

Istanbul Technical University - Science and Letters Faculty  
Mathematics Engineering Program



## Graduation Project

Student Name: Arif Çakır

Student Number: 090190355

Course: MAT4091E

Advisor: Prof. Dr. Atabey Kaygun

Submission Date: May 16, 2023

## Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Word Embedding Techniques</b>	<b>3</b>
2.1	Term Frequency-Inverse Document Frequency . . . . .	4
2.2	Word2Vec . . . . .	4
2.2.1	Skip-Gram Algorithm . . . . .	5
2.2.2	Continuous Bag of Words . . . . .	5
2.3	Global Vectors for Word Representation . . . . .	5
2.4	Bidirectional Encoder Representations from Transformers . . . . .	5
<b>3</b>	<b>Dataset</b>	<b>5</b>
3.1	Information About the Dataset . . . . .	5
3.2	Preparation of the Dataset . . . . .	6
<b>4</b>	<b>Applying Models to the Dataset</b>	<b>6</b>
4.1	TF-IDF . . . . .	7
4.2	Skip-Gram Algorithm . . . . .	7
4.3	Continuous Bag of Words . . . . .	7
<b>5</b>	<b>Analysis and Visualization</b>	<b>8</b>
<b>6</b>	<b>Conclusion</b>	<b>8</b>

## 1 Introduction

Language is a fundamental aspect of human communication and one of the most common ways to pass on information and culture throughout history. Due to its nature of encapsulating information, language is studied by linguists throughout the years. Thanks to this scientific foundation, machine learning, and computer science are also advanced upon this topic. Natural Language Processing (NLP) is the subfield of machine learning that studies the processing of human language by computers. One of the most popular NLP techniques in recent years is Word Embedding, which represents words as vectors in semantic space that allows applying mathematical operations on them to analyse. Word Embedding can be used on various subjects such as text classification, translation, and sentiment analysis with promising results. It is aimed in this paper to study several Word Embedding techniques and their applications. The aim is to create a model by applying various Word Embedding models and with the help of this model, describing the relationships and similarities of languages.

The rest of the paper is organised as follows: In Section 2, information about Word Embedding in general and several Word Embedding techniques are provided. Word Embedding techniques touched on in the paper are Term Frequency (TF), Inverse Document Frequency (IDF), Word2Vec, Global Vectors for Word Representation (GloVe), Bidirectional Encoder Representations from Transformers (BERT). This is followed by Section 3 which introduces the dataset Divine Comedy by Dante Alighieri. After that, several Word Embedding models are applied to the dataset in Section 4, and the outputs of those models are analysed in Section 5. Finally paper ends with some discussion and a conclusion in Section 6.

## 2 Word Embedding Techniques

Word Embedding is one of the most popular natural language processing techniques due to its vector representation of the words. As Agarwal stated, capturing semantic meaning of the words in a vector of text is the ambition of the word embedding techniques (Agarwal,

2022). With respect to that, some of the most popular word embedding techniques will be studied in this section. Those techniques are TF, which counts rarity of words; IDF, which counts rarity of words; Word2Vec, which uses cosine similarity; GloVe, which captures co-occurrence of words; and BERT, family of masked-language models introduced by Google.

## 2.1 Term Frequency-Inverse Document Frequency

Term Frequency (TF) is a word embedding technique which counts the occurrence of the words in a document. TF can be shown as following:

$$TF(i) = \frac{\log(Frequency(i, j))}{\log(TotalNumber(j))}$$

where  $Frequency(i, j)$  is the frequency of a word that occurred in a  $j$  word document and  $TotalNumber(j)$  is the total number of the words in the document.

On the other hand, Inverse Document Frequency (IDF) is practically the inverse of the TF. In this method, algorithm relies on the information that gained from the words which are rarely used. IDF can be shown as following:

$$IDF(i) = \log\left(\frac{TotalNumber(j)}{Frequency(j, i)}\right)$$

where  $Frequency(i, j)$  is the frequency of a word that occurred in a  $j$  word document and  $TotalNumber(j)$  is the total number of the words in the document.

TF-IDF mainly shows the degree of relevancy of word  $i$  in the document  $j$ . As Kırık and Güran stated, TF-IDF does not capture semantic relationship between words, and accepts them as independent values (Kırık and Güran, 2021). Due to TF-IDF's lack of capturing semantic relationship of the words, TF-IDF mainly used to detect stop words.

## 2.2 Word2Vec

Another word embedding model is Word2Vec, which consists of 2 versions: Skip-Gram Algorithm and Continuous Bag of Words. Word2Vec uses cosine similarity metric to measure similarity of the words. If cosine value of two words is 0, then words do not hold similarity. If cosine value of two words is 1, then the words are overlapping.

### **2.2.1 Skip-Gram Algorithm**

One of the Word2Vec model architectures is Skip-Gram Algorithm...

### **2.2.2 Continuous Bag of Words**

The algorithm that does exact opposite of the Skip-Gram Algorithm is called Continuous Bag of Words (CBOW)...

## **2.3 Global Vectors for Word Representation**

Global Vectors for Word Representation (GloVe)

## **2.4 Bidirectional Encoder Representations from Transformers**

Created by researchers at Google in 2018, Bidirectional Encoder Representations from Transformers (BERT) is a family of masked-language models...

# **3 Dataset**

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

## **3.1 Information About the Dataset**

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis

egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

### 3.2 Preparation of the Dataset

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

## 4 Applying Models to the Dataset

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan

eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

#### 4.1 TF-IDF

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

#### 4.2 Skip-Gram Algorithm

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

#### 4.3 Continuous Bag of Words

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

## 5 Analysis and Visualization

## 6 Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.