim of this research is to evaluate the VADER sentiment analyzer's performance with English language punctuation marks.

2. RELATED WORK

VADER separates itself from others in wording that it is touchier to assume articulations in online media settings, especially when overseeing web-based media messages and film reviews [4]. Another advantage of VADER is that it pro-vides information about the motivation and cynicism score and how positive, neutral, or negative an evaluation is. The positive, neutral and negative probabilities add up to 1. In addition, the compound score is a very useful metric in case we want a single measure of sentiment. Typical threshold values are represented as positive: compound score>=0.05 neutral: compound score between -0.05 and 0.05 negative: compound score<=-0.05 these are the most useful metrics for multidimensional [5]. VADER does not just ascribe a score to words; it also examines other linguistic and grammatical varieties, such as punctuation, capitalization, and the utilization of emoticons [6]. It was built by examining and selecting features from three preset lexicons: Linguistic Inquiry and Word Count (LIWC), Affective Norms for English Words (ANEW) and General Inquirer (GI). [7]. VADER centers around the words utilized in the sentence and afterward allocates a score to each word depending on the word dictionary [8]. VADER's author distinguished five heuristics dependent on linguistic and grammatical signals to pass on changes to feeling power that go past the pack of-words model.

The heuristics incorporate medicines for; (1) Punctuation (for example, number of '!'s); (2) capitalization (for example, 'I HATE YOU' is more extraordinary than 'I hate you'); (3) degree modifiers (for example, 'The service here is extremely good is more extraordinary than 'The service here is good); (4) constructive conjunction 'yet' to move the polarity; (5) tri-gram assessment to distinguish negation (for example 'The food here isn't actually all that great'') [9]. Punctuation is fundamental, and is utilized to pass on and explain the importance of written language.

It is such basic imprints as the full stop or the comma, and the more perplexing ones of semicolons and hyphens. Misunderstanding punctuation can change the whole meaning of a sentence [10] The main purpose of this research is to analyze the VADER sentiment analyzer that how VADER behaves with a sentences that contains a punctuation marks at the end of a sentence and by doing this whether the polarity scores of a sentences increases, decreases or remains same as without use of punctuation mark. Different text pre-processing techniques for correlating the sentiment scores of Twitter text with Bit coin prices during the COVID-19 pandemic have been created. The effect of numerous pre-processing functions, features, and time lengths of data on the correlation results have been investigated. Out of 13 procedures, that splitting sentences, removing Twitter explicit tags, or their combination generally expand the correlation of sentiment scores and volume polarity scores with Bit coin prices. Selecting the optimum pre-processing strategy would prompt machine learning prediction models to accomplish better accuracy as matched to the real prices [11]. In huge foundations, a pool of resource usage data is produced. This information can be cost-effectively used to comprehend the learning approach of students by teachers. Subsequently, hence, the purpose is to apply sentiment analysis on it for predicting the use of books and resources that would help the qualitative up gradation of the library. The data evaluated for the renewal of books and resources in the present work. The results obtained show that VADER sentiment algorithm was suitable in considerate the attitude and approach of students in the learning process [12]. VADER has an immense scope from analyzing the attitude of the person based on his tweet, to predicting the stock prices. But this field is pretty challenging. It is not easy to make a machine understand what accurately the person is saying. Two diverse methods have been utilized in sentiment analysis and compared i) VADER-Valence Aware Dictionary for sEntiment Reasoning ii) LSTM model (Long Short-Term Memory). VADER uses a lexicon-based approach, where the lexicon contains the intensity of all the sentiment showing words. The intensities are realized, the sentiment score is calculated and based on this sentiment score, the review is classified as either positive or negative. LSTM networks are very effective for sequential data like texts because they can relate the context of the sentence very well. preference of LSTM over RNN is higher as LSTM supports Long-term dependency which will help us predict our reviews better.[13] Most previous studies were concerned with to binary classification, a multi-classification system for analyzing tweets have been utilized to classify tweets related to the 2016 US elections. He results indicated that the VADER Sentiment Analyzer was an effective choice for sentiment analysis classification using Twitter data [14]. Different models for sentiment analysis reaching the efficiency of almost 85%-90%. But still need to emphasis on constructing models that have the competences to read between the lines, have the capabilities to understand human slangs and most importantly sarcasm [4].

3. MATERIALS AND METHODS

In this study, proposed architecture has been implemented using python programming language shown in Figure 1, at first step the dataset that consists of tweets have been read. In the second step Sentence Tokenizer that is in Natural Language Toolkit model (NLTK model) has been applied for pre-processing to split the whole paragraph into sentences. In the third step the VADER model has been applied. The sentiment analyzer VADER was executed on each sentence, and the polarity scores (positive, neutral, and negative) was saved without punctuation marks. After that, the fourteen punctuation marks were concatenated as given below; Exclamation (!), Comma (,), Full Stop (.), Question Mark (?), Round Brackets (), Curly Brackets {}, Square Brackets [], Colon (:), Apostrophe (‘), Dash (-), Hyphen (--), Semi-Colon (;), Slash (/), Quotation Mark (“ ”) at the end of the sentence and then polarity scores (positive, neutral, and negative) was saved.

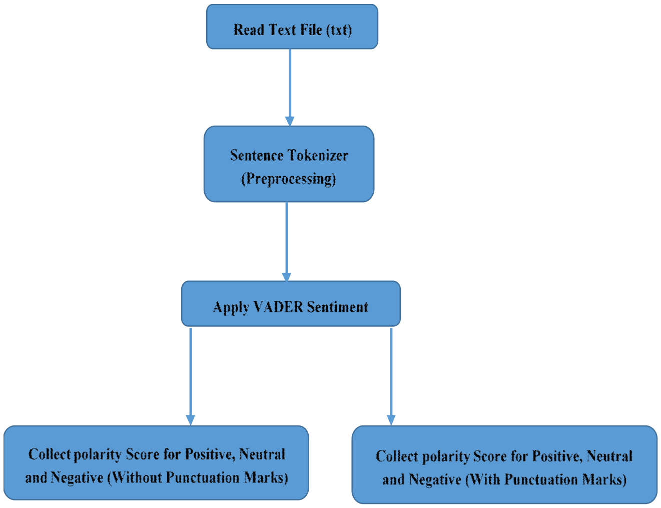

Read Text File (txt)
Sentence Tokenizer (Preprocessing)
Apply VADER Sentiment
Collect polarity Score for Positive, Neutral and Negative (Without Punctuation Marks)
Collect polarity Score for Positive, Neutral and Negative (With Punctuation Marks)

Figure 1. Architecture of the proposed method

3.1. SENTIMENT ANALYSIS:

- Import modules
- Open the input file (txt) and read contents from the text file
- Apply sentence tokenizer from NLTK model (Pre-processing)
- Run the sentiment analysis function to each sentence and save the polarity score for Positive, Neutral, and Negative. (Sentence without punctuation marks)
- Concatenate each punctuation mark individually at the end of the sentence and save the polarity scores for Positive, Neutral, and Negative (sentence with punctuation marks)
- Collect all the scores from both (sentence with punctuation marks) and (sentence without punctuation marks)
- Evaluate the difference in polarity for Positive, Neutral, and Negative in excel file
- Visualize the results.

3.1.1. IMPORT MODULES:

1. From vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
2. import nltk
3. from nltk.tokenize import sent_tokenize
4. from matplotlib import pyplot as plt

3.1.2. Open Input File (tweets.txt)

1. with open("tweets.txt") as file:
2. sentence = file.read()

3.1.3. Apply sentence tokenizer from NLTK model (Pre-processing)

1. text =sent_tokenize(sentence)

3.1.4. Run the sentiment analysis function to each text and save the polarity score for positive, neutral and negative (sentence without punctuation marks). For sentence in text:

1. object = SentimentIntensityAnalyzer()
2. dict = object.polarity_scores(sentence)
3. print(dict['pos']*100,"\t",dict['neu']*100,"\t",dict['neg']*100)

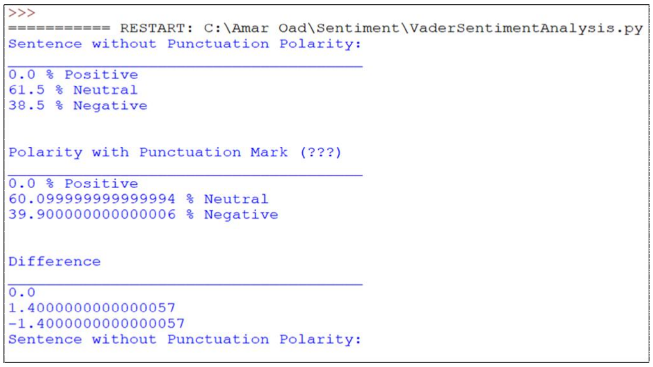
3.1.5. Concatenate each punctuation mark individually at the end of a sentence and save the polarity scores for positive, neutral and negative (sentence with punctuation marks). For sentence in text:

1. #print("Sentence: -",sentence,"\n")
2. sentence1=sentence.replace(".", "???")
3. object = SentimentIntensityAnalyzer()
4. dict1 = object.polarity_scores(sentence1)
5. print(dict1['pos']*100, "\t", dict1['neu']*100, "\t",dict1['neg']*100)

3.1.6. Pre-processing

The pre-processing mentions the changes applied to the data before sending it to the algorithm. Here we have used Sentence Tokenizer from NLTK Module to split the paragraph into sentences, and the rest of the symbols have been replaced manually.

4. RESULTS



```
>>>
=========== RESTART: C:\Amar Oad\Sentiment\VaderSentimentAnalysis.py
Sentence without Punctuation Polarity:
________________________________________
0.0 % Positive
61.5 % Neutral
38.5 % Negative


Polarity with Punctuation Mark (???)
________________________________________
0.0 % Positive
60.099999999999994 % Neutral
39.900000000000006 % Negative


Difference
________________________________________
0.0
1.4000000000000057
-1.4000000000000057
Sentence without Punctuation Polarity:
```

Figure 2. Positive, Neutral, and Negative Polarity with and without punctuation marks (???) and difference while executing python program.

Table 1. Average Positive, Neutral, and Negative Polarity Scores with punctuation marks.

S.no	Punctuation Mark	Sentence with Punctuation Mark		
		Average Positive Polarity	**Average Neutral Polarity**	**Average Negative Polarity**
01	**Exclamation (!)**	19.82760633	71.43542146	8.73636573
02	**Comma (,)**	18.4708199	73.392039	8.137033625
03	**Full Stop (.)**	18.4708199	73.392039	8.137033625
04	**Question Mark (?)**	19.35551205	72.119154	8.525441425
05	**Round Brackets ()**	18.4708199	73.392039	8.137033625
06	**Curly Brackets {}**	18.4708199	73.392039	8.137033625
07	**Square Brackets []**	18.4708199	73.392039	8.137033625
08	**Colon (:)**	18.4708199	73.392039	8.137033625
09	**Apostrophe (‘)**	18.4708199	73.392039	8.137033625
10	**Dash (-)**	18.4708199	73.392039	8.137033625
11	**The Hyphen (--)**	18.4708199	73.392039	8.137033625
12	**Semi Colon (;)**	18.4708199	73.392039	8.137033625
13	**Slash (/)**	18.4708199	73.392039	8.137033625
14	**Quotation Mark (“ ”)**	18.4708199	73.392039	8.137033625

Table 2. Average Positive, Neutral, and Negative Polarity Scores without punctuation marks.

S.no	**Sentence Without Punctuation Mark**		
	Average Positive Polarity	**Average Neutral Polarity**	**Average Negative Polarity**
01	18.47068939	73.39160909	8.13760172
02	18.47068939	73.39160909	8.13760172
03	18.47068939	73.39160909	8.13760172
04	18.47068939	73.39160909	8.13760172
05	18.47068939	73.39160909	8.13760172
06	18.47068939	73.39160909	8.13760172
07	18.47068939	73.39160909	8.13760172
08	18.47068939	73.39160909	8.13760172
09	18.47068939	73.39160909	8.13760172
10	18.47068939	73.39160909	8.13760172
11	18.47068939	73.39160909	8.13760172
12	18.47068939	73.39160909	8.13760172
13	18.47068939	73.39160909	8.13760172
14	18.47068939	73.39160909	8.13760172

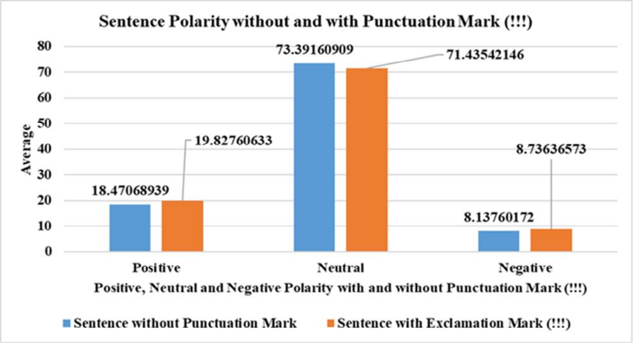


Figure 3. Sentence Polarity without and with Punctuation Mark (!!!)

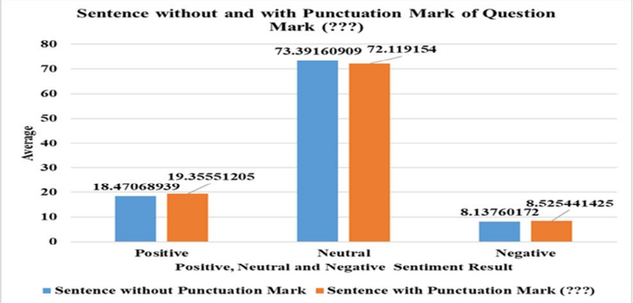
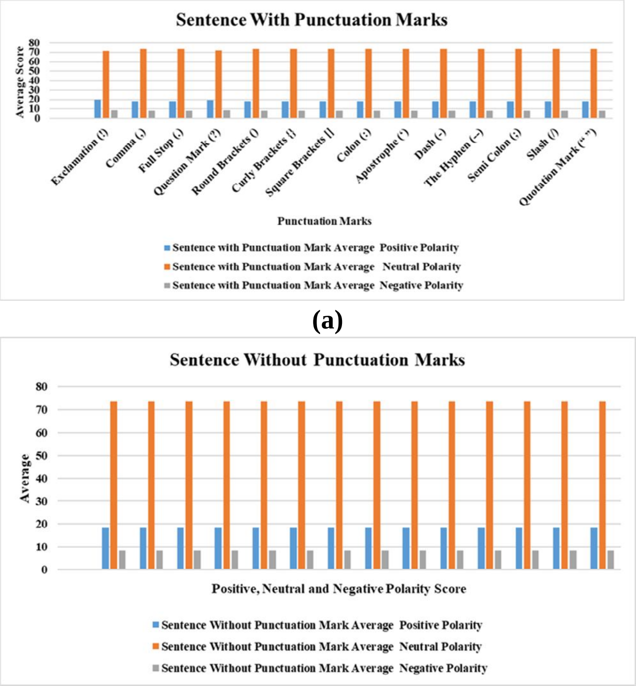


Figure 4. Sentence without and with Punctuation Mark of Question Mark (???)



(a)

(b)

Figure: 5 (a) Representing Sentences Average Positive, Neutral and Negative Polarity without Punctuation Marks, and **(b)** Representing Sentences Average Positive, Neutral and Negative Polarity without Punctuation Marks.

Table 3. Difference between the Average Positive, Average Neutral and Average Negative Polarity Scores for each Punctuation and explanation.

S.no	Punctuation Mark	Difference			Explanation
		Average Positive Polarity	Average Neutral Polarity	Average Negative Polarity	
01	Exclamation (!)	1.356916935	-1.95618762	0.59876401	By Concatenating (!) Exclamation Punctuation the both Average Positive Polarity increases by 1.356 and average Negative Polarity by 0.59876401 and Average Neutral Polarity decreased by -1.95618762.
02	Comma (,)	0.000131	0.00043	-0.00057	Minor difference has been created by using (,,,)
03	Full Stop (.)	-0.00224	-0.0024	0.004642	Minor difference has been created by using (…)
04	Question Mark (?)	0.884822662	-1.27245509	0.387839705	By Concatenating (???) Question Mark Punctuation Average Positive Polarity has been increased by 0.884822662, Average Negative Polarity has been increases by 0.387839705 and Neutral Polarity has also been decreases by -1.27245509
05	Round Brackets ( )	-0.00224	-0.0024	0.004642	Minor difference has been observed by using (())
06	Curly Brackets { }	-0.00224	-0.0024	0.004642	Minor difference has been observed by using ({})
07	Square Brackets [ ]	-0.00224237	-0.00239968	0.004642104	Minor difference has been observed by using ([])
08	Colon (:)	-0.00224	-0.0024	0.004642	Minor difference has been observed by using (:)
09	Apostrophe (‘)	-0.00224	-0.0024	0.004642	Minor difference has been observed by using (‘)
10	Dash (-)	-0.00224	-0.0024	0.004642	Minor difference has been observed by using (-)
11	The Hyphen (--)	0.000130508	0.000429909	-0.00056809	Minor difference has been observed by using (--)
12	Semi Colon (;)	-0.00224	-0.0024	0.004642	Minor difference has been observed by using (;)
13	Slash (/)	-0.00224	-0.0024	0.004642	Minor difference has been observed by using (/)
14	Quotation Mark (“ ”)	-0.00224	-0.0024	0.004642	Minor difference has been observed by using (“ “)

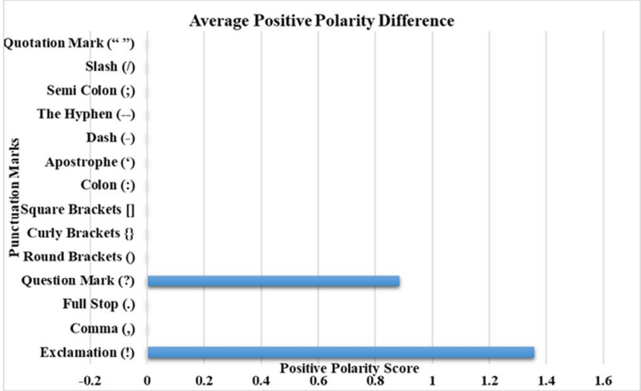


Figure 7. Representing Average Positive Polarity Difference in Sentence with Punctuation Marks, where it is seen that The Question Mark and Exclamation Marks have increased the polarity score. Rest of punctuation marks did not show any changes.

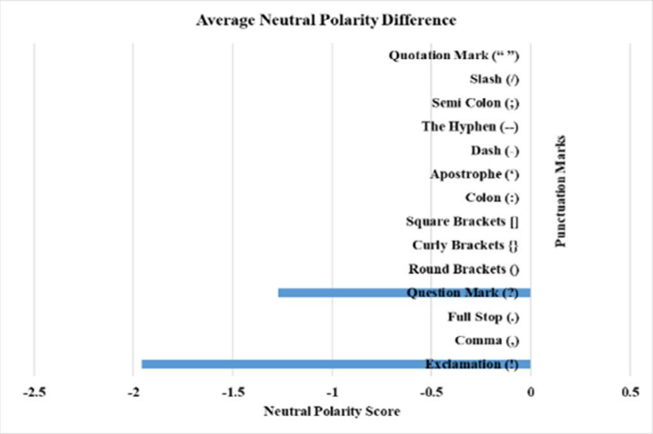


Figure 8. Representing Average Neutral Polarity Difference in Sentence with Punctuation Mark, where it is seen that by adding Question Mark and Exclamation Mark the neutral polarity score decreases and rest of punctuation marks did not show any changes.

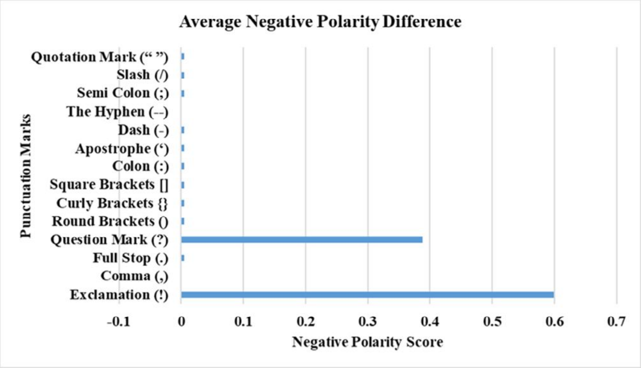


Figure 9. Representing Average Neutral Polarity Difference in Sentence with Punctuation Marks, where it is seen that by adding Question Mark and Exclamation Mark the negative polarity increases and rest of punctuation marks shows a little change.

5. DISCUSSION

The above results showed that the concatenation of Exclamation (!) maximizes the Average Positive Polarity by 1.356, average Negative Polarity by 0.59876401, and minimizes the Average Neutral Polarity by -1.95618762. By adding Punctuation marks individually at the end of the sentence, the Average Positive Polarity, and Average Negative Polarity increases while the Average Neutral Polarity decreases as shown in Figure: 3 and Figure:4. For Round Brackets ( ), Curly Brackets { }, Square Brackets [ ], Colon (:), Apostrophe (‘), Dash (-), Semi-Colon (;), Slash (/) and Full Stop (.) the Average Positive Polarity and Average Neutral Polarity decreases and Average Negative Polarity increases. Moreover, for Hyphen (--) and Comma (,), the Average Positive and Neutral polarity increases, and Average negative polarity decreases as shown in Table: 3. The difference of all punctuation marks is shown in Figure 8, Figure 9, and Figure 10.

Sentence without and with punctuation marks showing the Average Positive Polarity, Average Neutral Polarity, and Average Negative Polarity are presented in Figure 5(a) and (b).

6. CONCLUSIONS

The present study found that the Exclamation (!) and Question (???) marks increased the Positive and Negative Polarity scores and decreased Neutral polarity score. In contrast, other –punctuation marks increased the Positive and Neutral Polarity scores and decreased Negative Polarity scores, which means that after concatenating punctuation marks at the end of a sentence can make some in in polarity scores. For future work, VADER can be analyzed using Uni-Gram, Bi-Gram, and Tri-Gram to observe the polarity differences.

APPENDIX

Positive Difference= Average Positive Score of Sentence with Punctuation Mark – Average Positive Score of Sentence without Punctuation Mark.

Neutral Difference = Average Neutral Score of Sentence with Punctuation Mark – Average Neutral Score of Sentence without Punctuation Mark

Negative Difference= Average Negative Score of Sentence with Punctuation Mark – Average Negative Score of Sentence without Punctuation Mark

ACKNOWLEDGEMENT

This work is idea of authors and supervisor. Authors thanks to the reviewers and editor of journal for constructive comments.

REFERENCES

1. Tian, H.; Gao, C.; Xiao, X.; Liu, H.; He, B.; Wu, H.; Wang, H.; Wu, F. Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020.
2. Sheela, P, S.; Immaculate, I. Sentiment Analysis of Online Product Reviews Using DLMNN and Future Prediction of Online Product Using IANFIS. *J. Big Data* **2020**, *7*, 1–6, doi:10.21203/rs.2.19872/v1.
3. Sadr, H.; Solimandarabi, M.N.; Teshnehlab, M. A Novel Deep Learning Method for Textual Sentiment Analysis. *arXiv* **2021**, 1–13.
4. Vipul Kumar, C.; Ashish, B.; Amita, G. Twitter Sentiment Analysis Using Vader. *Int. J. Adv. Res. Ideas Innov. Technol.* **2018**, *4*, 485–489.
5. Bonta, V.; Kumaresh, N.; Janardhan, N. A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. *Asian J. Comput. Sci. Technol.* **2019**, *8*, 1–6.
6. Alsaeedi, A.; Khan, M.Z. A study on sentiment analysis techniques of Twitter data. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 361–374, doi:10.14569/ijacsa.2019.0100248.
7. Ribeiro, F.N.; Araújo, M.; Gonçalves, P.; André Gonçalves, M.; Benevenuto, F. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Sci.* **2016**, *5*, 1–29, doi:10.1140/epjds/s13688-016-0085-1.
8. Gonçalves, P.; Dalip, D.H.; Costa, H.; Gonçalves, M.A.; Benevenuto, F. On the combination of “off-the-shelf” sentiment analysis methods. *Proc. ACM Symp. Appl. Comput.* **2016**, *04-08-Apri*, 1158–1165, doi:10.1145/2851613.2851820.
9. Pano, T.; Kashef, R. A complete vader-based sentiment analysis of bitcoin (BTC) tweets during the ERA of COVID-19. *Big Data Cogn. Comput.* **2020**, *4*, 1–17, doi: 10.3390/bdcc4040033.
10. Deo, G.S.; Mishra, A.; Jalaluddin, Z.M.; Mahamuni, C.V. Predictive Analysis of Resource Usage Data in Academic Libraries using the VADER Sentiment Algorithm. In Proceedings of the Proceedings - 2020 12th International Conference on Computational Intelligence and Communication Networks, CICN 2020; 2020; pp. 221–228.
11. R. Adarsh, Ashwin Patil, Shubham Rayar, K.M.V. Comparison of VADER and LSTM for sentiment analysis. *Int. J. Recent Technol. Eng.* **2019**, *7*, 540–543.
12. Elbagir, S.; Yang, J. Twitter sentiment analysis using natural language toolkit and Vader sentiment. In Proceedings of the Proceedings of the International MultiConference of Engineers and Computer Scientist; 2019; Vol. 2239, pp. 12–16.