# THE STATE OF INFODEMIC ON TWITTER

BRANDON ATTAI, KELTEN FALEZ , TAHSIN CHOWDHURY

# DATA COLLECTION POLITIFACT

## Use Twitter API

- Tweepy
- Set up account
- Set access token
- Authenticate

## Query

- Use verified article titles as search query.
- Limited to tweets in English only.
- Work around the APIs rate limit.

## Inspect Response

- Pulled tweets were relevant to the titles.
- All tweets were very recent.

## Parse Response

- Response is a massive JSON object with lots of metadata.
- Use Tweepy methods to access and save the ones relevant to the project.

## Prepare Dataframe

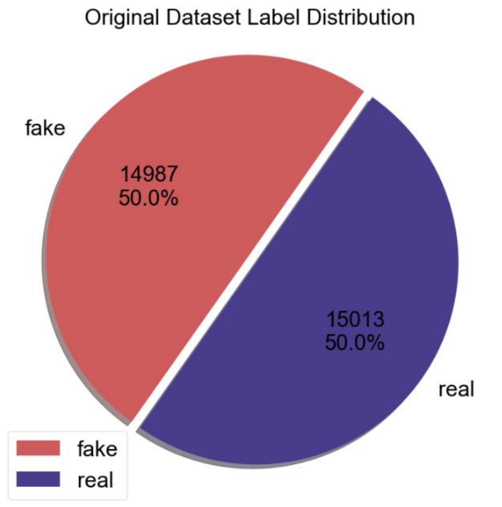
- Use the saved data from the response to build a dataframe.
- Save it as csv.

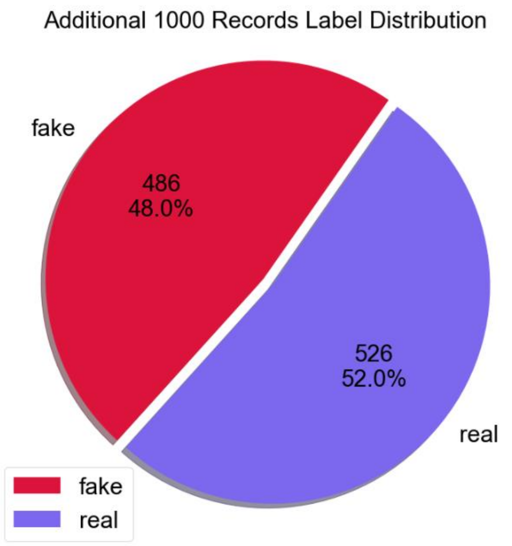
# LABELLING METHODOLOGY

- Does the tweet agree/disagree with well-known guidelines set by WHO, CDC, etc?
- Verify claim using fact-checking websites such as Politifact, Snopes, Healthfeedback.org
- Check account info such as account creation date, account handle, number of followers, replies etc.

# LABEL DISTRIBUTION

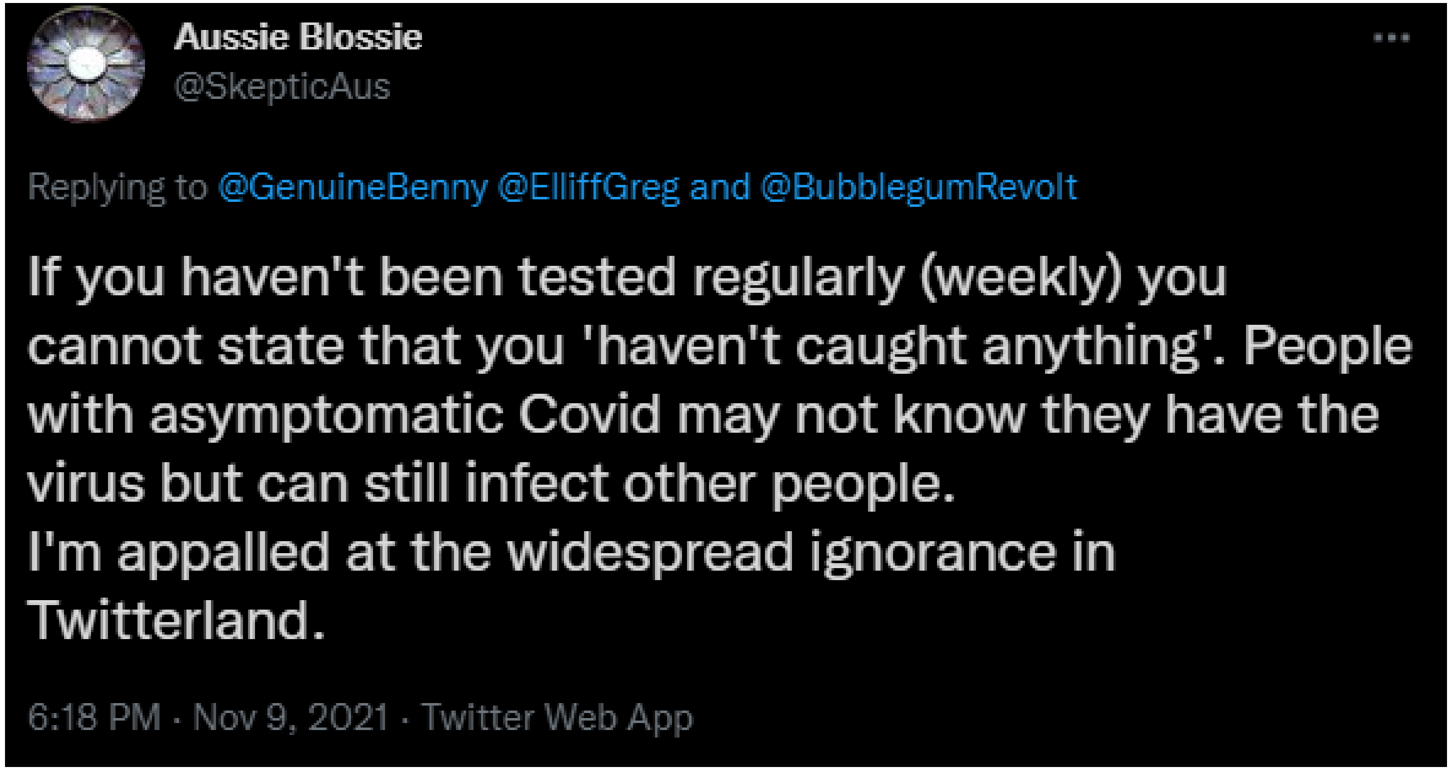
~ APPROX. 50-50 SPLIT



# LABELLING EXAMPLES - REAL

- Says asymptomatic COVID patients may be unaware they are sick and still infect other people.


Aussie Blossie
@SkepticAus

Replying to @GenuineBenny @ElliffGreg and @BubblegumRevolt

If you haven't been tested regularly (weekly) you cannot state that you 'haven't caught anything'. People with asymptomatic Covid may not know they have the virus but can still infect other people.
I'm appalled at the widespread ignorance in Twitterland.

6:18 PM · Nov 9, 2021 · Twitter Web App

https://twitter.com/SkepticAus/status/1458242576822116358

# ACCOUNT INFO – REAL

- Account over 7 years old and active.
- Handle not automatically generated.


https://twitter.com/SkepticAus

# FACT CHECK - REAL

- Politifiact is a well-respected fact-checking website.

POLITIFACT
The Poynter Institute

## COVID-19 can be transmitted by people without symptoms

### IF YOUR TIME IS SHORT

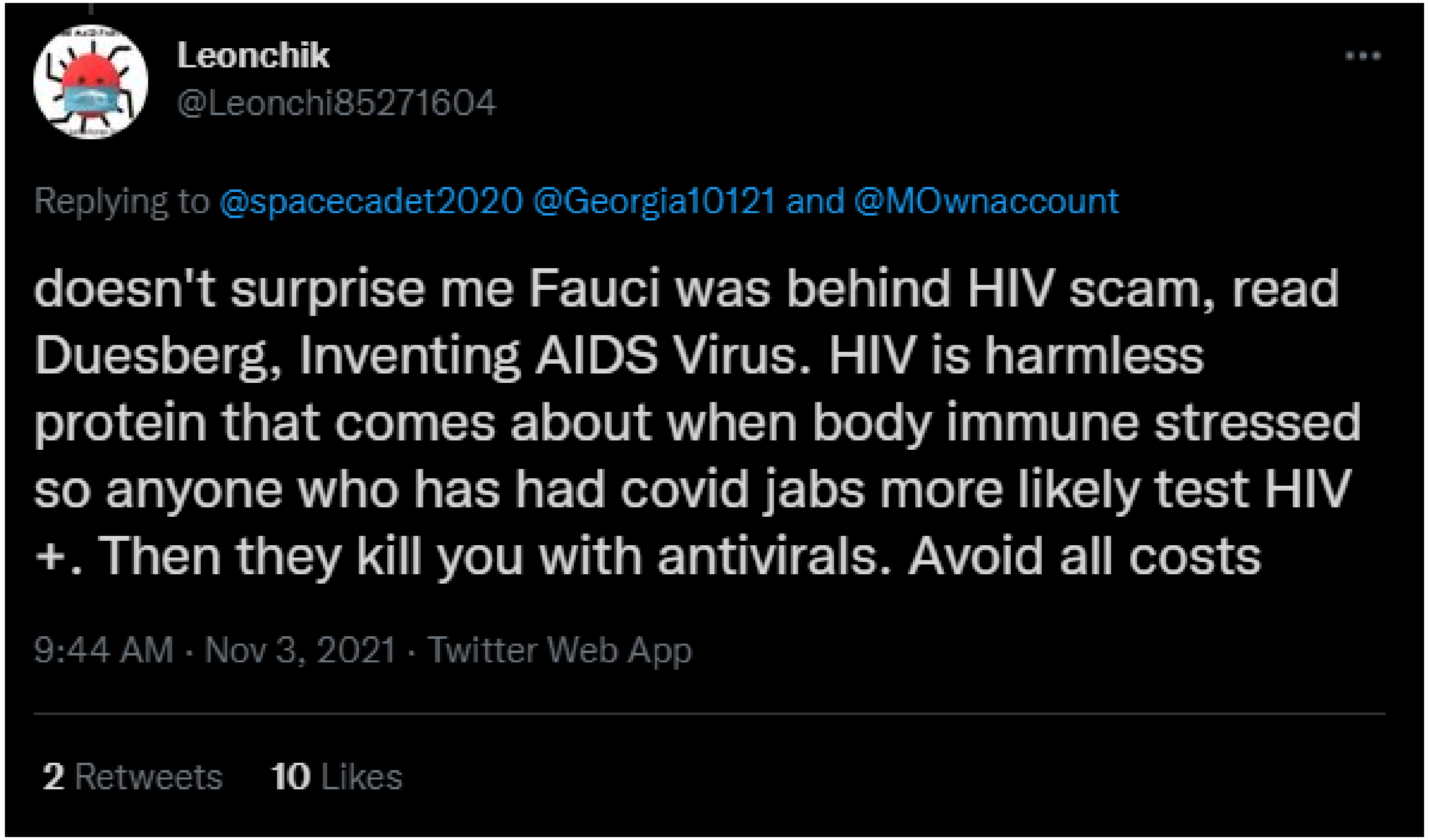
- Multiple studies have concluded that individuals who test positive for COVID-19 can transmit the virus to others, even if they show no symptoms.
- There is no consensus estimate on how frequently asymptomatic people transmit the virus to others.

**See the sources for this fact-check**

https://www.politifact.com/factchecks/2021/may/17/instagram-posts/covid-19-can-be-transmitted-people-without-symptom/

# LABELLING EXAMPLES - FAKE

- Says Dr. Fauci "invented" AIDS.
- Says Covid vaccines can lead to being HIV positive.




https://twitter.com/Leonchi85271604/status/1455923876983099398

# ACCOUNT INFO – FAKE

- No profile photo.
- Very newly created account.
- Automatically generated user handle.
- Low follower count.


https://twitter.com/Leonchi85271604

# FACT CHECK - FAKE

- Snopes.com is a well-respected fact checking website.
- Politifiact is a well-respected fact-checking website.

SAVE OUR SNOPES

Snopes

Search Snopes.com

Become a Member   Submit a Topic   Shop   Latest   Top   Fact Checks   Collectic

News › Medical

## Fauci's Guinea Pigs? Smear Campaign Rehashes 1980s HIV Clinical Drug Trial

Social media posts falsely claim that Dr. Anthony Fauci "murdered disabled children" in pursuit of an AIDS vaccine in the 1980s.

By Dan Evon

Published 27 October 2021, Updated 3 November 2021

https://www.snopes.com/news/2021/10/27/fauci-aids-drug-trial-on-kids/

POLITIFACT

The Poynter Institute

IF YOUR TIME IS SHORT

- An attempted COVID-19 vaccine that contained a fragment of an HIV protein was dropped because it led to some false-positive HIV test results.
- Researchers said there was no possibility the vaccine caused HIV infection and routine follow-up tests on trial participants confirmed no HIV virus present.

https://www.politifact.com/article/2021/jul/01/fact-checking-tiktok-video-nixed-covid-19-vaccine-/

# QUALITY OF DATA LABELLING

- Inter-Rater Reliability Method used to judge the quality of labelling: **Percentage Agreement**.
  - Add the number of agreements between each rater pair for each sample and then calculate the mean of that for the whole dataset.

**60%**

- Disagreements were resolved by accepting the label that was the majority.

# DATA PREPROCESSING

- Social Media Data
  - Social media presents a unique challenge, since a lot of content is text, as such NLP-based systems are required to work with social data and bring out insights.
  - The volume of data generated on Twitter in 2020:
    - Average of 6,000 tweets per second.
    - Approximately 350,000 tweets sent per minute.
    - Approximately 330 million active users per month
  - Social media presents a unique challenge, due to the large amount of textual data. Social platforms are the largest generators of unstructured natural language data.

https://www.statista.com/topics/737/twitter/#dossierKeyfigures

# UNIQUE CHALLENGES TO SOCIAL MEDIA TEXT DATA

| Standard Language | Social Media Language |
| --- | --- |
| Single Language | No Grammar |
| Single Script | Non-Standard Spelling |
| Formal | Special Characters (eg. Hashtags, emojis etc.) |
| Grammatically Correct | Constantly Evolving Vocabulary |
| Few or no spelling errors | Length of Text |
| Few non-textual elements such as emoticons, images etc. | Highly Informal |

Text data from social media is highly informal compared to text data from standard sources such as books etc. Special user defined functions and libraries were used in the preprocessing stages to specifically handle tweet data to remove noise and unnecessary information.

# PRE-PROCESSING PIPELINE

## Replacing Emojis

- Demoji Toolkit
- Replaces emojis with word representation
- ☺ - becomes 'smiley face'

## Remove URLs

- URLs add noise to the tweet data
- Twitter API has feature for obtaining URLs in tweet metadata
- Regex created to remove all URLs

## Digits and Punctuation Removal

- Regex to remove digits from the document
- Regex to remove punctuation such as capitalization, apostrophes etc.

## Spell Correction

- Replace incorrectly spelled words with correctly spelled words
- SpellChecker Toolkit
- Eg. 'yessssssssss' -> 'yes'

# PRE-PROCESSING PIPELINE (CONTINUED)

## Stop Word Removal

- Removal of stop words within the 'english' word set
- Removal of individual letters and words that are less than 3 characters

## Stemming

- Snowball stemmer toolkit used
- Stem the words in each tweet

## Tokenization

- Twokenize toolkit used designed for social media text deta from Twitter
- Tokenize each tweet

## Evaluation

- Dataset evaluated to ensure the steps were correctly done before next processing stages.

# EXAMPLES OF UDFS

```python
def replace_emojis(text):
  emojis = demoji.findall(text)
  for k, v in emojis.items():
      text = text.replace(k,v)
  return text
```

```python
@udf(returnType=Types.ArrayType(Types.StringType()))
def stem_tokenize_stopwords(text):
    return [stemmer.stem(word) for word in twokenize.tokenizeRawTweetText(text)
                  if word not in stopwords.words("english") and len(word) > 2]
```

```python
def fix_spellings(text):

    spellchecker = SpellChecker()
    text = re.sub(r'(.)\1+', r'\1\1', text)
    words = word_tokenize(text)
    misspelt = spellchecker.unknown(words)
    corrections = {k:None for k in words}
    for w in misspelt:
        corrections[w] = spellchecker.correction(w)
    for k,v in corrections.items():
        if v != None:
            text = re.sub(k, v, text) # replacement operation taking place here

    return text
```

# FEATURES GENERATED TO DEVELOP THE MACHINE LEARNING MODELS

- The Data Set contained a combination of string and numerical data types.
- Original paper used the content column for classification.
- Our expansion on the paper will include using feature selection
- Schema of the preprocessed data:

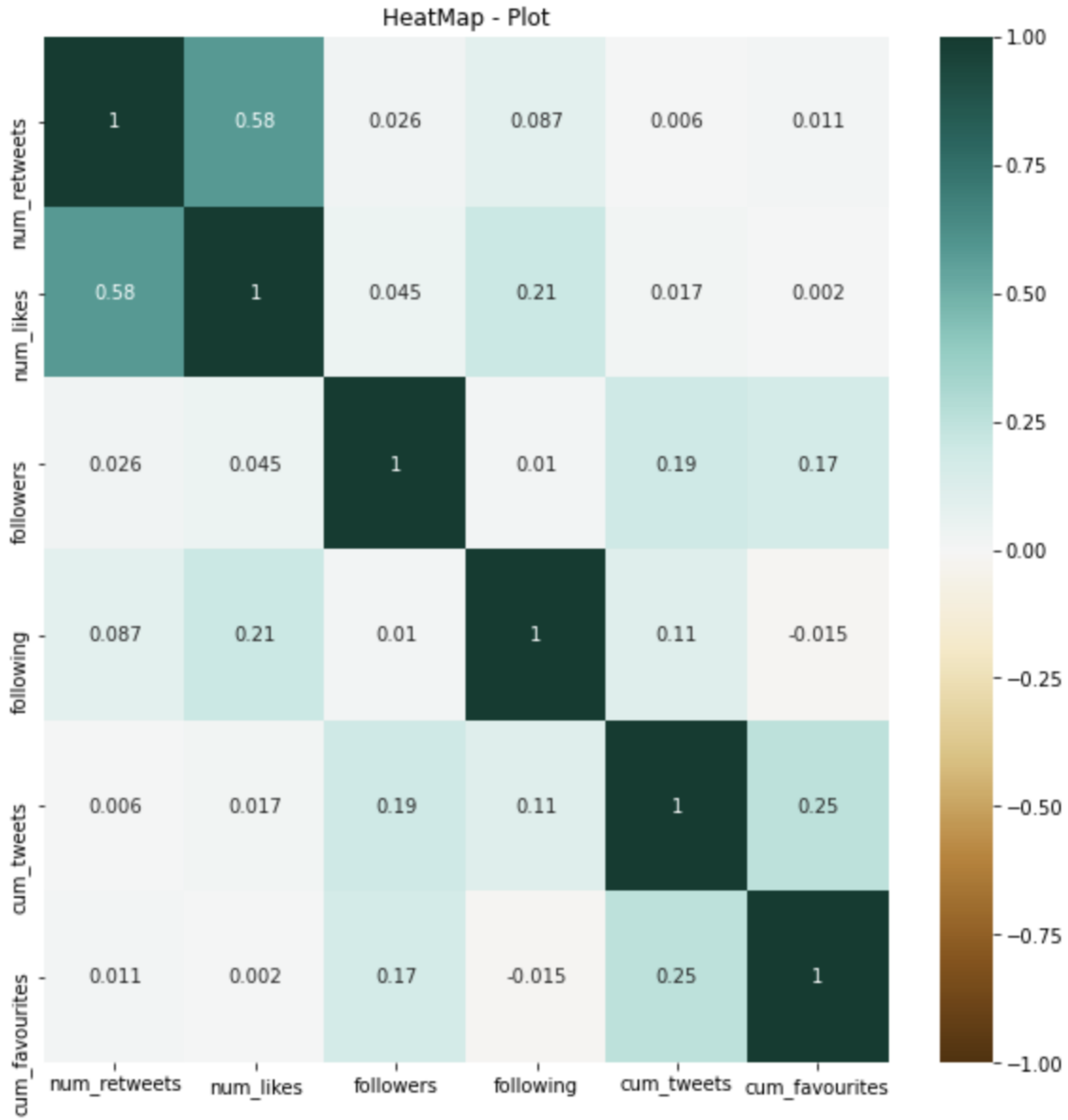
```
root
 |-- source: string (nullable = true)
 |-- content: string (nullable = true)
 |-- num_retweets: double (nullable = true)
 |-- num_likes: double (nullable = true)
 |-- url: string (nullable = true)
 |-- tweet_date: string (nullable = true)
 |-- screen_name: string (nullable = true)
 |-- name: string (nullable = true)
 |-- bio: string (nullable = true)
 |-- creation_date: string (nullable = true)
 |-- followers: double (nullable = true)
 |-- following: double (nullable = true)
 |-- cum_tweets: double (nullable = true)
 |-- cum_favourites: double (nullable = true)
 |-- label: string (nullable = true)
```

# QUALITY OF THE FEATURES

- Feature Engineering:
  - Preprocessed data needs to be fed into the machine learning model.
  - Capture the characteristics of the text into a numeric vector that can be understood by the ML algorithms.
  - Goal is to find the features that are most expressive of the data that will be useful for classification

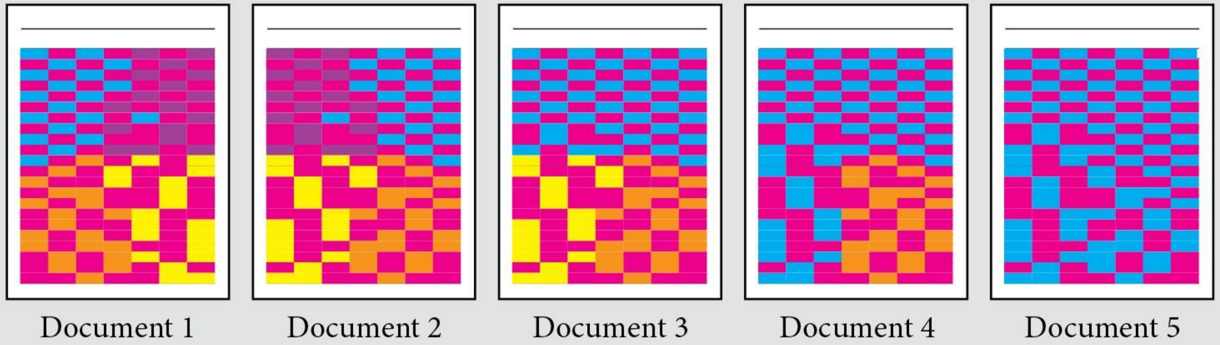
  Our Approach to Feature Engineering and Feature Selection:
- Original Paper's Approach
  - For the 'content', which contained the tweet text, used top 2000 features of TF-IDF to represent the words as vectors.
- Our Approach to Feature Engineering and Feature Selection:
  - Exploratory Data Analysis referred to from the original paper on the original data set to find the most impactful features contributing to the model.
  - Vectorize the features so it can be used in the ML.

# CORRELATION HEATMAP FOR NUMERICAL FEATURES



# TF-IDF (TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY)

- Used as a weighting factor for features
  - Weight increases as word frequency in document increases.
    - Offset by the number of times the word appears in the entire dataset / corpus.
      - Helps remove the importance of common words i.e. "The"
- **TF** = Term frequency (The number of times the term appears in the given body of the tweet)
- **IDF** = log (Total number of documents, n / number of documents in the dataset that contain a term)

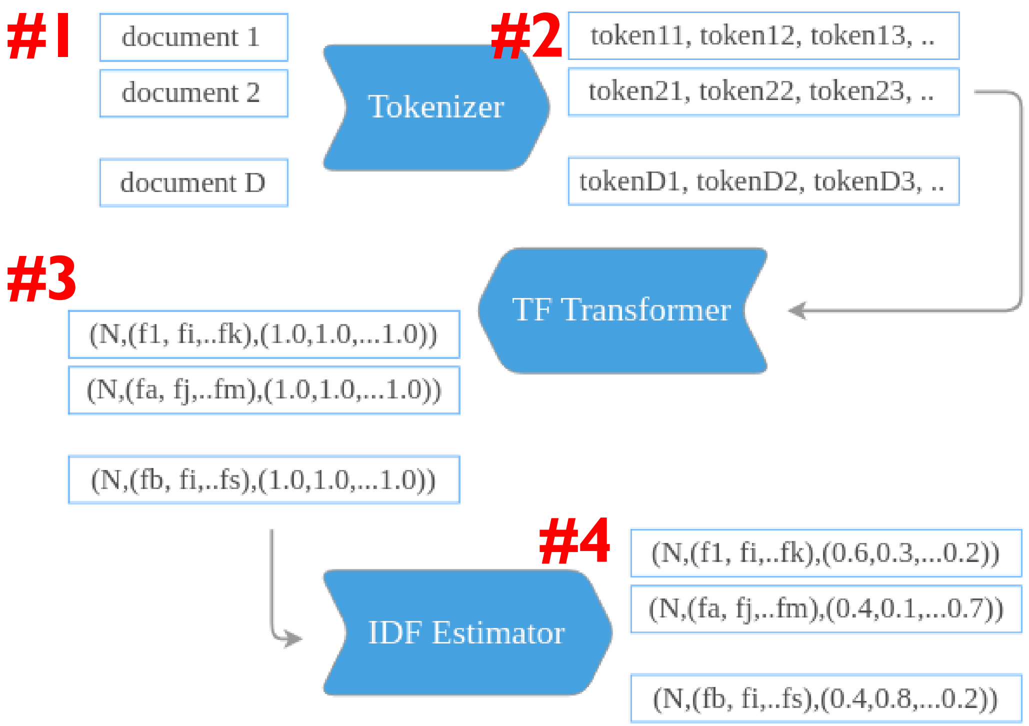



Corpus TF-IDF Values

| Documents | Terms | | | | |
|---|---|---|---|---|---|
| 1 | 0 | .05 | .027 | .012 | 0 |
| 2 | 0 | .05 | .027 | .012 | 0 |
| 3 | 0 | 0 | .027 | .012 | 0 |
| 4 | 0 | 0 | 0 | .012 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 |

# TF-IDF SPARKNLP IMPLEMENTATION

- **# 1:** Documents 1 – D: Tweet bodies from the dataset (corpus)
- **# 2:** Tweets are tokenized using Twokenizer input into TF Transformer
- **#3:** TF are obtained in the format:
  - *(Number of features(non-zero feature values)(feature importance))*
- **# 4:** Feature vectors (from HashingTF) are input into IDF Estimator
  - Each column is scaled




(**N**,(fa, fj,..fm),(1.0,1.0,...1.0)) : Number of features considered.
(N,**(fa, fj,..fm)**,(1.0,1.0,...1.0)) : Those features with non-zero value present in a document.
(N,(fa, fj,..fm),**(1.0,1.0,...1.0)**) : Values representing the importance of features in the document.

# TF-IDF - EXAMPLE

Number of features considered

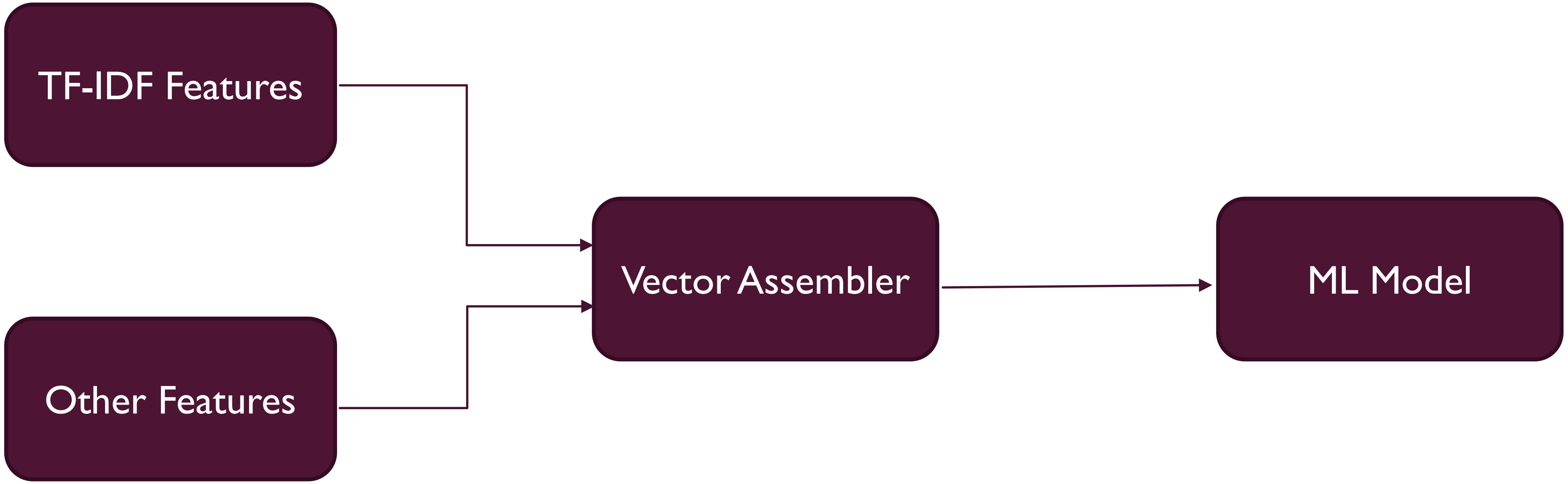
Features with non-zero value present in this tweet

```
content_tokens                                                 |tf_idf_features
[coronavirus, report, dead, wuhan, china, geller, report, news]|(2000,[102,302,395,433,505,714,821],[0.3364722366212129,0.8472978603872037,0.8472978603872037,0.8472978603872037,1.6945957207744073,0.5596157879354227,1.252762968495368])
```

Tokens from first tweet in the corpus

Values representing the importance of the features in this tweet.

# PIPELINE USING FEATURES




```
#Create a vector of the feature vectors

temp_va = VectorAssembler(inputCols=['token_count','num_retweets_scaled',
                                     'num_likes_scaled',
                                     'followers_scaled',
                                     'following_scaled',
                                     'cum_tweets_scaled',
                                     'tf_idf_features'],outputCol='features_vec')
final_df = temp_va.transform(df_with_features)
```

# THANK YOU