S L U J S T
Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 27-33
https://doi.org/10.56471/slujst.v4i.266
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec

Haisal Dauda Abubakar[1], Mahmood Umar[2]*, Muhammad Abdullahi Bakale[3]
[1]Dept. of Computer Science, Jigawa State Colledge of Education, Gumel, Nigeria.
[2]Dept. of Computer Sceince, Faculty of Science, Sokoto State University, Sokoto, Nigeria.
[3]Dept. of General Studies, Colledge of Agricuture and Animal Science Wurno, Sokoto, Nigeria.
abubakar1986@graduate.utm.my[1], mahmoodumar24@gmail.com[2]*, bakalemuhammad@gmail.com[3]

Abstract

In Sentiment Analysis, there are three (3) approaches namely, machine learning, lexicon-based and ruled based approaches. This study investigates on machine learning approaches which involves text vectorization or word embedding- an essential step in natural language processing tasks since most machine learning algorithms work with numerical input. Text vectorization involves the representation or mapping of words or documents of a corpus to numerical vectors of numbers or real numbers. There are several approaches in the literatures on document/text representation, however this study will focus on three (3) commonly used ones viz; Bag of words, TF-IDF, word2vec and doc2vec, and try to identify the reason behind that for review and recommendation to the researchers in hurry. Review of this study shows that TF-IDF feature vector representations generally outperforms other two (2) vectorization methods word2vec and doc2vec, specifically in book review sentiment classification. And therefore recommended for future studies in book review data set.

Keywords: *Sentiment Classification, Text Vectorization Methods-Bag of Words, Tf-Idf, Word2vec and Doc2vec.*

1. Introduction

Representation of textual contents in formats understandable by computer programs and machine learning algorithms is a vital step in sentiment identification in texts and is formally known as vectorization-the transformation or encoding of texts into numerical vectors for machine learning. In this study, three popular vectorization methods namely; TF-IDF, word2vec and doc2vec has been adopted regards sentiment classification. Review vectorization or word embedding is the representation or mapping of words or documents in a corpus to numerical vectors of real numbers. It's a subset of natural language processing, namely language modeling, that aims to represent words or documents using usable numerical representations. Most machine learning algorithms, which are increasingly being utilized in text categorization and sentiment analysis research, are designed to analyze numerical inputs, therefore this stage is critical. There are numerous approaches to representing documents/texts in the literature; however, only three widely used methods are explored in this study, namely, term frequency, inverse document frequency, word2vec, and doc2vec are the approaches.

The term frequency–inverse document frequency (abbreviated as tf-idf) method is widely used because it can reveal the relative importance of terms in a document or corpus. The curse of dimensionality, data sparsity, and the difficulty to capture semantic links between words in a document are all noteworthy limitations. Word2Vec is a type of neural network language model that uses the capabilities of neural networks to learn distributed representations of words to solve the problem of semantic relatedness and the curse of dimensionality. The term frequency–inverse document frequency (abbreviated as tf-idf) method is widely used because it can reveal the relative importance of terms in a document or corpus. It has a number of flaws, including the curse of dimensionality, data sparsity, and the difficulty to acquire data semantic relationships between words in a document.

Word2Vec belongs to a class of neural network language models that address the problem of semantic relatedness of words and the curse of dimensionality by exploiting the power of neural networks to learn their distributed representations. Word2Vec is the inspiration for a lot of recent word and document embedding techniques. Word2vec models are shallow neural networks with one hidden layer that map words to a lower-dimensional 53 vector space after being trained on a large input corpus. The dimension of the vector space in which closer and distant word vectors reflect similar and dissimilar words, respectively, is represented by the number of neurons in the hidden layer. As a result, many sentiment analysis and text classification tasks can use these word vectors as features. In word2vec, there are two types of neural network models: skip grams and continuous bag-of-words. The skip grams model predicts neighboring context words using a single-layer neural network given an input word. By combining several contextual terms, the continuous-bag-of-words model is similar to the skip-gram model.

Doc2vec, also known as paragraph vector, is a generalization of word2vec to accommodate sequences of words as in sentences, paragraphs, and documents. It sidesteps most of the weaknesses of bag-of-words-based models by representing a sequence of texts with a fixed-length dense feature vector while retaining the order and semantics of words.

2. Related Works

Text mining can be seen as knowledge extraction, or as text data mining (Team, 2019), as well as text mining as a method for knowledge discovery in databases (KDD). Text mining is the automated extraction of new (unknown) information from a variety of written resources by a computer. Text mining is not the same as online searchers looking for something that has previously been documented and written by someone else. It is clear from this that the problem is that all knowledge that isn't relevant to your needs should be ignored. The purpose of text mining is to find unknown information that no one knows about and hasn't been put down yet (Gupta, 2021).

Text Mining is an alteration in an area known as data mining, which tries to find interesting patterns in large databases (Yogapreethi & S, 2016). The processes of extracting interesting and non-trivial information and knowledge in unstructured texts, also called Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in texts (KDT). Text mining is a newly interdisciplinary area that focuses on the recovery of information, data mining, machine learning, statistics and computer linguistics.

2.1 Text Vectorization

Text vectorization or word embedding involves the representation or mapping of words or documents of a corpus to numerical vectors of numbers or real numbers (Prabhu, 2019). Since most machine learning algorithms deal with numerical input, it is a necessary step in machine learning-based natural language processing tasks and sentiment analysis. In the relevant literature, there are other ways for representing documents/text; however, the bag of words, TF-IDF, word2vec, and doc2vec embedding approaches are covered here because they are relevant to the topic at hand.

2.2 Bag of Word

The Bag of Words text vectorization model treats documents as a collection of words, regardless of the grammar or order in which the words are found; hence the name "bag of words." It treats each document as a set of numerical vectors with a fixed length (typically the number of unique words in the corpus) and each feature representing the frequency of occurrence of each word. (Huspi, Abubakar, & Umar, 2021).

For example the three sentences, Sent1, Sent 2 and Sent 3 would respectively be encoded as shown in Table 2.1 as conducted by (Abubakar, 2020):

Sent 1: “Shah is a good writer”
Sent 2: “To be a good writer, you need to practice”
Sent 3: “I enjoyed every bit of it”

2.3 Term frequency–inverse document frequency (TFIDF)

TFIDF is a scaled-down version of the popular bag-of-words game. The product of the word's frequency in the document (term frequency) and the logarithm of the division of the total number of documents in the corpus by the total number of documents in which the word I appears in the corpus represents each word I in a document (inverse document frequency) (Trstenjaka, Mikacb, & Donkoc, 2014). This means multiplying the number of times a word appears in a review text by the logarithm of dividing the total number of reviews in the corpus by the number of reviews in which the term appears when considering review texts as documents. Equation 1.1 is the mathematical formulation for TFIDF.

$$C(t,d) \times log\frac{N_d}{N_{dt}} \qquad 1.1$$

Where (t, d) represents the raw count of a word/term t in a document d, Nd represents the total number of documents in the corpus, and Nd, t represents the number of documents containing the term t. For TF-IDF vector representation, the example in Table 2.1 will be recreated as shown in Table 2.2 using the equation (1.1).

2.4 Word2vec

Word2Vec is part of a family of text vectorization techniques known as embedding, in which a shallow neural network is used to project words into a lower-dimensional numeric vector space depending on their linguistic context. It's a distributed vector representation of words based on the assumption that words with comparable meanings in the same context are represented similarly. As a result, word vectors with similar meanings are clustered together in the vector space. The challenges of high dimensionality and loss of word context that are unique to the bag of words and its n-gram versions are addressed by this family of vectorization approaches. Continuous bag of words (CBOW) and word2vec are two implementations of word2vec. (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and skip grams (T., Sutskever, & K., 2013). Each of this implementation is explained in what follows;

2.4.1 The continuous bag of words (CBOW) is a word2vec architecture that can predict a current or target word based on one or more context words. Each word in the vocabulary is mapped to a unique one-hot encoded vector and stored as columns in the matrix W in this architecture. For each unique word, just one position in each encoded vector of length equal to the vocabulary's size is set to 1, while the others are set to 0. Each word in column W is arranged so that it corresponds to its index or position vocabulary.

To forecast a specific target word, a combination of column vectors matching to a given group of words is used. Figure 2.1 depicts an excellent example of what is happening in CBOW.

2.4.2 Continuous Skip Grams: The CBOW model is the polar opposite of this word2vec architecture. As an example, the model is charged with predicting the context words for a predetermined window of context words given a word as input. In this structure, words that are close together are given more weight. As in CBOW, each word in the vocabulary is mapped to a column matrix W and one-hot encoded. The sole change, as shown in Figure 2.2, is in the model architecture and, as a result, problem formulation.

Table 2.1 *Bag of Word representation of Words (Abubakar, 2020)*

	Shah	Is	a	good	writer	To	Be	you	Need	Practice	I	Enjoyed	every	bit	of	It
Sent 1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
Sent 2	0	0	1	1	1	2	1	1	1	1	0	0	0	0	0	0
Sent 3	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1

Table 2.2 Term-Frequency Inverse-document frequency representation of words (Abubakar, 2020)

	Shah	is	a	good	Writer	to	Be	you	need	practice	I	Enjoyed	every	bit	of	it
Sent 1	0.48	0.48	0.18	0.18	0.18	0	0	0	0	0	0	0	0	0	0	0
Sent 2	0	0	0.18	0.18	0.18	0.45	0.48	0.48	0.48	0.48	0	0	0	0	0	0
Sent 3	0	0	0	0	0	0	0	0	0	0	0.48	0.48	0.48	0.48	0.48	0.48

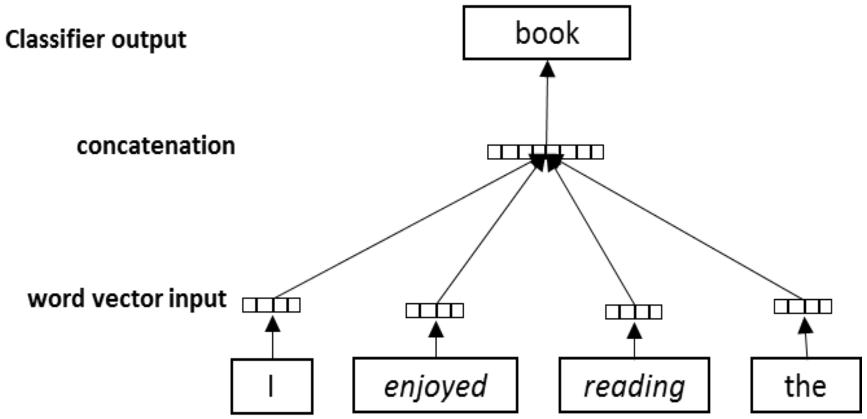

Figure 2.1 Continuous Bag of word model (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013)

2.5 Doc2vec

Doc2vec is an unsupervised learning neural-based approach that learns a numeric vector representation for varying lengths of sequences of sentences, such as paragraphs or a whole document, as described in the original literature (Q & T, 2014). Doc2vec is inspired by the distributed 35 word representation approach used in word2vec, in which word vectors are used to predict the next word in a given context of a sentence by predicting the next word in a document or paragraph given randomly dr. We interchangeably use the terms "document" and "paragraph" to refer to sentence sequences.

Every page in the corpus is mapped to a unique one-hot encoded vector as a column matrix P, in addition to mapping every word in the vocabulary to a unique one-hot encoded vector and storing it as a column matrix W as in word2vec. The document vectors are paired with word vectors or utilized separately, depending on the doc2vec model, to predict the next word or context word in randomly sampled fixed-length contexts from a predetermined window in the paragraph. The two doc2vec models, which are architecturally similar to word2vec, are as follows:

The Distributed Memory Model (DM) (shown in Figure 2.3) is similar to word2vec's CBOW, but with a document vector added. During back propagation neural network training, Word is concatenated with a document vector from P. While document vectors are shared across all documents, a paragraph vector is only shared across all contexts created from the same paragraph. The trained model can then be used to infer document vectors for fresh documents, which can then be fed into additional machine learning algorithms for prediction**.**

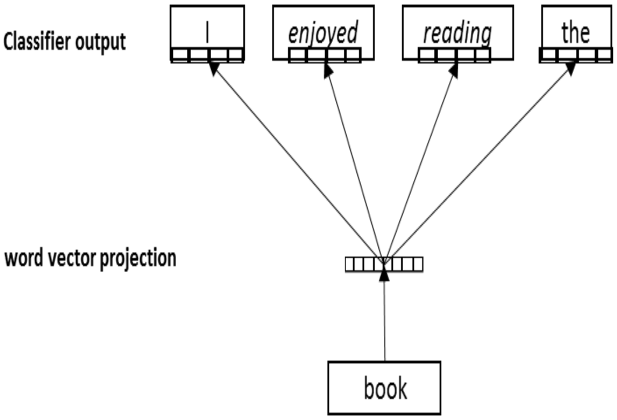

Figure 2.2 Skip Gram Model (T., Sutskever, & K., 2013)

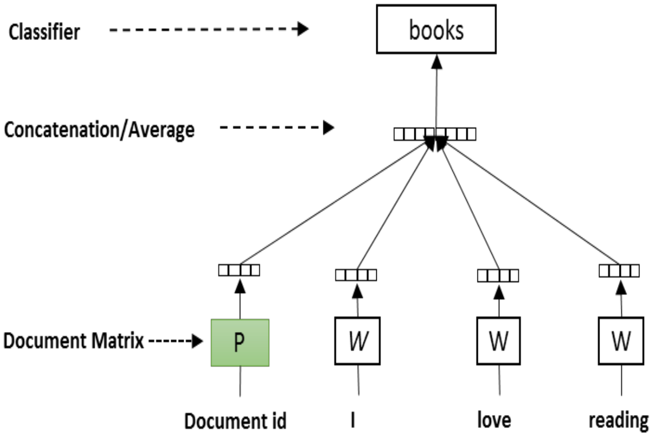

Figure 2.3 Distributed Memory Model (Q & T, 2014)

3. Materials and Methods

As mentioned in section 2. Methodology of this study deals with the gathering and reviewing of relevant literature and definition of the problems this research aims to address. 50 relevant literatures were searched based on the topics title: Sentiment Classification: Review of Word Vectorization Methods-Bag of Words, Tf-Idf, Word2vec and Doc2vec. The search was narrowed based on key words specifications. The keys words considered are: Sentiments analysis, words vectorization methods, bag of words, tf-idf, word2vec and vec2word.

4. Result and Discussions

Previous sentiment analysis research has used word/n-gram bags, word2vec, and doc2vec models to encode sentiment in written native languages. Each of these 40 techniques focuses on various levels of textual granularity that are encoded by generalizing some ideas from lower to higher levels of granularity. As a result, the abstract level of information is represented by the feature vectors created by each approach. Bag-of-words and n-grams, for example, use local representations to focus on words and short phrase sequences; word2vec focuses on the representative representation of words and phrases in text, while doc2vec focuses on words that relate their context to the entire document. The standard strategy is to examine each of these strategies separately or in comparison to two others. Rarely have all three methods presented been used in a single sentiment classification study.

5. Conclusion

Based on the reviews of different studies on the research topic (Sentiment Classification: Review of Word Vectorization Methods-Bag of Words, Tf-Idf, Word2vec and Doc2vec), result of this study shows that TF-IDF feature vector representations generally outperforms other two(2) vectorization methods word2vec and doc2vec, specifically in book review sentiment classification. This can be validated based on the study conducted by (Haisal A. D et al 2021) on the topic: *A Scheme of Pairwise Feature Combinations to Improve Sentiment Classification Using Book Review Dataset.* Combined scheme of TF-IDF-word2vec, TF-IDF-doc2vec, and doc2vecword2vec lead to improved sentiment classification of book reviews relative to single feature vectorization approaches. This study also reports from the previously mentioned author that, combination of TF-IDF-word2vec performed best compared to all other methods either combined or singly. Word level information from TF-IDF combined with contextual information from word2vec resulted in more informative feature vectors. Furthermore, the performance improvement cuts across the four considered evaluation metrics; classification accuracy, precision, recall and f1-score.

Acknowledgment

We want to acknowledge the effort of Dr. Sharin Hazlin Bint Huspi, Department of Computer Science, School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia for her contributions towards the success of this work.

References

Abubakar, H. D. (2020). *A Scheme Of Pairwise Feature Combinations To Improve Sentiment Classification Using Book Review Dataset* . Johor Bahru, Malaysia: Universiti Tecknologi Malysia Digital Librarry.

Al-Amin, M., Islam, M. S., & Uzzal, S. D. (2017). Sentiment Analysis of Bengali Comments with Word2vec and Sentiment information of words. *2017 International Conference on Electric, Computer and Communication Engineering(ECCE)* (pp. 186-190). United States: IEEE.

Bilgin, M., & Senturk, I. F. (2017). Sentiment analysis on Twitter data with semi-supervised Doc2Vec. *Conference: 2017 International Conference on Computer Science and Engineering (UBMK)* (pp. 661-666). UMBK: IEEE.

Chen, Q., & Sokolova, M. (2018, May 1). *Word2Vec and Doc2Vec in Unsupervised Sentiment Analysis of Clinical.* Retrieved from https://arxiv.org/abs/1805.0035: https://arxiv.org/abs/1805.0035

Demidova, L., Klyueva, I., Sokolova, Y., Stepanov, N., & & Tyart, N. (2017). Intellectual Approaches to the Improvement of the Classification Decision Quarlity on the Base of the SVM Classifier. *Procedia Computer Science*, 222-230.

Doaa, M. E.-D. (2016). Enhancement Bag-of-Words Model for Solving the Challenges of Sentiment Analysis. *International Journal of Advanced Computer Science and Applications (IJACSA), 7*, 1.

Gulenko, A., Wallschlager, M., Schmidt, F., Kao, O., & & Liu, F. (2016). Evaluating machine learning algorithms for anomaly detection in clouds. *Proceedings-2016 IEEE International Conference on Big Data*, (pp. 2716-2721).

Gupta, G. (2021, September 27). *introduction-to-data-mining-with-case-studies-by-gkgupta-11*. Retrieved from DIVINE VASTU: https://divinevastu.net/introduction-to-data-mining-with-case-studies-by-gkgupta-11/

Huspi, D. S., Abubakar, H. D., & Umar, M. (2021). A Scheme of Pairwise Feature Combinations to Improve Sentiment Classification Using Book Review Dataset. *International Journal of Innovative Computing, 12*(1), 1. Retrieved from . Retrieved from https://ijic.ut

Jiang, L., Wang, S., & Li, C. &. (2016). Structure extended multinomial naive Bayes. *Information Sciences*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their compositionality. *InProceedings* (p. 26). Curran Associates, Inc.

Nawangsaria, R. P., Kusumaningruma, R., & Wibowoa, A. (2019). Word2Vec for Indonesian Sentiment Analysis towards Hotel . *4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI)* (pp. 360–366). Indonesia: Elsevier.

Prabhu, T. N. (2019, November 11). *Understanding-nlp-word-embeddings-text-vectorization-1a23744f7223*. Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-nlp-word-embeddings-text-vectorization-1a23744f7223

Q, L., & T, M. (2014). Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (pp. 1188-1196). Beijing: Scientific Research.

Srujan, K. S., Nikhil, S. S., Rao, H. R., Karthik, K., & Harish, B. S. (2018). *Classification of Amazon Book Reviews Based on Sentiment Analysis.* Nature Singapore Pte Ltd: Springer.

T., i., Sutskever, & K., I. C. (2013). Distributed Representations of Words and Phrases and Their Compositionality. . *Proc of the 26th International Conference on Neural Information Processing Systems* (pp. 3111-3119). USA: Curran Associates Inc.,.

Team, D. F. (2019, October 14). *Data mining*. Retrieved from Data Flair: https://data-flair.training/blogs/text-mining/

Tripathy, A., Agrawal, A., & & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications,*.

Trstenjaka, B., Mikacb, S., & Donkoc, D. (2014). KNN with TF-IDF Based Framework for Text Categorization. *24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013* (pp. 356 – 1364). DAAM: Elsevier.

Yogapreethi, N., & S, M. (2016, August). A REVIEW ON TEXT MINING IN DATA. *International Journal on Soft Computing (IJSC), 7*, 2/3.

Zhang, X., & & LeCun, Y. (2015). Text Understanding from Scratch. *Text Understanding from Scratch*, 1–9.