TWEETING THE BLUES: DETECTING DEPRESSION AMONGST TWITTER USERS

CARTER, DAVID, JEREMY, AIDEN, MICHAEL

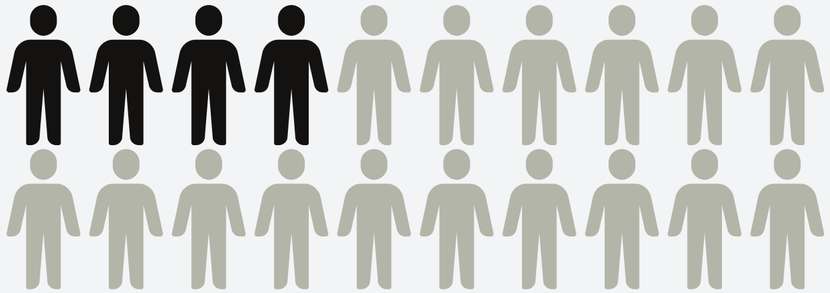
Table of Contents

THE PROBLEM

- 22% of U.S. Adults suffer from depression.
- 700,000 people die each year from suicide.
- Many individuals lack the resources for diagnosis and treatment, leaving their conditions unaddressed.

Depression in the United States Amongst Adults

22%



OUR DATA

Kaggle Dataset

- Collected in 2015 (Vijayshind Shinde) using Twitter API
- 20,000 data points
- Variables: Tweet text, user ID, follower count, friend count, number of statuses, and favorites
- 72 Users
- Label: 1 for depressed and 0 for not depressed.

Safa Dataset

- Collected in 2020 (Prof. Ramin Safa) using Twitter API
- 11 million data points
- Variables: Tweet text, follower count, friend count, number of statuses, favorites, and more
- 1123 Users
- Label: 1 for depressed and 0 for not depressed.

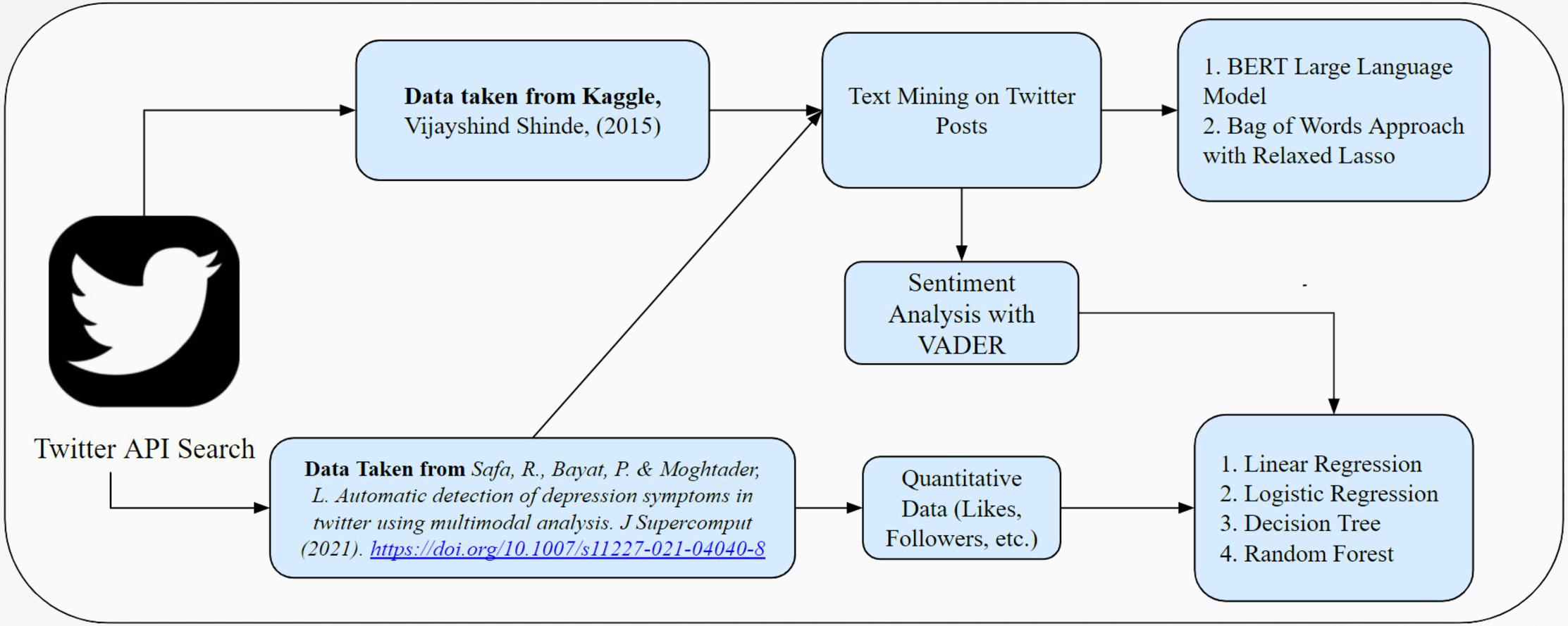
post_text	user_id	followers	friends	favorites	statuses	label
@more_than_meat @VeganChatRoom @VeganFoodRecipe I mean except for the ppl who kill the animals @ the slaughterhouse who are perfectly aware	49548465	677	1644	8375	12913	1
ZOMFG @lohanthanyfreak FAVED LIKE 8 OF MY TWEETS HOLY SKCLDNDWL I OBSESS not like I know her irl or something like that	3015971504	914	489	8717	3359	1
@evvok wheres my girlfriend	490044008	1849	561	8229	86352	0
PLEASE RT: @Mhousebrewery @Sentequk @fgclothingltd @jonblower @Minchinologist @high5itapp @Brano1979 @dcfc_flagman http://t.co/hQiaUvc8ud	1629737760	2377	738	3189	5624	1
So I told Zayin to rebel against the Center.. he actually did it the absolute madman hahahahahahaha!	762433972273950725	20	0	0	2143	0

post_text	label
@bevllisario anch'io 😭	1
@lhixcollins It sucks bc Biden legit tried to so many times and just. Kept. Getting talked through. I wish they had cut trump's fuckin mic. I'm interested for the town hall debate that's coming up.	0
@comp1120 @OmegaPSICrush @ABC Ummm…do some research (on non-rightwing sites) and get back to us.	0
the promo editing team are the true goats of this show.	0
@gbemsabiola I just remember her bouncy blond hair and pretty makeup!	1
i have literally have no idea what's on this test https://t.co/G33XRnHHYJ	1
@ShelbyKluver I hate the iPhone 12 design I'm getting the iPhone 11 Pro	1
@erieyel BET.	0
@Di_719 Amen and thank you for reminding me!🙏😊	0
Day 140: Paul Gardner portrayed by Jared Leto https://t.co/CLlCau2zVo	1
@Rollinintheseat @carbsley ""I was going on a date & her father said 'I want her home by 9:15.' I said, middle of September? That's cool."" - Steven Wright	1

METHODOLOGY & EDA

Preparing the dataset

METHODOLOGY



EXPLORATORY DATA ANALYSIS (EDA)

Variables

Followers - the number of followers a user has

Favorites - total count of times a user has liked a specific tweet indicating activity

Friends - the number of friends an account has

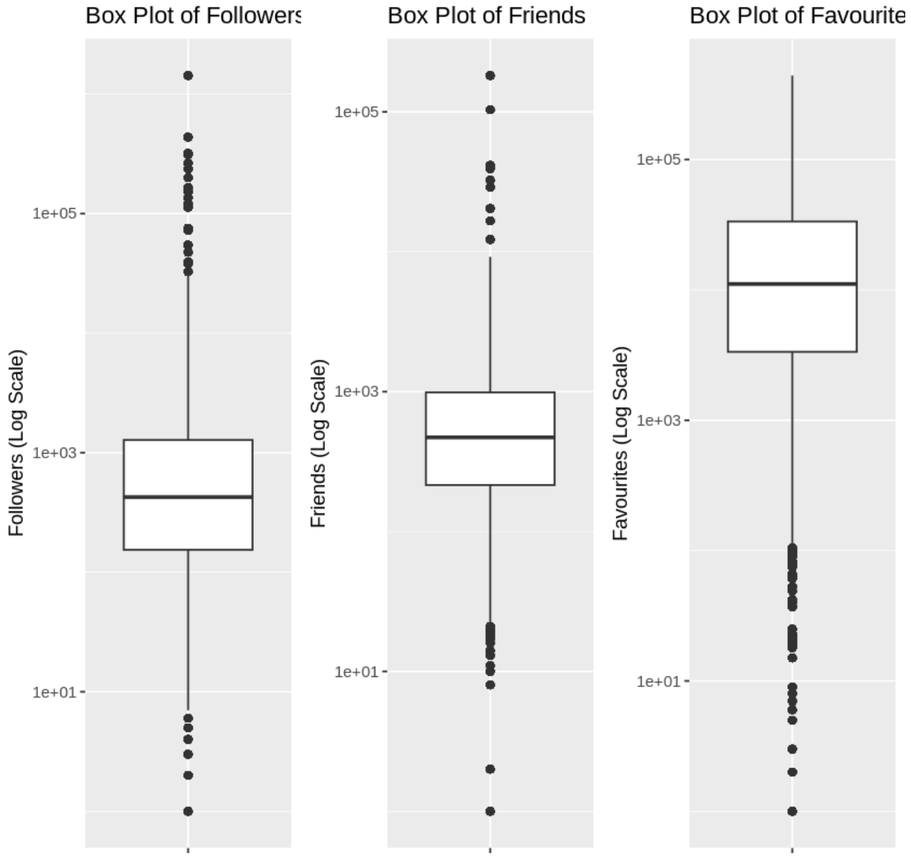
Sentiment - A score between -1 and 1 labeling the connotation of the text, with 1 being postive and -1 being negative

Summary Statistics

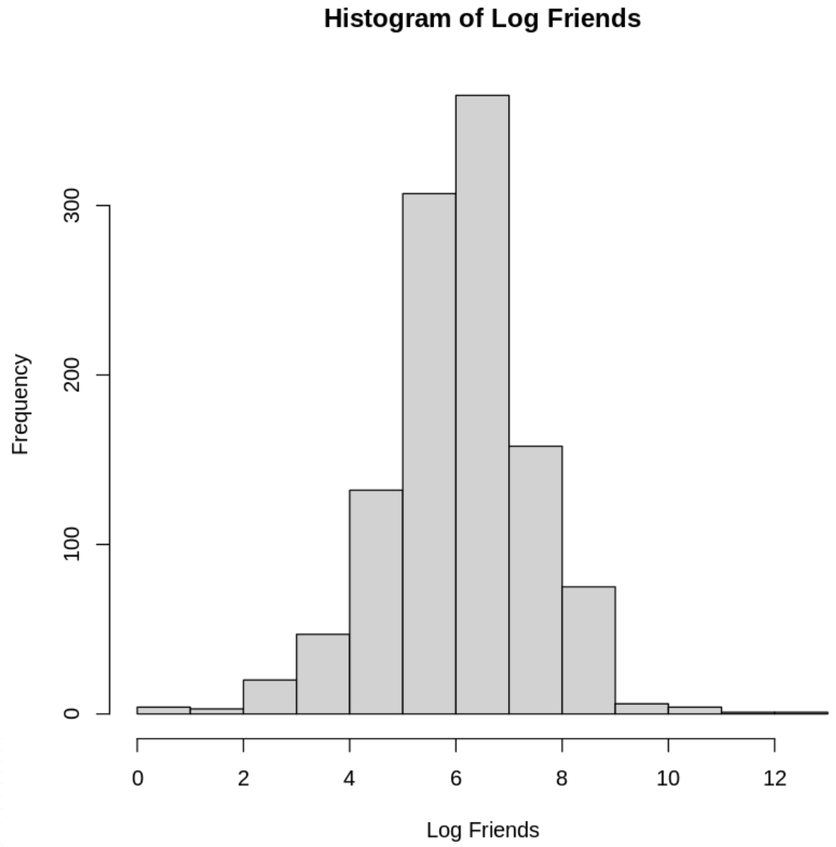
Statistic	Mean	St. Dev.	Min	Max
label	0.492	0.500	0	1
followers	5,232.726	49,174.850	0	1,429,030
friends	1,288.732	6,699.477	0	181,822
favorites	27,733.150	46,752.070	0	441,317
statuses	25,395.580	67,056.790	25	1,527,058
sentiment	0.106	0.117	−0.348	0.736

EDA

Box Plots

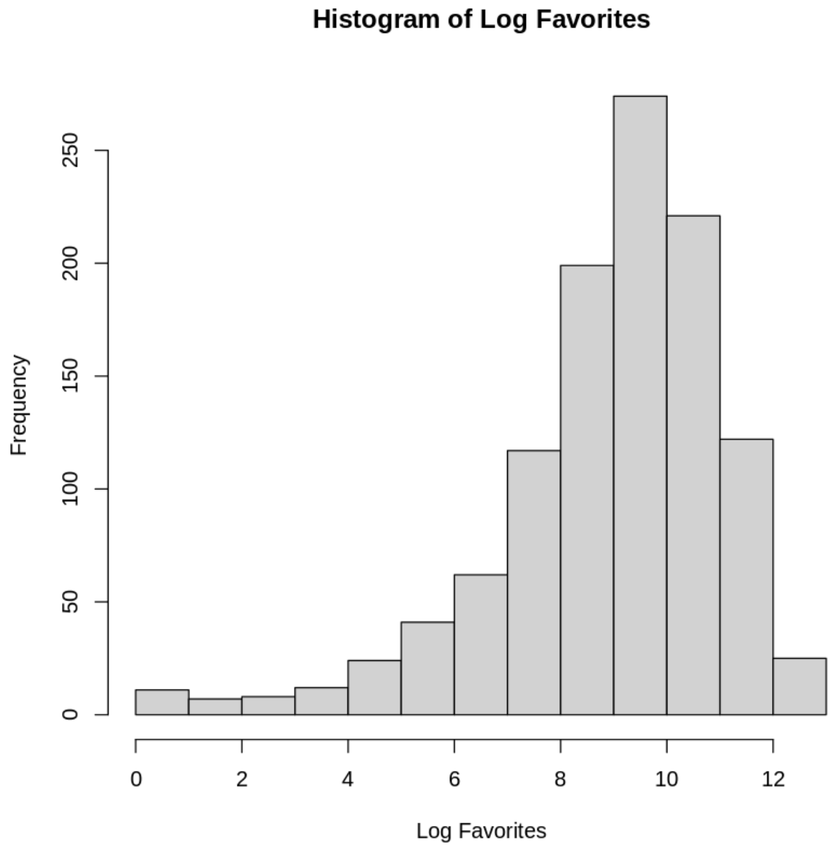


Histogram of Friends

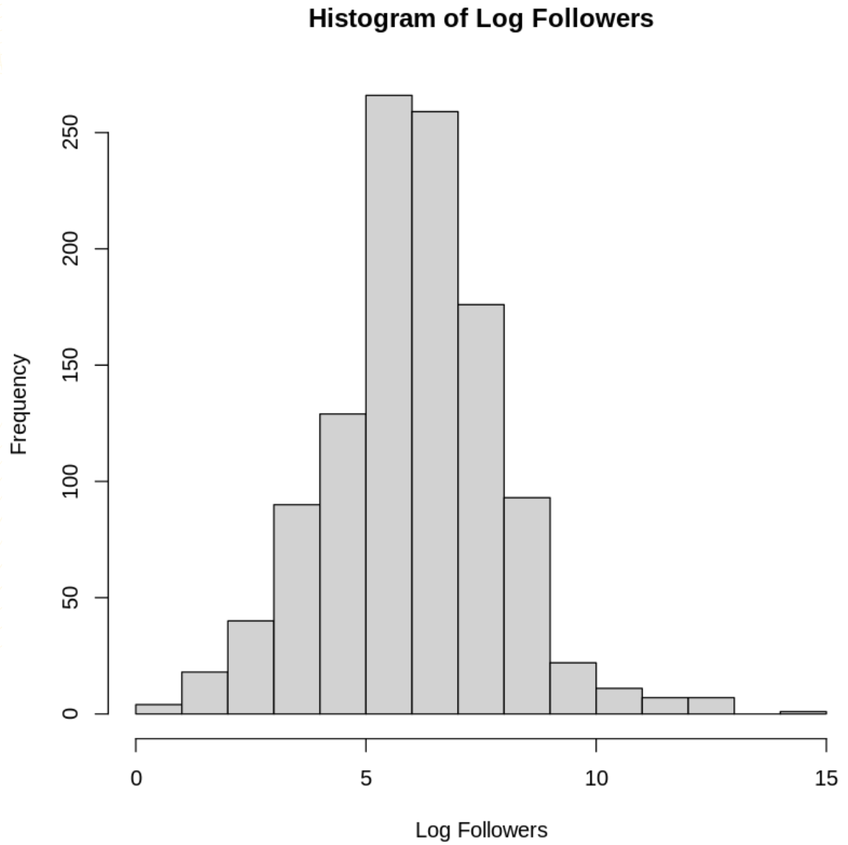


EDA

Histogram of Favorites



Histogram of Followers



PART 1: ANALYSIS OF VARAIBLES

Modeling Depression to Followers, Friends, Favorites, and Sentiment

Linear Probability Model

	Dependent variable:
	label
log_followers	-0.038***
	(0.012)
log_friends	-0.021
	(0.017)
log_favorites	0.047***
	(0.009)
sentiment_score_average	-0.083
	(0.141)
Constant	0.439***
	(0.083)
Observations	899
R2	0.041
Adjusted R2	0.037
Residual Std. Error	0.491 (df = 894)
F Statistic	9.608*** (df = 4; 894)
Note:	*p<0.1; **p<0.05; ***p<0.01

Inverse Relationship

Low R2 Term indicates the potential to include more variables

Linear Model Accuracy	59.82%

Logistic Model

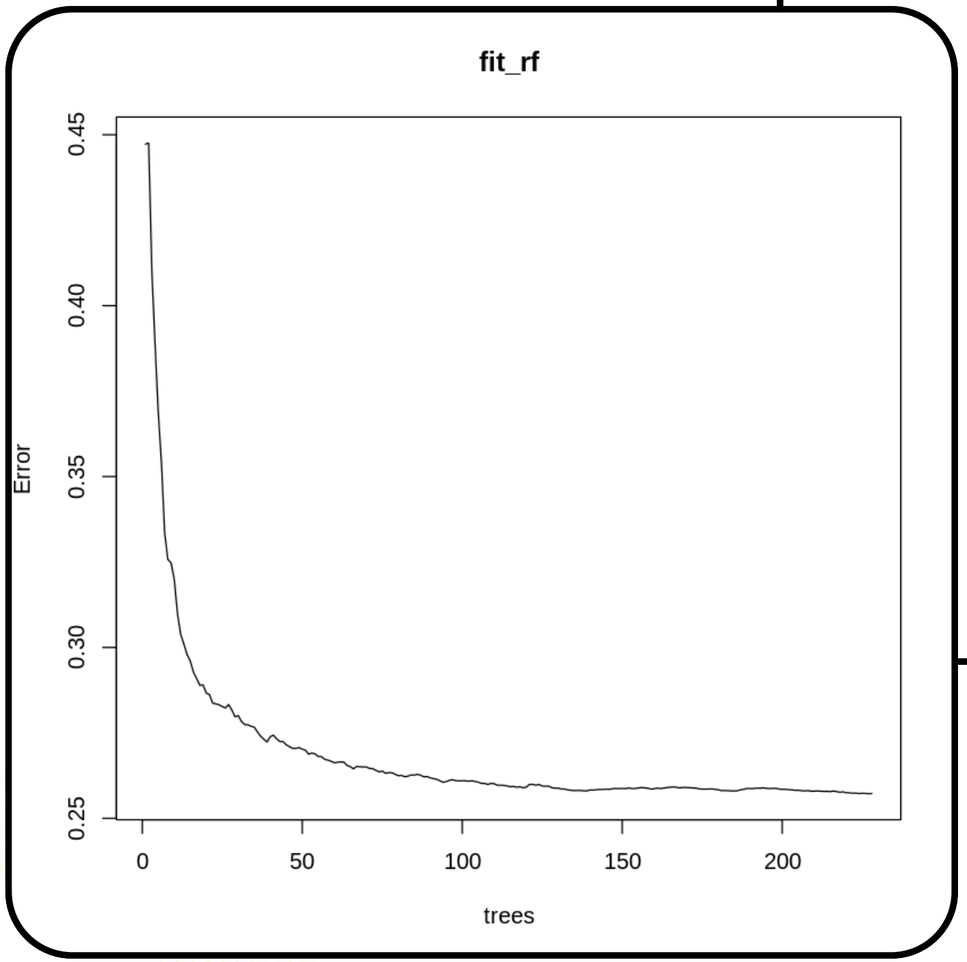
	Dependent variable:
	label
log_followers	-0.173***
	(0.053)
log_friends	-0.085
	(0.074)
log_favorites	0.207***
	(0.041)
sentiment_score	-0.352
	(0.595)
Constant	-0.287
	(0.357)
Observations	899
Log Likelihood	-603.533
Akaike Inf. Crit.	1,217.066

Note: *p<0.1; **p<0.05; ***p<0.01

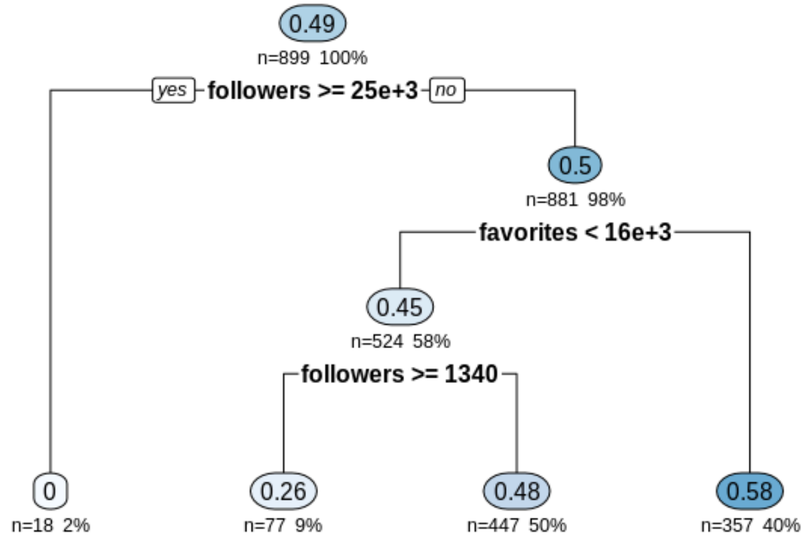
Logstic Model Accuracy	58.93%

Decision Tree & Random Forest

Error Graph for Random Forest



Decision Tree Diagram



Decision Tree Accuracy	52.23%
Random Forest Accuracy	60.27%

Comparison of Models

```
lm_test_preds <- ifelse(predict(fit_lm_1000, data.testset[, -"label"], type="response") > 0.5, 1, 0)
lm_test_accuracy <- mean(lm_test_preds == data.testset$label)
```

Model	Testing Accuracy
Linear Model	59.82%
Logistic Regression	58.93%
Decision Tree	52.23%
Random Forest	60.27%

PART 2: TEXT MINING

Using post_text to predict depression

Bag of Words

Process

01

DATA-CLEANING

- Corpus
- Removing & Cleaning

02

N-GRAMS

- Concatenation

03

WORD FREQUENCY TABLE

- Reduce bag size (>1%)
- Document Term Matrix

04

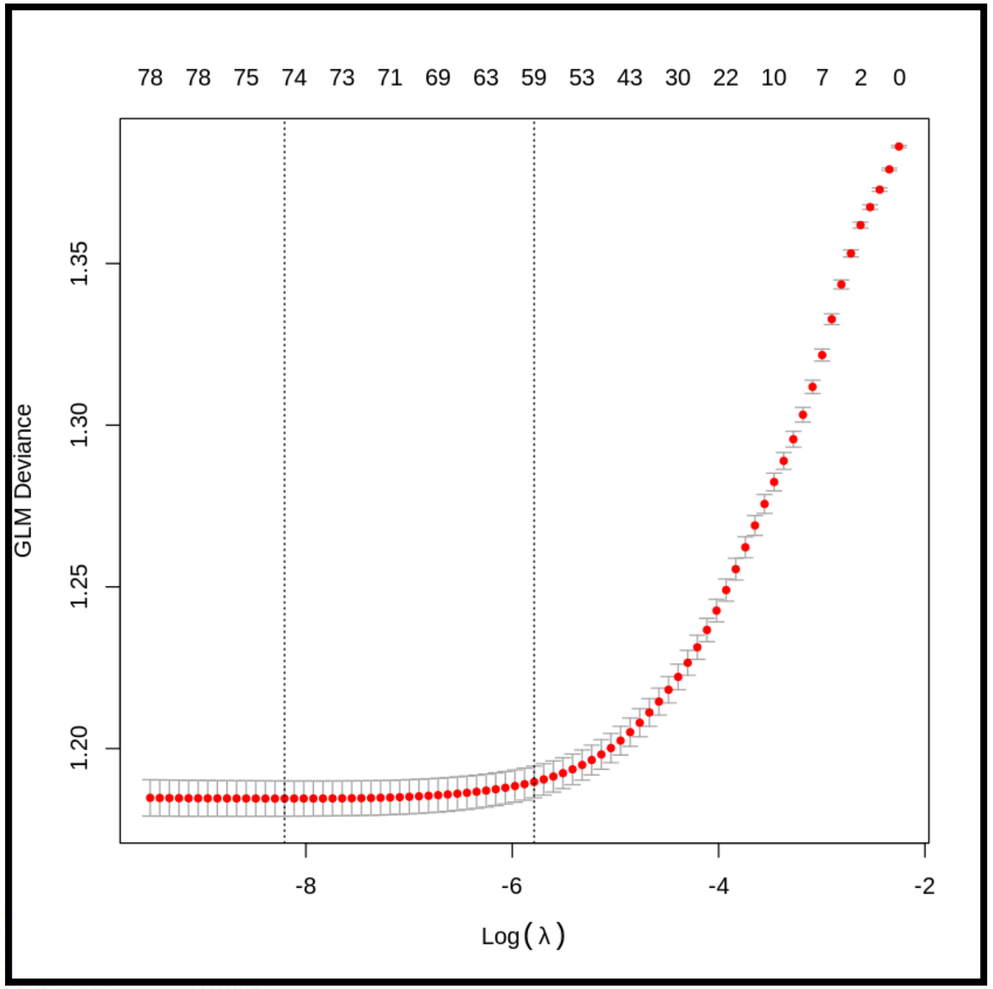
ANALYSIS

- LASSO
- Relaxed LASSO
- Word-Cloud

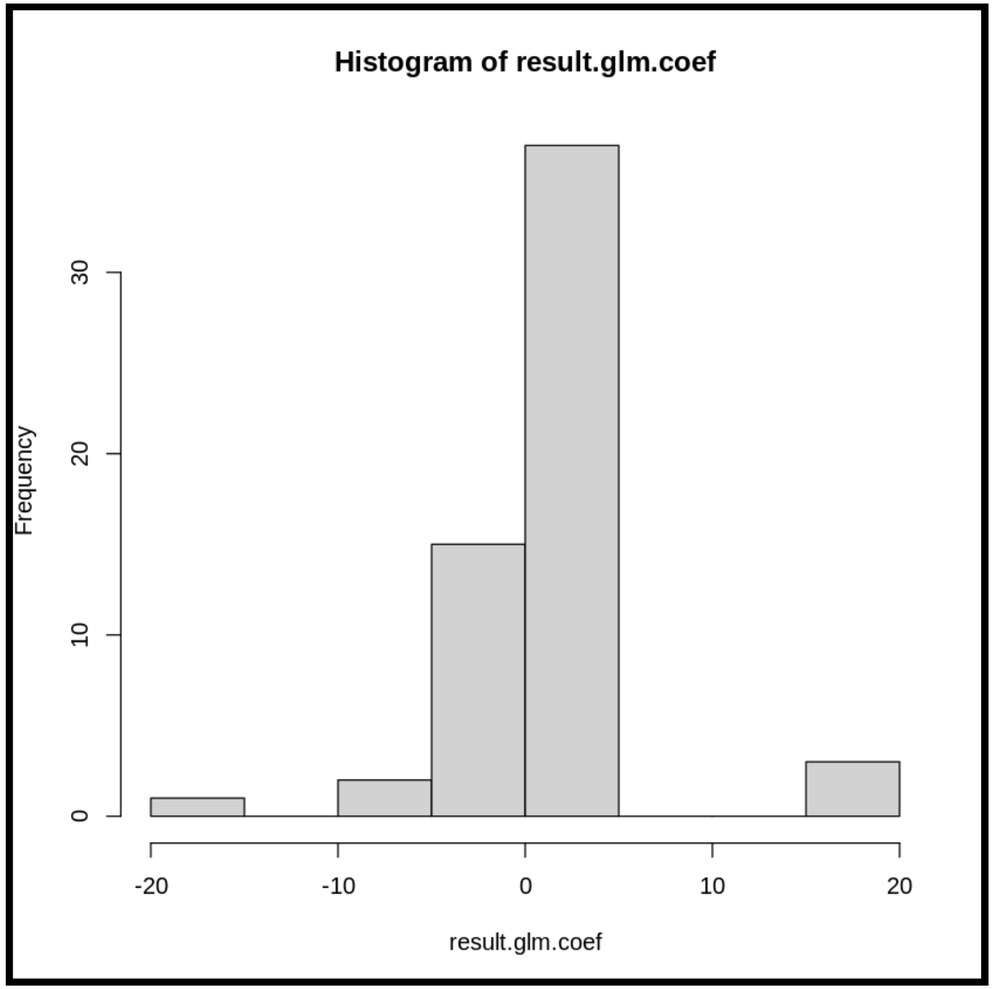
Bag of Words

Kaggle Dataset (2015)

LASSO Plot



Relaxed LASSO

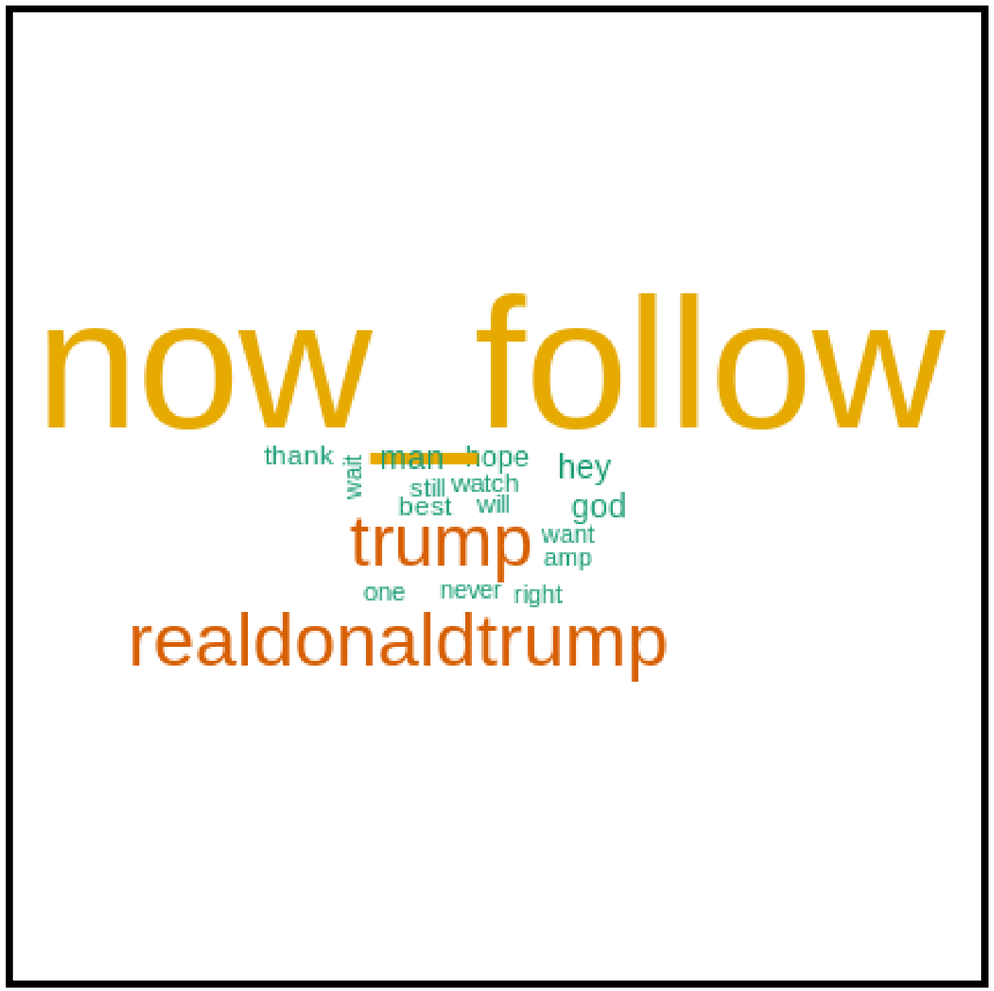


Bag of Words

Kaggle Dataset (2015)

+ Correlation

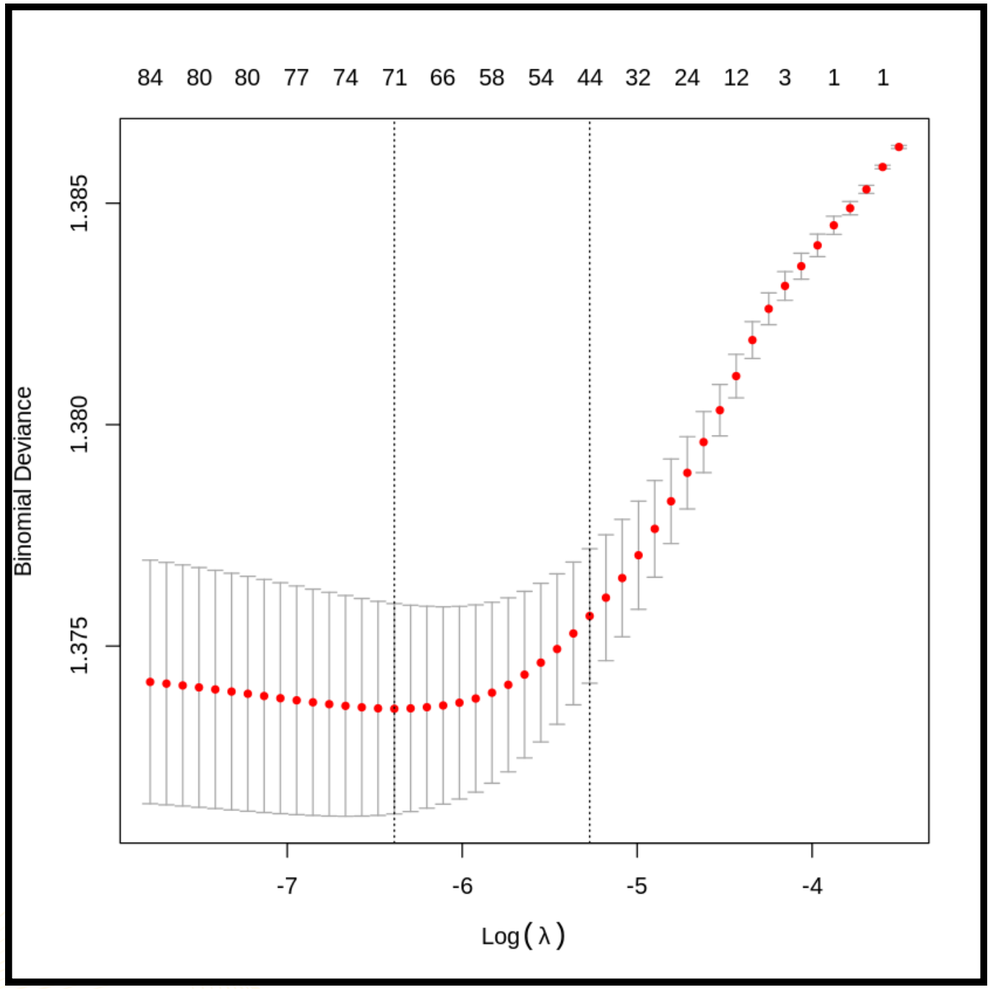
– Correlation



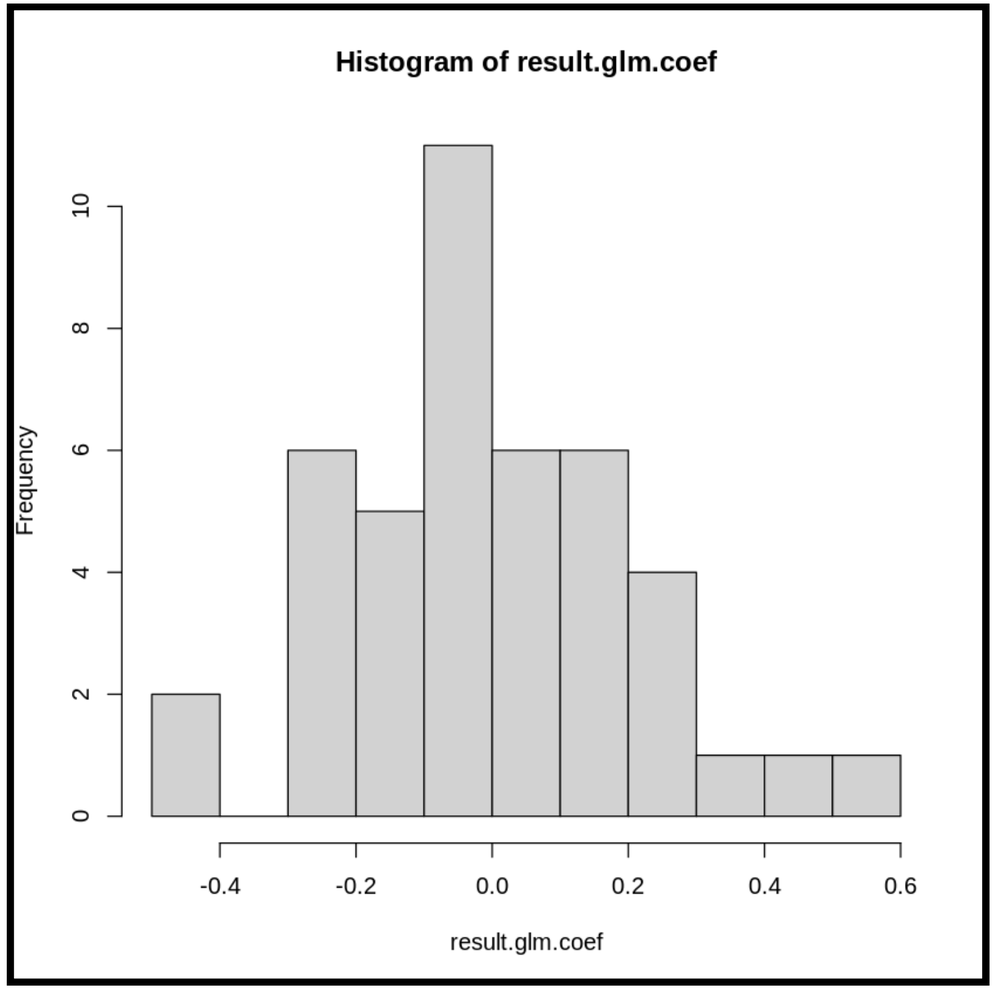
Bag of Words

Safa Dataset (2020)

LASSO Plot



Relaxed LASSO

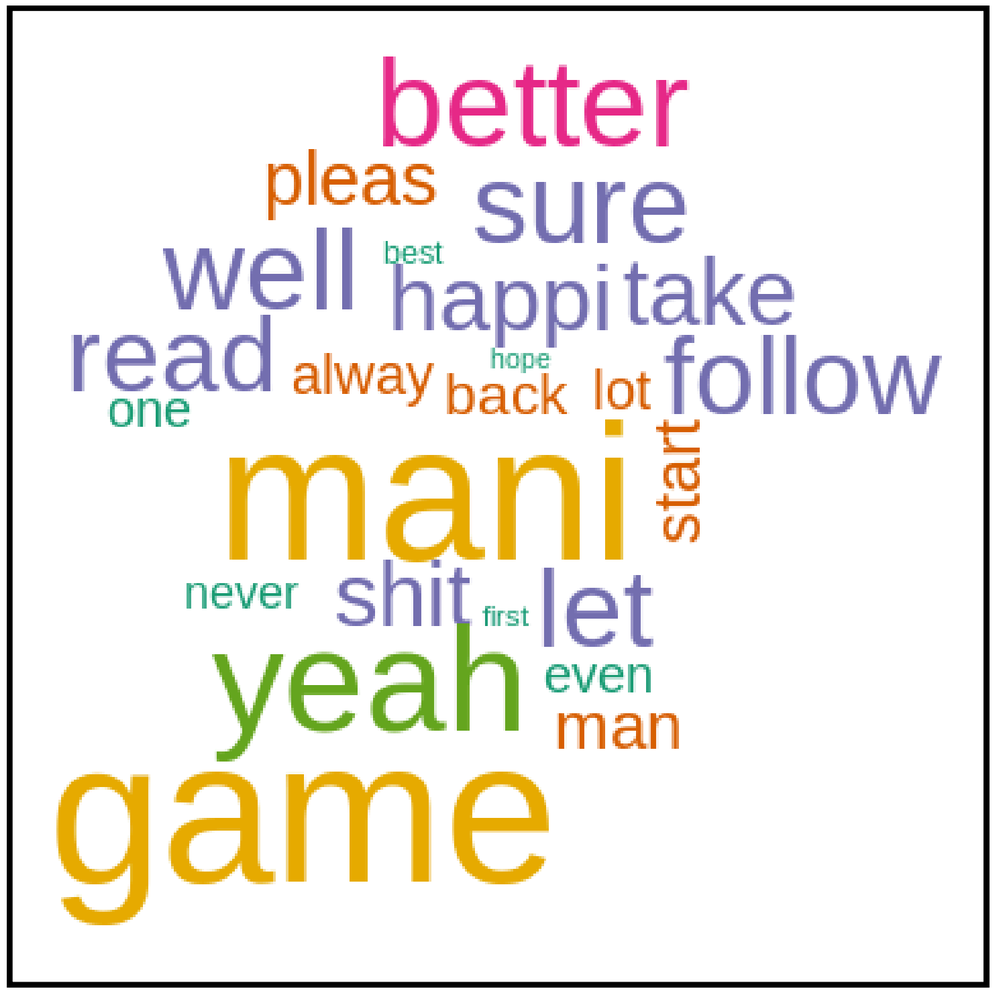


Bag of Words

Safa Dataset (2020)

+ Correlation

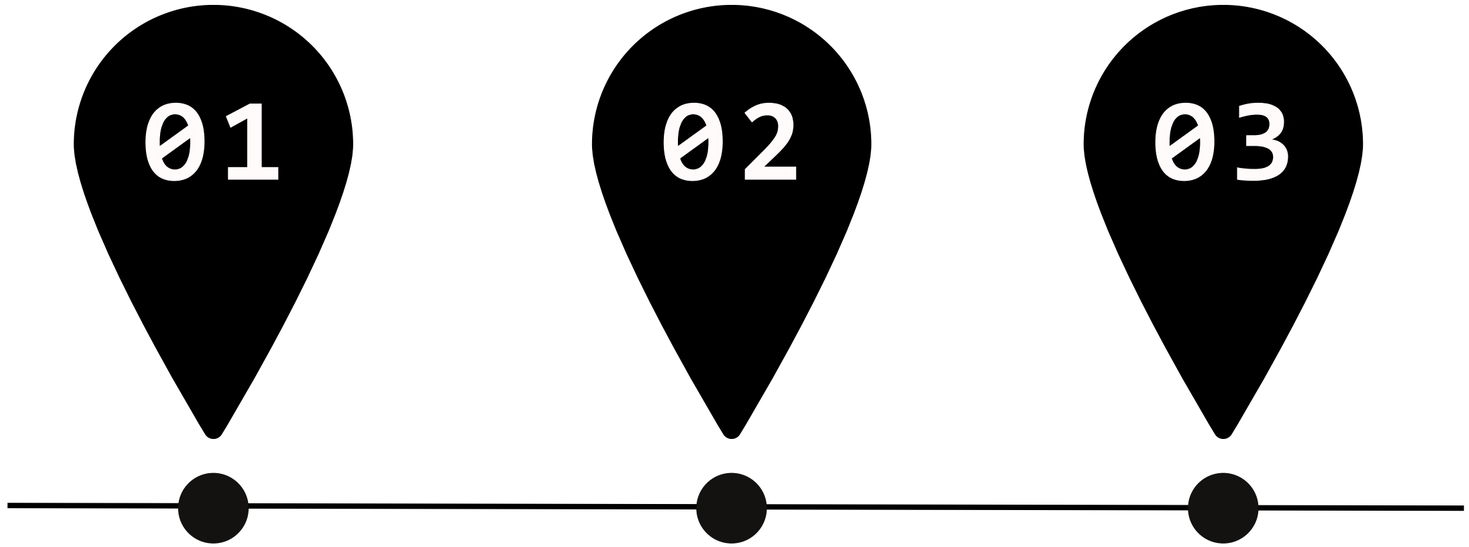
– Correlation



BERT

Bidirectional Encoder Representations from Transformers

Process



DATA-PREPROCESSING

- Select text and labels
- Tokenize text for input

HYPERPARAMETER SEARCH

- Batch size
- Number of epochs
- Learning rate

TRAIN AND EVALUATE MODEL

- Find best model based on accuracy

BERT

Bidirectional Encoder Representations from Transformers

Tokenizer

Classification Layer

(num_labels=2)

- **0** (Not Depressed)
- **1** (Depressed)

AIMH/mental-bert-base-cased · Hugging Face

We're on a journey to advance and democratize artificial intelligence through open source and open science.

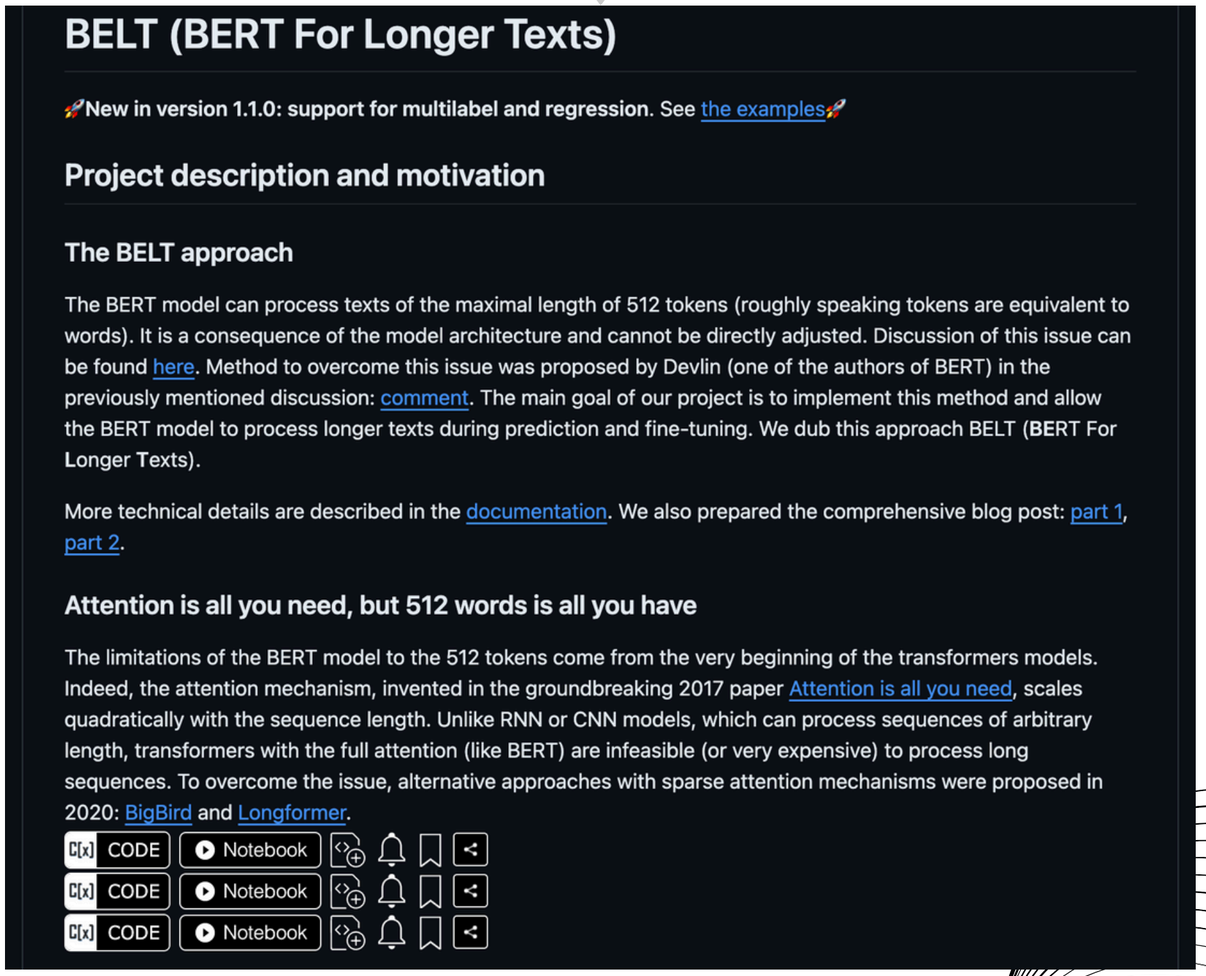
huggingface

BELT

BERT For Longer Texts

https://github.com/mim-solutions/bert_for_longer_texts

Concatenate Tweets



BELT (BERT For Longer Texts)

🚀**New in version 1.1.0: support for multilabel and regression**. See the examples🚀

Project description and motivation

The BELT approach

The BERT model can process texts of the maximal length of 512 tokens (roughly speaking tokens are equivalent to words). It is a consequence of the model architecture and cannot be directly adjusted. Discussion of this issue can be found here. Method to overcome this issue was proposed by Devlin (one of the authors of BERT) in the previously mentioned discussion: comment. The main goal of our project is to implement this method and allow the BERT model to process longer texts during prediction and fine-tuning. We dub this approach BELT (**BE**RT For Longer Texts).

More technical details are described in the documentation. We also prepared the comprehensive blog post: part 1, part 2.

Attention is all you need, but 512 words is all you have

The limitations of the BERT model to the 512 tokens come from the very beginning of the transformers models. Indeed, the attention mechanism, invented in the groundbreaking 2017 paper Attention is all you need, scales quadratically with the sequence length. Unlike RNN or CNN models, which can process sequences of arbitrary length, transformers with the full attention (like BERT) are infeasible (or very expensive) to process long sequences. To overcome the issue, alternative approaches with sparse attention mechanisms were proposed in 2020: BigBird and Longformer.

C[x] CODE Notebook

C[x] CODE Notebook

C[x] CODE Notebook

Model Comparison

Model	Testing Accuracy
Kaggle LASSO	64.01%
Kaggle LASSO with N-grams	63.76%
Kaggle Relaxed LASSO	64.31%
Kaggle Relaxed LASSO with N-grams	64.07%
Kaggle BERT	**92.76%**
Safa LASSO	54.25%
Safa LASSO with N-grams	54.05%
Safa Relaxed LASSO	53.05%
Safa Relaxed LASSO with N-grams	53.40%
Safa BERT	64.45%
Safa BELT	**92.92%**

FINAL RESULTS

Our findings

LIMITATIONS

The Kaggle dataset only had data of 72 unique user IDs, which heavily hinders our ability to predict depression based on a wholistic view of the user.

LIMITED USER DATA

The labels for depression were obtained through self-reporting, which can be biased and inconsistent, hindering the data's reliability and validity. These can introduce errors into the model, affecting its accuracy.

SELF-REPORTING

The two datasets only had followers, friends, likes, favorites, and post text as their predictors. Many other factors that can impact predictions such as age, income, and location were not considered.

LACK OF PREDICTORS

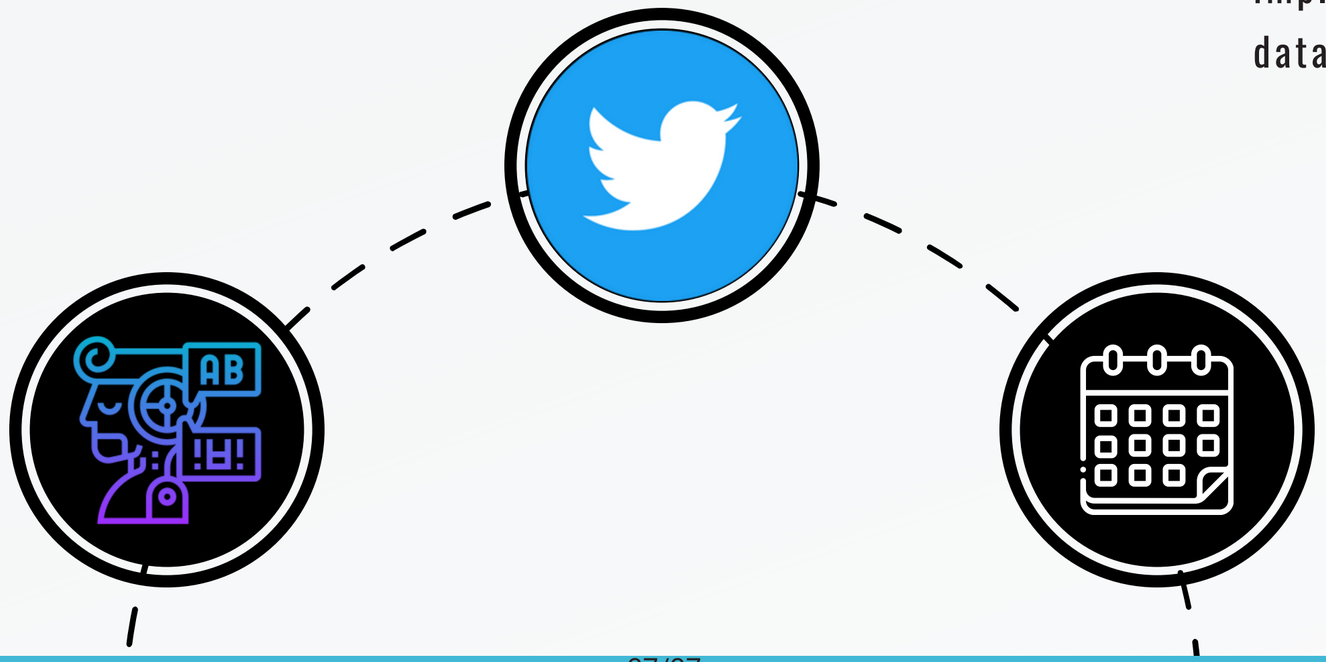
IMPLICATIONS

Predictive Factors

Number of favorites
(+ Correlation)
Number of followers
(- Correlation)

Model Results

Models had higher accuracy on the 2015 dataset compared to the 2020 dataset



Future Directions

1. Train models on more data
2. Create a model that implments both quantative data and natural langauge