A Comprehensive Survey on Text Corpus Representation using Vectorization Techniques

Authors

Aisha Umar Musa[1], Department of computer science
Dr. Ibrahim Abdullahi[2], Department of computer science
Professor A.E okeyinka[3], Department of computer science
Dr Adamu isah[4], Department of computer science
Dr. Adamu I abubakar[5], Department of computer science

Abstract: *- natural language processing is a method under artificial intelligent to teach or train machine to accept, understand, analysis and process human spoken word. However, vectorization or feature engineering is a very important stage in the natural language pipeline. It's the act of converting or representing textual information using numerical digit. However, this paperwork tends to comprehensively survey commonly used vectorization approach, by explicitly stating there most adopted usage, advantage (merits) and disadvantage (drawbacks). Based on the findings its identified that most vectorization approach has unique problems they solve. Thus, the NLP problems as hand determine the vectorization approach*

Index Terms: vectorization, features, pipeline, textual, NLP, artificial, intelligent, machine.

INTRODUCTION

Artificial intelligent thus AI is a broad term indication the application of a machine or computer to behave intelligently with or without human intervention [1]. Artificial intelligent can also be expressed or described as the application of computer software consisting of statistic or mathematical formular, used in processing input information and give result in a predefined manner[2]. Considering the development in world wide web, which as open the door to different means of sharing various form of information across different platform. This includes video, image, audio and event textual data. However, based on all this multimedia content mentioned textual data seem to be the most widely adopted means of communication, hence this attract the focus of many researchers in developing various automated techniques that can be used in understanding and analyzing contextual data

[3]. Natural Language processing is a sub discipline in the are field of Artificial. intelligent. It tends to fill the gap between human spoken language and computers. With natural language processing thus NLP, computers or machine are capable of understanding, analyzing and processing human language. The availability of huge data as increasingly call for the demands of developing natural language processing techniques [4]. Further, this generation produce information of billions of bytes every day and shared across the globe for various reasons. The call for the need of high computational power in other of uncover meaningful insight in textual information is essential. The advent of natural language processing in the field of artificial intelligent as proven many possibility of achieving better interactions or transparency between machines and human [3].

Furthermore, the conversion of textual information or data into vector representation is the approach use by NLP developers to interact with machine, and with the vector or mathematical representation machine can easily process task, understand and solve problem easily. Various approaches as been introduce by researcher, which range from sophisticated or simple approach to more complex methods for textual represent. One of the simple and straight forward approach of representing information is by mapping individual word or vocab to a unique identity number. The process of vectorization have high important in textual data preprocessing, the act enable system to better understand the context of a text data by converting it to a numerical representations [3]. Vectorization is something called feature extraction which is also define as the conversion of textual data into format that can be easily process and recognize by machine. It's required that text is converted into vector for machine learning process to be much easier. Many vectorizations approach has been developed which include, one hot encoding, hashing vectorizer, TFIDF-Vectorizer, Count Vectorizer, word2vec and the likes [5].

RELATED WORK

A. Survey Approaches

The survey paper work a comprehensive survey will be carryout on tradition and modern vectorization or feature extraction approach. We are going to employ contextual based capturing as an approach. Various vectorization techniques as the proposed by researcher in other to efficiently represent word, document contextual information using vectors (thus, list of digits used in representing a sing word, sentence, or even an entire document). However, this survey work will would consider the merit (advantage), limitation (drawbacks) and the most frequently usable aspect of all vectorization approach (thus, from transitional method to the modern vectorization approach). The traditional approach includes the one hot encoding, count vectorizer, bag of words, and TF-IDF Vectorizer while the modern approach includes Word2vec, Skip Gram, Continuous Back of Word, doc2vec, sentient2Vec, transformer etc.

B. Classification of Vectorization Techniques

In natural language processing vectorization or feature extraction is considered as a very import stage for solving any machine leaning problem. Irrespective of how good the algorithm used for modelling if poor features are feed into the machine algorithm, the result will be poor as well. The conversation of textual data or document into numerical format is also called text representation. However, the mathematic depiction of image content, video content and speech signal are easily understood by machine and straightforward, but text data are not straightforward thus conversion are necessary in other to be easily process by machine. Vectorization is been an active research area for the past year. This text representation or vectorization are categorically classified into four

1. The Basic Vectorization Approach
2. The Distributed Approach
3. The Universal Approach
4. And finally, the Hand-Crafted Approach

This rest of this section will comprehensively explain and describe this congeries along

with related paper that adopt each approach or techniques.

1. Basic Vectorization Techniques

Considering basic vectorization where text is directly converted mapped to unique ID number, they are the simplest form of representing text and also easy to implement. In basic approach a N dimensional vector size is use to represent sentence.

a. **One Hot Encoding**: - in one hot encoding, every individual word (W) found in a corpus is assign a unique numeric value call ID, this numeric value ranges from 1, to the size of the vector. In one hot encoding each vector represents each word, and in each vector contain 0 and 1 (thus 0 value for none occurrence of word, while 1 denote the location where a particular occurred) [6].

Considering the paper work of Rodriguez et al, 2018 [7], titled *beyond One-hot encoding: lower dimensional target embedding.* The researcher stated that target encoding play crucial role when learning CNN algorithm, and the use of one hot encoding is the most prevalent strategy due to its simplicity. This widespread encoding approach assumes a label space that is flat. Thus, it does not consider the rich mapping or relationship that exist between labels used during training. However, the researcher converts the target variable embedding into low dimensionality space which drastically improve the speed of convergence, while trying to conserver or preserve the accuracy. In the research work they adopted a normalize eigen representation of the class manifold which encode the variable with little or minimum information loss, and at the same time improve on the accuracy.

b. **Bag of Words: - is** defined as classical representation of textual data which is mostly used in natural language processing. The overall idea behind this approach is to collective represent bag or collection of word without considering the order of the word. Similarly, to BOW is also represent word using a unique ID (set of numeric value representing each word in a document).

(Irawaty et al, 2020) works on a paper work title: *vectorization comparison for sentiment analysis on social media YouTube: a survey.* In the paper work the researcher identifies various continuous bag of word vectorization approach such as Count Vectorizer, Hashing Vectorizer, TF-IDF vectorizer and the like. I the research paper CountVectorization is refers to as a technique used in collecting of textual document in a token count of matrix [5]. This approach not only provide simple way to numerically represent textual document, but also help in encoding new sample document using the predefine vocabulary.

Based on the paper work of Kumar et al., 2020) titled "Vectorization of Text Document for Identifying Unifiable News Article". The researcher tries to identify the effectiveness of text vectorization method, such as the bag of word, tf-idf score, document embedding and word embedding etc. the TF-IDF approach is refers to as Term frequency inverse document Frequency, which is also one of the most commonly used approach in NLP to convert text data into matrix vector representation [3].

2. Distributed Vectorization Techniques

This technique tries to solve the challenge face in the basic vectorization approach. These techniques employed low dimensional way of representing contextual data. This approach also uses neural network architecture to embed data, which is refers to as word Embedding [8].

a. Word Embedding

The word embedding approach map vectors coming from the distributional representation to the vector from the distributed representation. Word embedding is an approach used in capturing the distributional similarity between words [9].

Based on the paper work of (Dang et al., 2020) titled *sentiment analysis based on Depp learning: a comparative analysis.* The researcher indicated various word to vector approach such as Continuous Bag of Word, Skip-Gram, and Fast Text. This approaches are implemented using Glove, Genism [9].

Furthermore, both models are based on probability occurrence of words in proximity to other words. The Skip-gram approach is capable of predicting surrounding word for a given text input. The CBoW approach converge faster to that of Skip-Gram [9].

Based on the paper work of (Alhiwaratkun et al., 2018) titled *Probabilistic Fastest for Multi-Sence Word Embeddings.* The researcher introduces probabilistic fast text which tends to model word embedding capable of capturing many words sense, uncertainty information and sub word structure. however, the researchers adopted Gaussian's mixture density to represent individual words[10].

FINDINGS

i. One Hot Encoding

Usage: the simplest approach in representing textual data. Using in vectorizing textual corpus.

Advantage: it is very simple to implement, can be easily understood.

Limitation: it is inefficient to be computational stored, cannot represent text of sparse length, and it cannot handle out of word problem.

ii. Count Vectorization

Usage: for building a vocabulary of know distinct word and also use to encode new document using existing vocabulary

Advantage: it has fix length encoding for any textual document length, and also easy to implement

Limitation: size of vector increases along with vocabulary size, and cannot handle out of word embedding.

iii. TF-DF Vectorization

Usage: it can be used in search engines to capture document relevance based on a giving query.

Advantage: the score of **computed** IDF is based on vocabulary and this enable dynamic understanding in a changing copra.

Limitation: the vectors of the features is sparse and of high dimension.

iv. Skip-Gram Vectorization Embedding

Usage: it is used to predict surrounding words, for a given word or center word.

Advantage: it represents data using low dimensionality

Limitation: cannot handle out of word problem. And require high computational power

v. CBOW Vectorization Embedding

Usage: used in predicting center word, given the contextual information (thus, surrounding words).

Advantage: low dimensional data and rich in contextual information. approach converge faster to that of Skip-Gram

Limitation: cannot handle out of word vectors.

vi. Fast Text Vectorization Embedding

Usage: they are used to developed a character level embedding.

Advantage: it can handle out of word embedding.

Limitation: only perform efficient with character level matching. Thus, cosine similar between character are closer to that of words

RECOMMENDATION

It is recommended that situation was contextual information are not essential, basic vectorization approach should be implemented due to the nature of their simplicity. However, for NLP application such as entity name recognition, document ranking and the likes, distributed approach are more appropriate. Hence given that they are high computation power

CONCLUSION

in conclusion no correct way of vectorizing, hence the problem or task at hand determined the best vectorization to be adopted. This leads to the development of various vectorization or feature extraction approach been designed every day.

FURTHER STUDIES

Interested researcher can look into the effect of hybridizing this approach. Which will probably solve the limitation identify in using a single vectorization approach

Reference

[1] A. N. Ramesh, C. Kambhampati, J. R. T. Monson, and P. J. Drew, "Artificial intelligence in medicine," vol. 44, no. 0, pp. 334–338, 2004, doi: 10.1308/147870804290.

[2] T. Nadarzynski, O. Miles, A. Cowie, and D. Ridge, "Acceptability of artificial intelligence (AI) -led chatbot services in healthcare : A mixed-methods study," vol. 5, pp. 1–12, 2019, doi: 10.1177/2055207619871808.

[3] A. K. Singh and M. Shashi, "Vectorization of Text Documents for Identifying Unifiable News Articles," vol. 10, no. 7, pp. 305–310, 2019.

[4] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural Language Processing Advancements By Deep Learning: A Survey," pp. 1–23, 2020, [Online]. Available: http://arxiv.org/abs/2003.01200.

[5] I. Irawaty, R. Andreswari, and D. Pramesti, "Vectorizer Comparison for Sentiment Analysis on Social Media Youtube : A Case Study," pp. 69–74, 2020.

[6] C. Seger, "An investigation of categorical variable encoding techniques in one-hot and feature hashing An investigation of categorical variable encoding

techniques in machine learning : binary versus one-hot and feature hashing," 2018.

[7] P. Rodr and S. Escalera, "Beyond One-hot Encoding : lower dimensional target embedding," pp. 1–15, 2018.

[8] R. A. Stein, "An Analysis of Hierarchical Text Classification Using Word Embeddings ☆," 2018.

[9] A. C. Study, "Sentiment Analysis Based on Deep Learning :," 2020.

[10] A. G. Wilson, "Probabilistic FastText for Multi-Sense Word Embeddings," 2018.