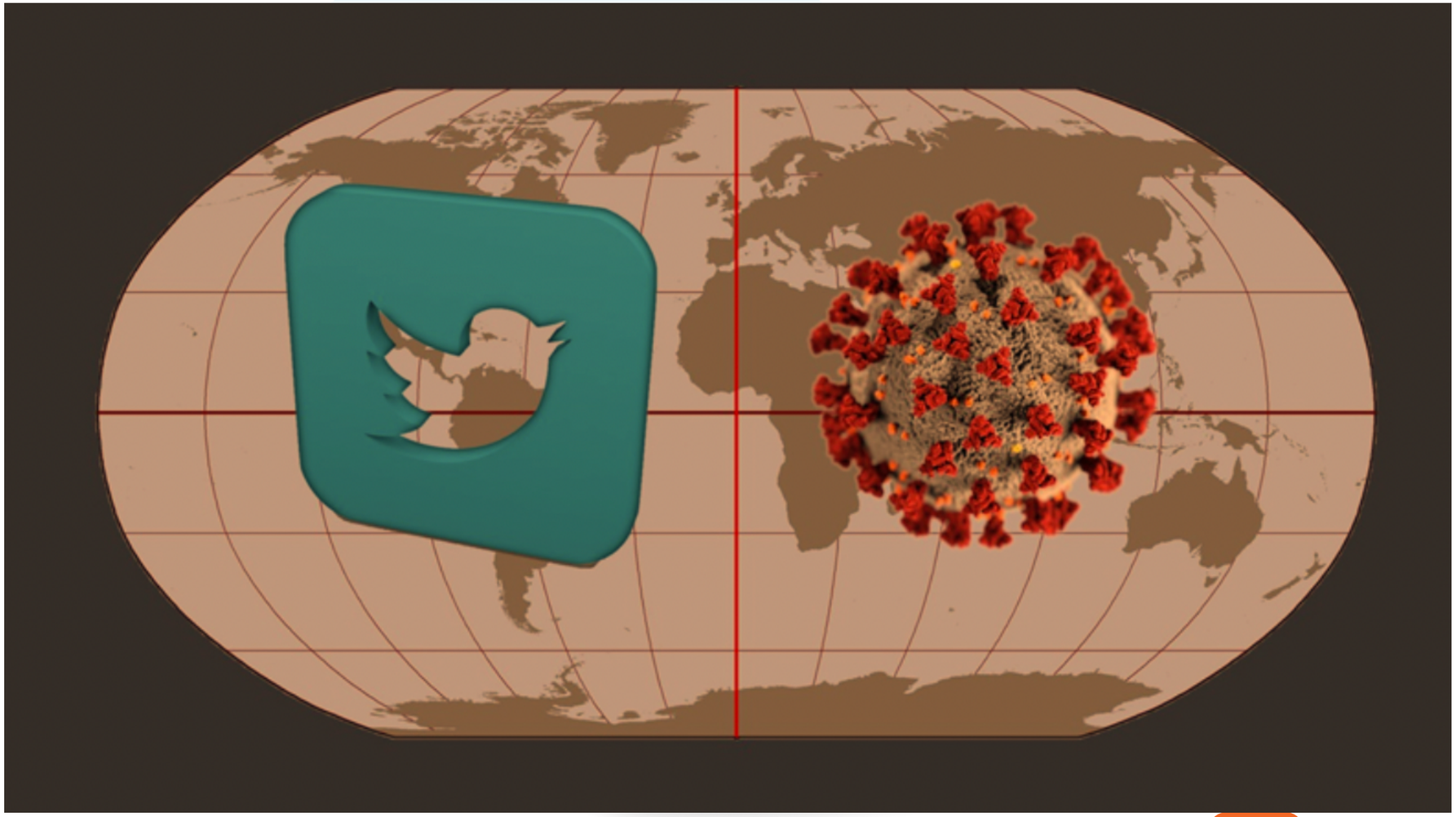
Jeremias Talks:

Predictive Analysis: Covid-19 Tweets



By: Jeremias Campos,
Data Scientist


COVID-19 will reshape our world.
We don't yet know when the crisis will end.
But we can be sure that by the time it does,
our world will look very different.
JOSEP BORRELL

Table of Contents

Introduction

- Business Problem
- Data Sources & Methods

Business Problem

- Our team was hired by unnamed **Twitter** executive to create a model which automatically classifies if a tweet is related to covid.

- This shall allow them to connect their users with covid-19 resources developed by official health organizations.

Data Understanding

METHODS:

- Natural Language Processing (NLP)

Data Sources:

- Covid-19 related dataset:
  - ~45,000 tweets

- Non-covid 19 dataset:
  - ~ 1.6 million tweets
    - Took sample of ~45,000 tweets

Use Cases of NLP

SOURCE

Metrics

Our priority is finding covid related tweets.

Hypothesis:

- H0 - A tweet is related to covid.
- HA - The tweet doesn't' have covid related information.

- **Recall** - Focused on finding covid related tweets
- **Accuracy** - How accurate our results are considering false identified tweets

Research

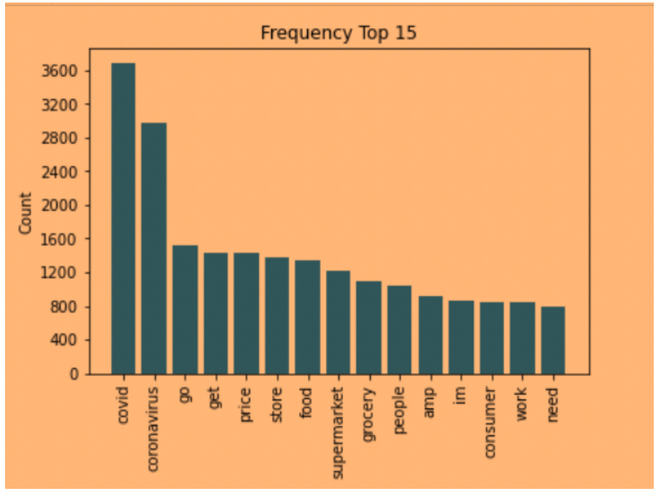
- Raw Data
- Modeling
- Model Results

Steps Before Model Begins

1. Preprocessing
    a. English words only
    b. Removal
        i. number
        ii. punctuations
        iii. mentions
        iv. hashtags
        v. urls & html links
    c. lowercase words
    d. Lemmatizer
2. Set up vectorizers for modeling process
3. Modeling

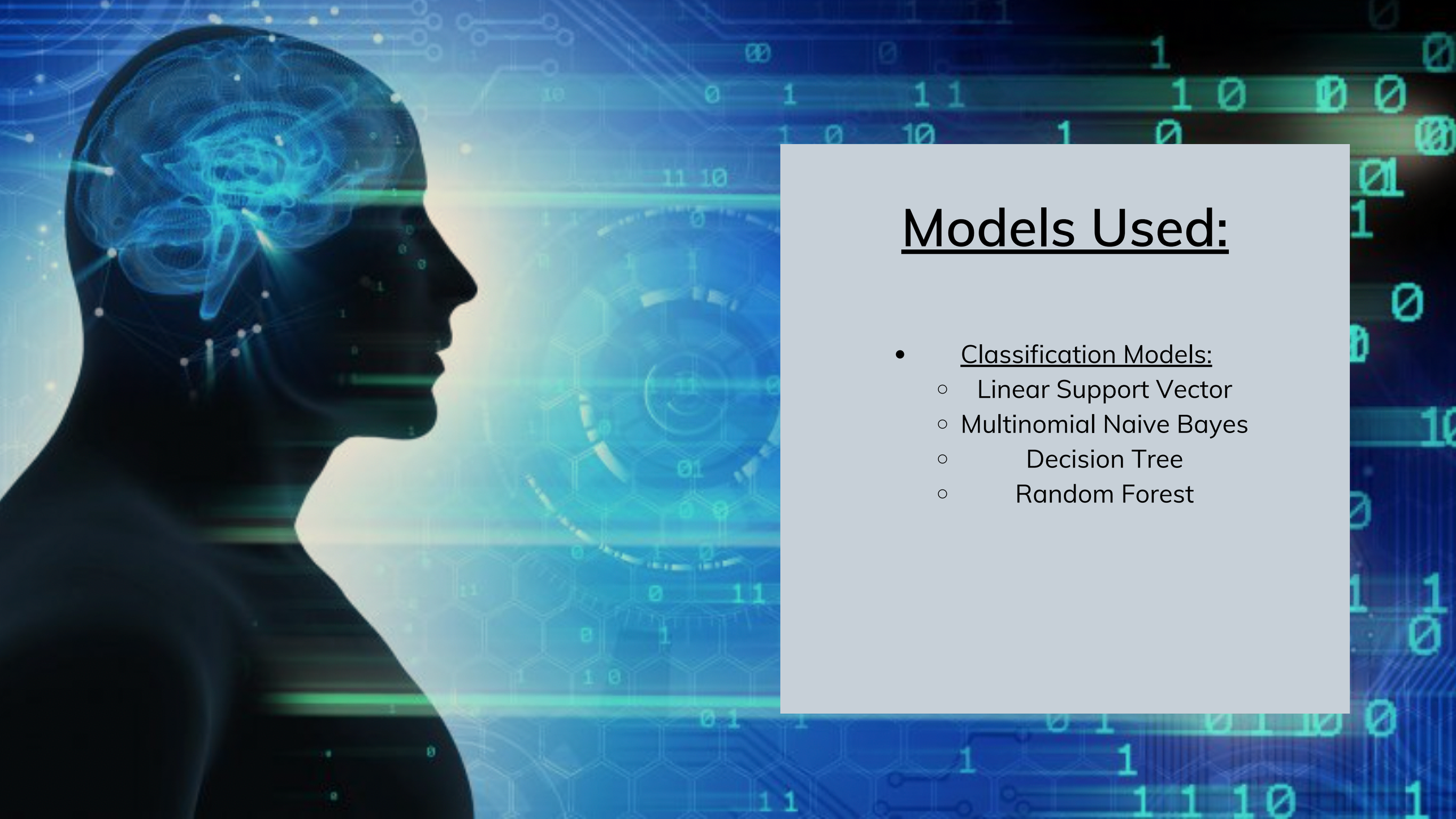
Covid Related Data

- ~45,000 tweets
- Main 2 words found in the tweets:
  - covid
  - coronavirus



Non-covid Related Data

- ~45,000 tweets
- Main 2 words found in the tweets:
  - covid
  - coronavirus



Models Used:

- Classification Models:
  - Linear Support Vector
  - Multinomial Naive Bayes
  - Decision Tree
  - Random Forest

Model Results!

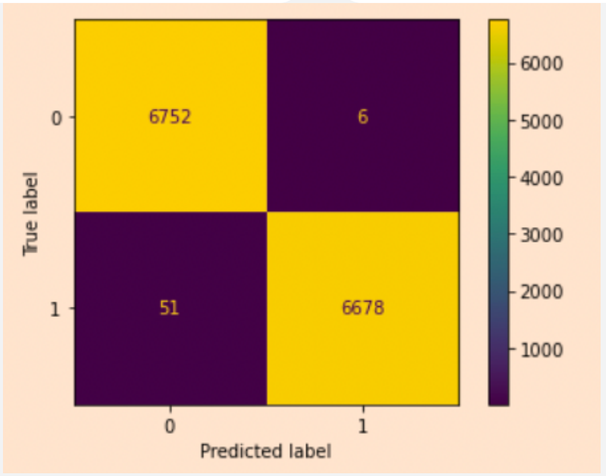
Random Tree Classifier was our best model!

RECALL

~99%

ACCURACY

~99%



Finalize

- Recommendations
- Next Steps
- Q & A?

Recommedations

#1

Additional preprocessing steps can be used for count vectorizer + tf-idf to attempt to get better metrics.

#2

Additional token patterns can be examined for count vectorizer + tf-idf to attempt to get better metrics.

Next Steps

- Extend the model to different trending health issues.
- Extend model to different languages.

Q & A:

Gracias for joining today's presentation!

JEREMIAS CAMPOS

- DATA SCIENTIST
- Email: JEREMIASCAMPOS3@GMAIL.COM
- LINKEDIN: /JEREMIASCAMPOS
- GITHUB: DATAJCAMPOS
- MEDIUM: @JEREMIASCAMPOS3