Mini Project 3

Johan Ng

Business Scenario

- Client: Political News agency

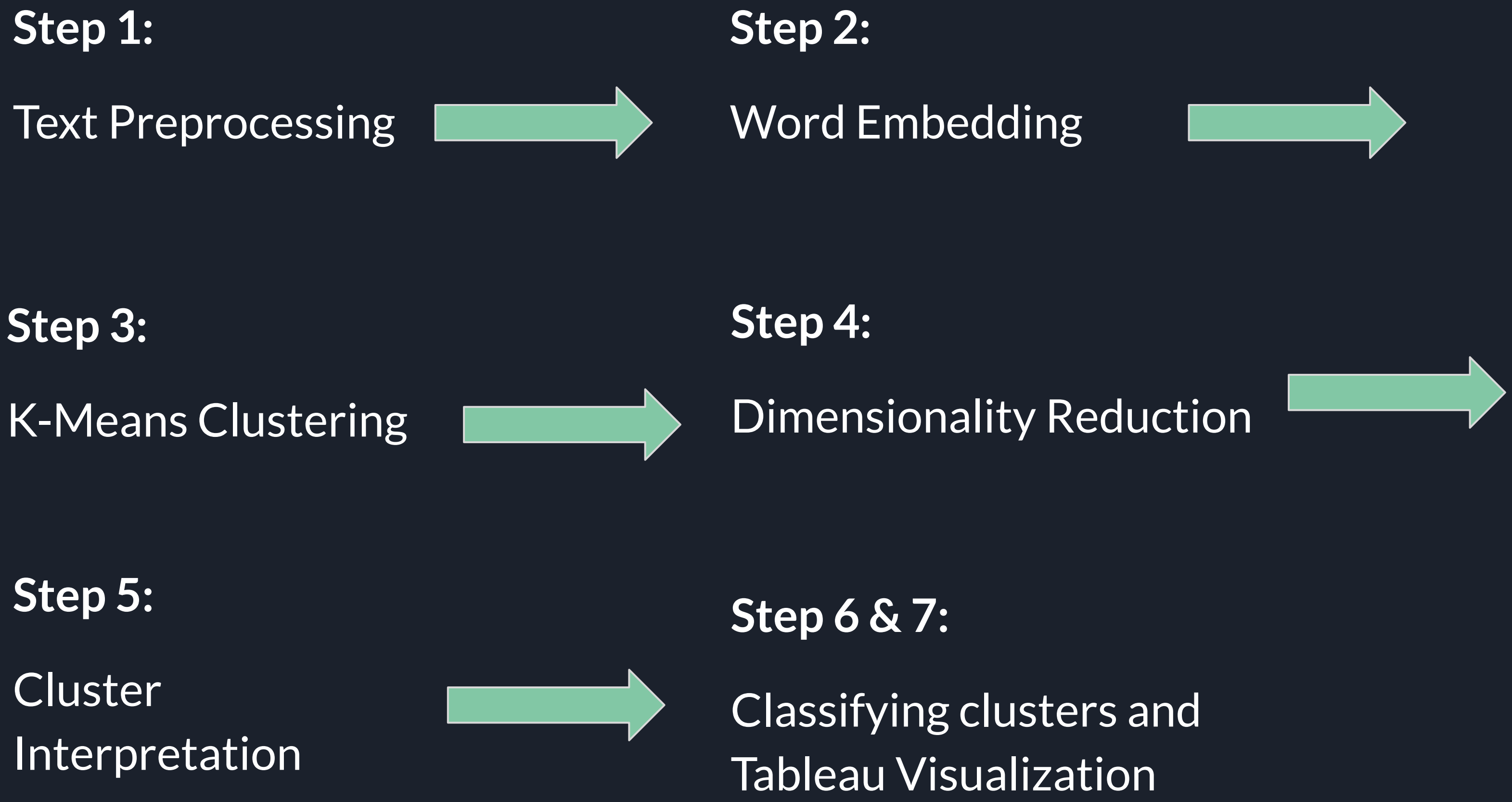
- Wants to classify fake news based on the headline of the article
  - Flagged as reliable through news headlines

- Provided a dataset of political news headlines from various news agency

Dataset

News Agency	Total Amount of Titles (43,424)
CNN	**1945**
Washington Post	**3047**
Reuters	**11,150**
CNBC News	**9703**
Gateway Pundit	**8615**
InfoWars	**397**
Huffington Post	**424**
LA Times	**295**
Others	**6834**

Approach to News Classification



Step 1:

Text Preprocessing

Step 2:

Word Embedding

Step 3:

K-Means Clustering

Step 4:

Dimensionality Reduction

Step 5:

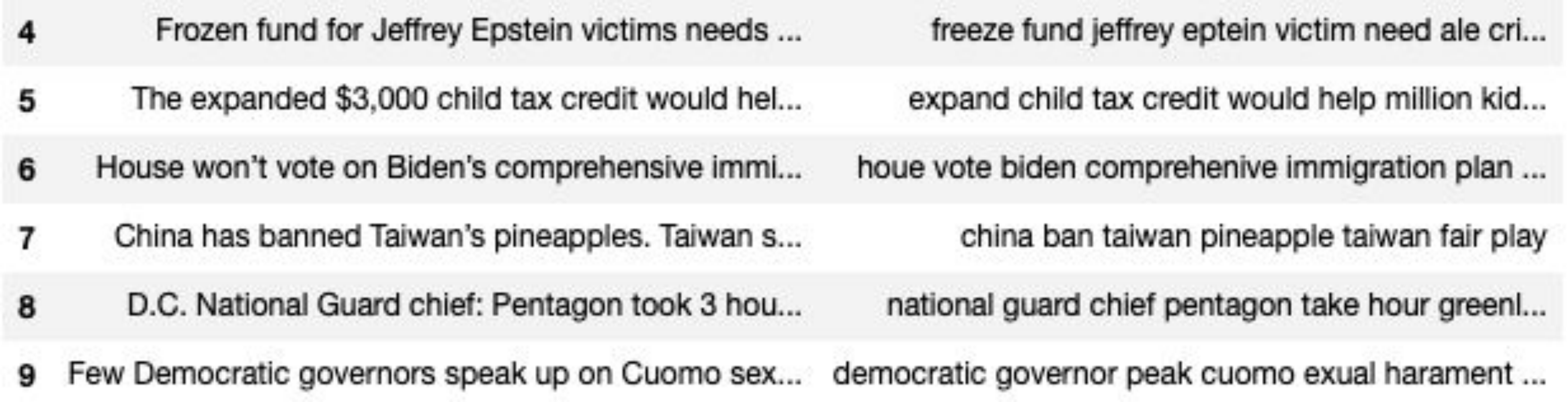
Cluster Interpretation

Step 6 & 7:

Classifying clusters and Tableau Visualization

Step 1: Pre-Processing Text

- Lower Case
- Remove numbers, punctuation, stopwords
- Tokenize & Lemmatize
- Bigram Phraser



	Original	After
4	Frozen fund for Jeffrey Epstein victims needs ...	freeze fund jeffrey eptein victim need ale cri...
5	The expanded $3,000 child tax credit would hel...	expand child tax credit would help million kid...
6	House won’t vote on Biden’s comprehensive immi...	houe vote biden comprehenive immigration plan ...
7	China has banned Taiwan’s pineapples. Taiwan s...	china ban taiwan pineapple taiwan fair play
8	D.C. National Guard chief: Pentagon took 3 hou...	national guard chief pentagon take hour greenl...
9	Few Democratic governors speak up on Cuomo sex...	democratic governor peak cuomo exual harament ...

Step 2: Word Embedding

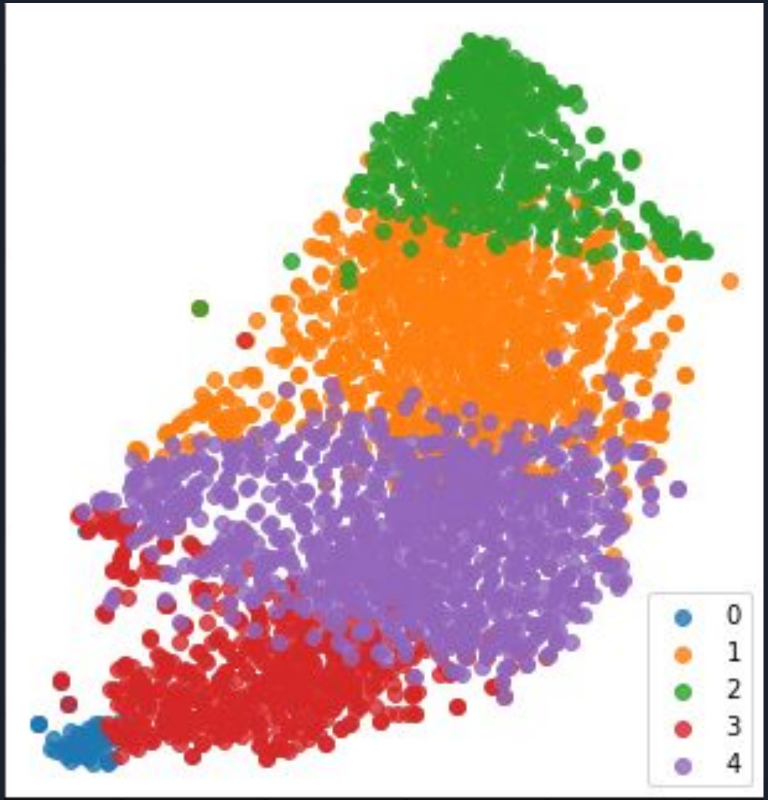
- Convert to Word Vectors (300 dimensions)
- Remove Low Frequency words (<20)
- **3031 Vocab Words**

Step 3: K-Means Clustering

- Clustering Word Vectors

- **5 Clusters**

Step 4: Dimensionality Reduction



Step 5: Interpretation of Clusters

What is Fake News?

- Misleading Information
- Damaging reputation

Defining Political Fake News by Headlines

- Biased
- Emotionally-Driven

Biased

Flip-Flop Fauci Claims Trump’s “Denial and Lack of Facts” Contributed To High COVID Death Toll

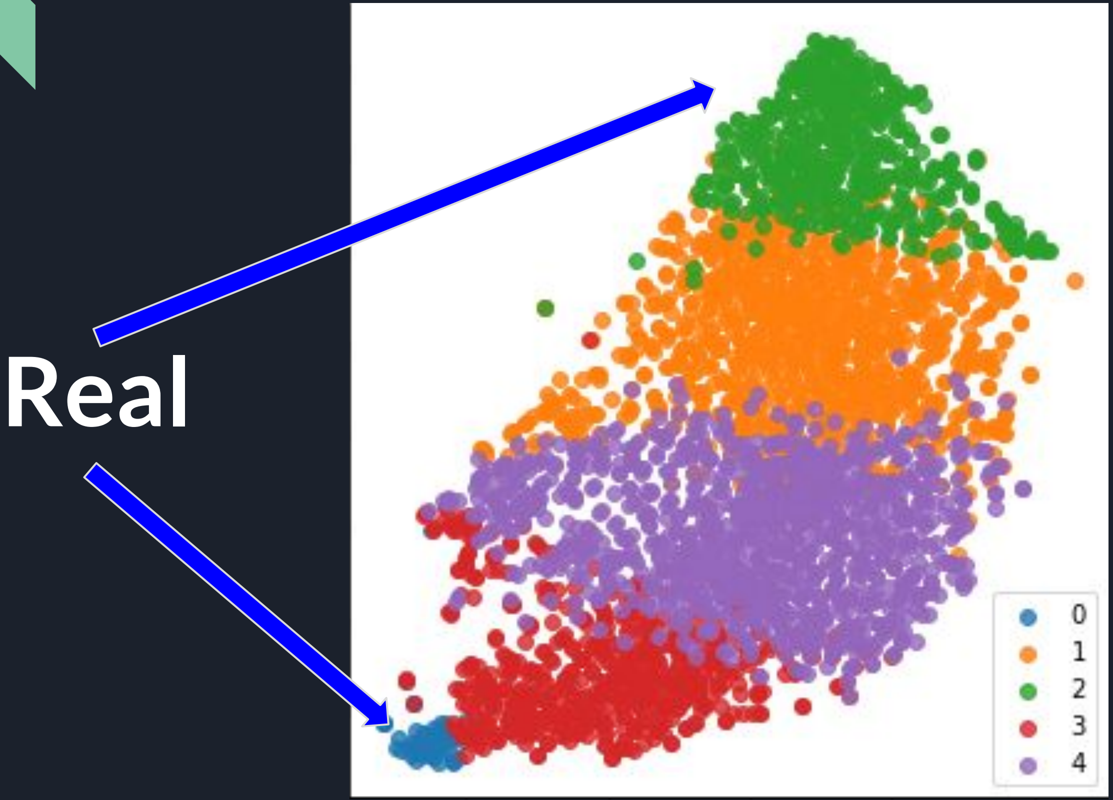
Emotionally Driven

Why Is FDA Playing Politics with New Warnings on Hydroxychloroquine? Why Would they Put American Lives in Danger?

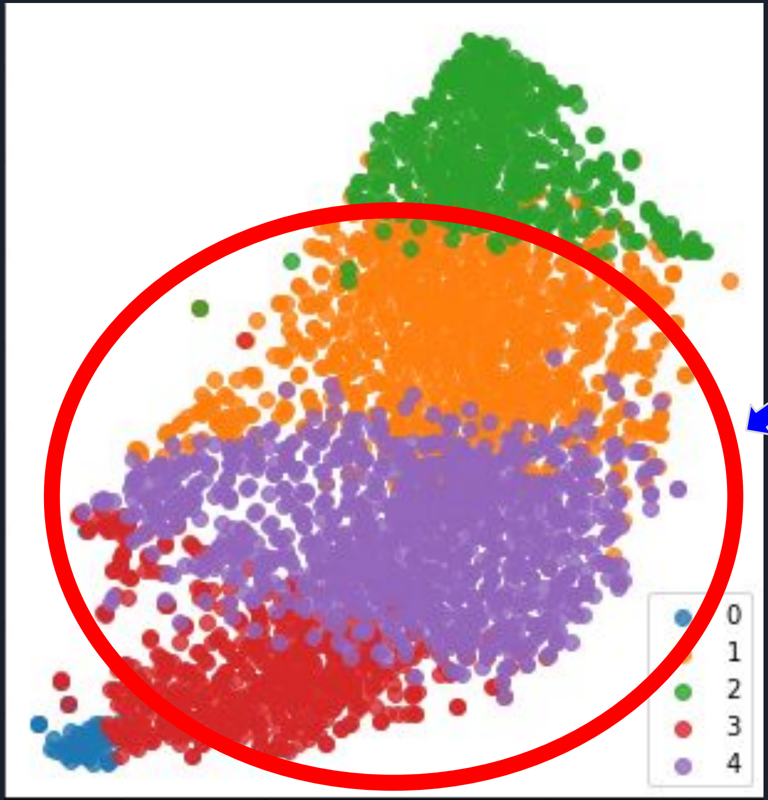
Keywords Identified (Examples)

- Leftist
- Crook Hillary
- Illegal Alien
- Racist
- Liberal Media
- Deplorable
- Swamp
- Commie
- Evil
- Hysterical

Step 6: Classifying the Clusters



Step 6: Classifying the Clusters



Potential Fake

Step 7: Tableau Visualization

Summary

- Model is quite accurate at detecting fake news

- Focus on being more neutral-sounding when deciding on the headline

- Additional Corpus is required for more accurate classification