International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958(Online), Volume-9 Issue-4, April 2020

Business Sentiment Quotient Analysis using Natural Language Processing

Syed Salim, Madhu B K

Abstract: ***Online business has opened up several avenues for researchers and computer scientists to initiate new research models. The business activities that the customers accomplish certainly produce abundant information /data. Analysis of the data/information will obviously produce useful inferences and many declarations. These inferences may support the system in improving the quality of service, understand the current market requirement, Trend of the business, future need of the society and so on. In this connection the current paper is trying to propose a feature extraction technique named as Business Sentiment Quotient (BSQ). BSQ involves word2vec[1] word embedding technique from Natural Language Processing. Number of tweets related to business are accessed from twitter and processed to estimate BSQ using python programming language. BSQ may be utilized for further Machine Learning Activities.***

Keywords : Word2vec, Business Sentiment Quotient, Natural Language Processing.

I. INTRODUCTION

The machine learning activities basically require the samples for clustering and classification [2]. The sample is a collection of one or more feature values. Each feature of the sample defines the new dimension of the sample. Natural Language Processing (NLP) has several features namely LDP [3], TF-IDF [4] and son on. Each additional feature may enhance the property of the sample and lead to new derivative of research. The classification mainly depends on the type of features and number of right samples involved in the process. The accuracy and efficiency of the classification is directly proportional to the applicability of the features considered for the classification. Hence, there will be always a wide scope is given for the new feature extraction techniques. This paper is trying to make one such attempt to propose new feature extraction technique called as Business Sentiment Quotient(BSQ).

Revised Manuscript Received on April 18, 2020.
* Correspondence Author
Syed Salim*, Research Scholar, Department of Computer Science & Engineering, Vidya Vikas Institute of Engineering & Technology, Mysuru, Visvesvaraya Technological University, Belagavi, India, Email: syed.vviet@gmail.com
Dr. Madhu B K, Professor and Head, Department of Computer Science & Engineering, Vidya Vikas Institute of Engineering & Technology, Mysuru, Visvesvaraya Technological University, Belagavi, India, Email: prof.madhubk@gmail.com

II. LITERATURE SURVEY

A. Feature and Feature extraction process

Feature is a one of the major component corresponding to definition of sample identity. Various experiments takes place in order to address the definition of new entities in research developments. Author [5] has extracted feature from stereo for stereo matching applications. Here author has used Feature pyramid net to accommodate more amount of information and restore the property. Author[6] has tried to extract non-negative matrix factorization (NMF) on hyper-spectral image (HIS). Author [7] has introduced incremental gradient descent DBFE (IGDDBFE) to cover more amount of information for Guassian Maximum Likelihood Classifier. Author [8] has proposed a feature extraction technique based on wavelet packet and Common Space Pattern to strengthen correlational properties in the feature. This short span of survey indicates that the importance of feature extraction in research work.

B. Clustering and Classification

Clustering and Classification are the major machine learning approaches to group and label the samples. Author [9] has extracted features based on word embedding techniques then classified using FastText, XGBoost, CNN and other Classification techniques. Author found FastText as efficient in Hierarchical Text Classification. Author [10] has collected survey on different types of text feature extraction and selection for text classification.

III. MATHODOLOGY

Figure 1 is the methodology of the proposed research work. This methodology mainly focuses on the collection of the tweets and applies word2vec process against the identified business keywords.

A. Collection of tweets from Twitter

Twitter has provided a platform for researchers and scientists to create number of analytical applications which may be helpful for society. In this regard Python's tweepy[11] package deals with accessing of these tweets into the process.

B. Construction of business keyword repository

Tweets related to business may contain the positive or negative opinions about the products. Analysis of these tweets is possible through the identification and context relation estimation of its containing keywords.

Table I shows such predictive keywords required to estimate cosine similarity estimations.

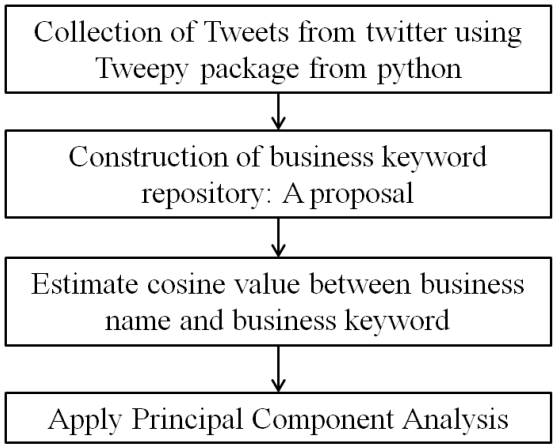

Fig. 1.Methodology of the proposed feature extraction technique

Table-I: Positive and Negative Business Sentiment Keyword Repository

SN	Business sentiment
1	Applicable, good, advantageous, acceptable, effective, less cost, affordable, Cost effective
2	Not happy, Not interested, Bad, useless, not applicable

C. Estimate Cosine value between Business name and Business sentiment keyword

Word2vec model follows two architectures that is CBOW (Continuous Bag of Words) and Skip-gram. CBOW is the process of predicting the keyword based on the two or more context words. Similarly, skip-gram is the process of predicting bag of words based on the given single keyword. Figure 2 shows the CBOW and Skip-gram models

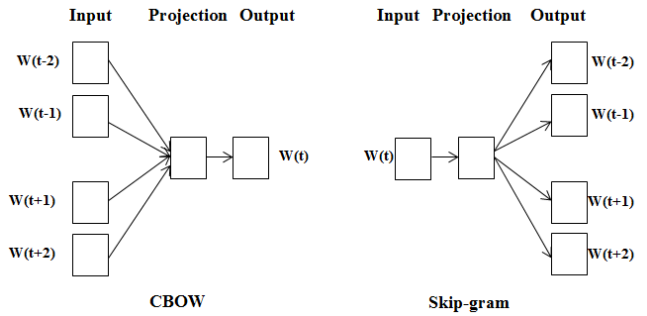

Fig. 2.Architectures of CBOW and Skip-gram

The proposed paper uses only CBOW cosine similarity estimations. Algorithm 1 produces the Business Sentiment Quotient

Algorithm 1: Business Sentiment Quotient

1. Choose the tweets by considering hashtag as Business and selected product.

2. for each tweet

i. Estimate cosine similarity between each identified keyword as mentioned in the Table I

ii. Apply PCA[12] on only selected positive or negative keyword cosine similarity estimations to generate either Business Positive Sentiment Quotient or Business Negative Sentiment Quotient.

IV. RESULTS AND DISCUSSIONS

The proposed methodology is implemented in python using tweepy package. Figure .3 shows certain tweets selected from twitter through python program.

Fig. 3.Filtered Tweets under Business and Oil Industries

Table II shows the Cosine Similarity values of all the positive keywords corresponding to the selected tweets. Similarly, Table III defines Cosine Similarity values on negative keywords.

Table-II : Cosine Similarity values of positive keywords

Tweets	good	accep table	less cost	More useful	afford able
STweet1	0	0.6	0	0.3	0
STweet2	0	0.81	0	0	0.27
STweet3	0.23	0	0.7	0	0
STweet4	0.72	0	0..34	0	0
……					

Table-III : Cosine Similarity values of negative keywords

Tweets	Not happ y	Not interest ed	Bad	Useles s	Not applic able
STweet5	0	0.43	0	0.23	0
STweet6	0.45	0.32	0	0	0
STweet7	0	0	0.67	0	0.55
STweet8	0	0	0.37	0.62	0
……					

Table IV shows the PCA estimations, as per the survey PCA is mainly used for dimensionality reduction. PCA estimation considers only non-zero cosine similarity values of the Table II and Table III.

Table-IV: PCA estimations of both Table II and Table III features into PSQ and NSQ

	PSQ		NSQ
STweet1	0.92	STweet5	0.11
STweet2	0.55	STweet6	0.16
STweet3	0.19	STweet7	0.78
STweet4	0.21	STweet8	0.82
……			

V. CONCLUSION

Research in data science mainly focuses on deriving new dimension of the samples. Invention of new feature or property can describe the sample and may improve the cluster and classification of the sample. In this regard, the current paper has tried to propose a methodology to derive two features namely Business Positive Sentiment Quotient (BPSQ) and Business Negative Sentiment Quotient (BNSQ). These features are certainly helpful while identifying samples through its behavioural properties in Natural Language Processing

ACKNOWLEDGMENT

Authors would like to thank the Management, Principal and all the teaching, Non-Teaching faculty members of Vidya Vikas Institute of Engineering and Technology, Mysuru, India for supporting and encouraging our research work.
And we gratefully thank the Visvesvaraya Technological University, Jnana Sangama, Belagavi for encouragement and support extended to this research work.

REFERENCES

1. Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
2. Alpaydin, Ethem. Introduction to machine learning. MIT press, 2020.
3. Jelodar, Hamed, et al. "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey." Multimedia Tools and Applications 78.11 (2019): 15169-15211.
4. Yun-tao, Zhang, Gong Ling, and Wang Yong-cheng. "An improved TF-IDF approach for text classification." Journal of Zhejiang University-Science A 6.1 (2005): 49-55.
5. L. Yang, J. Zhang and Y. Yang, "A feature extraction technique in stereo matching network," 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 2019, pp. 393-396.
6. C. Ji, M. Ye, H. Lu, F. Yao and Y. Qian, "Feature Extraction of Hyperspectral Imagery Based on Deep NMF," IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 2019, pp. 1092-1095.
7. S. Woo and C. Lee, "Incremental Feature Extraction Based on Gaussian Maximum Likelihood," 2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), JeJu, Korea (South), 2019, pp. 1-4.
8. G. Feng, L. Hao and G. Nuo, "Feature Extraction Algorithm based on CSP and Wavelet Packet for Motor Imagery EEG signals," 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), Wuxi, China, 2019, pp. 798-802.
9. Stein, Roger Alan, Patricia A. Jaques, and Joao Francisco Valiati. "An analysis of hierarchical text classification using word embeddings." Information Sciences 471 (2019): 216-232.
10. Deng, X., Li, Y., Weng, J. et al. Feature selection for text classification: A review. Multimed Tools Appl 78, 3797–3816 (2019). https://doi.org/10.1007/s11042-018-6083-5
11. Roesslein, Joshua. "Tweepy. An easy-to-use Python library for accessing the Twitter API." Github repository at http://github. com/tweepy/tweepy (2009).
12. Wold, Svante, Kim Esbensen, and Paul Geladi. "Principal component analysis." Chemometrics and intelligent laboratory systems 2.1-3 (1987): 37-52.

AUTHORS PROFILE

Syed Salim, is pursuing his PhD degree in Computer Scinece & Engineering under the guidance of Dr. Madhu B K from Visvesvaraya Technological University, Belagavi, He has done his M.Tech in Computer Netwroks & Engineering, MBA in Marketing & B.E in Informtion Science & Engineering from VTU Belagavi.Currently he is working as Assistant Professor in the Department of Computer Science & Engineering at Vidya Vikas Institute of Engineering & Technology Mysuru, which is affiliated to Visvesvaraya Technological University, Belagavi, Currently, he is working on Business Sentiment Analysis, Machine Learning using Python programming language. Salim has published several papers in National and International publications.

Dr Madhu B K, is at Technocrat with exposure to academics, Software Industry and Research. He is M.Tech and PhD in Computer Science & Engineering, with specific research interest in the field of Agile Software Testing. He is Advisor and Consultant for Performance Testing, Optimizing SDLC and Implementation of CMMI Levels. Member of Vocational Training Group. He is also Visiting Professor, Guest Faculty and Editorial Member and Reviewer of IEAE, IJCEE, GJCSM, JIS, IEEE International Conference and published 40+ research papers in IEEE and many reputed journals. He has 2 students completed PhD under his guidance and presently guiding 2 PhD research scholars. At present he is Professor and Head of CSE and with additional charges as Head, Placements at Vidya Vikas Institute of Engineering & Technology Mysuru, which is affiliated to Visvesvaraya Technological University, Belagavi, India.