**Question:** Does COVID affect post trends and engagements from mental illness related subreddits?

**Hypothesis:**

**Project Outline**

1. Descriptive Statistics
   1. Mean mode
   2. 9 subreddits explanation
   3. Post engagement/trend
   4. Sample size
   5. Central tendency , dispersion
2. Correlation Analysis
   1. Correlation heatmap
   2. Calculation- log transformation
3. Data Visualization
   1. descriptive statistics
   2. correlation coefficient
   3. Topic modeling
   4. Time series
4. NLP - topic modeling – BERT, LDA (Current Process)
5. Time series - ARIMA (Current Process)
   1. Which data for time series?
6. Statistical testings
   1. T- test
   2. F1 score
7. Conclusion

**Project Literature Review - Previous works**

<https://ieeexplore.ieee.org/abstract/document/8250563>

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/<https://dl.acm.org/doi/pdf/10.1145/3446999.3447012>

<https://www.jmir.org/2020/10/e22635/>

<https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/>

Bert, lda - doc of numbers, latest model - LLM compact models, embedding → LDA

Are people just comfortable about mental illness, mental health

How does that change? - does they , using topic modeling

<https://www.psychiatry.org/news-room/apa-blogs/poll-american-workers-are-increasingly-comfortable>

## Reference

<https://ir.cs.georgetown.edu/resources/smhd.html>

<https://www.hindawi.com/journals/cin/2022/7893775/>

(process of bert)

Huggingface - transformer Pytorch or tensorflow implement of BERT

Hugging face advantage:

* Ease of use, extensive documentation, large community support
* Provides a wide range of pre-trained models, including BERT and variants
* Easily fine-tuning models on custom datasets
* Flexible and customizable, allowing users to easily modify and experiment with different architectures, tokenizers, and hyperparameters

Tensorflow BERT

* Seamless integration with other tensorflow-based models and tools
* Strong support for deployment in production environments, esp. If your existing infrastructure is built around tensorflow
* Tensorflow ecosystem may offer additional resources and tutorials specific to bert implementation

Preprocessing:

* Feature extraction
* Informal way : hashtags, special characters, needless words
* Substituting jargon and emojis (emojipedia)

Tokenization

* Stemming , lemmatization
* TreebankWordTokenizer - NLTK
* Bag of words
* Removing stopwords
* Conversion to lowercase
* POS (part of speech)- using NLTK
* Word2vec

BERT modeling

* Using hugging face pytorch

Classifier models

* CNN , LTSM

BERT

Embedding layer

Num of dropout, epochs – compare f1 score

CT BERT

Embedding layer

Num of dropout, epochs – compare f1 score

Proposed model <https://www.sciencedirect.com/science/article/pii/S1877050923000716>

1. LDA for topic modeling
2. BERT sentiment classification

<https://ieeexplore.ieee.org/abstract/document/9558988>

1. LDA - probability based topic modeling
2. Bert - transformer package

<https://www.hindawi.com/journals/cin/2022/7893775/>

Mental illness social media