Variance in data length and the accuracy of defining a word in context using BERT.

How does the length of data affect a BERT model's accuracy to define a word in context?

Computer Science

Word Count:

Table of Contents

[Introduction: 2](#_Toc78984529)

[Machine Learning: 2](#_Toc78984530)

[Natural Language Processing: 2](#_Toc78984531)

[BERT: 3](#_Toc78984532)

[Part of Speech Tagging: 3](#_Toc78984533)

[Transformers: 4](#_Toc78984534)

[The Experiment: 4](#_Toc78984535)

[Analysis: 4](#_Toc78984536)

[Conclusion: 4](#_Toc78984537)

[Further Study: 5](#_Toc78984538)

[Works Cited: 5](#_Toc78984539)

[Works Consulted: 5](#_Toc78984540)

[Appendix: 5](#_Toc78984541)

ASSESMENT CRITERIA

A: Focus and method

* Choice of Topic
* RQ
* Methodology and sources utilized

B: knowledge and understanding

* Context
* Communication

C: Critical Thinking

* Research
* Analysis
* Discussion and evaluation

D: Presentation

* Structure
* Layout

E: Engagement

* Research focus
* Planning and process

# Introduction:

WRITE THIS AT THE END.

Context

* What key aspects can you discuss to ensure you’ve provided some context underpinning your research question?

# <https://arxiv.org/abs/2005.07503>: COVID-Twitter-BERT

* + It would be impossible to analyze this with people, too much information. BERT on the other hand could analyze millions of tweets etc much quicker, tackling some big dangerous issues such as this. Natural Language Processing is essential in our modern world with so much information recorded, and social media and other forms of mediums are growing wildly. This can not only help police these sites but also help with building advertisement profiles of people, understanding search results more etc. This brings along ethical considerations though, including who should be allowed to use these models to interpret so much data? Is it ok if twitter/google/etc has access but the rest of the people don’t? monopoly? Should it be okay that an algorithm and a model can decide our freedom on the internet? Ban us? Etc.

Outline of Argument

* What features, aspects, factors, theories and so forth will your essay utilize in order to arrive at a conclusion?

Scope

* What authors, scientists, case studies, theories and so on have been consulted to answer your research question?

Worthiness

* Why is your research question worthy of investigation?

Methodological approach and/or source material you’ve used to arrive at your conclusion.

In the modern age, Natural Language Processing ([NLP](#_Natural_Language_Processing:)) is used daily. From translators to speech recognition and spell checkers, every application uses some form of NLP to interpret and utilize human language.

One path that has been taken to allow algorithms and machines to better comprehend speech is machine learning (ML).

# Machine Learning:

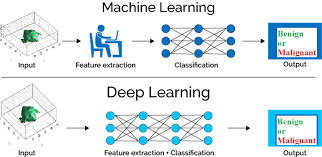
# Natural Language Processing:

Encompassing many different areas of study including [NLP](#_Natural_Language_Processing:), machine learning is defined as “the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data.”. Machine learning has seen exponential growth in use recently and for good reason:

Paragraph:

* Thesis
  + Establish what point or issue your paragraph is going to develop. This should be written in the form of a statement that is relevant to the overall question or topic being addressed.
* Development
  + Explain in greater detail what is meant by the thesis
* Evidence
  + Evidence
* Balance
  + Offer an argument opposite to the one you are making to show you have looked at the issue from more than one angle, but make sure you counter-argue so that you have still made your point. (this can be its own paragraph also)
* Analysis
  + How does this information gathered relate to my question?
  + What answer (even partial) does the source provide?
  + Link to academic theories or approaches
  + Provide evidence to support points and highlight relationship between them.

Loss function, accuracy



* Transformer for NLP in Machine Learning
* Sequence to Sequence
* LSTM vs Transformers
  + “Recurrent Networks were, until now, one of the best ways to capture the timely dependencies in sequences. However, the team presenting the paper proved that an architecture with only attention-mechanisms without any RNN (Recurrent Neural Networks) can improve on the results in translation task and other tasks! One improvement on Natural Language Tasks is presented by a team introducing BERT: [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](https://arxiv.org/abs/1810.04805).”
* <https://www.youtube.com/watch?v=S27pHKBEp30>
* Statistical approaches to NLP

“Natural language processing (NLP) is the study of mathematical and computational modeling of various aspects of language and the development of a wide range of systems. These include spoken language systems that integrate speech and natural language; cooperative interfaces to databases and knowledge bases that model aspects of human-human interaction; multilingual interfaces; machine translation; and message-understanding systems, among others. Research in NLP is highly interdisciplinary, involving concepts in computer science, linguistics, logic, and psychology. NLP has a special role in computer science because many aspects of the field deal with linguistic features of computation and NLP seeks to model language computationally.” https://www.jstor.org/stable/2879169

# BERT:

VERY GOOD EXPLANATION: <https://www.youtube.com/watch?v=xI0HHN5XKDo>

How it works + how it uses vectors

“BERT outperforms previous methods because it is the first unsupervised, deeply **bidirectional** system for pre-training NLP.”

<https://medium.com/dataseries/why-does-xlnet-outperform-bert-da98a8503d5b>

<https://searchengineland.com/faq-all-about-the-bert-algorithm-in-google-search-324193>:BERT algorithm in Google search

BERT is a deeply bidirectional, unsupervised language representation, pre-trained using only a plain text corpus. BERT takes into account the context for each occurrence of a given word

“The BERT loss function takes into consideration only the prediction of the masked values and ignores the prediction of the non-masked words. As a consequence, the model converges slower than directional models, a characteristic which is offset by its increased context awareness.” <https://towardsdatascience.com/breaking-bert-down-430461f60efb>

<https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>

<https://github.com/negedng/bert-embedding#egg=bert_embedding>

### Part of Speech Tagging:

“Parts of speech tagging simply refers to assigning parts of speech to individual words in a sentence, which means that, unlike phrase matching, which is performed at the sentence or multi-word level, parts of speech tagging is performed at the token level.” <https://stackabuse.com/python-for-nlp-parts-of-speech-tagging-and-named-entity-recognition>

<https://www.youtube.com/watch?v=fv6Z3ZrAWuU>: good video, possibly more good videos explaining other concepts

“As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial taskspecific architecture modifications.”

“The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training.”

<https://arxiv.org/pdf/1810.04805.pdf>

# Transformers:

What is a transformer?

How does it work?

Benefits, disadvantages to it vs LSTM?

* LSTM vs Transformers
  + “Recurrent Networks were, until now, one of the best ways to capture the timely dependencies in sequences. However, the team presenting the paper proved that an architecture with only attention-mechanisms without any RNN (Recurrent Neural Networks) can improve on the results in translation task and other tasks! One improvement on Natural Language Tasks is presented by a team introducing BERT: [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](https://arxiv.org/abs/1810.04805).”

# Dimensionality Reduction:

# The Experiment:

Evaluate the experiment

* How were the results achieved?
* How was your data analyzed?
* How reliable (or valid) was the chose method or source material?
* Was there another approach that could (or needed to) have been followed?

To test the accuracy of a BERT model when defining words in context, an adapted piece of code was used. The code first imports the data from a CSV file stored as a GitHub Gist. Then generates BERT embeddings for every token of the sentence. BERT embeddings are,,,. Tokens are,,,. After the embeddings have been generated the embeddings for the word being tested are selected and saved as a NumPy array. Next, the data is reduced to 2 Dimensions. Once it is reduced to 2D it can be plotted on a graph. Finally, to calculate the accuracy a K-NNC model is implemented with LOOCV.

This is the base code that is used for every dataset. However, to create different datasets based on length of sentence a different piece of code is used. The main code is then run for different data lengths and the results are stored manually as seen in the table and graphs below.

The embeddings were created using a BERT model (the only way to get BERT embeddings) and so is therefore obviously reliable. However, the method of getting an accuracy value could be unreliable. Although this is a readily used model it could in theory maybe not be the best model for this experiment

# Analysis:

Evaluation:

* some words have lower accuracies possibly because the meanings are more similar to each other than other words, it’s even hard for humans to tell the difference.
  + Also, possible that BERT uses context occurrence to find the meaning therefore when words are used together often, such as date (person) and date (like a dinner reservation). They would have similar words around them and therefore training data would mean BERT would find it harder to tell them apart.
* How accurate is the accuracy function?
* What is the difference in the way they group together? Try to add contour lines
* Could have used more data however I had to collect the sentences and sort by myself, I might have also made a mistake.

# Conclusion:

What solution(s) has your research indicated?

What has been learned from the results or data?

Where there any apparent contradictions that your research indicated? How have you accounted for these?

What, if any, were the limitations of research approach or methodology chosen?

Are there any unanswered questions? Do they fall within the remit of your research question? If yes, how are you planning to answer them?

ANSWER THE RESEARCH QUESTION

Reiterate the key findings and main points developed in the body of the essay and provide a resolution to your research question.

# Further Study:

Graphical user interface, text, application

Description automatically generated

# Works Cited:

# Works Consulted:

# Appendix: