iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6040

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Semantic Text Similarity

T. Keerthana[1], B. Meghana Reddy[2], A. Archana Reddy[3]
[1, 2, 3]*Student, Department of Information Technology, Vardhaman College of Engineering, Hyderabad, Telangana, India.*

Abstract: Semantics has been, in recent years, an open area of study in the field of information retrieval. Considering supervised classification of text, which is the main focus of this work, semantics is involved in different stages of processing of text: during the indexing stage, during the training phase, and via class prediction step. New foundational similarity measurements from text to text may replace classical similarity measures commonly used with other decision-making classification methods. Here is a brand-new method to test semantic resemblance using the Corpus-based Approach which is sponsored for a text as a brand-new feature that summarizes semantic resemblance between concepts representing the pair-to-pair compared text documents.

I. INTRODUCTION

In natural language processing or plagiarism control of software, the fundamental challenge is to find the significance of the text. For various purposes-Plagiarism control, information collection, and machinery for answering questions-semantic similitude is important. Nevertheless, semantics relation is difficult for computers. As a consequence of the advances in machine learning, machines in text semantics are being developed and various algorithms are being proposed. Semantic theoretical research does not require training the machine, it implies understanding.

The idea is that the two text files will be taken and the words in the sentences separated. The next move is to find clues in descriptive data based on data that is already accessible on the computer until the words are separated. For instance, if you find the word 'happy,' then all the related words are searched in the database for our data to match. To date, the text semantic algorithms have been taken into account, and further research is underway.

Millions of small text messages are published every day on social media: own research shows that every tweet has one to thirty words. We need sufficient information retrieval algorithms to tap into this stream of minimal text fragments. Corpus-based represents an example for traditional and popular texts like news items to be compared [1] and [2]. [2]. It relies on word overlap to look for similarities, but Corpus-Based algorithm often does not work in very short texts, in which word overlap is rare. This is why we need phrase representations that understand more than words.

In 2013, Makarov et al published three papers on distributed word embeddings, which culminated in Google's program word2vec, to capture semantic similarities between the words [3], [4], [5]. Since then, scientists have extensively used embedding in natural language processing to develop state-of-the-art algorithms, such as speaking part [6], finishing sentence [7], the hashtag prediction [8], etc. Nevertheless, there is a lack of analysis and insight into the way that embedding is efficiently incorporated into one sentence that incorporates most of its semantic information

Many authors agreed to use the multi-layer sensing [6], [11], the classification[12] or trim a text in a fixed-length [11] as thebasis for a mean or maximal Le and Makarov's paragraph Vector algorithm — also known as paragraph2vec — is a strong tool for identifying correct vectors of sentences, paragraphs, and variable-length documents [13]. The algorithm tries, simultaneously via the method word2vec, to find embedding for individual words and paragraphs.

However, it is understood beforehand that paragraphs are compiled. This means that it takes additional training to find the vector representation of a new and potentially unknown sentence which, after all, is far larger than the number of differences. Consequently, Paragraph2vec is not an appropriate candidate to be included in a medium like social media, for example. Further research is, therefore, necessary to derive optimal phrases based on word integration. Through observing and comparing the output of multiple word combinations in a short texts match task, we come to a new methodology, whereby both Corpus-Based and word embedding signals combined.

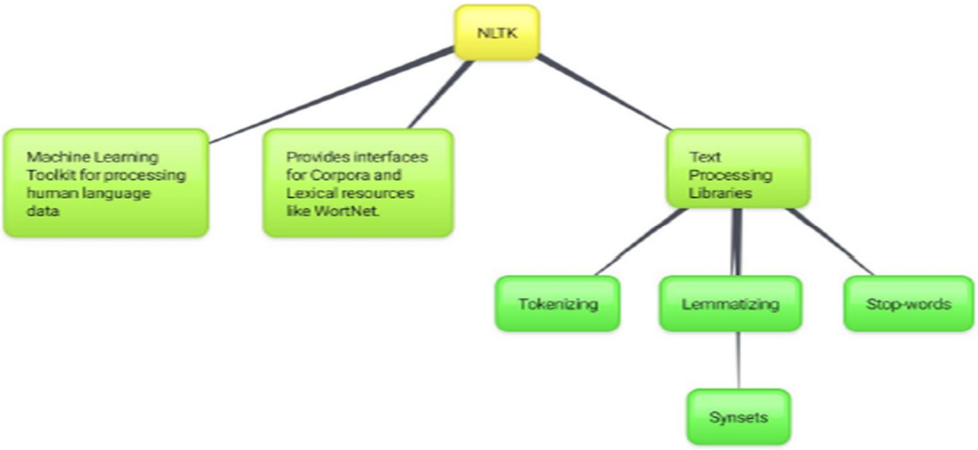
Having the effect of frequent words — i.e., for all considered fragments; we show how to word embedding is combined into a new vector representation in this paper. These leads in contrast to traditional Corpus-based techniques or basic heuristic approaches to combine word embedding, to a major increase in the efficacy of semantically related short text excerpts being found. Our approach is an initial step towards a hybrid approach that integrates word embedding into a distributed representation with Corpus-Based information for a short fragment.

II. LITERATURE SURVEY

The comparability between long texts is assessed in comprehensive literature or objects, but fewer works related [Hatzivassiloglou et al. 1999; Landauer and Dumais 1997; Maguitman et al. 2005; Meadow et al. 2000]. For similarity comparison between phrases or short texts [Foltz et al. 1998. 1998]. Related studies can be categorized in approximately four major categories: Processes of the text focused on the co-occurrence/vector, corpus-based processes. In information retrieval (IR) systems [Meadow et al. 2000], the software most used are vector-based approaches for database models.

A document is defined as a word about an input query. List and questions in the text are compared to similar documents The Salton and Lesk 1971 Database by a similarity metric. One-word extension Methods of cooccurrence contribute to methods used in text mining and discussion agents generally [Corley and Mihalcea] 2005). 2005]. It is presumed that more similar papers have been made. In general, having more terms. Finally, the sentence description is not very effective, since the vector dimensions ina short text or a sentence are very large, the resulting vectors therefore would have ingredients. Hybrid approaches apply both [Turney 2001] and corpus-based interventions. Semantic similarity of the word to assess the similitude of text. Curds et al. [2006] propose a joint method for the semantical similarity measurement of texts by using information deriving from the similarity of Words of the part. Specifically, they are using two corpus-based measures PMIIR) (Turney 2001) [Turney 2001] Latent Semantic Analysis (see Section 3.2) and LSA) [Landauer et al. (details). 1998] and six steps focused on information [Conrath and Jiang 1997; Chodorov 1998; Lesk 1986; 1986; Wu and Palmer 1994; Resnik 1995]. Word semanticity semblance and combined the results to indicate how these measures can be used to derive a metric similarity between text and text. You assess your Method for the identification of paraphrases. This method 's key downside It measures the similarity between the terms of eight different methods, It's not effective computationally. Li et al. [2006] suggest a different hybrid method that results in a similar text the knowledge found in the comparable texts is semantic and syntactic. You're dynamically the suggested approach shapes a combined word collection with only the separate terms Sentence words in pairs. For each sentence, the WordNet lexical database [Miller et al .1993] is used to derive a raw semantic vector. For every sentence, a word order vector is generated with information from the index of lexical. Since every word in a sentence makes a different contribution to by using information content derived from a corpus, the meaning of the whole sentence is weight Combining the crude semantic A semanticity vector is obtained with the information content of the corpus Two sentences for each. Semantic similarity is determined based on Two vectors of the week. The similarity of order is calculated with both order Vectors. Finally, the simulation of the sentence comes from the combination of semantical similarity and sequence. Such two-hybrid acts [Li et al. 2006; Mihalcea 2006; et al. 2006] fail to take the string similarity into account, which in certain cases plays an important role. We discuss the importance of string similarity next to the book functional methods attempt to describe a word using a collection of predefined methods Features, features. A trained classifier provides a similarity between two texts. Nonetheless, it is difficult to find useful features and get values for these features from sentences.

III. IMPLEMENTATION



Supervised Learning Algorithm Corpus-based Approach

1) *S1* - A cemetery is a place where dead people's bodies or their ashes are buried.
2) *S2* - A graveyard is an area of land, sometimes near a church, where dead people are buried.

A. Tokenizing

Dividing the body of text into sentences and phrases. Tokenization is a tokenizing of two kinds of phrases and verbs. This is one-sentence punctuation and is not being carried out for the tokenization of a sentence.

```
['A', 'cemetery', 'place', 'dead',
'people', "'s", 'bodies', 'ashes',
'buried', '.']
['A', 'graveyard', 'area', 'land',
',', 'sometimes', 'near', 'church'
, ',', 'dead', 'people', 'buried',
'.']
```

B. Stop Words

Often some specific terms which seem worthless when choosing documents that conform to a user's needs are fully omitted from the language. The key technique for creating a "stop list" is to sort words by selection frequency, and then to use the most frequenting terms, often hand-filtered for their semantic contents concerning the context of the documents being indexed, to be used as a stop list, whose members are then disregarded during indexing. The figure gives an example of a stop list. The figure gives an example of a stop list. The number of postings a program will store is substantially lowered by the use of a stop list. Stop words from the text you wish to process can be filtered. There is no universal stop word list in NLP, but the NLTK module includes a stop word list.

C. Lemmatizing

Lemmatization aims to reduce the shape of a term to a common basis, often by the derivatives. Lemmatization usually refers to careful use of the word's vocabulary and morphology to delete inflictive endings only and to return the basic or dictionary of a word known as the lemma.

```
['A', 'cemetery', 'place', 'dead', 'people
', "'s", 'bodies', 'ashes', 'buried', '.']


['A', 'graveyard', 'area', 'land', ',', 's
ometimes', 'near', 'church', ',', 'dead',
'people', 'buried', '.']
```

If faced with the token, 'saw ' lemmatization will see or see whether the token was used as a noun or as a substantive. Lemmatization usually collapses the lemma's various inflexible forms only. Then, you cannot look at your root stem in a dictionary, that means the word you end up with; you can look up a lemma, however. Lemmatize is the only thing that needs to be noticed is that lemmatize is part of the voicemail parameter, 'pos'. It not given the default is, 'noun.

```
['A', 'cemetery', 'is', 'a', 'placing', '
where', 'dead', 'people', "'s", 'body', '
or', 'their', 'ash', 'are', 'buried']


['A', 'graveyard', 'is', 'an', 'area', 'o
f', 'land', ',', 'sometimes', 'near', 'a'
, 'church', ',', 'where', 'dead', 'people
', 'are', 'buried', '.']
```

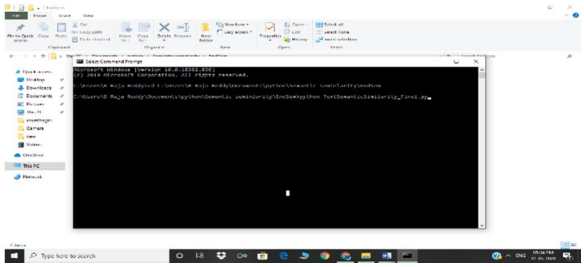
D. Synsets

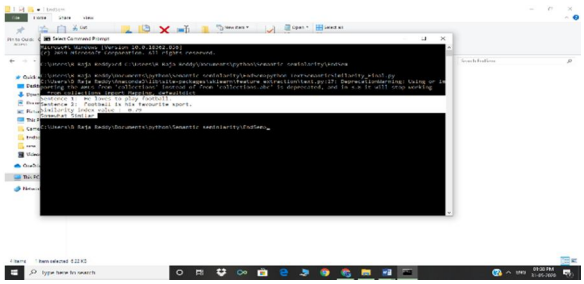
WordNet is an English lexical database and a part of the NLTK corpus. We can use WordNet to find terms, synonyms, antonyms, and more alongside the NLTK module. Synonyms are a word with a similar meaning; hence a group of synonyms is a synonym. The lemmas are synonyms, so the antonyms to the lemmas can be identified using antonyms. Then, by using the Wu and Palmer approach for semantic interrelation, we can also easily compare the similarity between the two words and they're tense.

```
word1 :  ['Area']
word2 :  ['Place']
Similarity index : 0.9333333333333333
```

```
word1 :  ['Bull']
word2 :  ['Hair']
Similarity index : 0.5333333333333333
```

IV. RESULT





V. CONCLUSION

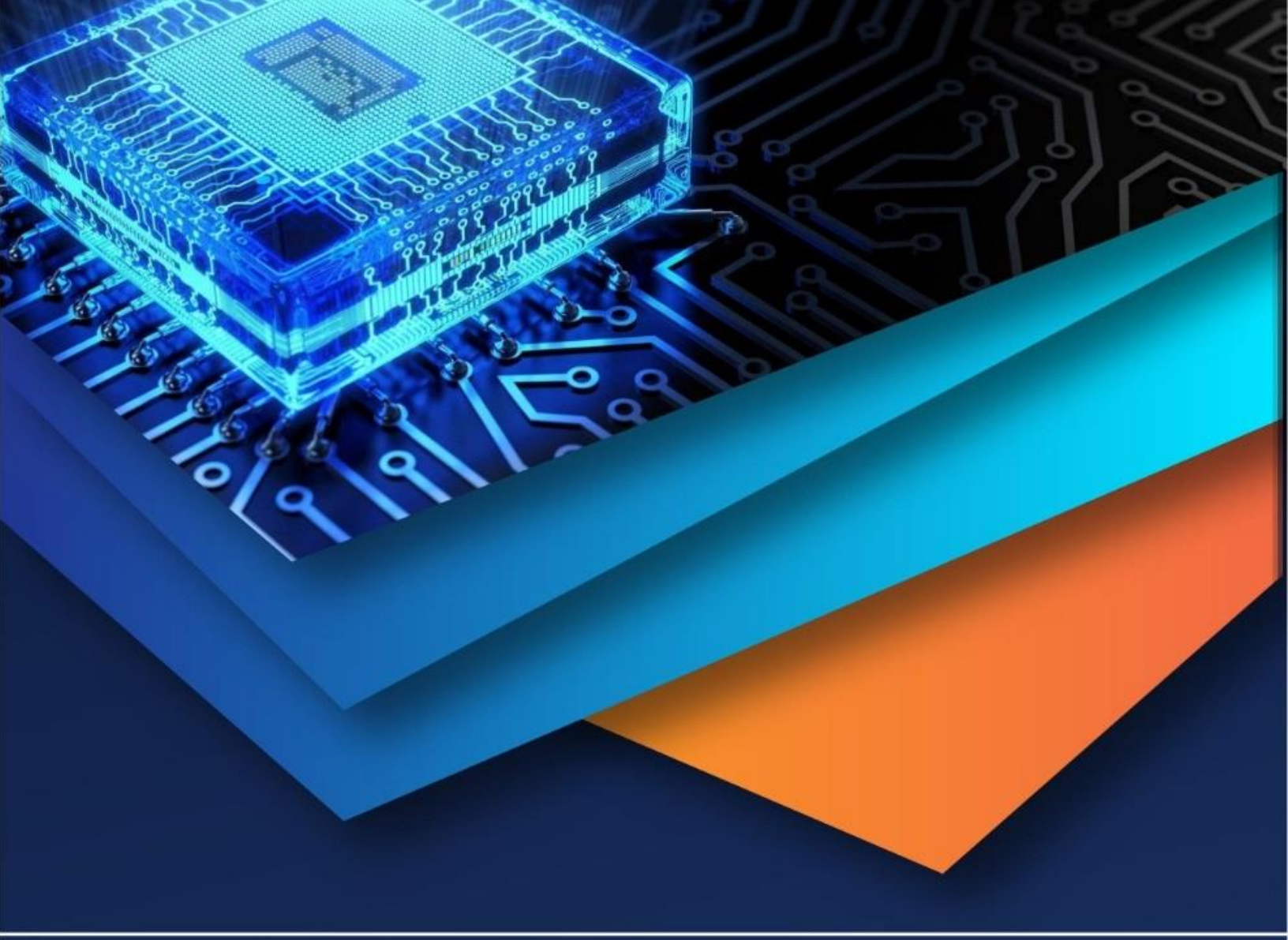
Two phrases/sentences are taken for semanticization. They are two different, different, or somewhat similar sentences. The collection of stop words is then described in English. We get S1 = fI, given, card, garden, and S2 = fIn, garden, gave, cards, after removal of the special characters and punctuation, and after removing all words and lemmatization. We find the synonyms of the lemmatized words which are called the synsets after lemmatization. Instead, the first S1 word is compared to all the S2 terms and it will start iterative, finding the similitude index in S2 for each of the terms. We find the mean of the calculated similarity indexes, so, we use machine learning to analyze semantic similitude. The phrases with a similarity index that is less than 0.65 are marked, 'Not Similar,' whereas the phrases with 0.65 to 0.8 are marked, 'Somewhat Similar' and of 0.8 are marked 'Similar.

REFERENCES

[1] T. graph," Text semantic analysis and semantic graph", Stackoverflow.com, 2017. [Online]. Available: http://stackoverflow.com/questions/35666726/text-semantic-analysisand- semantic-graph. [Accessed: 27- Feb- 2017].

[2] " How to do Semantic Keyword Research Using NLP and Text Analysis - AYLIEN", AYLIEN, 2017. [Online]. Available: http://blog.aylien.com/how-to-do-semantic-keyword-research-using-nlpand/. [Accessed: 01- Mar- 2017].

[3] " Understanding Semantic Analysis (And Why This Title is Totally Meta) - Boom train", Boom train, 2017. [Online]. Available: https://boomtrain.com/understanding-semantic-analysis/. [Accessed: 01- Mar- 2017].



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY