A COMPARATIVE STUDY OF EMBEDDING-BASED TEXTUAL SEMANTIC SIMILARITY TECHNIQUES FOR URDU LANGUAGE

NAZISH KHURSHEED
01-241221-003

A thesis submitted in fulfillment of the
requirements for the award of the degree of
Master of Science (Software Engineering)

Department of Software Engineering

BAHRIA UNIVERSITY ISLAMABAD

FEBRUARY 2024

APPROVAL FOR EXAMINATION

Scholar's Name: Nazish Khursheed Registration No. 01-241221-003

Program of Study: MS (Software Engineering)

Thesis Title: A Comparative Study of Embedding-Based Textual Semantic Similarity Tasks for Urdu Language

It is to certify that the above scholar's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for examination. I have also conducted plagiarism test of this thesis using HEC prescribed software and found similarity index that is within the permissible limit set by the HEC for the MS degree thesis. I have also found the thesis in a format recognized by the BU for the MS thesis.

Principal Supervisor's

Signature: _____

Date: _____

Name: _____

AUTHOR'S DECLARATION

I, Nazish Khursheed hereby state that my MS thesis titled "A Comparative Study of Embedding-Based Textual Semantic Similarity Tasks for Urdu Language" is my own work and has not been submitted previously by me for taking any degree from this university Bahria University Islamabad or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw/cancel my MS degree.

Name of scholar: Nazish Khursheed (01-241221-003)

Date: _____

PLAGIARISM UNDERTAKING

I, Nazish Khursheed, solemnly declare that research work presented in the thesis titled "A Comparative Study of Embedding-Based Textual Semantic Similarity Tasks for Urdu Language" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and Bahria University towards plagiarism. Therefore I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis evenafter award of MS degree, the university reserves the right to withdraw/revoke my MS degree and that HEC and the University has the right to publish my name on the HEC/University website on which names of scholars are placed who submitted plagiarized thesis.

Scholar / Author's Sign: _____

Name of the Scholar: Nazish Khursheed (01-241221-003)

DEDICATION

To my beloved mother and father

ACKNOWLEDGEMENT

I would start by thanking ALLAH Almighty, with gratitude for giving me strength in every aspect of life and helping me in this thesis as well.

I wish to express my sincere appreciation to my thesis supervisor, Dr. Raja Suleman, for his support, guidance, and valuable feedback throughout this thesis. His expertise and encouragement have been instrumental in shaping my research and helping me to overcome the challenges that I faced.

I want to acknowledge my parents, who supported me and. This thesis work is dedicated to my parents, who have been a constant source of support during the challenges of life and prayed for my success. I am truly thankful for the support in every aspect whether that is financial, emotional, or mental.

Lastly, I would like to acknowledge the contributions of all the participants who took part in my study, without whom this research would not have been possible.

ABSTRACT

Textual Semantic Similarity (TSS) evaluates the degree to which two sentences or short texts are semantically proportional to one another. It plays an increasingly important role in tasks such as machine translation, information retrieval and textual forgery detection. TSS is one of the significant problems in the field of Natural Language Processing (NLP). Text reuse and plagiarism detection are famous examples of TSS. TSS could be found several levels, for example, word, sentence, and document level. Existing approaches have relied upon word and sentence level embedding for various languages (English, Arabic, Hindi, Turkish, etc.) to retrieve similarity index. Our research focuses on studying the existing approaches and comparing these for Urdu TSS tasks by using Word, Sentence and Document-level embedding respectively.

Keywords: Word2Vec, Sen2Vec, Doc2Vec, Semantic Similarity

TABLE OF CONTENTS

CHAPTER **TITLE** **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NLP	-	Natural Language Processing
TSS	-	Textual Semantic Similarity
CBOW	-	Continuous Bag-of-Words
SG	-	Skip-Gram
BERT	-	Bidirectional Transformer
OSAC	-	Open Source Arabic Corpus
EGY	-	Egyptian dialect
MSA	-	Modern Standard Arabic
CNN	-	Convolutional Neural Network
TF-IDF	-	Term Frequency-Inverse Document Frequency
SWN	-	SentiWordNet

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

The term NLP stands for Natural Language Processing is a branch of Artificial Intelligence that enables computers to generate, manipulate and understand human language as humans do in their normal life [1]. The fact is computers only understand and work on its own language that is 0's and 1's.To solve this problem the term NLP came into existence in computer science under the domain of Artificial Intelligence [3] [6]. Within very limited time NLP made tremendous progress in understanding and analyzing data in various languages apart from English [2]. One critical aspect of NLP is measuring textual semantic similarity, which plays a pivotal role in several applications, including information retrieval, machine translation, text summarization, and recommendation systems etc. [2][8][13]. The term Textual Semantic Similarity (TSS) means two texts are different in context but having the same meaning e.g. احمد سکول میں پڑھنے جاتا ہے۔ and احمد سکول میں پڑھتا ہے۔,both the sentences have different wordings but exactly having the same meaning. Word2Vec, Sen2Vec, and Doc2Vec are all popular models used in natural language processing (NLP) for generating distributed representations of words, sentences, and documents, respectively.

1.1. Word2Vec

- Model aims to learn continuous vector representations of words from large corpora of text.
- Word2Vec represents each word in a fixed-size vector space, where the position of a word is learned based on its context in the corpus.

- The key idea is that words appearing in similar contexts will have similar vector representations.
- Word2Vec has two primary architectures: Continuous Bag of Words (CBOW) and Skip-gram. CBOW predicts the current word given a context, while Skip-gram predicts the surrounding words given a current word. Both models are trained using a neural network framework [18]. Traditional techniques for representing words in natural language processing (NLP) often rely on one-hot encoding, which leads to high-dimensional and sparse representations. For distributed representations to address this issue, where each word is represented by a dense vector in a continuous vector space Word2Vec should be train to maximize the average log probability of predicting context words given the current word (for Skip-gram) or predicting the current word given context words (for CBOW) [2][5].

Word2Vec can be used for both words and phrases as vectors in the same continuous vector space, the model naturally captures the compositionality of language. The compositionality property enables the model to accurately predict the meaning of complex phrases and sentences based on the meanings of their constituent words and phrases [19].

1.2. Sen2Vec

- Sen2Vec extends the idea of Word2Vec to generate vector representations for entire sentences or phrases. Sen2Vec learns high-quality representations for entire sentences [20].
- Instead of treating each word in isolation, Sen2Vec considers the entire sentence as a context for generating the vector representation [15].
- Sen2Vec embeddings are computationally efficient to compute and store compared to other more complex models. This efficiency makes them practical for real-world applications and large-scale NLP systems. Methods like averaging or pooling are often used to combine word vectors into a single vector representation for the sentence [21] [8].

• Sen2Vec offers a powerful and flexible approach for representing sentences in semantic similarity tasks, providing benefits such as semantic preservation, contextual understanding, generalization, efficiency, interpretability, and ease of use [6][8].

1.3. Doc2Vec

- Doc2Vec, also known as Paragraph Vector, is an extension of Word2Vec [22].
- It is designed to learn fixed-length vector representations for variable-length pieces of text, such as documents or paragraphs.
- Doc2Vec incorporates the idea of paragraph-level context along with word-level context.
- Similar to Word2Vec, Doc2Vec can use both CBOW and Skip-gram architectures, with an additional vector representing the entire document.
- Doc2Vec is useful for tasks like document retrieval, topic modeling, and clustering. Doc2Vec model is used to maximize the average log probability of predicting words in the context of a paragraph given its paragraph vector. The paragraph vector is trained to predict target words in the context of the paragraph. The Paragraph Vector model is scalable and can be applied to large corpora of text data. Moreover, the learned representations generalize well to unseen texts and tasks, making them suitable for a wide range of natural language processing applications. [14] [7] [17].
- Doc2Vec model offers versatile capabilities for representing and analyzing documents in various natural language processing tasks, ranging from classification and clustering to information retrieval and machine translation. Its ability to capture semantic information makes it a valuable tool for a wide range of text analysis applications [23].

1.2. Research Gap

As TSS problems has been addressed for different languages for example, English, Spanish, Turkish, however, TSS of document-level embedding for one of the widely spoken but under resourced language i.e., Urdu has not been inscribed. In our study, we will address this gap for Urdu TSS by applying textual similarity estimation techniques for semantic search on Urdu documents and evaluate the performance of these approaches.

1.3.Problem Statement

Word and Sentence-level embeddings have been used in Textual Semantic Similarity tasks to assign similarity scores to short texts in Urdu language. These techniques only cater to the word and sentence-level scope respectively. Such techniques are constrained in their ability to capture the overall semantics on a document-level. Short texts are documents which are composed of multiple words and sentences that collectively provide meaning to text. A document-level embedding technique is needed to assign semantic similarity scores for short texts in Urdu language.

1.4.Research Questions

RQ 1: What are the different text embedding techniques?

RQ 2: How do different embedding techniques work?

RQ 3: How can text embedding techniques be applied for textual semantic similarity tasks?

RQ 4: How do different text embedding techniques compare to each other?

1.5.Research Objectives

The objective of this study is to compare state-of-the-art word, sentence and document-embedding feature extraction techniques (Word2Vec, Sen2Vec and Doc2Vec) and to detect features that can be used to find TSS for Urdu language at different levels.

1.6. Contribution of the study

Although several methods have been used that measure the textual similarity of sentences, words or documents, but we are using Doc2Vec, Word2Vec and Sen2Vec model on same datasets and eventually compare their results to view clear picture that which model is performing best for low resource language i.e. Urdu. The reason to use these models is, as they are state-of- the art models .This study will provide the complete guidelines for new researchers who want to do TSS tasks on Urdu language regarding which model is performing best for Urdu in NLP relevant to TSS tasks. These models have proven to be effective in capturing semantic meanings of words, sentences, and documents in a continuous vector space, enabling various downstream NLP tasks.

1.7.TSS Tasks

TSS tasks we are considering in our study for Urdu language are;

- Cosine-similarity between two words (how much two words are similar to each other).
- Top similar words (on the basis of title word).
- Semantic search for top similar news (on the basis of given title)

1.8. Outline of this thesis

The organization of this paper is as follows: **Chapter 1** “Introduction” section includes the introduction of the study. **Chapter 2** “Literature Review” section summarizes the related works on embedding techniques. **Chapter 3** “Research Methodology” section explains the whole methodology of this study. **Chapter 4** “Results and Evaluation” section shows the results obtained. Finally, **Chapter 5** “Conclusion” section highlights the key findings and conclusion of this study.

CHAPTER 2

LITERATURE REVIEW

This section gives an overview of the prior research work that has been done in this area. Textual semantic similarity is contributory in talking the special demanding situations faced with the aid of low-aid languages like Urdu. It improves the distinction and efficiency of numerous herbal language processing programs, making digital content more available, applicable, and useful for Urdu knowing communities while contributing to the conservation and observe of the language.

Xin Tang et al., [1] proposed a solution to improve the multilingual semantic textual similarity in low-resource languages by using a shared sentence encoder. This shared encoder, help to attain numerous sentence representations for a sentence in special language-precise semantic space, and make use of them in an ensemble model for higher overall performance in similarity assessment. Experimental outcomes display that their version continually beats most advanced non-MT tactics, and even reach the same performance brand new MT- based totally methods in Spanish mission. Its miles noteworthy that their framework is an established approach to construct multilingual sentence illustration requiring no language unique prepossessing and handmade capabilities.

Ahmed Al-Ani et al., [2] used a useful technique in NLP is word embedding. Word2vec could translate the late Egyptian dialect into Modern Standard Arabic. We can train the Word2Vec model on monolingual data, which overcomes the parallel data problem. Word2vec is based on Continuous Bag-of-Words (CBOW) and Skip-gram. While Skip-gram guesses the context using the dispersed representation of the input word, the CBOW combines the distributed representation of the target word's surrounding terms to try to forecast it. Additionally, Word2vec has proven that it is capable of capturing semantic differences between MSA and EGY without the use of

rules. Although only a small sample of data was used to evaluate the model, it is anticipated that it will also work well with larger samples.

Derry Jatnikaa et al., [5] proposed that Word2Vec can quickly generate word vector representations leveraging architectural techniques like Continuous Bag of Words (CBOW) and Skip Gramm, which may be employed in a broad range of language processing tasks. Similar words are frequently grouped together in blocks and share similar vector values. As a consequence, Word2Vec may determine the value of word similarity from the training of a big corpus. The number of instances a word shows up in the database determined by the window size and vector dimensions used affects the Word2Vec model's similarity value. The generated term has less context if the window size and dimensions vector size are too small. Regarding the larger window size and vector dimension, the probability that the pair will appear increases the more context theword is produced.

Adnen Mahmoud et al., [4] proposed Arabic detection of plagiarism is a challenging task due to the diversity of the Arabic language's features, which include its productivity, linguistic, and expressive structure. On the other hand, a word's ability to have multiple lexical categories in different scenarios permits us to have multiple meanings for the word, which alters the purpose of the sentence. Arabic paraphrase detection in this context enables estimating the degree to which a suspicious Arabic text and basic Arabic text are comparable based on their circumstances. Based on a combination of different Natural Language Processing NLP techniques, such as the TF-IDF method to enhance the recognition of words that are extremely descriptive in each sentence as well as distributed word vector representations using the word2vec algorithm to reduce computational complexity and to maximize the possibility of predicting words according to the context in which they would be used, authors proposed a semantic textual similarity approach for paraphrase identification in Arabic texts. In the end, the Open Source Arabic Corpus (OSAC) was used to evaluate the suggested technique, and the results were encouraging. Apart from the encouraging outcomes they were able to achieve with their suggested strategy, they want to make a number of changes to it in the future. One of these is the usage of a CNN (convolutional neural network) to enhance their method's ability to recognize patterns of statistical significance in the context of phrases.

Karlo Babic et al., [3] concentrated on quantifying the similarity of short texts in terms of semantics. The Word2Vec model, its expansion using NASARI word sense embeddings provided by the Babelfy system, the FastText model, and the classic TF-IDF as the baseline are all tested and compared. When determining the degree of similarity between two brief texts, they mix these representational methods with traditional centroid- and BM25-derived approaches and their adaptations. Word2Vec, NASARI+ Word2Vec, and FastText are algorithms based on deep learning and neural networks that beat the standard TF-IDF model, according to evaluation results using the SICK and Lee datasets. It appears that the core Word2Vec model performs better than its extension despite their attempt to enhance it by including the NASARI dataset as an outside source of knowledge. It is clear that the semantics offered by NASARI do not boost performance in the findings as was expected. The NASARI dataset may not be complete, which could be the cause. They consequently anticipate that this expanded representational model, NASARI + Word2Vec, will perform better with the improved NASARI and Babelfy. This is still an unanswered query that needs to be explored further. All of these findings suggest that there is opportunity for advancement and that novel approaches for determining how comparable short texts are can be defined.

Qufei Chen et al., [14] proposed unsupervised sentiment analysis, a branch of machine learning and natural language processing, uses data without sentiment labelling. Unsupervised sentiment analysis is a part of sentiment analysis that is becoming more and more significant. Sentiment analysis is a cutting- edge examination of unstructured text on its own. They used cutting- edge methods to apply Word2Vec and Doc2Vec unsupervised machine learning models to find the stated feelings in scientific and medical literature. The results of Word2Vec's unsupervised sentiment analysis of the obesity data match those of SentiWordNet. The data obtained from Word2Vec and SentiWordNet outcomes and the Doc2Vec findings somewhat agree. The Welch's test on the same data showed an important distinction among the most favorable and unfavorable SWN sentiment evaluation. The most favorable and most negative SWN evaluations on the Reuters data were not significantly different according to the Welch test, and the Word2Vec and Doc2Vec results did not match SWN's. They draw the conclusion that Word2Vec demonstrated a more trustworthy sentiment analysis than Doc2Vec in the unsupervised sentiment analysis of medical texts. The Welch's test significant results can be used as a gauge of how well Word2Vec

and SWN results match up. It is necessary to do a more thorough analysis of Doc2Vec's efficiency.

Iqra Muneer et al., [10] proposed standard techniques for Cross-Lingual Text Reuse Detection X-TRD and large benchmark corpora are generated for English-Urdu language pair at sentence level obtaining the best results at binary and ternary level. Shahzad Nazir et al., [11] presented research on Urdu Word embedding by using dataset of different categories using word2vec model the results were then compared with state-of-the-art techniques that outperformed. Ihsan et al., [9] conducted comprehensive research on Roman Urdu and Urdu Language for product review in terms of classification techniques, feature extraction and preprocessing the research work was compared with the previous research that uses either a lexicon-based approach or machine learning to find the sentence's polarity.

Sajadul Hassan et al., [7] proposed a Word2vec model used for Urdu word embedding using a skip-gram which is unsupervised neural network model that performs both semantic and syntactic analysis on set of information. W. Wang et al., [16] performed Chines text Keyword extraction that is based on the Doc2Vec word vector and TextRank, as only the internal representation of the text is consider that comes up with most frequent word but this is not a good approach, so a Doc2Vec is consider that represent the vector representation of a word from a training dataset, Doc2Vec method is an improved one as compare to the word2vec that neglects the text information or order of the word in the document. Doc2Vec has two models Distributed Memory (DM) and Distributed Bag of Word (DBOW) that improves the performance for word vector representation. Akef et al., [17] trained a Doc2Vec model on a Persian poems dataset that calculates the cosine similarity of the sentences, verses with highest cosine similarity are consider as a correct answer that improves the performance over 6% as compared to previously used benchmarks.

Santillan et al., [13] generated poems using transformers that are coupled with doc2vec embedding a cosine similarity score is used that choses the best output result based on similarity factors doc2vec uses DBOW algorithm during training the use transformer captures the style of a poem in training set that captures the style of a poem and produce output that belongs to a particular poet similarly cosine similarity ensure good cohesion between output and input.

Wang et al., [16] proposed a hybrid semantic representation with prior external and internal knowledge for word similarity a set of related word is constructed and vector against each set of words is obtained based on small chines dataset using CBOW and GloVe embedding models that increases the stability of the similarity results.

Rahman et al., [12] used the clusters of words for examine the relational similarity of words using word2vec model and performed both extrinsic and intrinsic evaluations by using skip-gram SG and Continuous bag of word (CBOW) techniques of word2vec for Bangla language that gives the best performance.

D. Verma et al., [15] determines the semantic similarity between two small paragraphs by using three similarity functions Euclidean Similarity function, Manhattan Similarity function and the Cosine similarity function out of these Manhattan Similarity function outperformed as compared to other two. Compact the computational complication and the data sparsity problem using word2vec algorithm. The obtained vectors averaged thereafter to make a sentence vector representation (Sen2vec). Then, they applied Convolutional neural network CNN model with different statistic regularities for document modeling and semantic similarity measurement. Compared to word2vec model, they examined the performance of Sen2vec and CNN models for sentence modeling and similarity computation.Sen2vec method was able to bridge lexical gaps and information limit by the use of the average of all word vectors representations and CNN was beneficial to imprisonment more related information and calculate the degree of semantic understanding.

Table 2-1 Reviewed Research Work

Ref.	Year	Techniques Used	Results

[1]	2018	Researchers used multilingual semantic textual similarity in low-resource languages by using a shared sentence encoder.	This framework is an established approach to construct multilingual sentence illustration requiring no language unique prepossessing and handmade capabilities.
[2]	2011	Word2Vec Model	Word2vec has proven that it is capable of capturing semantic differences between MSA and EGY without the use of rules. Although only a small sample of data was used to evaluate the model, it is anticipated that it will also work well with larger samples.
[5]	2019	Word2Vec Model	The larger window size and vector dimension, the probability that the pair will appear increases the more context the word is produced.
[4]	2021	TF-IDF ,CNN & Word2Vec model	Enhanced the recognition of words that are extremely descriptive in each sentence. Word2vec algorithm reduced computational complexity and maximized the possibility of predicting words according to the context in which they would be used.
[3]	2020	Word2Vec, NASARI+ Word2Vec, and FastText algorithm	They conclude that this expanded representational model, NASARI + Word2Vec, will perform better with the improved NASARI and

			Babelfy.
[7]	2021	Doc2Vec Model ,TextRank	Doc2Vec method is an improved one as compare to the word2vec that neglects the text information or order of the word in the document.
[12]	2020	Skip-gram & CBOW	Both models shown more accuracy than previous studies on Bangla language.
[15]	2020	Sen2Vec Model ,CNN	Researchers compared word2vec model, they examined the performance of Sen2vec and CNN models for computation.Sen2vec method was able to bridge lexical gaps and information limit by the use of the average of all word vectors representations and CNN was beneficial to imprisonment more related information and calculate the degree of semantic understanding.
[16]	2020	CBOW and GloVe embedding models.	These models increased the stability of the similarity results for Chinese language.
[17]	2020	Doc2Vec Model	Model on a Persian poems dataset calculates the cosine similarity of the sentences, verses with highest cosine similarity are consider as a correct answer that

			improves the performance over 6% as compared to previously used benchmarks.
[11]	2022	Word2Vec model	Research on Urdu Word embedding by using dataset of different categories using word2vec model the results were then compared with state-of-the-art techniques that outperformed.
[13]		Doc2vec model	Model captures the style of a poem and produce output that belongs to a particular poet, similarly cosine similarity ensure good cohesion between output and input for English language.

CHAPTER 3

RESEARCH METHODOLOGY

3.1. Introduction

This chapter outlines a research framework that utilizes word embedding techniques for TSS tasks exclusively for Urdu language. There are various steps involved in our research process. Initially data is collected, then pre-processed to refine it for models. Then data partitioning into training dataset and testing dataset. Then, model training and finally human evaluation performance.

3.2. Research Design

By nature, the present research is an experimental study and exploratory in nature. It investigates the accuracy of Word2Vec, Sen2Vec and Doc2Vec models to perform TSS tasks at different levels for Urdu language exclusively. Results shown to participant to evaluate the accuracy, what it’s actually for human.

3.3. Data Collection Method

Experiments have been performed on already available datasets i.e. “Reuse Corpus” and “One million Urdu news” datasets for short similar texts and for news of different categories like entertainment, supports, politics etc. respectively. Links for datasets are;

https://data.mendeley.com/datasets/834vsxnb99/3
https://www.research.lancs.ac.uk/portal/en/datasets/urdu-short-text-reuse-corpus-ustrc(5a509221-a313-4c5a-8fde-f0924977701a).html

3.4. Sampling Method

Cluster sampling technique was used for the selection of the sample. In this study, short text dataset have five thousand seven hundred and seventy seven records while another have one lack eleven thousand eight hundred and sixty records. The dataset with 5,777 records was divided into 6 clusters, and then 4 clusters were randomly selected having almost 950 records in each cluster. Random values from selected clusters have shown in Table 3.1.

Table 3.1 Values from small dataset

Clusters	**Values (text on these indices)**
1	32,74,214,246,313,595
3	708,988,1043,2008,2500
5	4507,5001,5608,5768
4	3501,3008,2992,2896,3054

The dataset “One million Urdu news” with 111,860 was divided into 6 clusters having 20,000 records in each, and then 4 clusters were randomly selected. Table 3.2 shows the random values (text on these indices) from selected clusters.

Table 3.2 Values from large dataset

Clusters	**Values (text on these indices)**
1	599,503,490,101,7500,8900,9501,9642,9925
4	54007,66001,42000,66002,51002,53301
3	24000,29997,30000,36000,42000,48000,40001,39902
5	92000,72004,95000,80100,90011,86032,82220

3.5. Experimental Procedure

Word2Vec, Sen2Vec and Doc2Vec models have been implemented by using Gensim library in Python. These models have been implemented individually at different levels for both the datasets to investigate which method performing best, either on all levels or any individual level regarding to the short and long text as well. Word2Vec model have been applied by using both Skip-gram and CBOW techniques on word ,sentence and on document level as well for both the datasets to evaluate whether it performs well only on word level or it can perform better for sentence and document level as well no matters what is the length of the text.

Similarly we performed experiments by using Sen2Vec model by training both Word2Vec and Doc2Vec models at sentence level, as Gensim library does not include Sen2Vec model specifically. All similar experiments have been performed by using

Doc2Vec model as well. We had vector _size = 100 for small dataset and vector_size = 500 for large dataset, while performing experiments for all the models.

Experiments have been performed on all test cases to evaluate the results. Then accuracy of these models have been calculated by using Logistic Regression. Our benchmark is human so a human evaluation framework have been designed to compare the accuracy of models with human evaluation. Results obtained from experiments for both the datasets have been shared with human by hiding the models' name. Both the datasets have been provided to human as well.

3.6. Human Evaluation Procedure

First of all we provided clear guidelines to our participant on how to identify keywords, similar news, and estimate cosine similarity. (We provided the datasets and our test samples to our benchmark). The participant of our study to evaluate the results is a 19th grade Army Officer, Lt.Col Sardar Suhail Khan serving as Commandant in Defense Battle School (DBS) Sialkot. Table 3.3 shows the whole procedure of human evaluation.

Table 3.3 Human Evaluation Procedure

TSS Tasks	Procedure	Description	
Extracting Top Ten Matching Keywords	Read and Analyze	Our participant should read the provided datasets thoroughly. Our participant should list these keywords.	At the end human will rate the accuracy of models according to the
Identifying Similar News	Read and	Participant should search the dataset for	

	Analyze	news articles that he find similar to the title given. Our participant should select and rank the top ten news articles that are most similar to the title.	rating scale. <u>Rating Scale:</u> 1. Very Poor 2. Poor 3. Average 4. Good 5. Excellent
Cosine Similarity	Read and Analyze	Our participant will rate the similarity shown by all models.	

3.7. Variables of the Study

Two types of variables have been used in our experimental study.

- Independent Variable
- Dependent Variable

The independent variable is the variable that is manipulated or controlled, while the dependent variable is the variable that is observed or measured in response to changes in the independent variable. These variables are essential components of research design and analysis, helping researchers understand causal relationships and make informed conclusions about their hypotheses. In this study “Length of the text “is independent and “Models’ accuracy” is dependent variable respectively.

3.8. Data Analysis Method

The data was analyzed by our benchmark by providing him results got from Word2Vec, Sen2Vec and Doc2Vec experimental models which performed on TSS tasks i.e. (i) search top ten similar words, (ii) find cosine similarity between two words and (iii) Search top ten news similar to the title given. Result sheet along with both the datasets were provided to human without mentioning models' names to avoid biased rating from our participant. Rating scale is;

1. Very Poor
2. Poor
3. Average
4. Good
5. Excellent

Human read, analyzed the results of all models and assigned the rating to the models according to their performance, then we compared the accuracy shown by models and what it actually for humans.

CHAPTER 4

RESULTS AND EVALUATION

4.1. Experimental Results

Accuracy of Word2Vec, Sen2Vec and Doc2Vec for the small dataset is being calculated by Logistic Regression, results shown by Logistic Regression for small dataset are in table 4.1.

Table 4.1 Models' accuracy for small dataset

TSS Tasks on Small Dataset	Model output			
	Word2Vec(Skip - gram) (accuracy)	Word2Vec CBOW) (accuracy)	Sen2Vec (accuracy)	Doc2Vec (accuracy)
SS score between two words	0.0104	0.0104	0.0104	0.0281
Semantic search based on a single index	0.7395	0.7395	0.0104	0.0025
Top similar words	0.5104	0.5104	0.0129	0.0026

Table 4.1 shows Word2Vec model with skip-gram and CBOW technique performing same for all TSS tasks, Sen2Vec performed little better than Doc2Vec for top ten similar words and to search top ten similar news while Doc2Vec model shown highest accuracy for cosine similarity task, whereas Word2Vec model has highest accuracy for two TSS tasks i.e. semantic search for top ten news and top ten similar words as well. These results have been visually represented in Figure 4-1.

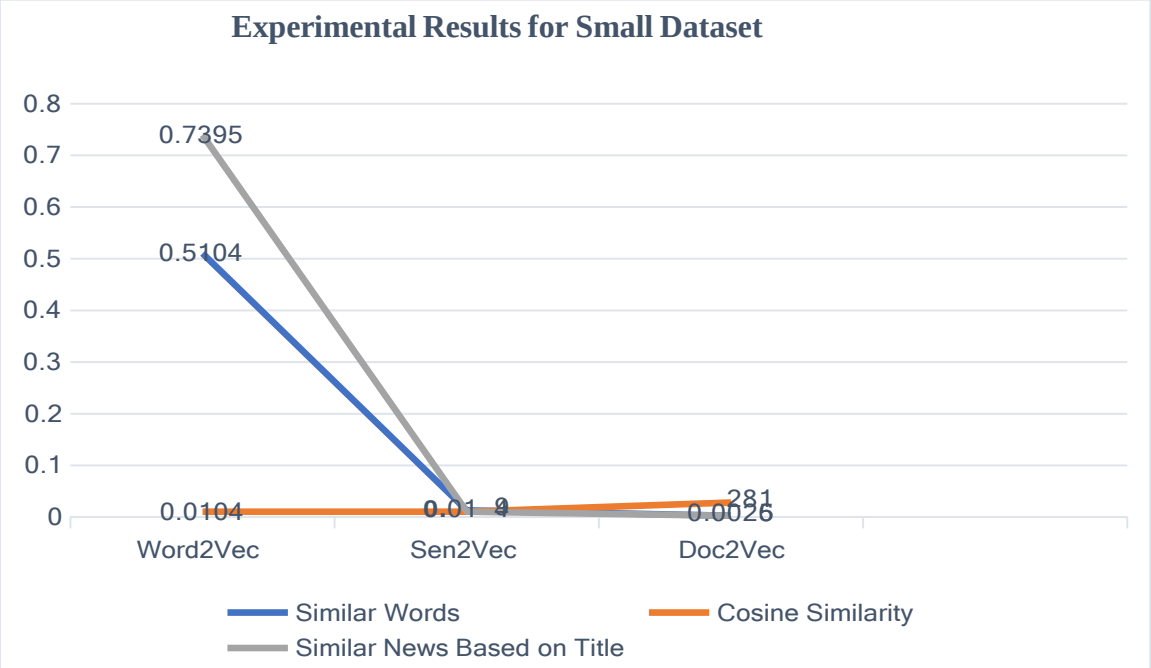

Figure 4-1 Experimental Results for Small Dataset

In figure 4-1 x-axis represents models and y-axis represents accuracy of the models.

Accuracy of Word2Vec, Sen2Vec and Doc2Vec for the large dataset which has 111,860 records is also calculated by Logistic Regression, results shown by Logistic Regression in table 4.2.

Table 4.2 Models' accuracy for large dataset

TSS Tasks on Large Dataset	**Model output**			
	Word2Vec(Skip-gram) (accuracy)	**Word2Vec CBOW) (accuracy)**	**Sen2Vec (accuracy)**	**Doc2Vec (accuracy)**
SS score between two words	0.7401	0.7311	0.4210	0.5462
Semantic search based on a single index	0.7445	0.6425	0.3451	0.4215
Top similar words	0.6211	0.5426	0.5421	0.5322

Table 4.2 shows Word2Vec model with skip-gram and CBOW technique performing almost same for only one TSS tasks i.e. cosine similarity between two words while CBOW technique has performed better than skip-gram technique on large dataset for rest of two TSS tasks.Sen2Vec model shown less accuracy than Word2Vec and Doc2Vec model for all the TSS tasks, whereas Doc2Vec model shown accuracy better than Sen2Vec model for two TSS tasks i.e. semantic search for similar news and cosine similarity task. From the table 4.2, we found Word2Vec model's accuracy is highest (with Skip-gram technique) on large dataset for TSS tasks as compare to Sen2Vec and Doc2Vec model.

Figure 4-2 represents experimental results for large dataset.

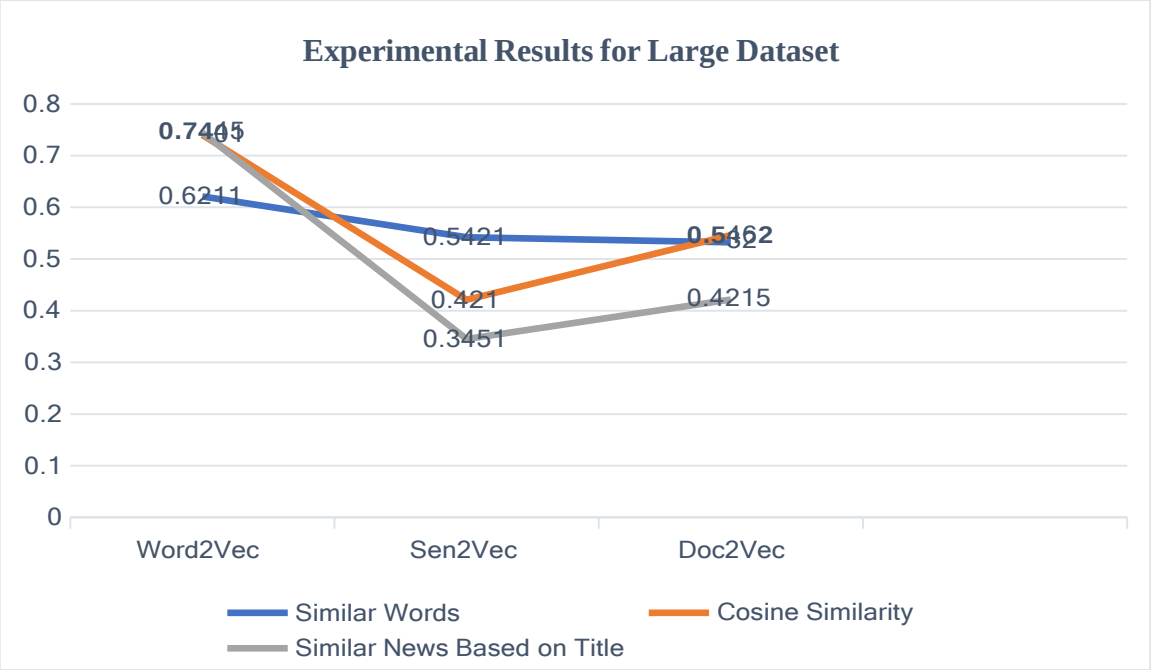

Figure 4-2 Experimental Results for Large Dataset

In figure 4-2 x-axis represents models and y-axis represents accuracy of the models.

4.1.1. Experimental Findings

We found Word2Vec model is performing same with both the techniques i.e. skip-gram and CBOW on small dataset. Experimental results also shown highest accuracy of Word2Vec model as compare to Sen2Vec and Doc2Vec model. In the case of large dataset, we found Word2Vec model is performing better with skip- gram technique than CBOW technique on large dataset for all TSS tasks.Sen2Vec and Doc2Vec models shown less accuracy than Word2Vec model for large dataset as well.

4.2. Human Results

Results obtained from the experiments have been shared with human without showing models' names to evaluate the accuracy of models and to rate them accordingly, rating scale was mentioned earlier in chapter three. Table 4.3 shows results got from human for small dataset.

Table 4.3: Human evaluation results for small dataset

TSS Tasks on small Dataset	**Human Output**			
	Word2Vec (accuracy)	**Sen2Vec (accuracy)**	**Doc2Vec (accuracy)**	**Rating Scale**
SS score between two words	4	2	2	1-5
Semantic search based on a single index	3	5	4	1-5
Top similar words	4	2	2	1-5

4.2.1. Findings for Small Dataset

According to human evaluation Word2Vec model is performing best for two TSS tasks i.e. for cosine similarity and top ten similar words, whereas Sen2Vec showed excellent results for semantic search based on given title (find the top ten similar news).Doc2Vec model performed good for semantic search based on given title but poor progress have been shown for rest of two TSS tasks. Similarly experimental results for large dataset have been shared with our participant.

Table 4.4 showing human evaluation results for large dataset.

Table 4.4 Human evaluation results for large dataset

TSS Tasks on large Dataset	Human Output			
	Word2Vec (accuracy)	Sen2Vec (accuracy)	Doc2Vec (accuracy)	Rating Scale
SS score between two words	4	1	3	1-5
Semantic search based on a single index	4	1	2	1-5
Top similar words	4	4	2	1-5

4.2.2. Findings for Large Dataset

Human rated Word2Vec a best model as compare to Sen2Vec and Doc2Vec model as it has shown good results for all TSS tasks on large dataset, whereas Sen2Vec shown very poor performance for two tasks and good for only one task same Doc2Vec did for large dataset, as it shown average results for only one task and poor results for rest of the two tasks.

4.3. Accuracy Comparison for Small Dataset

Now we will compare the accuracy shown by models and what it actually for humans for both datasets. Table 4.5 shows the whole information regarding small dataset.

Table 4.5 Accuracy Comparison for Small Dataset

TSS Tasks on Small Dataset	**Model Output**				**Human Output**			
	Word2Vec(Skip-gram) (accuracy)	Word2Vec(CBOW) (accuracy)	Sen2Vec (accuracy)	Doc2Vec (accuracy)	Word2Vec (accuracy)	Sen2Vec (accuracy)	Doc2Vec (accuracy)	Rating Scale
SS score between two words	0.0104	0.0104	0.0104	0.0281	4	2	2	1-5
Semantic search based on a single index	0.7395	0.7395	0.0104	0.0025	3	5	4	1-5
Top similar words	0.5104	0.5104	0.0129	0.0026	4	2	2	1-5

4.3.1. Comparison Findings for Small Dataset

Experimental results shown Word2Vec model has highest accuracy for two TSS tasks i.e. semantic search for top similar news and top ten similar words as well,

Sen2Vec performed little better than Doc2Vec for top ten similar words and to search top ten similar news while Doc2Vec model shown highest accuracy for cosine similarity task.

On the other hand human assigned highest rating to Sen2Vec for semantic search of top similar news and Word2Vec got second highest rating for the rest of two TSS tasks. Figure 4-3 shows comparison findings for small dataset.

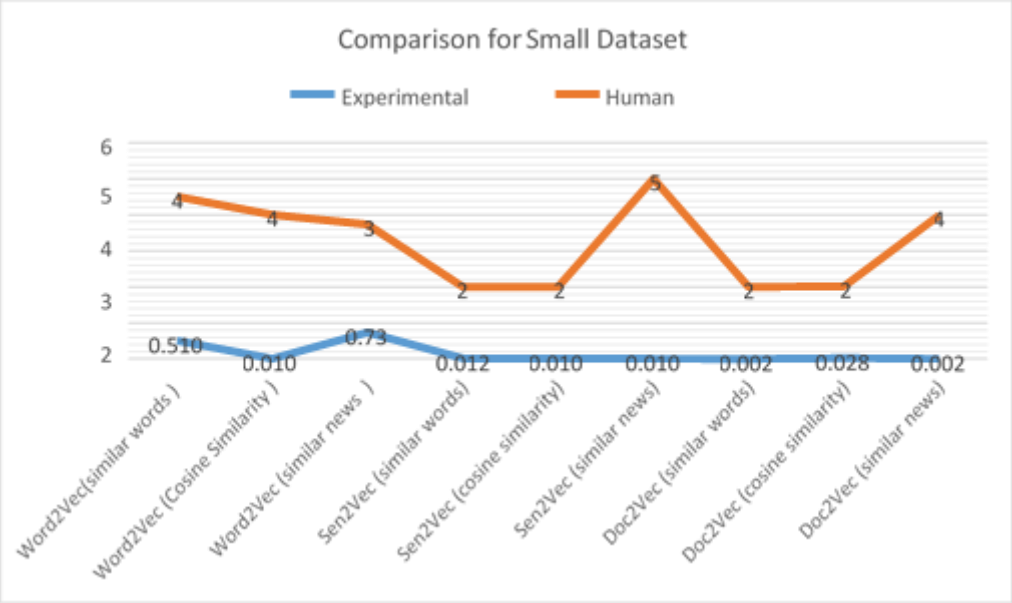

Figure 4-3 Accuracy Comparison for Small Dataset

In Figure 4-3, X-axis represents Models for each TSS tasks and Y-axis represents accuracy of the models.

4.4. Accuracy Comparison for Large Dataset

Accuracy comparison for large dataset regarding three TSS tasks both from experimental point of view and human point of view is shown in table 4.6.

Table 4.6 Accuracy Comparison for Large Dataset

TSS Tasks on Small Dataset	**Model Output**				**Human Output**			
	Word2Vec(Skip-gram) (accuracy)	Word2Vec(CBOW) (accuracy)	Sen2Vec (accuracy)	Doc2Vec (accuracy)	Word2Vec (accuracy)	Sen2Vec (accuracy)	Doc2Vec (accuracy)	Rating Scale
SS score between two words	0.7401	0.7311	0.4210	0.5462	4	1	3	1-5
Semantic search based on a single index	0.7445	0.6425	0.3451	0.4215	4	1	2	1-5
Top similar words	0.6211	0.5426	0.5421	0.5322	4	4	2	1-5

4.4.1. Comparison Findings for Large Dataset

In the case of large dataset, experimental results shown that Word2Vec model is performing better with skip-gram technique than CBOW technique for all TSS tasks.Sen2Vec and Doc2Vec models shown less accuracy as compare to Word2Vec model.

As per human rating, Word2Vec got good performance rating for all three tasks as compare to Sen2Vec and Doc2Vec model.Sen2Vec model got good performance only in search of top similar words rest it had very poor performance, whereas Doc2Vec model got poor performance rating in two tasks and average rating in only one task i.e. cosine similarity. Figure 4-4 shows comparison findings for large dataset.

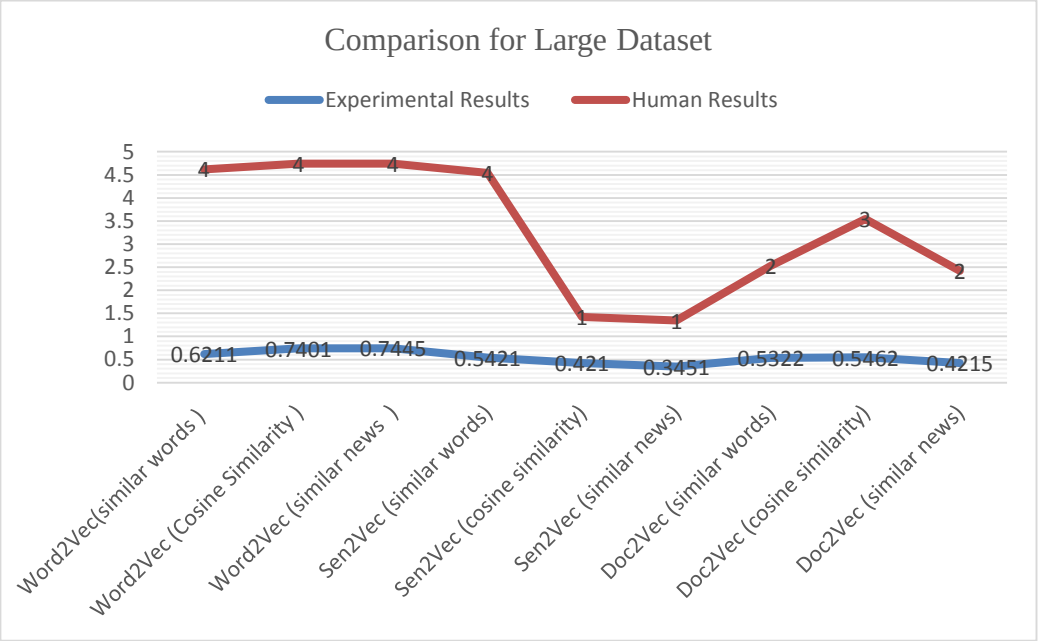

Figure 4-4 Accuracy Comparison for Small Dataset

In Figure 4-4, X-axis represents Models for each TSS tasks and Y-axis represents accuracy of the models.

CHAPTER 5

CONCLUSION

This chapter discusses the conclusion of our work on the accuracy performance of Textual Semantic Similarity techniques i.e. Word2Vec, Sen2Vec and Doc2Vec. In addition, we discussed the key findings of this study. This experimental study was conducted to assess the performance of Word2Vec, Sen2Vec and Doc2Vec modelindividually at different levels regarding TSS tasks for low resource language Urdu. This study is helpful for those who are interested to do TSS tasks for Urdu language without wasting time, as it provides clear direction for researchers, that which model they should use according to their requirement at different levels. In this study independent variable was the length of the document while dependent variable was accuracy of the models.

5.1. Conclusion

Study concluded that Word2Vec model is the state-of the-art model to perform TSS tasks for Urdu language both for small and large datasets as compare to Sen2Vec and Doc2Vec model. For excellent results from small dataset to search top similar news Sen2Vec model is recommended by training Doc2Vec model at sentence level, as Gensim library does not include Sen2Vec separately like Word2Vec and Doc2Vec models are available in Gensim library. For rest of the TSS tasks Word2Vec model is the best choice no matters what the size of dataset. Vector_size and skip-gram technique are the key factors for model’s best performance.

5.2. Key Findings

Study evaluated few key points.

- Vector_size is one of the most important parameter regarding models' performance. For small dataset we had vector_size=100 and vector_size=500 for large dataset. For best results adjust vector_size accordingly.
- Word2Vec model performed best with skip-gram technique as compared to CBOW technique for both datasets.
- In case of Sen2Vec model, for best results train Doc2Vec model at sentence level when need to perform semantic search for top similar news exclusively from small dataset.

REFERENCES

[1] X. Tang, 'Improving multilingual semantic textual similarity with shared sentence encoder for low- resource languages', 2018, ArXiv preprint arXiv: 1810.08740.

[2] A.Ahmaed & Al-Ani, 'Translating dialectal Arabic as low resource language using word embedding', *In International Conference Recent Advances in Natural Language Processing, RANLP,* 2017

[3] K.Babić, F.Guerra, & A. Meštrović, ' A comparison of approaches for measuring the semantic similarity of short texts based on word embeddings', *Journal of Information and Organizational Sciences,* 2020,44(2), 231-246.

[4] A.Mahmoud, & M. Zrigui, 'Semantic similarity analysis for paraphrase identification in Arabic texts', *In Proceedings of the 31st Pacific Asia conference on language, information and computation* (pp. 274-281), 2017.

[5] D.Jatnika, M.Bijaksana, & A.Suryani, 'Word2vec model analysis for semantic similarities in English words', 2019, Procedia Computer Science, 157, 160- 167.

[6] I. Khan, A. Khan, M.Alam, F. Subhan, & M.Asghar, 'A review of Urdu sentiment analysis with multilingual perspective: A case of Urdu and roman Urdu language', 2021.

[7] S. Hassan, H. Kumhar, M. Kirmani, & J. Sheetlani,' Word Embedding Generation for Urdu Language using Word2vec model',2021.

[8] D.Verma, & S. Muralikrishna,' Semantic similarity between short paragraphs using Deep Learning', *In 2020 IEEE International Conference on Electronics, Computing and Communication Technologies* (CONECCT) (pp. 1-5). IEEE.

[9] A.Al-Ani, & A.Almansor, 'Translating dialectal Arabic as low resource language using word embedding', *In International Conference Recent Advances in Natural Language Processing,* RANLP. S. Jack Damico, N. Müller, M. J. Ball, and P. A. Prelock, 'The Handbook of Language and Speech Disorders, Second Edition. Edited Autism Spectrum Disorders', 2021.

[10] I.Muneer, & R.Nawab, 'Cross-lingual text reuse detection at sentence level for English–Urdu language pair', *Computer Speech & Language*, 75, 101381, 2022.

[11] S.Nazir, M. Asif, S. Sahi, S. Ahmad, Y. Ghadi, & M.Aziz, 'Toward the development of large-scale word embedding for low-resourced language', 2022, IEEE Access, 10, 54091-54097.

[12] R.Rahman,' Robust and consistent estimation of word embedding for Bangla language by fine-tuning word2vec model', *In 2020 23rd International Conference on Computer and Information Technology (ICCIT)* (pp. 1- 6). IEEE, 2020.

[13] M. Santillan, & A.Azcarraga,'Poem generation using transformers and Doc2Vec embeddings*', In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7).* IEEE.

[14] Q.Chen, & M. Sokolova, 'Specialists, scientists, and sentiments: Word2Vec and Doc2Vec in analysis of scientific and medical texts', 2021, SN Computer Science, 2, 1-11.

[15] Y.Wang, J. Liu, K. Wang, K., & F. Yin,' A Hybrid Semantic Representation with Internal and External Knowledge for Word Similarity', *In 2020 3rd International Conference on Artificial Intelligence and Big Data* (ICAIBD) (pp. 264-268). IEEE.

[16] W. Wang, & S.Yu,' Chinese text keyword extraction based on Doc2vec and TextRank', *In 2020 Chinese Control and Decision Conference* (CCDC) (pp. 369-373). IEEE.

[17] S.Akef, M. Bokaei, & H. Sameti,' Training Doc2Vec on a Corpus of Persian Poems to Answer Thematic Similarity Multiple-Choice Questions', In *2020 10th International Symposium on Telecommunications (IST)* (pp. 146-149). IEEE.

[18] T. Mikolov, K. Chen, G. Corrado, & J. Dean, ' Efficient Estimation of Word Representations in Vector Space', *arXiv preprint arXiv:1301.3781*, 2013.

[19] T.Mikolov, I.Sutskever, K.Chen, G.Corrado, & J. Dean, 'Distributed Representations of Words and Phrases and their Compositionality', 2013, *Advances in Neural Information Processing Systems*, 3111-3119.

[20] M.Pagliardini, P.Gupta & M. Jaggi, ' Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features', *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 528-540.

[21] W.Nie, J. Zhang, &Z. Li,' Learning Sentence Representation with Neural Network for Semantic Similarity', *IEEE Access*, 7, 11615-11622 ,2019.

[22] Q.Le, & T. Mikolov, T,' Distributed Representations of Sentences and Documents', *Proceedings of the 31st International Conference on Machine Learning* (ICML), 1188-1196, 2014.

[23] J.Lau, , & T. Baldwin,' An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation', *arXiv preprint arXiv:1607.05368,*2016.

APPENDIX A

Experimental Results from Reuse Corpus

Top ten similar words		**Human Rating (1-5)**		**Human Rating (1-5)**		**Human Rating (1-5)**
وائٹ واش	کلاسیفکیش میچز سیریز میچوں میچ زیمبیا ڈے دو ٹیسٹ سیمی		**Fail** (not present in vocabulary)		**Fail**	
آؤٹ	بولڈ نقصان رنز ڈھیر وکٹیں سکی سکور وکٹ سکے بنا		بولڈ ڈھیر' سکور گیندوں چوکوں آوٹ' چھکوں ملہ گیند ئوٹ		بولڈ آوٹ آئوٹ بدر ڈھیر رخصت ہیراتھ چلتا پویلین وکٹ	
نااہلی	صوبائی بندی عراقی برادری سلطانہ راج پناہ رفیق رابطے سنوڈن		**Fail**		کمرہ کیاگیا ہاؤس نگ پروجیکٹر میڈیا، اعلانیہ ملازمت غداری بیرسٹر	
سیریز	میچ ,میچوں ,میچز ,ٹیسٹ ,ڈے		ڈے' میچوں' ون' ٹوینٹی' ٹیسٹ		میچوں میچز آسٹریلیا ون	

	زیمبیا, ون, فائنل, آسٹریلیا, ایونٹ,		میچز میچ' آسٹریلیا' دسمبر پہلا'		ایونٹ ٹونٹی میچ ٹیم بھارت ڈے	

Cosine Similarity

	"روز, ڈے"	0.7826		0.6708		0.1308
	ونگلش,انگلش	0.7932		0.0000		0.0000
	ہوائی, فضائی	0.8921		0.0000		0.0000
	پہاڑ, کوہ پیماہ	0.7812		0.0000		None
	بلے باز, کھلاڑی	0.8545		0.0000		0.0000
	اہلکار, پولیس	0.4232		0.0000		0.0000

Similar News		Human rating
افغانستان	پہلے روز جب کھیل ختم ہوا تو پاکستان نے 4 وکٹوں پر 261 رنز بنا لئے تھے۔ یونس خان 133 اور اسد شفیق 55 رنز پر کھیل رہے تھے۔ وفاقی حکومت نے خصوصی خط کے ذریعے چیئرمین سی ڈی اے سمیت تمام وزراء اعلیٰ، چیف سیکرٹریز، وزیراعظم آزادکشمیر، اور وزیر اعلی گلگت بلتستان کوآگاہ کیاکہ آئندہ ماہ تقریبات یوم آزادی کو بھرپور طریقے سے منانے اور افواج پاکستان کی خدمات بالخصوص آئی ڈی پیزکے ساتھ اظہاریکجہتی کو یقینی بنانے کیلئے خصوصی ہدایات جاری کی گئیں۔, Similarity: 0.633763313293457 Title: پہلے روز جب کھیل ختم ہوا تو پاکستان نے 4 وکٹوں پر 261 رنز بنا لئے تھے۔ یونس خان 133 اور اسد شفیق 55 رنز پر کھیل رہے تھے۔ میڈیا رپورٹ کے مطابق جشن آزاد ی کے موقع پر وزیراعظم نواز شریف نے فیصلہ کیا ہے کہ وہ یوم آزادی پاک فوج اور شمالی وزیرستان کے متاثرین کیساتھ منائیں گے اور ان کیساتھ آج کا دن گزاریں گے جبکہ اس موقع پر کور کمانڈر پشاور لیفٹیننٹ جنرل خالد ربانی اور وزیراعظم نواز شریف آئی ڈی پیز کیمپ میں پرچم کشائی کریں گے تاکہ ملک و قوم کو یہ پیغام دیاجاسکے کہ حکومت اور فوج کے درمیان کوئی خلیج نہیں اور حکومت کو شمالی وزیرستان کے متاثرین کا قوم کیلئے گھروں کوچھوڑنے کی قربانی کا احساس ہے۔, Similarity: 0.6198327541351318 Title: پہلے روز جب کھیل ختم ہوا تو پاکستان نے 4 وکٹوں پر 261 رنز بنا لئے تھے۔ یونس خان 133 اور اسد شفیق 55 رنز پر کھیل رہے تھے۔ میڈیا رپورٹ کے مطابق جشن آزاد ی کے موقع پر وزیراعظم نواز شریف نے فیصلہ کیا ہے کہ وہ یوم آزادی پاک فوج اور شمالی وزیرستان کے متاثرین کیساتھ منائیں گے اور ان کیساتھ آج کا دن گزاریں گے جبکہ اس موقع پر کور	

	کمانڈر پشاور لیفٹیننٹ جنرل خالد ربانی اور وزیراعظم نواز شریف آئی ڈی پیز کیمپ میں پرچم کشائی کریں گے تاکہ ملک و قوم کو یہ پیغام دیاجاسکے کہ حکومت اور فوج کے درمیان کوئی خلیج نہیں اور حکومت کو شمالی وزیرستان کے متاثرین کا قوم کیلئے گھروں کوچھوڑنے کی قربانی کا احساس ہے., Similarity: 0.6198327541351318 Title: پہلے روز جب کھیل ختم ہوا تو پاکستان نے 4 وکٹوں پر 261 رنز بنا لئے تھے۔ یونس خان 133 اور اسد شفیق 55 رنز پر کھیل رہے تھے۔ بھارتی ریاست تلنگانہ کے وزیر اعلیٰ کی بیٹی اور ضلع ناظم آباد سے بھارتی ایوان زیریں لوک سبھا کی رکن کے کویتا نے دو ٹوک الفاظ میں کہا ہے کہ جموں و کشمیر اور تلنگانہ بھارت کاحصہ نہیں ہیں اور بھارت نے ان دونوں علاقوں پر زبردستی قبضہ کررکھا ہے ., Similarity: 0.6126382350921631 Title: پہلے روز جب کھیل ختم ہوا تو پاکستان نے 4 وکٹوں پر 261 رنز بنا لئے تھے۔ یونس خان 133 اور اسد شفیق 55 رنز پر کھیل رہے تھے۔ وفاقی حکومت نے خصوصی خط کے ذریعے چیئرمین سی ڈی اے سمیت تمام وزراء اعلی ،چیف سیکرٹریز،وزیراعظم آزادکشمیر،اور وزیر اعلی گلگت بلتستان کوآگاہ کیاکہ آئندہ ماہ تقریبات یوم آزادی کو بھرپور طریقے سے منانے اورافواج پاکستان کی خدمات باالخصوص آئی ڈی پیزکے ساتھ اظہاریکجہتی کویقینی بنانے کیلئے خصوصی ہدایات جاری کی جائیں ., Similarity: 0.6125782132148743 Title: وزیراعظم کے مشیر برائے قومی سلامتی و امور خارجہ سرتاج عزیز نے وزیراعظم کے خصوصی نمائندہ کی حیثیت سے اتوارکوکابل کا ایک روزہ دورہ کیا۔ انہوں نے وزیراعظم محمد نواز شریف کی طرف سے افغان صدر محمد اشرف غنی احمد زئی کو دورہ پاکستان کے لئے باضابطہ دعوت دی۔ صدر اشرف غنی نے دعوت کی تعریف کی اور کہا کہ وہ جلد پاکستان کا دورہ کریں گے۔ عمران خان کے بیٹوں کو بھی آزادی مارچ میں شامل کرنے کا فیصلہ, Similarity: 0.6111899614334106 Title: پہلے روز جب کھیل ختم ہوا تو پاکستان نے 4 وکٹوں پر 261 رنز بنا لئے تھے۔ یونس خان 133 اور اسد شفیق 55 رنز پر کھیل رہے تھے۔ گذشتہ برسوں میں پاکستان نے دہشت گردی کے خلاف جنگ میں جو رقم خرچ کی ہے وہ امریکہ کے ساتھ ایک معاہدے کے تحت پاکستان کو واپس کی جاتی رہی ہے۔ بی بی سی کے مطابق مانا جا رہا ہے کہ افغانستان سے امریکی فوج کی واپسی کے بعد شاید یہ کولشن سپورٹ فنڈ جاری نہیں رہے اور پاکستان اسے جاری رکھنے کی کوشش کر رہا ہے., Similarity: 0.6070142984390259	
فوج	وزیراعظم نواز شریف ایک روزہ دورے پر گذشتہ روز شمالی وزیرستان کے صدر مقام میران شاہ پہنچے۔ آئی ایس پی آر کے مطابق آرمی چیف جنرل راحیل شریف نے وزیراعظم نواز شریف کا استقبال کیا۔ وزیراعظم کو آپریشن ضرب عضب پر بریفنگ دی گئی۔ **عسکری قیادت** نے بریفنگ کے دوران بتایا کہ آپریشن ضرب عضب کے نتیجے میں دہشت گردوں کی کمر ٹوٹ چکی ہے اور میرانشاہ، میر علی، بویا اور دتہ خیل سمیت 90 فیصد علاقے کلیئر ہو چکے ہیں۔ عمران خان کے بیٹوں کو بھی آزادی مارچ میں شامل کرنے کا فیصلہ, Similarity: 0.7170335054397583	

	Title: وزیراعظم نے لوگوں کو مخاطب کرتے ہوئے کہا کہ یہ آپ کا اور پوری قوم کا فرض ہے کہ وہ ایسے عناصر کو نوٹس لیں جو عوام کی ترقی وخوشحالی برداشت نہیں کرسکتے۔ عوام کسی کو دھرنوں کی اجازت نہ دے۔ سیلاب کی صورتحال کا ذکر کرتے ہوئے وزیراعظم محمد نواز شریف نے کہا کہ دریائے چناب میں سیلاب کی سطح انتہائی بلند ہے جبکہ دیگر دریاؤں میں بھی یہی صورتحال جس سے قیمتی جانوں کے نقصان کے علاوہ فصلوں کو بھی نقصان پہنچا ہے، یہ صورتحال غیر متوقع تھی کیونکہ حالیہ بارش سے 20 تیس سالہ ریکارڈ ٹوٹا ہے تاہم حکومت امدادی سرگرمیوں کے حوالے سے بھرپور اقدامات کررہی ہے۔ وزیراعظم نے کہا کہ سیلاب سے متاثرہ مختلف علاقوں میں آبادی کو کامیابی سے دیگر محفوظ مقامات پر منتقل کیا گیا۔ حکومت اور مسلح **افواج ریسکیو اور امدادی سرگرمیو**ں میں پوری طرح سے مصروف عمل ہیں۔ وزیراعظم نے کہا کہ دھرنوں سے کہیں زیادہ ضروری ہے کہ ایسی صورتحال میں حکومت کے ساتھ ان کوششوں میں ہاتھ بٹایا جائے۔ قبل ازیں وزیراعظم نے وزیر دفاع خواجہ محمد آصف کے ہمراہ علاقے کا دورہ کیا اور ضلع میں حالیہ بارشوں اور سیلاب سے ہونے والے نقصان کا فضائی جائزہ لیا۔ اس موقع پر علاقے کے منتخب نمائندے، ضلعی انتظامیہ کے افسران اور دیگر اعلیٰ حکام بھی موجود تھے۔ عمران خان کے بیٹوں کو بھی آزادی مارچ میں شامل کرنے کا فیصلہ, Similarity: 0.6891546845436096 Title: ڈپٹی اٹارنی جنرل خواجہ سعید ظفر نے پیش ہوکر کہا کہ اٹارنی جنرل سلمان اسلم بٹ ملک سے باہر ہیں لہذا مقدمے کی کارروائی ملتوی کی جائے۔ درخواست گزار گوہر نواز سندھو نے اپنے دلائل میں سابق وزیر اعظم یوسف رضا گیلانی اور محمد اظہر صدیقی کے کیسز کے فیصلو ں کا حوالہ دیا تو ڈپٹی اٹارنی جنرل نے آیت اللہ، ڈاکٹر عمران اور لیاقت حسین کے کیسوں کا حوالہ دیتے ہوئے کہا کہ 18 ویں ترمیم سے پہلے عدالت ان آئینی مندرجات کی تشریح کر چکی ہے تاہم آج ایک بارپھران آئینی شقوں کی تشریح لازم ہوچکی ہے۔ ان کاکہناتھا کہ یوسف رضا گیلانی کی نا اہلی کا معاملہ وزیر اعظم نواز شریف کے کیس سے مختلف تھا، سابق وزیراعظم یوسف رضا گیلانی کو عدالت نے سزا دے دی تھی۔ عمران خان کے بیٹوں کو بھی آزادی مارچ میں شامل کرنے کا فیصلہ, Similarity: 0.6771401166915894 Title: چینی صدر پاکستان کیلئے سرمایہ کاری کے متعدد منصوبے لے کر آرہے تھے لیکن دھرنوں کی وجہ سے ان کا دورہ پاکستان منسوخ ہو گیا ہے۔ حکومت ہر سیلاب سے متاثرہ لوگوں کی ہر ممکن امداد کرے گی۔ انہوں نے انتظامیہ کو بھی ہدایت کی کہ وہ متاثرہ علاقوں میں ریلیف کے کام کو مزید تیز کریں اور متاثرین کی امداد کیلئے ہرممکن اقدامات کریں۔ وزیراعظم کو بتایا گیا کہ سیلاب سے وسیع اراضی پر کھڑی فصلوں کو نقصان پہنچا ہے۔ وزیراعظم نے طوفانی بارشوں اور سیلاب سے جانی و مالی نقصان کی آئندہ چوبیس گھنٹے میں رپورٹ طلب کر لی ہے۔ وزیراعظم نے ہدایت کی کہ بارشوں اورسیلاب کی وجہ سے جاں بحق ہونے والوں کے لواحقین کو پانچ لاکھ روپے کی امداد کے علاوہ زرعی اجناس اور دیگر ضروریات زندگی کا سامان فوری فراہم کی جائیں۔ سمبڑیال سے نامہ نگار کے مطابق وزیراعظم نے وفاقی وزیر دفاع خواجہ	

	محمد آصف کے ساتھ سمبڑیال کے علاقہ بیلہ میں دریائے چناب کے سیلابی پانی سے متاثرہ علاقوں بھگل غربی، بھگل شرقی، رندھیر، دوڑ، بکھڑیوالی، جمال پور، پیرکوٹ، حسین پور، چاؤ کے کلاں، چاؤکے خورد، رانا بہرام، نوگراں، کوٹ دینہ، لالے والی، کلووال، دوبرجی چندا سنگھ سمیت دیگر علاقوں کا فضائی جائزہ لیا، سیالکوٹ انٹر نیشنل ایئرپورٹ پر انتظامیہ کو ہدایت دیتے ہوئے کہا کہ متاثرین سیلاب کے نقصان کا جائزہ لے کر فوری رپورٹ پیش کی جائے تاکہ نقصان کا ازالہ کیا جا سکے۔ انہوں نے کہاکہ حکومت متاثرین کی مکمل بحالی تک چین سے نہیں بیٹھے گی۔ نواز شریف عمران خان کے بیٹوں کو بھی آزادی مارچ میں شامل کرنے کا فیصلہ, Similarity: 0.6743393540382385 Title: چینی صدر پاکستان کیلئے سرمایہ کاری کے متعدد منصوبے لے کر آرہے تھے لیکن دھرنوں کی وجہ سے ان کا دورہ پاکستان منسوخ ہو گیا ہے۔ حکومت ہر سیلاب سے متاثرہ لوگوں کی ہر ممکن امداد کرے گی۔ انہوں نے انتظامیہ کو بھی ہدایت کی کہ وہ متاثرہ علاقوں میں ریلیف کے کام کو مزید تیز کریں اور متاثرین کی امداد کیلئے ہرممکن اقدامات کریں۔ وزیراعظم کو بتایا گیا کہ سیلاب سے وسیع اراضی پر کھڑی فصلوں کو نقصان پہنچا ہے۔ وزیراعظم نے طوفانی بارشوں اور سیلاب سے جانی و مالی نقصان کی آئندہ چوبیس گھنٹے میں رپورٹ طلب کر لی ہے۔ وزیراعظم نے ہدایت کی کہ بارشوں اورسیلاب کی وجہ سے جاں بحق ہونے والوں کے لواحقین کو پانچ لاکھ روپے کی امداد کے علاوہ زرعی اجناس اور دیگر ضروریات زندگی کا سامان فوری فراہم کی جائیں۔ سمبڑیال سے نامہ نگار کے مطابق وزیراعظم نے وفاقی وزیر دفاع خواجہ محمد آصف کے ساتھ سمبڑیال کے علاقہ بیلہ میں دریائے چناب کے سیلابی پانی سے متاثرہ علاقوں بھگل غربی، بھگل شرقی، رندھیر، دوڑ، بکھڑیوالی، جمال پور، پیرکوٹ، حسین پور، چاؤ کے کلاں، چاؤکے خورد، رانا بہرام، نوگراں، کوٹ دینہ، لالے والی، کلووال، دوبرجی چندا سنگھ سمیت دیگر علاقوں کا فضائی جائزہ لیا، سیالکوٹ انٹر نیشنل ایئرپورٹ پر انتظامیہ کو ہدایت دیتے ہوئے کہا کہ متاثرین سیلاب کے نقصان کا جائزہ لے کر فوری رپورٹ پیش کی جائے تاکہ نقصان کا ازالہ کیا جا سکے۔ انہوں نے کہاکہ حکومت متاثرین کی مکمل بحالی تک چین سے نہیں بیٹھے گی۔ نواز شریف عمران خان کے بیٹوں کو بھی آزادی مارچ میں شامل کرنے کا فیصلہ, Similarity: 0.6743393540382385	
تفصیلات کے مطابق	. عمران خان کے بیٹوں کو بھی آزادی مارچ میں شامل کرنے کا فیصلہ, Similarity: 0.6058599948883057 Title: جمعرات کو آئی سی سی کی طرف سے جاری کردہ بیان کے مطابق سالانہ سیمینار میں ایلیٹ پینل آف آئی سی سی ایمپائرز کے تمام 12 اراکین نے شرکت کی۔ ورکشاپ میں ایلیٹ پینل آف آئی سی سی کے سات میچ ریفریز، انٹرنیشنل پینل آف آئی سی سی ایمپائرز کے تین اور چار ایمپائر کوچز شامل تھے۔ عمران خان کے بیٹوں کو بھی آزادی مارچ میں شامل کرنے کا فیصلہ, Similarity: 0.5847100019454956 Title: وزیراعظم نواز شریف ایک روزہ دورے پر گذشتہ روز شمالی وزیرستان کے صدر مقام میران	

	شاہ پہنچے۔ آئی ایس پی آر کے مطابق آرمی چیف جنرل راحیل شریف نے وزیراعظم نواز شریف کا استقبال کیا۔ وزیراعظم کو آپریشن ضرب عضب پر بریفنگ دی گئی۔ عسکری قیادت نے بریفنگ کے دوران بتایا کہ آپریشن ضرب عضب کے نتیجے میں دہشت گردوں کی کمر ٹوٹ چکی ہے اور میرانشاہ، میر علی، بویا اور دتہ خیل سمیت 90 فیصد علاقے کلیئر ہو چکے ہیں۔ عمران خان کے بیٹوں کو بھی آزادی مارچ میں شامل کرنے کا فیصلہ, Similarity: 0.5229158997535706 Title: آسٹریلیا کے خلاف سیریز کیلئے پاکستان کی ون ڈے اور ٹی ٹونٹی کرکٹ ٹیمیں یو اے ای پہنچ گئیں پاکستان کی ٹی ٹونٹی اور ون ڈے کرکٹ ٹیمیں آسٹریلیا کے خلاف ٹی ٹونٹی اور ون ڈے کرکٹ سیریز کھیلنے کیلئے یو اے ای پہنچ گئیں۔ عمران خان کے بیٹوں کو بھی آزادی مارچ میں شامل کرنے کا فیصلہ, Similarity: 0.5011909604072571 Title: ڈی جی آئی ایس پی آرکے مطابق آرمی چیف نےکور ہیڈ کوارٹرز کراچی کادورہ کیاجہاں کور کمانڈرلیفٹیننٹ جنرل نوید مختار اور ڈی جی رینجرز سندھ میجر جنرل بلال اکبر نے ٹارگٹڈ آپریشن کے حوالے سے بریفنگ دی اورشہرمیں امن وامان کی تازہ ترین صورتحال سے آگاہ کیاجبکہ اس موقع پرڈی جی آئی ایس آئی لیفٹیننٹ جنرل رضوان اختربھی موجودتھے .جنرل راحیل شریف نے کراچی آپریشن میں پیشرفت پراظہاراطمینان کرتے ہوئے دہشت گردوں کیخلاف کارروائیاں مزید تیز کرنے کی ہدایت کی اورکہاکہ ملکی معیشت کی بحالی اورعوام کی خوشحالی کیلئے کراچی میں امن وامان اوّلین ترجیح ہے ،آپریشن کے باعث ٹارگٹ کلنگ اور بھتہ خوری کے واقعات میں کمی ہوئی ہے ، عمران خان کے بیٹوں کو بھی آزادی مارچ میں شامل کرنے کا فیصلہ, Similarity: 0.48905420303344727 Title: وزیراعظم محمد نواز شریف نے پیر کوآزادکشمیر کے علاقہ راولاکوٹ کا دورہ کیا اور حالیہ موسلادھار بارشوں کے باعث ہونے والے نقصانات اور امدادی کارروائیوں کا خود جائزہ لیا۔ آزاد جموں و کشمیر کے صدر سردار یعقوب خان، وزیراعظم آزاد و جموں کشمیر چوہدری عبدالمجید، اور سابق وزیراعظم آزاد جموں و کشمیر راجہ فاروق حیدر ان کے ہمراہ تھے۔ راولاکوٹ پہنچنے کے فوراً بعد انہیں چیف سیکرٹری آزادجموں و کشمیر نے بارشوں اور لینڈ سلائیڈنگ سے ہونے والے نقصانات سے آگاہ کیا۔ عمران خان کے بیٹوں کو بھی آزادی مارچ میں شامل کرنے کا فیصلہ, Similarity: 0.4862251281738281 Title: پاکستان کرکٹ بورڈ (پی سی بی) نے مایہ ناز آل راؤنڈر شاہد آفریدی کو قومی ٹی ٹونٹی کرکٹ ٹیم کا آئی سی سی ورلڈ ٹی ٹونٹی کرکٹ ٹورنامنٹ 2016ء تک کپتان مقرر کردیا، مصباح الحق کو اگلے سال ورلڈ کپ تک ون ڈے اور ٹیسٹ ٹیم کا کپتان برقرار رکھنے کے فیصلے کی توثیق کی گئی ہے۔ عمران خان کے بیٹوں کو بھی آزادی مارچ میں شامل کرنے کا فیصلہ, Similarity: 0.480886429548263555 Title: پہلے روز جب کھیل ختم ہوا تو پاکستان نے 4 وکٹوں پر 261 رنز بنا لئے تھے۔ یونس خان 133 اور اسد شفیق 55 رنز پر کھیل رہے تھے۔ راولپنڈی سٹیڈیم میں راولپنڈی ریمز اور پشاور پینتھرز کی ٹیموں کے درمیان میچ کھیلا جائے گا ، ملتان سٹیڈیم میں پاکستان واپڈا	

	اور زرعی ترقیاتی بینک لمیٹڈ کی ٹیموں کے درمیان ٹکراؤ ہو گا ، لاہور مین ایکل سی سی اے میں لاہور لائنز اور ملتان ٹائیگرز کی ٹیمیں مد مقابل ہونگی، فیصل آباد اقبال سٹیڈیم میں نیشنل بینک آف پاکستان اورسوئی نادرن پائپ لائن لمیٹڈکی ٹیموں کا آمنا سامنا ہوگا جبکہ قذافی سٹیڈیم لاہور میں یونائٹڈ بینک لمیٹڈ اور پورٹ قاسم کی ٹیموں کے درمیان میچ کھیلا جائے گا۔, Similarity: 0.4726278483867645	
سری لنکا	not present in vocabulary)	

APPENDIX B

Experimental Results from One Million Urdu News

Top ten similar words		Human Rating (1-5)		Human Rating (1-5)		Human Rating (1-5)
تاریخ	0.7797 قیمتوں 0.7611 قمیت 0.7168 قیمتیں 0.6864 نرخ 0.6618 قمت 0.6396 نرخوں 0.6158 کھپت 0.6086 بوری 0.6051 قدر 0.5768 پیسے		داستانیں 0.3984' 'تاریخیں 0.3941 بےتاریخ' 0.3749 دنیا' 0.3628 **1150000** **0.3462** دنیائے' 0.3448 انعامی' 0.3427 پالیسی' 0.3417 ریخ', 0.3394 داستان' 0.3355 ,		کلبیورپی' 0.3790 جیتیٹائمز' 0.3772 روایت 0.3578 داستانیں' 0.3517 مدت" 0.351 , تاریخینپی' 0.34143 دنیا, 0.3396 تاریخیں', 0.3327 زبانکوئی 0.3321 ' ویدربھا' 0.3309	
"ٹیسٹ"	Similarity: 0.5272 سٹریلیا 0.5065 اسٹریلیا 0.4923 انگلینڈ 0.4723 ڈے		اسٹریلیا" 0.4900 سٹریلیا' 0.4866 میچ' 0.4812 'انگلینڈ', 0.4727		انگلینڈ 0.5118 اسٹریلیا 0.5019 میچ 0.4993 سٹریلیا 0.4884 پاکستان , 0.4832	

	0.4494 زمبابوے 0.4366 شارجہ 0.4305 میچ 0.4259 ٹسیٹ 0.4112 کیویز 0.4085 ٹوئنٹی		'پاکستان' 0.4664 'ڈے' 0.4569 انٹرنیشنل' 0.4410 زمبابوے, 0.4272 روزہ, 0.4267 'کیویز 0.4244		مصباح 0.4483 ٹوئنٹی 0.4428 انٹرنیشن 0.4230 ڈے 0.4158 ایشز 0.41503	
جوان	Similarity: 0.4783 بوڑھے 0.4352 بوڑھا 0.4177 ذہین 0.4033 جنگجو 0.3950 تندرست 0.3948 معذور 0.3811 سپاہی 0.3810 رسیدہ 0.3807 باہمت 0.3748 پیما		بوڑھے' 0.4345 بوڑھا 0.4166 جنگجو 0.3791 معذور 0.3617 جوانوں 0.3565 گائینلیکن 0.3558 توانا 0.3537 متحرک 0.3507 ذہین 0.3490 سوار 0.3479		بوڑھا 0.3894 بوڑھے'' 0.3784 ہزارفٹ 0.3521 جودہ 0.3502 مغرور'' 0.3452 گائینلین 0.3443 جوانوں 0.3431 گونگے 0.3401 سپاہی 0.3361 لڑکیاں' 0.3301	
ممالک	Similarity: 0.8719 ملکوں 0.6363 خطوں 0.6323 ریاستوں 0.5260 شعبوں		ملکوں' 0.8758 خطوں 0.5707 ریاستوں 0.5653 شہروں 0.5552 شعبوں		Fail	

	0.5174 معیشتوں 0.5107 ثقافتوں 0.5104 ریاستیں 0.4995 قوموں 0.4947 شہروں 0.4926 مذاہب		0.5251 ملک 0.5122 ٹیموں 0.5054 'معیشتوں 0.4945 علاقوں, 0.4813 صوبوں' 0.4624			

Cosine Similarity		Human Rating (1-5)		Human Rating (1-5)		Human Rating (1-5)
اجلاس ، قرضوں	0.2613		0.0000		0.2968	
"پاکستان " "اسلام باد	0.6241		0.0000		0.0000	
روپے , نوٹ	0.8257		0.0000		0.2344	
فیشن ویک ,	0.4121		0.0000		0.0000	
پوزیشن , پہلی	0.2956		0.0000		-0.0667	

Title	Results	Human Rating
ڈبلیو ڈبلیو ای چیمپئن شپ	بروک لیسنر سے جڑے دلچسپ حقائق ڈبلیو ڈبلیو ای چیمپئن کو چوروں نے لوٹ لیا کا بہترین ریسلر کس کو قرار 2017 دیا گیا البرٹو ڈیل ریو ڈبلیو ڈبلیو ای چھوڑنے کے لیے تیار اے جے اسٹائلز اچانک ڈبلیو ڈبلیو ای چیمپئن ...کیسے ڈبلیو ڈبلیو ای میں طویل عرصہ بادشاہت قائم ...رکھن سال بعد رومن رینز کا خواب پورا رومن رینز کی معطلی کے بعد ڈبلیو ڈبلیو ای ...میں و	

	رومن رینز کی ہفتوں بعد ڈبلیو ڈبلیو ای میں واپسی رومن رینز اور سیٹھ رولنز کو سابق دوست سے شکست بروک لیسنر ڈبلیو ڈبلیو ای کے چند بڑے سپر اسٹارز اے جے اسٹائلز دنیا بھر میں معروف ریسلر اور ڈبل موجودہ ڈبلیو ڈبلیو ای چیمپئن اے جے اسٹائلز کو ایک معروف ریسلر نے اپنے حالیہ کردار سے دلبرداش جان سینا کی واپسیڈبلیو ڈبلیو ای میں لگ بھگ ماہ ڈبلیو ڈبلیو ای میں طویل عرصہ بادشاہت قائم رکھن ڈبلیو ڈبلیو ای کے مقبول ترین ریسلرز میں سے ایک ڈبلیو ڈبلیو ای ورلڈ ہیوی ویٹ چیمپئن شپ کے لیے ریسل مینیا کے قریب تے ہی ڈبلیو ڈبلیو ای ورلڈ ہ ڈین امبروز نے سیٹھ رولنز اور رومن رینز کے چیلن	
1992 کا ورلڈ کپ	انڈر 19 کرکٹ ورلڈ کپ قومی ٹیم شرکت کیلئے لے جوہان کے راوانا ورلڈ کپ ہاکی ٹورنامنٹ سے شروع ہوگا ئی سی سی نے مینز کرکٹ ورلڈ کپ سپر لیگ کا باضاب کرکٹ ورلڈ کپ 2019 کا کانٹ ڈان شروع اسٹریٹ چلڈرن ورلڈ کپ 2018 پاکستان سیمی فائنل ایشیا کپ کرکٹ کی تاریخ پر ایک نظر عالمی کپ ہاکی ٹورنامنٹ کافائنل کھیلاجا رہاہے قومی ہاکی ٹیم نے ورلڈ کپ 2018 کیلئے کوالیفائی پاکستان ہاکی ٹیم نے ورلڈ کپ کیلئے کوالیفائی کر انٹرنیشل ہاکی فیڈریشن کا چیمپیئنز ٹرافی ختم کر	
عظیم شاعر	شاعر ناصر کاظمی کا شمار اردو کے صف اول کے لاہورموسیقار ماسٹر عاشق شعرا حسین انتقال کرگئے عزیز میاں قوال کوہم سے بچھڑے ہوئے چودہ برس بیت اردو کے عظیم شاعر جگر مراد بادی کا کلام گا کر شہنشاہ غزل استاد مہدی حسن مرحوم کی اج 93 ویں س انقلابی شاعر دانشور فیض احمد فیض کی 107 ویں سا اردو کی مشہور ناول نگار بانو قدسیہ کی پہلی برسی عظیم شاعر احمد ندیم قاسمی کو بچھڑے 13 برس بیت گئے مایہ ناز شاعر ناصر کاظمی کی 91ویں سالگرہ منائی معروف قوال عزیز میاں کا اج 76 واں یوم پیدائش ہے	

ٹی 20 میچ	ورلڈ ٹی ٹوئنٹیہندوستان کو پہلے میچ میں شکست پاکستان اور انگلینڈ کے درمیان دوسراون ڈے کل کھ ایشیا کپ ٹی ٹوئنٹی یو اے ای کا ٹاس جیت کر بیٹن پاکستان اور انگلینڈ کے درمیان خری ون ڈے کل دبئ ورلڈٹی20 افغان ٹیم کا سری لنکا کے خلاف ٹاس جیت ورلڈ ٹی 20 افغانستان اور سری لنکا اج ٹکرائیں گے دوسرا ٹی ٹوئنٹی پاکستان نے جنوبی افریقا کو6 رن پاکستان اور جنوبی افریقہ ویمن کرکٹ ٹیم سیریز ا ورلڈ ٹی ٹوئنٹی میں دو میچز کھیلے جائیں گے پاکستان کرکٹ ٹیم تیسرا پریکٹس ون ڈے لیسٹر شائر	
شاہد افریدی کو ٹی ٹوئنٹی ٹیم کا کپتان	بابر اعظم ٹی ٹوئنٹی کے بعد قومی ون ڈے ٹیم کے کپتن سرفراز احمدکو ٹیسٹ ٹیم کی کپتانی بھی سونپ دی گئی پی سی بی کا اظہرعلی کوون ڈےٹیم کا نیاکپتان مقرر نئے ایک روزہ کپتان کے لیے اظہر علی کا نام تجویز شاہد افریدی کو کپتان بنانے کا فیصلہ خوش ائند ہ پہلا ٹی ٹوئنٹیپاکستان کاسٹریلیا کیخلاف ٹاس جیت زمبابوے کیخلاف ٹی ٹونٹی سیریز قومی اسکواڈ کا ا انگلینڈسےواحدٹی20میچ پاکستان ٹیم سےمتعلق مشاور دورہ بھارت پاکستان ٹیم کا اعلانشاہد فریدی ون ڈے شاہد فریدی کیلئے ایک خری موقع	
کورونا وائرس	ماڈل اداکارہ صحیفہ جبار خٹک بھی کورونا پوزیٹو ماڈل اداکار صحیفہ جبار بھی کورونا وائرس میں مبتلا کورونا وائرس کا ٹیسٹ مثبت انے کی بات جھوٹی ہے معروف اداکار طارق ملک کورونا وائرس میں مبتلا بالی وڈ گلوکارہ کانیکا کپور نے متعدد وی وی ئی پاکستان میں کورونا وائرس کے 2ہزار 50نئے کیسز عالمی شہرت یافتہ افریقی جیز گلوکار مینو دیبینگ کورونا وائرس کی وجہ سے اولمپکس مقابلے منسوخ ہو معروف اداکار صغیر الہی کھچی کورونا وائرس سے چل روبینہ اشرف اور سکینہ سمو میں کورونا وائرس کی	

Title	Results	Human Rating
ڈبلیو ڈبلیو ای چیمپئن شپ	0.9669473 سندھ میں سی این جی اسٹیشنز بدھ کراچی کی صبح بجے سے 24 گھنٹوں کیلئے بند رہیں گے سوئی سدرن حکام کے مطابق سی این جی اسٹیشنزبدھ کی صبح بجے سے جمعرات کی صبح بجے تک بند رہیں گے. 0.9627335 کراچیسندھ میں گیس لوڈ مینجمنٹ کے تحت صبح8بجے سے سی این جی اسٹیشنز کو 24 گھنٹوں کے لئے بند کردیا جائے گاجو اتوار کی صبح بجے کھلیں گےکراچی سمیت سندھ بھر میں سی این جی فلنگ اسٹیشنز گیس لوڈ مینجمنٹ شیڈول کے تحت 8بجے سے بند ہوجائیں گے سی این جی ایسوسی ایشن کے مطابق24 گھنٹوں کی بندش کے بعد صوبے بھر کے تمام سی این جی اسٹیشنز اتوار کی صبح بجے دوبارہ کھول دیے جائیں گے 0.962425	

3:

Title	**Results**	**Human Rating**
ڈبلیو ڈبلیو ای چیمپئن شپ	کراچی09 نومبر 2018 پاکستان اسٹاک ایکسچینج میں کاروباری ہفتے کا خری روز ہنڈریڈ انڈیکس میں منفی رجحان جاری ہے پی ایس ایکس ہنڈریڈ انڈیکس 39 پوائنٹس کی کمی سے 41 ہزار 332 کی سطح پر پہنچ گیا ہے یہ بھی پڑھیے کراچیاسٹاک مارکیٹ میں اتار چڑھا جاری پاکستان اسٹاک ایکسچینج میں ہنڈریڈ انڈیکس میں 450 پوائنٹس کا اضافہ Similarity: 0.8333720564842224 پاکستان نے سٹریلیا کو دوسرے ٹی ٹوئنٹی میچ میں سنسنی خیز مقابلے کے بعد سپر اوور میں دس وکٹوں سے شکست دے کر سیریز میں فیصلہ کن برتری حاصل کرلی ہے خبر رساں ادارے Similarity: 0.8276383876800537 کراچی15 مئی 2019 کاروباری ہفتے کے تیسرے روز پاکستان اسٹاک ایکسچینج میں زبردست تیزی جاری ہے جہاں ہنڈریڈ انڈیکس میں چار سو سے زائد پوائنٹس کا اضافہ ہوا ہے تفصیلات کے مطابق کاروباری ہفتے کے تیسرے روز پاکستان اسٹاک ایکسچینج میں مارکیٹ کے غاز سے ہی انڈیکس میں زبردست تیزی دیکھی گئی اور ٹریڈنگ کے دوران ہنڈریڈ انڈیکس چار سو سے زائد پوائنٹس اضافے کے ساتھ چونتیس ہزار کی حد عبور کرگئی Similarity: 0.8226531147956848 لیسٹرشائر 92 نیوز لیسسٹر شائر کیخلاف دوروزہ پریکٹس میچ میں پاکستان نے وکٹوں پر 321 رنز پر اننگز ڈیکلیئر کر دی اظہر علی فخرزمان اور عثمان صلاح الدین نے نصف سنچریاں اسکور کیں ہدف کے تعاقب میں لیسٹر شائر کی بیٹنگ جاری ہے Similarity: 0.8165571093559265 کراچی 20 فروری 2020 پاکستان اسٹاک ایکسچینج میں اتار چڑھا جاری کاروبار میں 46 پوائنٹس کی کمی پاکستان اسٹاک ایکسچینج میں اتار چڑھا کاروباری ہفتے کے چوتھے روز ہنڈریڈ انڈیکس میں 46 پوائنٹس کی کمی دیکھی جا رہی ہے 46پوائنٹس کی کمی سے مارکیٹ 40ہزار 527 سطح پر ٹریڈ	

	کررہی ہے گذشتہ روز مارکیٹ 399 پوائنٹس کے اضافے سے40 ہزار 574 پر بند ہوئی Similarity: 0.8158262372016907 کراچیکراچی اسٹاک ایکسچینج میں کاروباری ہفتے کے غاز میں تیزی کا رجحان دیکھنے میں رہا ہے ٹریڈنگ کے دوران انڈیکس میں 117 پوائنٹس کا اضافہ ریکارڈ کیا گیا جس کے نتیجے میں کے ایس ای 100 انڈیکس 23546 کی سطح پر پہنچ گیا Similarity: 0.8140777945518494 دبئیئی سی سی اینٹی کرپشن یونٹ کو میچ فکسنگ میں ملوث پلیئرمل گیائی سی سی کے چیف ایگزیکٹیو ڈیوڈ رچرڈسن نے دعوی کیا ہے کہ میچ فکسنگ کی تحقیقات جلد مکمل ہوجائے گی Similarity: 0.8138026595115662 کراچیکراچی اسٹاک ایکسچینج100انڈیکس میں 85پوائنٹس کے اضافے کے بعد 100انڈیکس پہلی بار26ہزارپوائنٹس کی سطح عبورکرگیا 100انڈیکس 26ہزار46پوائنٹس کی بلندترین سطح پربند Similarity: 0.8128382563591003 کراچیکراچی اسٹاک ایکس چینج میں مندی کا رجحان ہے اور کے ایس ای 100 انڈیکس میں 248 پوائنٹس کی کمی دیکھی گئی ہےانڈیکس میں 248 پوائنٹسکی کمی کے بعد کے ایس ای 100 انڈیکس 17698پوائنٹس کی سطح پرگیا ہے Similarity: 0.8115747570991516 کراچی اج کراچی اسٹاک ایکس چینج مینابتدا میں تیزی کا رجحان ہے اور کے ایس ای 100 انڈیکس میں 280 پوائنٹس کا اضافہ دیکھا جا رہا ہےجس کے بعد انڈیکس 30 ہزار793 پوائنٹس کی بلند ترین سطح پرپہنچ گیاہے Similarity: 0.8114549517631531	