https://doi.org/10.51483/IJDSBDA.1.2.2021.23-30 ISSN: 2710-2599

SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Data Science and Big Data Analytics

Publisher's Home Page: https: / / www.svedbergopen.com /

Research Paper Open Access

Developing and testing a tool to classify sentiment analysis

Sameer Kumar Acharya[1*]

[1]Data Science Department, NMIMS University, Mumbai, India. E-mail: sameeracharya.nmims@gmail.com

Article Info

Volume 1, Issue 2, May 2021
Received : 23 December 2020
Accepted : 19 March 2021
Published : 05 May 2021
doi: 10.51483/IJDSBDA.1.2.2021.23-30

Abstract

The era has faced with explosive growth in data generation. Data generation has undergone a renaissance change. This availability of data has led a paradigm shift in the E-commerce sector; data is no longer a by-product of business activities, but are the asset to a company it helps in providing insights which are required in satisfying customers' needs. This paper provides an overview of sentiment analysis of product reviews based on different algorithms and its efficiency in determining positive from negative reviews based on N-gram, Bigram with the application of Count-Vectorizer and (Term Frequency-Inverse Document Frequency) (TFIDF) Matrix. Different classification models have been employed to check the prediction accuracy of the unlabeled text. Based on the above classification and tool has been developed which predicts the incoming reviews and classify its sentiment polarity.

Keywords: *Text mining, sentiments, K-Nearest Neighbor (KNN), Random forest, Multinomial Naïve Bayes, TFIDF, Count-Vectorizer*

1. Introduction

Internet and modern pieces of equipment mobiles, tabs have been the sheer force in the rise of social media which is used for information dissemination, communication through different technological platforms. User-generated content is the major component of social media, examples include web blogs, micro-blogs, Facebook, Twitter, Amazon reviews, online forums, Wikipedia, podcast, live streams, avatar-based virtual reality.

Social media is a very critical part of information dissemination among companies it helps to implement and use it as a marketing/branding tool. Social media is a platform with widespread adoption and unprecedented reach within the community, user, business, and government, etc. Everyone from researchers to application-based companies is interested in social media and has been skyrocketing in its application for different purposes. Business community/houses are tapping in social media content for its rich information. It is been used in the execution of marketing and branding strategies for innovation, product design and stakeholder relation for companies. For government and non-profit organization, it is mean of communication and sharing information to general people effectively. Sentiment analysis an attitude, judgment about an entity, feelings describe sentiment towards a product/entity.

Sentiment analysis is one major task of NLP (Natural Language Processing). Sentiment analysis of reviews from e-commerce in analyzed to capture authors feeling, emotion toward an entity. Millions of people share their opinion on an

* *Corresponding author: Sameer Kumar Acharya, Data Science Department, NMIMS University, Mumbai, India. E-mail: sameeracharya.nmims@gmail.com*

e-commerce platform for users to take an informed decision before buying/selling a product/service. Since the functionality of product/services cannot be determined without consuming, these opinions provide a platform to discuss and get insights about an entity/product before consuming/buying it. Opinion mining tries to capture this information by analyzing unstructured text data in form of reviews, comment. Sentiment analysis can be performed at a different granularity level.

Ex-1. An electronic product (innovative intelligent speakers) is recently launched by an e-commerce as an intelligent personal assistant, it sounds great, capable of voice interaction, alarms, streaming podcasts, provide weather and traffic condition on a real-time basis. However, the product has some technical issues and was updated and launched as a different product.

The sentiment expressed can be positive, negative and neutral with different intensities from 1-5 rating on Likert scale/stars in case of reviews as employed by many e-commerce sites.

- *Document-level:* Sentiment analysis on a document level (Lillian, 2002) used documents/reviews to find overall polarity as positive and negative which provides insights as in what was overall experience of the buyer of a product.
- *Sentence-level:* Sentiment analysis on sentences of the document analyzed on an individual sentence in the reviews to find polarity based on sentence per se.
- *Author-level:* Sentiment analysis on an individual author (Joshi, 2013) defines the preference of the person who authored the review can attain different polarities since reviews are subjective.

The sentiment expressed can be positive, negative and neutral with different intensities from 1-5 rating on Likert scale/stars in case of reviews as employed by many e-commerce sites.

2. Objective and gap analysis

2.1. Objective

The main objective of this project is to classify sentiment polarity based on text inputs, the algorithm employed in this research classifies positive from negative sentiment based on different algorithms like K-Nearest Neighbor (KNN), Naive Bayes classifier and Random Forest. Sentiment analysis tool has been developed where based on URL input the algorithm will fetch recent reviews from an e-commerce website and analyze the reviews to give polarity output as positive or negative. Further, a recommendation system will fetch similar products based on Euclidean distance among the products.

Several papers have suggested extensive use of sentiment analysis in different areas of research. It provides a polarity of a statement as positive or negative based on features present in the text. Even though these techniques have been employed extensively in NLP still it is difficult to determine whether the statement in positive or negative as facts are expressed on the products/entities functionality whereas opinions are about the attributes/properties of the entity. (Berman, 2017) mentioned opinions are completely subjective based on properties, past experience and general feeling towards an entity. Further based on different algorithms used in past maximum prediction accuracy was achieved with SVM and Multinomial Naive Bayes classifier, in this research the use of Multinomial Naive Bayes classifier gave an accuracy of 87%. Live data has been fetched and fed in the algorithm which gave 85-87% prediction accuracy.

2.2. Gap analysis

Language used can be complex as different special characters emoticons, native languages are being used to showcase the features and experience of a product, sarcastic remarks and phrases, online spammers post spam forums which have irrelevant or fake opinions, also the misclassification of reviews and rating used where customer/buyer is happy with the products and gives positive feedback but rates it below 3-star these conditions makes it difficult to provide the best classification of text as positive or negative. Most of the research papers suggest only a preliminary analysis of text classification has been done. In this research, an algorithm tool has been developed which can fetch live data from e-commerce websites (Reviews of products) the classifier algorithm will determine the polarity of the text as positive or negative.

2.2.1. Model flowchart

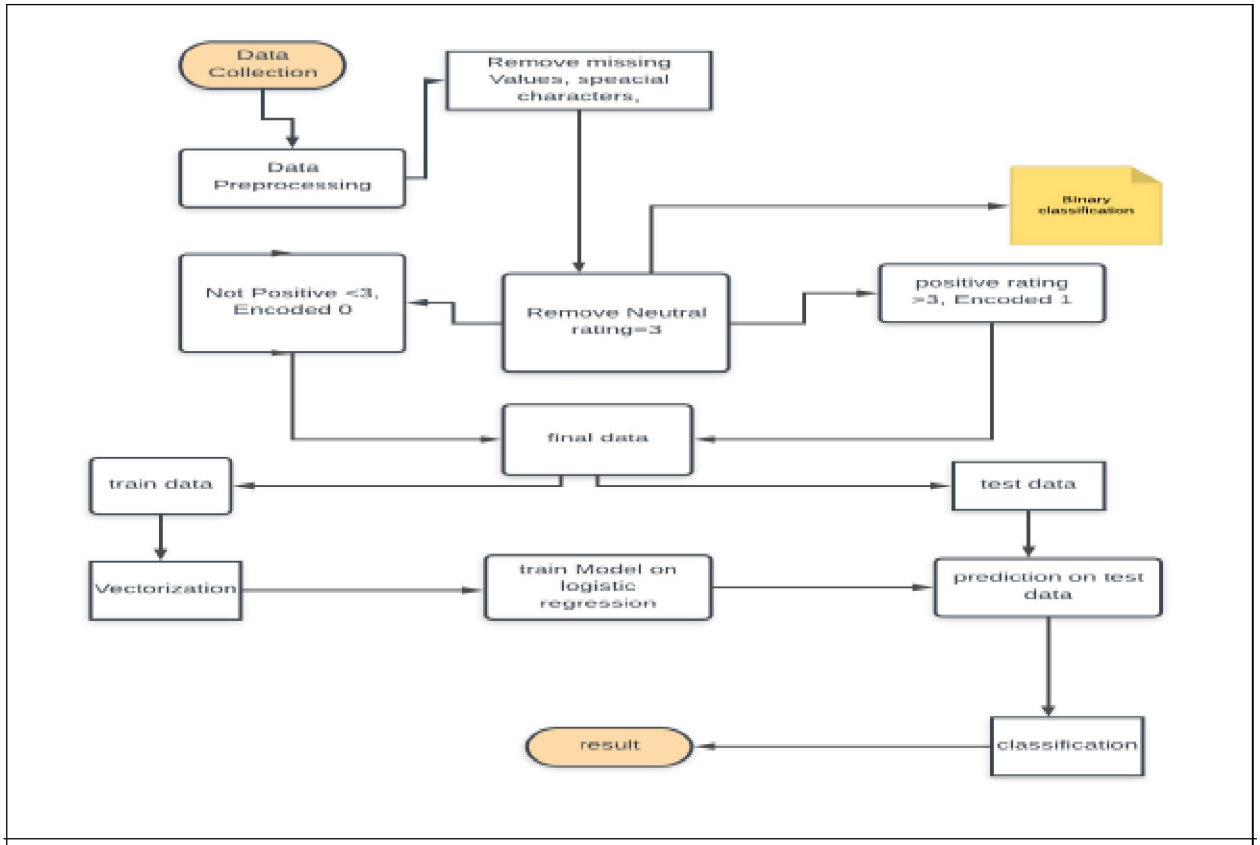

Figure 1: Classifier system design

3. Related works

Many research papers have been reviewed related to sentiment analysis and recommender system through different algorithms.

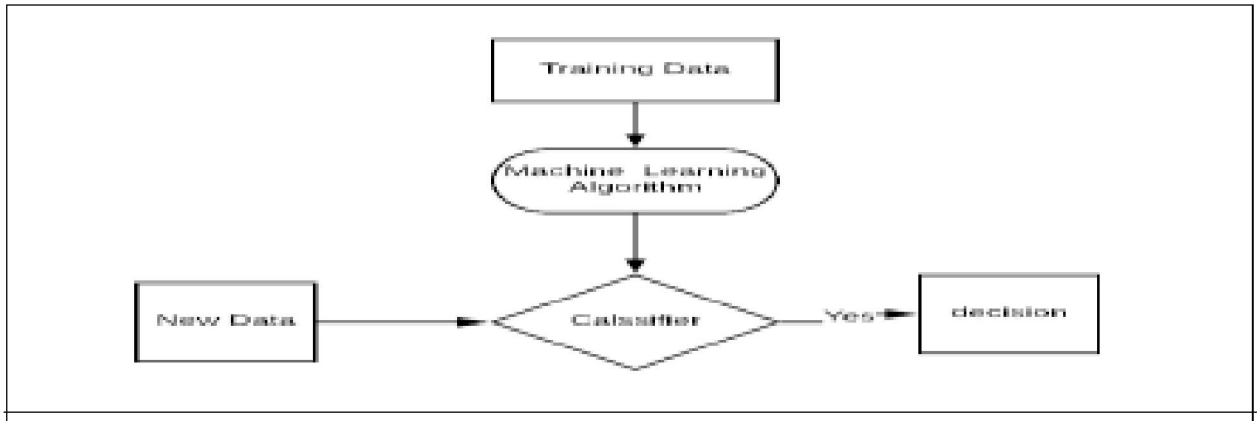

Figure 2: Machine learning technique

3.1. Naive Bayes classifier

A family of a probabilistic classifier based on Bayes probability theorem known for creating simple models in the field of classification with an assumption of no correlation between features. This classifier is simple yet powerful (Rish, 2001).

$$Posterior\ probability = \frac{conditional\ probability * prior\ probability}{evidence}$$

3.2. Random Forest

An ensemble method for classification which operates constructing a multitude tree during training of data and giving output as a mode of the class or means prediction (Ho, 1995), random forest corrects the over fitting of training data (Ho, 1998).

$$\hat{y}=\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{m}W_j(x_i,x')y_i=\sum_{i=1}^{n}\left(\frac{1}{m}W_j(x_i,x')\right)y_i$$

3.3. KNN- K Nearest Neighbor

KNN is lazy algorithms (Yong *et al*., 2009) it is one of the simplest algorithms which classify objects into predefine classes, and this algorithm does not require training data to perform classification (Guo *et al*., 2004). Training dataset can be used in the testing phase; this algorithm finds and classifies similar objects based on Euclidean distance from each other (Tan, 2005).

$$x(x,y)=\left(\sum_{i=1}^{n=1}|X_i=Y_i|\right)^{1/2}$$

3.3.1. Euclidean distance

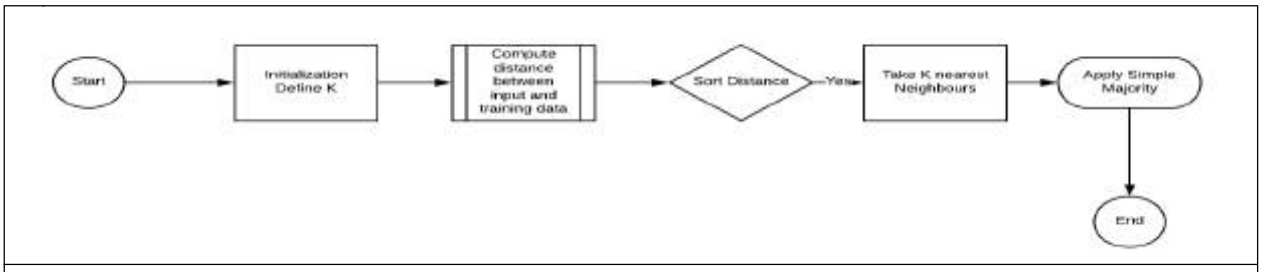

Figure 3: KNN

3.4. Natural Language Processing

Field of computer science that interacts between a computer and human language (natural), how a machine learns and analyzes a large amount of natural language data. It involves speech, language understanding and generation. It includes:

- Rule-based
- Statistical NLP

3.5. Word2Vec

Group of models employed to generate word embeddings, these models are 2-layer neural networks trained to construct linguistic contexts of words (Tomas Mikolov, 2013) word2vec produces a vector space from the input of corpus of text, with n-dimensions unique words in the corpus which is then assigned to the corresponding vector in space. Similar words in context are located in proximity to each other than words with a different context. Word2vec can be trained with negative sampling or hierarchical Soft-Max.

Table 1: Text categorization

Index	Description	References
N-gram	Character type pattern	(Caropreso, 2001)
Individual words	Lexical matching	(Salton, 1989) (Gerard Salton, 1986) (Salton, 1973)
Set of an individual words	Characterize co-occurrence of words	(Ho and Funakoshi, 1998) (Ho, 2000)
Multi-Words	Contextual information of words	(Li, 2008) (Papka and Allan, 1998)

Defines the predetermined assigned to the text documents where documents can be a text of reviews. The Kernel function is used for linearly inseparable problems (Aizerman, 1964). Considering multiclass problem n-gram and bigram methods have been employed/adopted.

3.6. TF-IDF

Inspired from IDF (Inverse Document Frequency), Proposed by 27. Sparck Jones (1972) and Sparck Jones, (2004.) The main intuition behind this method is an occurrence of term in many documents is not a very good discriminator, such a term should be given less weight then the term appeared in fewer documents. This method use term weight given to frequency of word appearance in the document. Tf-IDF defines the appearance of the word whether it is relevant in the document or not in the collection of document (Cliff, 2011).

4. Datasets

Datasets have been scrapped from Amazon.com for Echo-dot speakers from September 2017-October 2018, it consists of 119,486 reviews, it contains 96,775 positive and 22,741 negative reviews with ratings from 1 to 5, 1-2 star being negative rating, 3 neutral rating, and 4-5 star rating being positive.

5. Methodology

5.1. Semi-supervised

Labeled data and unlabeled data is feed into an algorithm to predict the accuracy of the model.

5.2. Data classification methodology

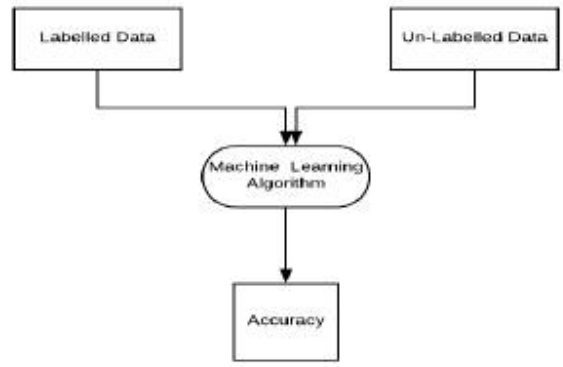

Figure 4: Semi-supervised learning

5.2.1. Sentiment analysis

Since the data set acquired is already labeled as positive or negative based on ratings by e-commerce during the investigation it was found 50% of data set is misclassified as positive or negative. Dataset was imbalanced positive to negative reviews were in a ratio of 4:1 use of text blob to predict the polarity of dataset predicted of 96,000 positive reviews 18,000+ reviews were misclassified as positive with negative reviews when checked with sentiment polarity from text-blob. Similarly in negative reviews of 21,810 we found 9,000 reviews as misclassified. To reduce data imbalance we have taken 10,905 data points from positive reviews and merge it with 10,905 negative classified reviews.

Post cleaning of data stop words and special characters were removed with lemmatization of NLP, the dataset was joined based on 'Author', 'Review description' and sentiment polarity 'Auto label'. 2,000 reviews were labeled by sentiment polarity and were divided in 70:30 ratios of control set and test set. Machine learning algorithms Naive Bayes, Random Forest and KNN algorithm were used to predict train and test accuracy of reviews data. Based on the prediction value of test data a better classifier was determined (multinomial Naive Bayes). An algorithm was generated which can classify live database on the classifier model. Model accuracy was tested.

6. Results and analysis

Sentiment Analysis: Dataset was split in a ratio of 70:30 TFIDF Vectorizer was used to offset the frequency of a word in the corpus. It converts text reviews into a sparse matrix. Different machine learning algorithms were used to train and test the sample data the accuracy achieved was.

Table 2: Algorithm result comparison

	Accuracy	
	Train set	**Test set**
Multinomial Naive Bayes	93.7	87.38
Random Forest	78.89	76.81
KNN	99.15	49.8

Based on the above results Multinomial Naive Bayes gave better results than Random Forest and KNN. The accuracy attained in the test dataset was 87.38.

Table 3: Confusion matrix

		Predicted Label	
		Negative	**Positive**
Sentiment-Polarity	**Negative**	2949	351
	Positive	475	2775

From above confusion matrix, it is determined the algorithm classified true result up to 87%.

N-grams, Bi-Grams in TFIDF Vectorizer, the model was unable to differentiate between positive and negative reviews

Reviews are treated the same by our model:

- not an issue, the product is nice = 1
- an issue, the product is not nice = 1

After extracting n-gram and bigram reviews were classified correctly

- not an issue, the product is nice = 1
- an issue, the product is not nice = 0

Predicting real-time reviews of a product from Amazon classified positive and negative polarity of the statement correctly. Data have been labeled as positive = 1, negative = 0.

	review_desc	pred_label
0	camera performance good battery life good char...	0
1	overall phone good nice front camera good self...	1
2	good product	1
3	xiaomi fooled u given money 64 bit hardware ca...	0
4	product good within price mobile front camera ...	1
5	nice purchase especially sd processor good pri...	1
6	y2 best phone price range.its potrait mode goo...	1
7	i using phone last month good well bad point d...	1
8	it 's even 5 day since purchased product i wou...	0
9	only speaker 's sound quality much good other ...	0

Real-time prediction of Xiaomi Redmi Note of total 10 reviews the classifier predicted nine reviews correctly one negative positive review was classified as negative because of words used 'Harassing'

Real-time prediction of Whirlpool 6.2 kg fully-automatic top loading washing machine reviews. The classifier predicted all reviews correctly.

	review_desc	pred_label
0	till person came whirlpool company install pro...	0
1	do purchase bangalore whirlpool guy come insta...	0
2	machine brilliant price range still giving 4 s...	0
3	v poor customer service whirlpool 3 time call ...	0
4	it expensive shown amazon app guy came install...	0
5	it working expected it fully automatic one n't...	1
6	excellent machine really good quality washing ...	1
7	a now issue going well.. the reason i purchase...	0

Positive Reviews: Analysis interpreted some words as positive like 'good', 'nice', 'mind-blowing', 'handy', 'love', and 'light'.

Negative Reviews: Analysis interpreted some words as negative 'disappointed', 'useless', 'terrible', 'annoying', 'poor', 'horrible', 'frustrating', 'bad', 'stupid', and 'disappointing'

Use of such negative words led to misclassification of positive sentiments as negative. This limitation was overcome by use of n-gram and bigram which reduced the misclassification.

7. Live reviews prediction

An interface model is being evaluated in this paper which can further be developed as a website or software to predict and analyze real-time reviews of a product/entity. X-path, CSS – selector from html of websites can be used to scrap reviews for real-time analysis with selenium. A website address with product page is being fed in the URL input, which opens the page and scraps first 100 recent reviews of the product and classifies it as positive or negative. Both businesses as well as the customer will benefit from this classification based on which user can decide the next step.

8. Conclusion

Sentiment analysis has its own challenges of misclassification which can be affected by the use of word in the sentence which can affect the polarity of the sentence. Prediction accuracy for classification with different algorithms gave the best result with Naive Bayes. Reviews prediction from e-commerce websites gave either positive or negative classification. This research determines the sentiment polarity of an entity from extracted reviews and used these algorithms to predict real-time reviews. Even though reviews give a better understanding of properties and features of an entity but still are difficult to analyze because of use of foreign languages, emotions, sarcasm, and pictorial representation such as emoticons etc. The accuracy of this model varies from 70-87% depending on the use of words in text.

9. Limitations

Since scrapping a data from a website without authentication is illegal, it is difficult to scrap reviews of product. Model developed only takes review of first page of a product from e-commerce since scrapping multiple pages lead to blocking of an IP address. Analyzing multiple data is very time consuming and process stops in between because of computational limitation.

Acknowledgment

I would like to thank my Project Guide, Dr Shailaja Rego, Associate Professor, NMIMS Mumbai, for the patient, guidance, encouragement and advice she has provided throughout the project. I have been extremely lucky to have a project guide who cared so much about my work, and who responded to my questions and queries so promptly. I would also like to thank Dr Shelja Jose for her support and encouragement and timely advice on subject matter.

References

Aizerman, A.B. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*. 821-837.

Ben, Schafer, J., and J. K. (1999). Recommender Systems in E-Commerce. *GroupLens Research Project*, MN 55455.

Berman, M. (2017, July 5). *Sentiment Analysis: Overview, Applications and Benefits*. Retrieved July 5, 2017, from growthaccelerationpartners: https://www.growthaccelerationpartners.com/blog/sentiment-analysis/

Caropreso, M.F. (2001). A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. *Semanticsscholar*. 385.

Cliff, G. (2011). Semantic Analysis: An introduction. New York Oxford University Press. p.17.

Ho, T.K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 10.1109/34.709601, 832 - 844.

Gerard Salton, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc. New York, NY, USA.

Guo, G., Wang, H., Bell, D., Bi, Y. and Greer, K. (2004). KNN Model-Based Approach in Classification. Researchgate.

Ho, T.K. (1995). Random decision forests. *IEEE Computer Society*. 278.

Ho, T.B. (2000). Non-hierarchical document clustering based on a tolerance rough set model. *International Journal of Intelligent Systems*. 199-212.

Ho, Tu Bao and Funakoshi, Kaname. (1998). Information retrieval using rough sets. *Journal of the Japanese Society for Artificial Intelligence*. 13(3), 424-433.

Joshi, S.M. (October 14-18, 2013). Sentiment aggregation using concept net ontology. IJCNLP, Sixth International Joint Conference on Natural Language Processing. 570-578.

Li, Y.J. (2008). Text document clustering based on frequent word meaning sequences. *Data & Knowledge Engineering*. 381-404.

Lillian, B.P. (02, 2002). Sentiment classification using machine learning techniques. *EMNLP*.

Papka, R. and Allan, J. (1998). Document classification using multiword features. *In Proceedings of the seventh international conference on information and knowledge management*. 124-131.

Rish, I. (2001, January). An Empirical Study of the Naïve Bayes Classifier. *ResearchGate*. 46.

Salton, G. (1973). On the specification of term values in automatic indexing. *Journal of Documentation*. 351-372.

Salton, G. (1989). *The transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

Sparck Jones, K. (2004). IDF term weighting and IR research lessons. *Journal of Documentation*. 521-523.

Sparck, Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*. 11-21.

Tan, S. (2005). Neighbor-weighted K-nearest neighbor for unbalanced text corpus, *Expert Systems with Applications*. 667-671. ACM Digital Library.

Tomas Mikolov, K.C. (2013). Efficient Estimation of Word Representations in. arXiv, 1301.3781v3.

Yong, Z., Youwen, L. and Shixiong, X. (2009). An Improved KNN Text Classification Algorithm Based on Clustering. *Journal of Computers*. 4(3), 230-237.

Cite this article as: Sameer Kumar Acharya (2021). Developing and testing a tool to classify sentiment analysis. *International Journal of Data Science and Big Data Analytics*. 1(2), 23-30. doi: 10.51483/IJDSBDA.1.2.2021.23-30.