**OPERATING SYSTEMS**

**PROJECT**

****

**TEXT SIMILARITY ALGORITHMS**

**NAME**

**RAJVANSH**

**ROLL NUMBER**

**2020UEA4513**

**SUBMITTED TO**

**SOMIYA MA’AM**

1. Text Analytics – An Introduction

Text analytics combines a set of machine learning, statistical and linguistic techniques to process large volumes of unstructured text or text that does not have a predefined format, to derive insights and patterns. It enables businesses, governments, researchers, and media to exploit the enormous content at their disposal for making crucial decisions. Text analytics uses a variety of techniques – sentiment analysis, topic modelling, named entity recognition, term frequency, and event extraction.[1]

The emotion communicated by the unstructured text is identified via sentiment analysis. Product reviews, customer interactions, social media posts, forum conversations, and blogs are all examples of input text. [1]

The 7 Basic Functions of Text Analytics [2]

1. **Language Identification** - The initial stage in text analytics is determining the language in which the material is written, English or Arabic Each language has its own quirks, therefore it's critical to understand what we're working with.
2. **Tokenization** - Now that we know what language the material is written in, we can divide it into sections. Tokens are the individual units of meaning that you are working with. This can include words, phonemes, or even entire phrases. Tokenization is the process of dividing text content into tokens.
3. **Sentence Breaking** - Now we have to identify where the sentences stop once you've recognised the tokens.
4. **Part of Speech Tagging** - The technique of identifying each token's part of speech inside a text and assigning the appropriate tag is known as part of speech tagging.
5. **Chunking** - A variety of sentence-breaking techniques is known as chunking. It separates a sentence into its constituent components.
6. **Syntax Parsing** - The syntax parsing sub-function can be used to ascertain a sentence's structure. Sentence diagramming is actually simply a fancy term for syntax parsing. However, it's an essential pre-processing step for sentiment analysis and other NLP features.
7. **Sentence Chaining** - Sentence chaining, often referred to as sentence relation, is the last stage in preparing unstructured text for further in-depth analysis. In reality, after links between phrases have been made, you may perform intricate analysis like contrasting emotion ratings and fast creating precise summaries of lengthy texts.

2. Natural Language Processing (NLP)

The field of computer science known as "natural language processing" (NLP) is more particularly the field of "artificial intelligence" (AI) that is concerned with providing computers the capacity to comprehend written and spoken words in a manner similar to that of humans. Automatically extract, categorise, and label pieces of text and speech data using computer algorithms combined with machine learning and deep learning models is known as statistical NLP. After doing so, each element's potential meaning is given a statistical probability score. Currently, deep learning models and learning methods based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) allow NLP systems to "learn" as they go along and extract ever-more-accurate meaning from massive amounts of unlabelled, unstructured text and speech data sets. [3]

Use cases of NLP

* **Text summarization - In order to summarise massive amounts of digital text and provide summaries and synopses for indexes, research databases, or busy readers who don't have time to read the complete text, text summarising employs NLP techniques.**
* **Chatbots - In order to reply with the proper action or helpful remarks, Alexa uses speech recognition to identify patterns in voice requests and natural language production. The same magic is worked by chatbots in response to text input.**
* **Machine translation - In terms of accuracy, machine translation technologies have made good progress. Text that has been translated into one language and then back into the original is a fantastic approach to evaluate any machine translation software.**

The current project intends to explore the implementation of different text matching algorithms on different defined test cases. In this context, two records are generated namely base file and response file wherein the base file refers to the text with which the text in response file is compared and accordingly the similarity score will be generated using various algorithms.

3. Methodologies used in the project

3.1 String to String Matching

This was our first approach towards the problem. Here we simply performed string to string matching. The response from the user is simply get compared with the base in our database. If the response is same as the base the score is increased.

In this method, the text comparison is completely dependent on the order of the words.

3.2 Jaccard Index Method

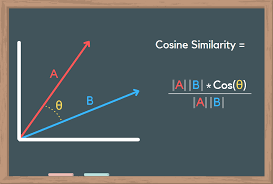
In this method the words can be rearranged in any order but it does not compare the meaning of two different text documents.

Its value drops even though the words have same meaning. [4]

3.3 Cosine Similarity Method

This was our third strategy for solving the issue. It calculates how similar two vectors in an inner product space are to one another. It establishes if two vectors are generally pointing in the same direction by calculating the cosine of the angle between them. In text analysis, it is frequently used to gauge document similarity.

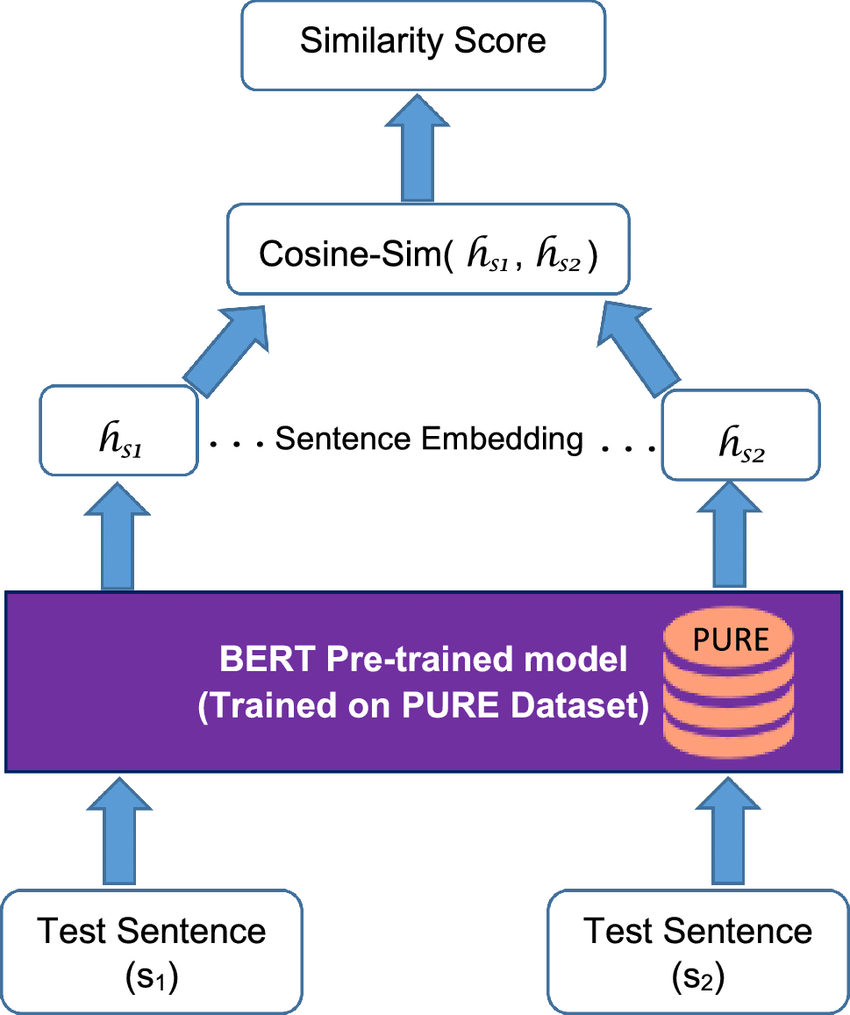
This is a improved version of jaccard index method, but it still lacks in the accuracy. [4]



3.4 Bert Model Method

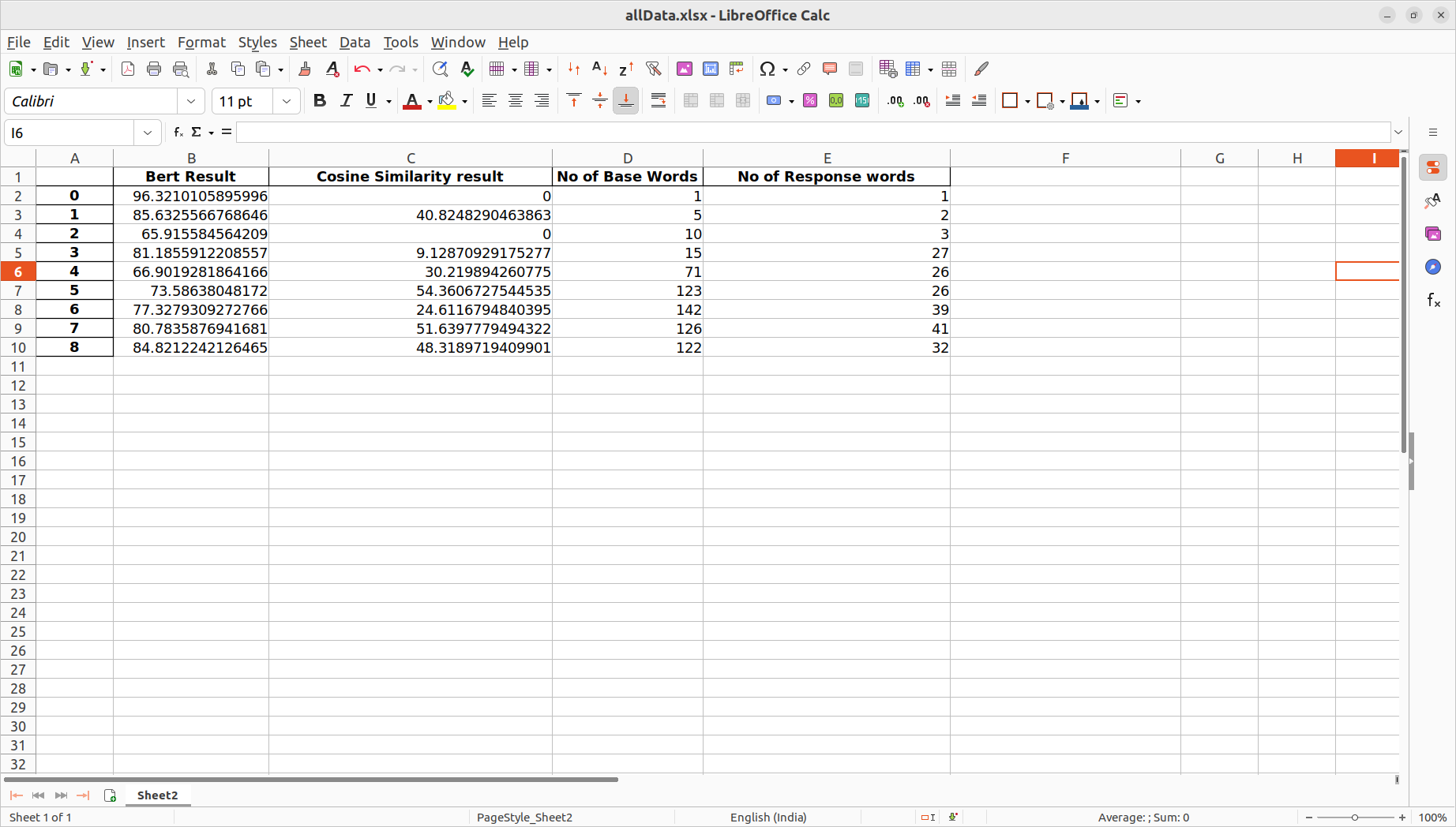
This was our final and fourth attempt to solve the issue. BERT is a free and open-source machine learning framework for dealing with natural language (NLP). BERT uses the surrounding text to provide context in order to assist computers grasp the meaning of ambiguous words in text. With the use of question and answer datasets, the BERT framework may be adjusted after being pre-trained on text from Wikipedia.

Bidirectional Encoder Representations from Transformers, or BERT, is a deep learning model that is based on Transformers. In Transformers, each output element is connected to each input element, and the weightings between them are dynamically determined based upon their relationship. [5]



4. Results and Discussion

In order to understand the implementation of above algorithms, 10 test cases were generated ; 5 of which had approximately same text count while rest 5 cases had significant difference in word count in order to further understand if word count had any impact on the similarity score.

****

For example if we take the test case 0,

Test Number - 0

Number Of Base Words - 1

Number Of Response Words - 1

Base - hi

Response - hey

Cosine Similarity Result - 0.0

BERT Similarity Result - 96.32101058959961

Here we can see that the word count difference is zero between the base and the response.

Cosine similarity result for this similarithy test is zero even though the words “hey” and “hi” have the same meaning. However, in case of BERT, the score is quite relevant with the text used for analysis.

Similarly, now if we see the test case 6,

Test Number - 6

Number Of Base Words - 142

Number Of Response Words - 39

Base - Astronomy is the science that studies the laws of the stars is a natural science that studies celestial objects and phenomena. It uses mathematics, physics, and chemistry in order to explain their origin and evolution. Objects of interest include planets, moons, stars, nebulae, galaxies, and comets. Relevant phenomena include supernova explosions, gamma ray bursts, quasars, blazars, pulsars, and cosmic microwave background radiation. More generally, astronomy studies everything that originates beyond Earth's atmosphere. Cosmology is a branch of astronomy that studies the universe as a whole. Astronomy is one of the oldest natural sciences. The early

civilizations in recorded history made methodical observations of the night sky. These include the Babylonians, Greeks, Indians, Egyptians, Chinese, Maya, and many ancient indigenous peoples of the Americas. In the past, astronomy included disciplines as diverse as astrometry, celestial navigation, observational astronomy, and the making of calendars.

Response - Astronomy is the study of everything in the universe that’s beyond our own planet’s atmosphere. The planets in our own solar system, our own star the sun, and the bright stars can all be seen with the naked eye.

Cosine Similarity Result - 24.61167948403948

BERT Similarity Result - 77.32793092727661

Here it can be seen that the difference between the word count between the base and the response is significant. Even though the word difference is quite high here, the Bert Model's output is again favourable with 77% score as compared to Cosine analysis with 24% score.

5. Conclusion and future scope

As per our understanding about the results being generated, it has been observed that BERT Model provides good similarity score. As we have already discussed, NLP includes BERT. And a significant portion of this is attributable to BERT's capacity to incorporate word meaning into tightly packed vectors. Because each value in the vector has a value and a purpose for being there, we refer to them as dense vectors. The majority of the values in sparse vectors, such one-hot encoded vectors, are 0, in contrast to this. Each encoder layer (there are several) generates a collection of dense vectors, and BERT is excellent at producing these dense vectors.

Further it is intended to explore the possibility of more accurate algorithms to generate text matching score efficiently. In this context, efforts will be to continue the work at CSIR-NPL in order to have a beneficial outcome.

References

[1] - <https://www.tibco.com/reference-center/what-is-text-analytics>

[2] - <https://www.lexalytics.com/blog/text-analytics-functions-explained>

[3] - <https://www.ibm.com/cloud/learn/natural-language-processing>

[4] - [https://towardsdatascience.com/3-basic-distance-measurement-in-text-mining-5852becff1d7#:~:text=Jaccard%20Similarity%20%3D%20(Intersection%20of%20A,so%20the%20score%20is%200](https://towardsdatascience.com/3-basic-distance-measurement-in-text-mining-5852becff1d7" \l ":~:text=Jaccard Similarity %3D (Intersection of A,so the score is 0)

[5] - <https://towardsdatascience.com/bert-for-measuring-text-similarity-eec91c6bf9e1>