

# **PREDICTING SIMILARITY IN SENTENCES THROUGH WORD EMBEDDING**

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE AWARD OF THE DEGREE  
OF

**MASTER OF TECHNOLOGY**

**IN**

**INFORMATION SYSTEMS**

Submitted By:

**MAYANK KUMAR MITTAL**

(2K17/ISY/10)

Under the supervision of

**Dr. Anil Singh Parihar**

(Associate Professor, Department of CSE)



**DEPARTMENT OF INFORMATION TECHNOLOGY**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi-110042

2017 - 19

## **CANDIDATE'S DECLARATION**

I, Mayank Kumar Mittal, Roll No. 2K17/ISY/10 student of M.Tech Information Systems, hereby declare that the project Dissertation titled “Predicting Similarity in Sentences through Word Embedding” which is submitted by me to the Department of Information Technology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi

Mayank Kumar Mittal

Date:

## **CERTIFICATE**

I hereby certify that the Project Dissertation titled “Predicting Similarity in Sentences through Word Embedding” which is submitted by Mayank Kumar Mittal, Roll No 2K17/ISY/10 Information Technology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

**Dr. Anil Singh Parihar**

Date:

**SUPERVISOR**

## **ACKNOWLEDGMENT**

I express my gratitude to my major project guide Dr. Anil Singh Parihar, Associate Professor, Department of CSE, Delhi Technological University, for the valuable support and guidance he provided in making this major project. It is my pleasure to record my sincere thanks to my respected guide for his constructive criticism and insight without which the project would not have shaped as it has.

I humbly extend my words of gratitude to other faculty members of this department for providing their valuable help and time whenever it was required.

Mayank Kumar Mittal

Roll No. 2K17/ISY/10

M.Tech (Information Systems)

E-mail: mkmittalofficial@gmail.com

## **ABSTRACT**

In the field of natural language processing, learning the context from a given sentence is a very important and challenging task. Which is great source for predicting the intention of user, this prediction will help to detect the fake NEWS, for creating more interactive artificial intelligent bot that will interact better respond better act better, for giving the better recommendations such as recommending music for that purpose word embedding in used to bridge the gap between the computing machine and the real world. In this research we have examined the various existing models i.e, regression models like multilinear regression, support vector machine, random forest, match LSTM to detect the similar sentences. Also, we compared their results based on accuracy achieved. Moreover, we proposed new model based on convolution neural network warping with time distributed layer which outperform with respect to other models from 77.67% to 83.72%.

## Table of Contents

<b>Title</b>	<b>Page no.</b>
CANDIDATE'S DECLARATION	ii
CERTIFICATE	iii
ACKNOWLEDGEMENT	iv
ABSTRACT	v
Figures and Tables	vii
<b>1. INTRODUCTION</b>	<b>1</b>
<b>1.1 Thesis Organization</b>	<b>3</b>
<b>2.LITERATURE SURVEY</b>	<b>4</b>
<b>2.1. Research and improvement of feature words weight based on     TFIDF algorithm</b>	<b>4</b>
<b>2.2. Inferring Affective Meanings of Words from Word Embedding</b>	<b>5</b>
<b>2.3. Two Improved Continuous Bag-of-Word Models</b>	<b>6</b>
<b>2.4. Weighted Word2vec Based On The Distance Of Words</b>	<b>7</b>
<b>2.5. ConceptVector: Text Visual Analytics via Interactive Lexicon     Building using Word Embedding</b>	<b>8</b>
<b>2.6. Vector Representation of Words for Sentiment Analysis Using     GloVe</b>	<b>8</b>
<b>3. TECHNIQUES &amp; METHODS</b>	<b>9</b>
<b>3.1 Data Analysis</b>	<b>9</b>
<b>3.2 Data Pre-Processing</b>	<b>9</b>
<b>3.3 Corpus Representation</b>	<b>10</b>

<b>3.4 Models</b>	<b>14</b>
<b>4. PROPOSED APPROACH</b>	<b>21</b>
<b>4.1 Data Preprocessing</b>	<b>21</b>
<b>4.2 Training And Prediction</b>	<b>21</b>
<b>5. EXPERIMENTAL RESULTS</b>	<b>24</b>
<b>5.1 Data Analysis</b>	<b>24</b>
<b>5.2 Evaluation</b>	<b>27</b>
<b>6. CONCLUSION AND FUTURE WORK</b>	<b>34</b>
<b>References</b>	

## List of Tables

<b>Table</b>	<b>Title</b>	<b>Page no.</b>
Table 1.1:	Dummy Featurised representation through word embedding	2
Table 2.1:	Comparison of TDIDF and Word Embedding	5
Table 5.1:	Evaluation of Different Techniques	29
Table 5.2:	Evaluation of Proposed Approach CoWe	33



## List of Figures

<b>Fig</b>	<b>Title</b>	<b>Page no.</b>
Figure 1.1:	Google query results	1
Figure 2.1:	Regression method for affective representation	6
Figure 2.2:	Continuous Bag of word network	7
Figure 3.1:	Word Vector Representation	11
Figure 3.2:	Embedding Matrix Generation	12
Figure 3.3:	SVR Hyperplane	15
Figure 3.4:	LSTM Architecture	16
Figure 3.5:	Match LSTM Architecture	17
Figure 3.6:	Convolution Neural Network 1 Dimensional	17
Figure 3.7:	Convolution Sliding Window Working	18
Figure 3.8:	Working of Rectified Linear Unit	20
Figure 4.1:	Proposed Prediction model (CoWe)	22
Figure 4.2:	Convolution Layer	22
Figure 5.1:	Duplicate graph visualization	24
Figure 5.2:	Number of occurrence of words	25
Figure 5.3:	Log-Histogram of sentence appearance counts	25
Figure 5.4:	Normalized Histogram of character count in sentences	26
Figure 5.5:	Normalized Histogram of character count in sentences	26
Figure 5.6:	Plotting word vectors in the 2 dimensional space	28
Figure 5.7:	50 Dimensional train loss vs valid loss CoWe	31

Figure 5.8: 100 Dimensional train loss vs valid loss CoWe	31
Figure 5.9: 200 Dimensions train loss vs valid loss CoWe	32
Figure 5.10: 300 Dimensions train loss vs valid loss CoWe	33