Determining Feature-Size for Text to Numeric Conversion based on BOW and TF-IDF

Hasan J. Alyamani

Department of Information Systems, Faculty of Computing and Information Technology in Rabigh (FCITR),
King Abdulaziz University, Jeddah 21589, Saudi Arabia

Summary
Machine Learning is the most popular method used in data science. Growth of data is not only numeric data but also text data. Most of the algorithm of supervised and unsupervised machine learning algorithms use numeric data. Now it is required to convert text data into numeric. There are many techniques for this conversion. Researcher confuses which technique is best in what situation. Here in proposed work BOW (Bag-of-Words) and TF-IDF (Term-Frequency-Inverse-Document-Frequency) has been studied based on different features to determine best method. After experimental results on text data, TF-IDF and BOW both provide better performance at range from 100 to 150 number of features.

Key Word: *Machine Learning, Supervised and Un-Supervised Learning, TF-IDF, BOW*

1. Introduction

Rapid growth of data in shape of comments, reviews or opinion become the most interesting field of interest for data science. Most of the data is written in natural language i.e. in text form. This data can change the mind of 80% people in any context by reading this text [1]. So, analysis is very hot issue related to this text. Most of the work has been done in this regard such as feeling of a person opinion about aspect [2][3][4]. Machine-learning and scored-based are two approaches for classification of text [5][6]. Algorithms based on supervised or unsupervised methods of machine learning uses training data while different attributes of an entity is used for other methods of learning [7]. Opinions can also be determined as positive or negative by using the predefined scores of the words [8][9]. Researchers has also done work by using combined approaches with SentiWordNet and lexical resources find out score of slang [10]. Extracting Sentiment Orientation of an opinion Lexicon based approach has been used for scoring[8][11][12]. To find the polarity of sentence, a predefined list of positive and negative words can also be considered [13][14]. Training matrix is used for sentiment analysis in based on random forest method i.e. sentiment analysis can be done by different methods [15][16][17][18], each method has improved accuracy with respect to previous one. Every day, huge amount of information is generating with heavy speed. This information is mostly unstructured and require lot of preprocessing. As whole finding of polarity of all unstructured data is very laborious, so this data is organized in different form of categories, this is extraction of aspect [12]. Now sentiment analysis based on particular aspects requires less efforts as compare to sentiment analysis of an object with respect to all aspects [19][20].

All above discussed work of literature has used text data. In these works, major laborious task was to convert text data into suitable numeric values. Here, two methods have been studied to convert text data into numeric vectors. After experiments on different feature sizes on publicly available dataset of positive/negative reviews, it is found that TF-IDF & BOW achieved 76% accuracy on 100 feature sizes.

2. Proposed Methodology

Complete proposed model is depicted in Fig-1. Figure is showing conversion of dataset into a form which is compatible for our models. First of all, Dependent and independent features will be extracted from text dataset. Label encoding will be done on preprocessed independent variable. In label encoding in Python, we replace the categorical value with a numeric value between 0 and the number of classes minus 1. If the categorical variable value contains 5 distinct classes, we use (0, 1, 2, 3, and 4). Dependent variable containing text will be converted through TF-IDF and BOW. Numerical Statistic method TF-IDF find important words based on frequency to construct a vector. In BOW, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity [21]. Vector can be created based on feature size. Feature size is needed when need to create a set (Vocabulary) that includes all the unique terms that you can find in all the sentences of your available dataset. Although text has limited maximum size, so feature

Manuscript received December 5, 2021
Manuscript revised December 20, 2021
https://doi.org/10.22937/IJCSNS.2022.22.1.39

should be not very large nor very small. Finally based on confusion matrix result, size of feature can be selected.

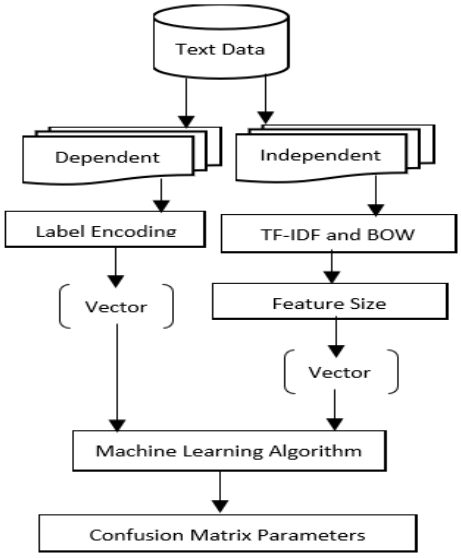

Fig-1: Proposed Work for Feature Size

3- Results

A dataset for sentiment analysis is downloaded from [22]. It has two column reviews and class containing 942 records. Sentences column contains the text of user opinion and class column has predefined class i.e. positive and negative. A sample consists of 10-reviews listing of the said datasets is presented in Table-1.

Table-1: Sample of Dataset

S.No	Sentences (Independent Variable)	Class (Dependent Variable)
S1	"took an hour to get our food only 4 tables in restaurant my food was look too worst"	Negative
S2	"the worst was the salmon sashimi"	Negative
S3	"also there are combos like a burger"	Positive
S4	"this was like the final blow"	Positive
S5	"i found this place by accident and i could not be happier"	Negative
S6	"seems like a good quick place to grab a bite of some familiar pub food"	Positive
S7	"overall i like this place a lot"	Positive
S8	"the only redeeming quality of the restaurant was that it was very inexpensive"	Positive
S9	"ample portions and good prices"	Positive
S10	"poor service"	Negative

Whole dataset has been applied on proposed model shown in Fig-1 using python. Here, processing of 10-text sentences will be shown to check the accuracy proposed model with different feature sizes. From dataset, sentences are independent variables and class is dependent variable. As sentences are in text format so, it has converted into a numeric value using BOW and TF-IDF. Result of numeric values from text dataset is shown in Table-2. This is the description of numeric values of 10-sentences based on feature size 5 (very very small).

Table-2: Numeric Values of Sample Dataset

S.No	Numeric Values from BOW	Numeric Values from TF-IDF
S1	[0 0 0 0 0]	[0. 0. 0. 0. 0.]
S2	[0 1 0 0 0]	[0. 1. 0. 0. 0.]
S3	[1 1 1 0 1]	[0.47000809 0.59050945 0.40565166 0. 0.51559453]
S4	[1 0 1 0 0]	[0.75703364 0. 0.6533759 0. 0.]
S5	[1 0 3 0 1]	[0.33506044 0. 0.86754566 0. 0.36755821]
S6	[1 0 0 0 0]	[1. 0. 0. 0. 0.]
S7	[0 1 0 0 0]	[0. 1. 0. 0. 0.]
S8	[2 1 1 0 0]	[0.79534396 0.4996277 0.34322026 0. 0.]
S9	[0 0 1 0 0]	[0. 0. 1. 0. 0.]
S10	[0 0 0 0 0]	[0. 0. 0. 0. 0.]

Dependent variables contain 'positive' and 'negative' labels, so it has been converted into simple encoded vector by using 'Label Encoding'. Result of first 10-records is shown in Table-5. Here 1 means positive and 0 means negative.

Table-3: Label Encoding on Dependent Variable

[0, 0, 1, 1, 0, 1, 1, 1, 1, 0]

This dataset has been split with 50% training and 50% testing data. Based on this feature size BOW achieved 61% and TF-IDF achieved 60% accuracy. Achieved accuracies based on different attempts is given below in Table-4.

Table-4: Accuracies Based on Different Feature Sizes

Feature Size	Accuracy (BOW)	Accuracy (TF-IDF)
5	61%	60%
10	64%	64%
20	65%	66%
100	78%	78%
200	74%	72%
300	69%	67%

From Table-4, it is observed that at feature size 100, accuracy of test data is 78%. Now, classification through BOW and TD-IDF based on feature size 100 is given below.

Table-5: Predicted Values of Sample Data

Sentence	Predicted Class of Sample Data (Bow)	Predicted Class of Sample Data (TF-IDF)	Actual Class of Sample Data
S1	0.45397688	0.4385167	0
S2	0.57271367	0.5483891	0
S3	0.06508475	0.2862084	1
S4	0.93353033	0.8592434	1
S5	0.64087673	0.6882219	0
S6	0.72411488	0.6636587	1
S7	0.76752427	0.9812035	1
S8	0.32345774	0.232341	1
S9	0.74057148	0.8735293	1
S10	0.32345774	0.232341	0

From Table-5, values is depicting that BOW and TF-IDF produces almost same predicted values where greater or equal to 0.5 values denotes 1 and less than 0.5 denotes 0. Mostly are the right prediction. Rest of confusion matrix based on different feature size using BOW and TF-IDF are shown in Table-6 and Table-7.

Table-6: Confusion Matrix Measures using BOW Method.

Parameters	Size-5	Size-10	Size-20	Size-100	Size-200	Size-300
precision	64	64	66	76	75	69
recall	62	64	66	76	75	69
f1-score	59	64	66	76	75	69

Table-7: Confusion Matrix Measures using TF-IDF Method.

Parameters	Size-5	Size-10	Size-20	Size-100	Size-200	Size-300
precision	62	65	66	77	73	67
recall	61	65	66	77	73	67
f1-score	59	65	66	77	73	67

4. Conclusion

Now a day there is lot of information available on internet, no one can read all huge amount of data. Everyone required a mechanism through which particular or document or sentences can be analyzed in short time. Machine learning algorithms are most important for such analysis. For these algorithms either supervised or unsupervised, input dataset should be in form of numeric values. Although BOW and TF-IDF provides a better platform for numeric conversion. But most of the time consumes on feature extraction with respect to size. Proposed works made different attempts on different sizes and found size of feature in range 100 to 150 is better for high accuracies. Based on this size 100, both methods achieved 78% accuracies. Rest of confusion matrix based on different attempts is given below in Fig-2.

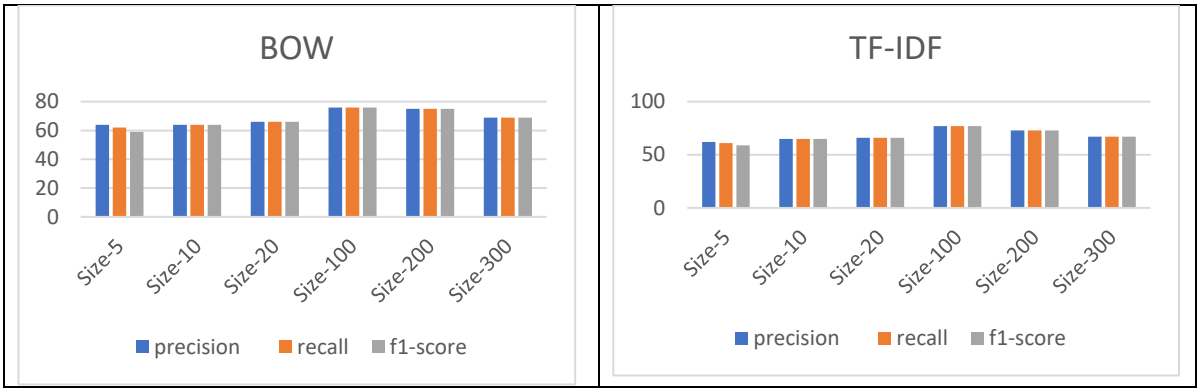

Fig-2: Comparison of Different Feature Sizes

By comparison of other measures of confusion matrix, it is concluded that feature size 100 is best by using BOW and TF-IDF. This accuracy 78% can also be increased by using preprocessing and feature extraction concepts. As purpose of proposed work is to determine the number of features, so rest of the concepts to increase the accuracy has been excluded.

References

[1] "3 Reasons Urgent Care Facilities Should Care About Online Reviews," 31 july 2017, 2017. https://resources.reputation.com/reputation-com-blog/3-reasons-urgent-care-facilities-should-care-about-online-reviews.

[2] B. Liu, Sentiment Analysis and Opinion Mining, vol. 5, no. 1. 2012.

[3] F. e-M. K. Khan, B.B. Baharudin, A. Khan, "Mining opinion from text documents," Adv. Res. Comput. Commun. Eng., vol. 3, no. 7, pp. 217–222.

[4]B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inf. Retr., vol. 2, no. 1–2, pp. 1–135, 2008, doi: 10.1561/1500000011.

[5] W. Y. and L. H. A. Wang S, Li D, Song X, "A feature selection method based on improved fisher's discriminant ratio for text sentiment classification," Expert Syst. Appl., vol. 38, no. 7, pp. 8696-8702., 2011.

[6] L. C. H. and C. H. Chen LS, "A neural network based approach for sentiment classification in the blogosphere.," J. Informetr., vol. 5, no. 2, pp. 313–322, 2011.

[7] T. Brychcín, M. Konkol, and J. Steinberger, "Machine Learning Approach to Aspect-Based Sentiment Analysis," in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 817–822.

[8] F. M. Kundi, A. Khan, S. Ahmad, and M. Z. Asghar, "Lexicon-Based Sentiment Analysis in the Social Web," J. Basic. Appl. Sci. Res, vol. 4, no. 6, pp. 238–248, 2014.

[9] S. Muhammad and F. Masud, "MMO: Multiply-Minus-One Rule for Detecting & Ranking Positive and Negative Opinion," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 5, pp. 122–127, 2016, doi: 10.14569/IJACSA.2016.070519.

[10] M. Z. A. Fazal Masud Kundi, Shakeel Ahmad, Aurangzeb Khan, "Detection and Scoring of Internet Slangs for Sentiment Analysis Using SentiWordNet Fazal," Life Sci. J., vol. 11, no. 1, pp. 66–72, 2014.

[11] J. Fiaidhi, O. Mohammed, S. Mohammed, S. Fong, and T. Kim, "Opinion Mining over Twitterspace : Classifying Tweets Programmatically using the R Approach," Digit. Inf. Manag. (ICDIM), Seventh Int. Conf. on. IEEE, pp. 313–319, 2012.

[12] A. Jeyapriya and C. S. K. Selvi, "Extracting aspects and mining opinions in product reviews using supervised learning algorithm," in 2nd International Conference on Electronics and Communication Systems, ICECS 2015, 2015, pp. 548–552, doi: 10.1109/ECS.2015.7124967.

[13] D. K. Kirange, R. R. Deshmukh, and M. D. K. Kirange, "Aspect Based Sentiment analysis SemEval-2014 Task 4,"

Asian J. Comput. Sci. {&} Inf. Technol., vol. 4, no. 8, pp. 72–75, Aug. 2014, doi: 10.15520/ajcsit.v4i8.9.
[14] Deepak Kumar Gupta and Asif Ekbal, "Supervised Machine Learning for Aspect based Sentiment Analysis," in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 319–323.
[15] T. Shaikh and D. Deshpande, "A Review on Opinion Mining and Sentiment Analysis," in IJCA Proceedings on National Conference on Recent Trends in Computer Science and Information Technology, 2016, no. 2, pp. 6–9.
[16] A. Alghunaim, M. Mohtarami, S. Cyphers, and J. Glass, "A Vector Space Approach for Aspect Based Sentiment Analysis," Proc. NAACL-HLT 2015, pp. 116–122, 2015.
[17] M. Cuadros, S. Sebastian, G. Rigau, E. H. Unibertsitatea, and S. Sebastian, "V3: Unsupervised Aspect Based Sentiment Analysis for SemEval-2015 Task 12," no. SemEval, pp. 714–718, 2015.
[18] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 Task 4 : Sentiment Analysis in Twitter," Proc. 11th Int. Work. Semant. Eval., vol. 3, no. 4, pp. 502–518, 2017.
[19] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, "Pulse: Mining Customer Opinions from Free Text," in Proceedings of the 6th international conference on Advances in Intelligent Data Analysis, 2005, pp. 121–132, doi: 10.1007/11552253_12.
[20] L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie review mining and summarization," in Proceedings of the 15th ACM international conference on Information and knowledge management - CIKM '06, 2006, pp. 43–50, doi: 10.1145/1183614.1183625.
[21] "https://en.wikipedia.org/wiki/Tf-idf." .
[22] "https://www.kaggle.com." .

Hasan J. Alyamani is a Chairman of Information Systems Department, King Abdulaziz University, Saudi Arabia. He received his B.Sc. (Computer Science) from Umm Al-Qura University, Saudi Arabia in 2006, MS (Computer Science) from The University of Waikato, New Zealand in 2012 and PhD (Computer Science) from Macquarie University, Australia in 2019. He has produced many publications in Journal of international repute and presented papers in international conferences.

Email: hjalyamani@kau.edu.sa