DataJCampos / **COVID-19-Tweets-NLP** Public

☆ **2** stars 0 **forks**

Starred ▾ Unwatch ▾

Code Issues Pull requests Actions Projects Wiki Security

main ▾ ···

DataJCampos updated readme ··· 1 minute ago **6**

.gitignore	yesterday
Covid-19 Tweets NLP.ipynb	12 minutes ago
Predictive Analysis Covid-19 Tweets .pdf	12 minutes ago
README.md	1 minute ago

README.md

Tweet Analysis Using Natural Language Processing Model (NLP)

By: Jeremias Campos

Introduction

A Twitter executive has hired me to develop an algorithm that can detect whether a tweet is related to covid or not.

Project Goal

For this project, I shall be combining 2 datasets including covid tweets and noncovid tweets. This will be able to help the Twitter team to identify the covid related tweets and allow them to connect their users with covid-19 resources developed by official health organizations.

To achieve this goal, we shall be using natural language processing (NLP).

Data Understanding

Covid-19 related dataset was collected from Kaggle, which consists of columns including information such as:

- COLUMN NAMES: Location, date of tweet, original tweet, sentiment (positive, negative, neutral), etc.
- total number of tweets: 44955 https://www.kaggle.com/datatattle/covid-19-nlp-text-classification

- We shall look specifically at the original tweet text, and develop a column with a “target” of 1 showing it is related to covid-19.

The non-covid 19 dataset was also collected from Kaggle, consisting of columns including:

- COLUMN NAMES: Target (negative, positive, neutral), ids, date, flag, user, and tweet
- total number of tweets: 1.6 million https://www.kaggle.com/kazanova/sentiment140
- The tweet column shall be used and a “target” column of 0 shall represent that the tweets are unrelated to covid.

Metrics

Our project will answer following question:

Can we predict tweets related to covid?

Hypothesis:

```
H0 - The tweet is related to covid.

HA - There is statisticaly significant proof that the tweet isn't related
to covid.
```

TP, TN, FP, FN definition

```
TP - we predicted covid tweet and it actually exist.

TN - we predicted that tweet isn't covid related and the tweet actually
isn't related to covid.

FP - We predicted covid tweet but it was not a covid tweet.

FN - We predicted that there is no covid tweet but it actually existed.
```

Metrics used

```
To compare models we will focus on 2 major metrics:

Recall - We will be focused to minimize FN.

Accuracy - how good we can predict TP and TN. General metrics that will
show model performance.
```

Data Preparation

1. Covid-19 dataset
    - Remove unrelated columns
    - Train/test sets already separated
    - Ended up with 44955 tweets
2. Noncovid-19 dataset
    - Create sample of 44955
    - Remove unrelated columns
3. Data Preprocessing
    - Remove unnecessary numbers, punctuations, etc.
    - Tokenization
    - Lower casing
    - Stop words removal
    - Stemming
    - Lemmatization
4. Create dataset with both
    - Concat test/train split
    - Train for both covid & non-covid dataset
    - Test for both covid & non-covid dataset
5. Vectorizer
    - Count Vectorizer
    - TF-IDF Vectorizer

*These vectorizers were selected to look further into training, validation, and ultimately our test set. Parameters were altered including min_df, max_features, etc. to increase performance of our models.

Modeling

After the data preparation was complete, 4 models were created including: - Linear Support Vector Classifier - Multinomial Naive Bayes - Decision Tree Classifier - Random Forest Classifier

As modeling began, we used both vectorizers with the training and validation set. The TD-IDF performed better than count vectorizer on the models so we utilized this to finalize our metrics. After we ran a cross validation on the all models using TF-IDF. Random tree classifier ended being the best model which ultimately was used to confirmed our metrics using the testing set.

Conclusions

Based on results our final model will be: "Random Forest Classifier using TF-IDF"

With the following parameters after tuning:

```
Accuracy - 0.99577

Recall - 0.99242
```

Because of the following reasons:

1. It has high accuracy and recall.

In a future project, additional preprocessing steps & token patterns can be used for count vectorizer + tf-idf to attempt to get better metrics.

In conclusion, this data tells us that twitter can use NLP to predict whether if a tweet is related to covid and be able to provide resources to those positing about it.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● **Jupyter Notebook** 100.0%