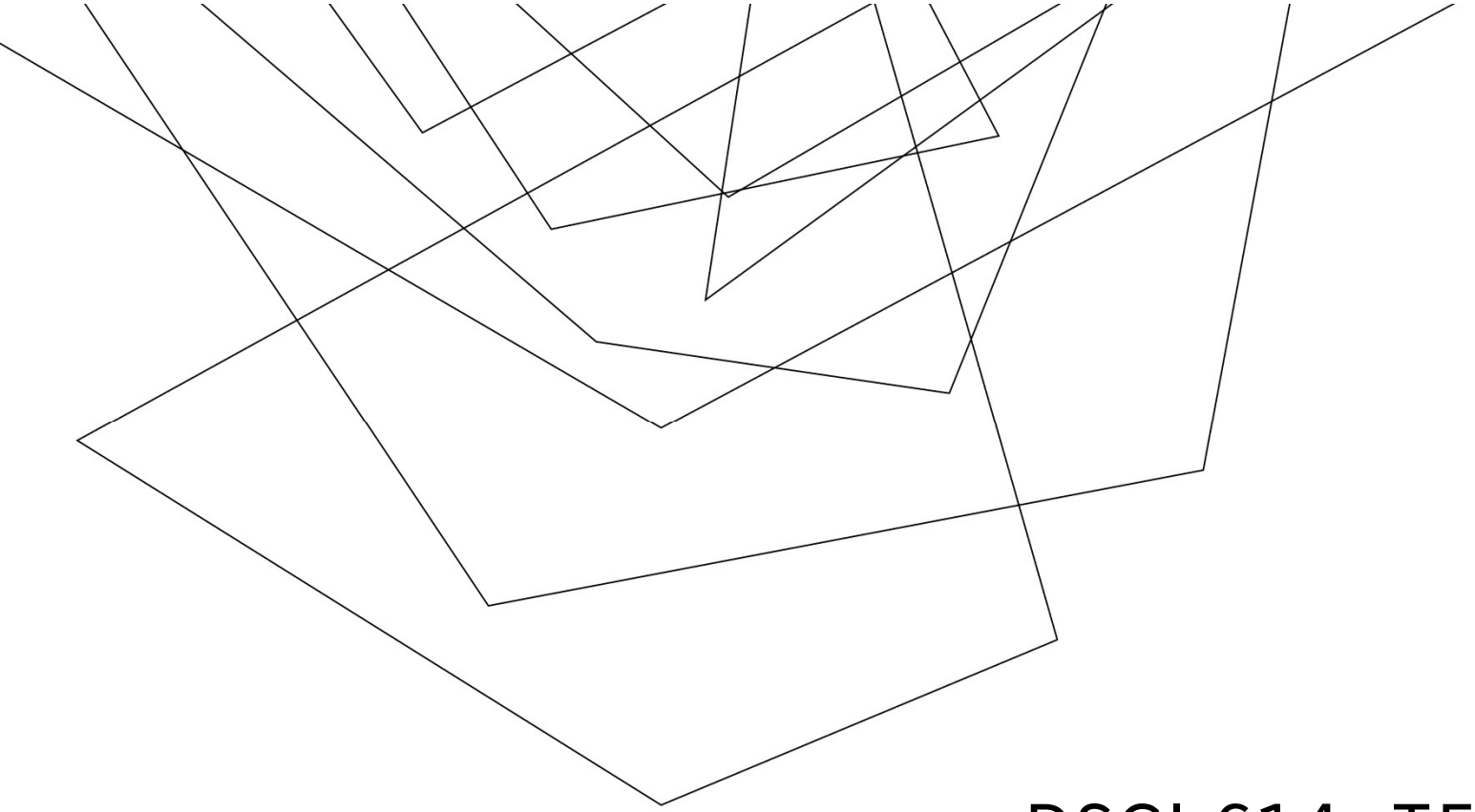
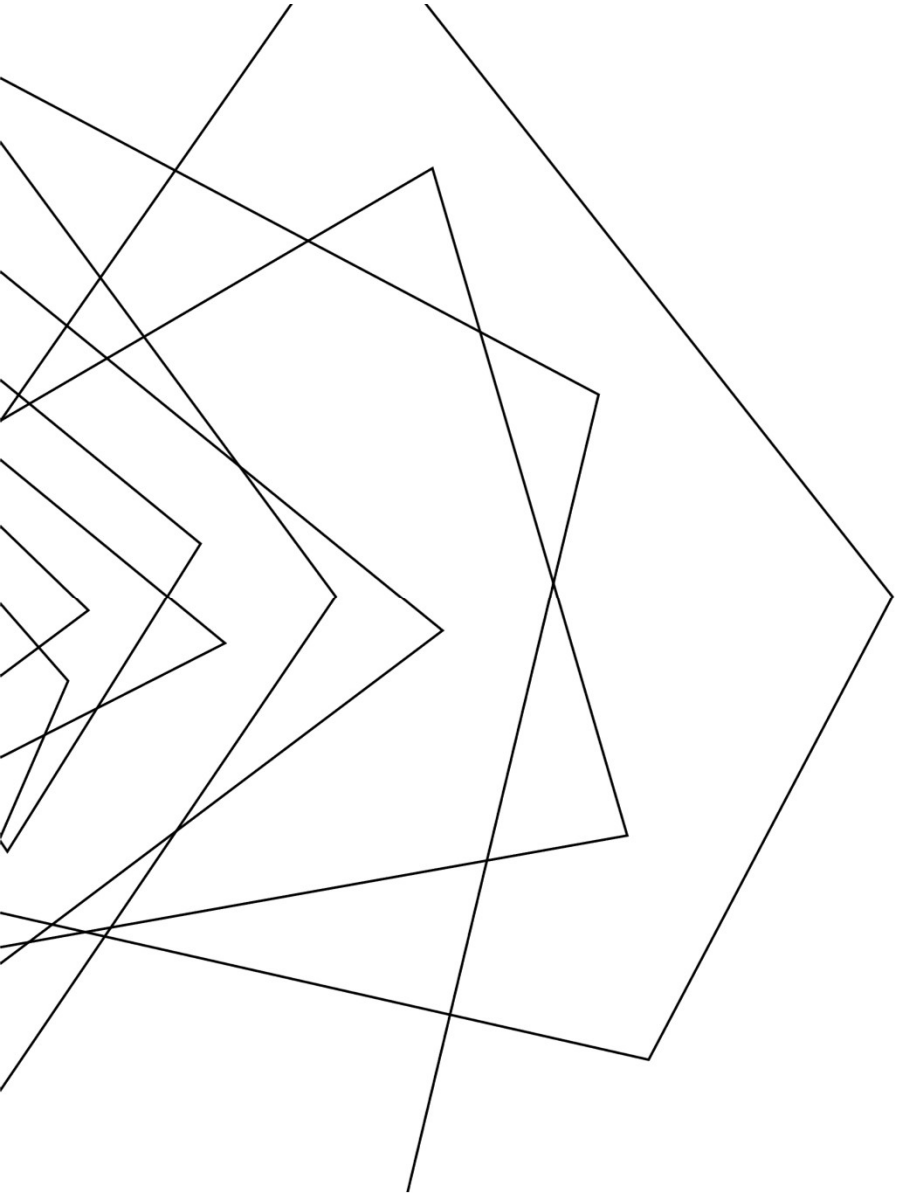


DSCI 614: TEXT MINING FINAL PROJECT

Lauren Schmiedeler



PROJECT STEPS

DATA CLEANING AND PREPROCESSING

Text Cleaning, Parts of Speech, Dependency Parser, Named Entity Recognition, Tokens, Lemmas

DATA VISUALIZATION

Sentiment, Length, Word Count, Top Unigrams and Bigrams, Top Positive Tokens

MODELING AND PREDICTION

Count Vectorizer, Tf-idf Vectorizer, Random Forest, Logistic Regression, Support Vector Classification, Prediction, Evaluation, Hyperparameter Tuning

TOPIC MODELING

Latent Dirichlet Allocation (LDA), Dimension Reduction

DATA CLEANING AND PREPROCESSING: OVERVIEW

This dataset is from Kaggle and includes tweets about coronavirus that have been pulled from Twitter and manually tagged.

Columns:

- **Tweet_texts** = Location + TweetAt + OriginalTweet
- **Sentiment**
  - Extremely Negative, Negative, Neutral, Positive, Extremely Positive

This dataset contains both a **training set** and a **test set**.

```
number of observations in the training set = 41157
number of observations in the test set = 3798
number of observations in the training set after cleaning = 41142
number of observations in the test set after cleaning = 3798
```

DATA CLEANING AND PREPROCESSING: TEXT CLEANING

Text Cleaning Steps:

- Remove
  - dates and times,
  - hyperlinks,
  - hashtags,
  - usernames,
  - special characters,
  - one and two letter words,
  - extra spaces.
- Convert to lowercase.
- Delete rows that contain empty strings or NaNs in the **Tweet_texts** column.

Before Cleaning:

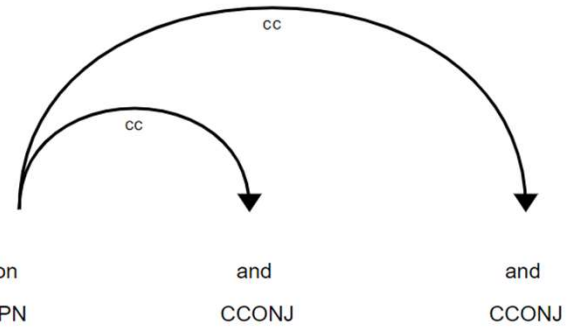
```
TWEET 0 : London 16-03-2020 @MeNyrbie @Phil_Gahan @Chrisitv https://t.co/iFz9FAn
2Pa and https://t.co/xX6ghGFzCC and https://t.co/I2NlzdxNo8

TWEET 1 : UK 16-03-2020 advice Talk to your neighbours family to exchange phone
numbers create contact list with phone numbers of neighbours schools employer ch
emist GP set up online shopping accounts if poss adequate supplies of regular me
ds but not over order

TWEET 2 : Vagabonds 16-03-2020 Coronavirus Australia: Woolworths to give elderl
y, disabled dedicated shopping hours amid COVID-19 outbreak https://t.co/bInCA9V
p8P
```

After Cleaning:

```
TWEET 0 CLEANED : london and and

TWEET 1 CLEANED : advice talk your neighbours family exchange phone numbers crea
te contact list with phone numbers neighbours schools employer chemist set onlin
e shopping accounts poss adequate supplies regular meds but not over order

TWEET 2 CLEANED : vagabonds coronavirus australia woolworths give elderly disabl
ed dedicated shopping hours amid covid 19 outbreak
```

DATA CLEANING AND PREPROCESSING: PARTS OF SPEECH, DEPENDENCY PARSER, NAMED ENTITY RECOGNITION

```
TWEET: london and and

PARTS OF SPEECH:
london = PROPN
and = CCONJ
and = CCONJ

DEPENDENCY PARSER:
```




```
NAMED ENTITY RECOGNITION:
london = GPE
```

GEOLOCATION, MONEY, and QUANTITY:

western canada your dreams breaking gas prices dropping tonight 9 cents **MONEY** putting price 1 10 9 litre **QUANTITY** metro van lowest 17 years reason because drop demand with people staying home energy sector will suffer

melbourne **GPE** australia **GPE** facebook live from joins speak about staying safe and well during the pandemic join the conversation

australia **GPE** everything but the socks pants and jocks that was wearing that particular day tonight 9pm

mid michigan **LOC** just accused doubling prices some cleaning supplies during the pandemic

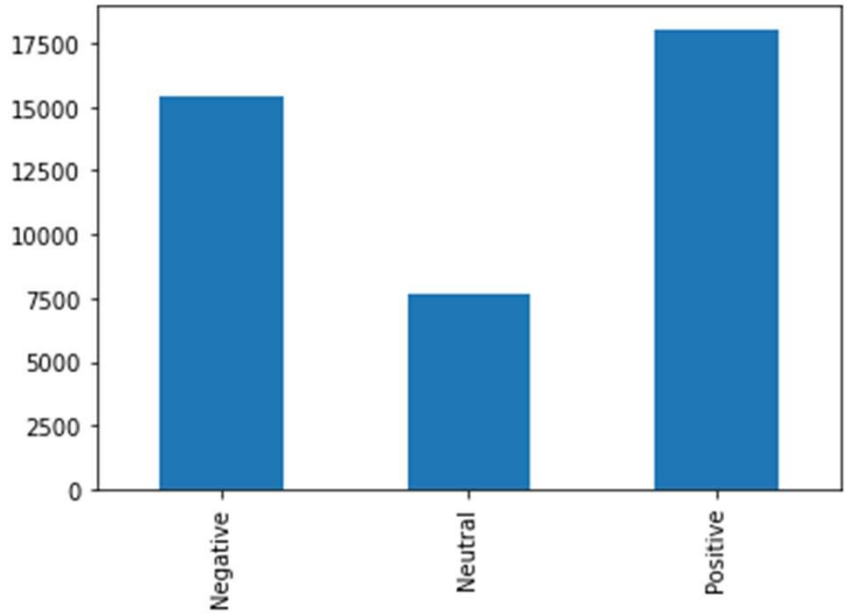
DATA CLEANING AND PREPROCESSING: TOKENS, LEMMAS

```
TWEET 0 TOKENS : [london, and, and]

TWEET 1 TOKENS : [advice, talk, your, neighbours, family, exchange, phone, numbers, create, contact, list,
with, phone, numbers, neighbours, schools, employer, chemist, set, online, shopping, accounts, poss, adequa
te, supplies, regular, meds, but, not, over, order]

TWEET 2 TOKENS : [vagabonds, coronavirus, australia, woolworths, give, elderly, disabled, dedicated, shoppi
ng, hours, amid, covid, 19, outbreak]

TWEET 3 TOKENS : [food, stock, not, the, only, one, which, empty, please, don, panic, there, will, enough,
food, for, everyone, you, not, take, more, than, you, need, stay, calm, stay, safe]

TWEET 4 TOKENS : [ready, supermarket, during, the, outbreak, not, because, paranoid, but, because, food, st
ock, litteraly, empty, the, serious, thing, but, please, don, panic, causes, shortage]
```

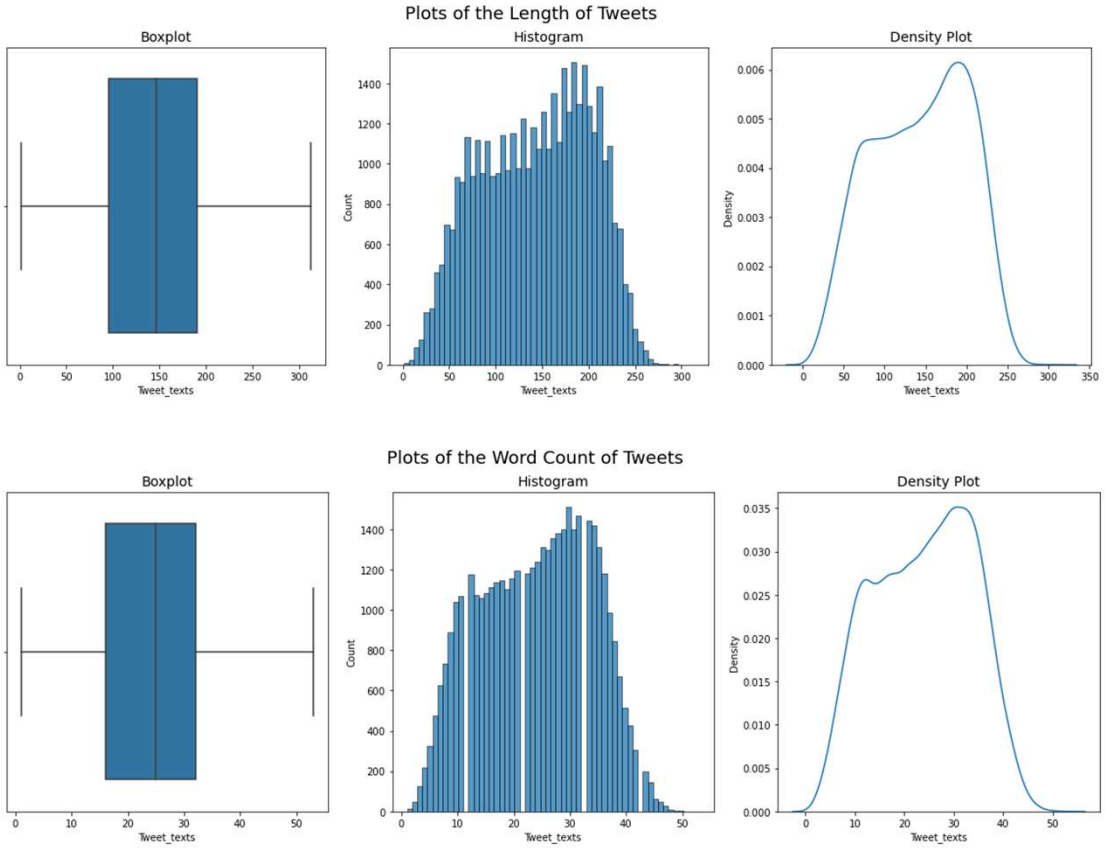
```
TWEET 0 LEMMAS : london and and

TWEET 1 LEMMAS : advice talk your neighbour family exchange phone number create contact list with phone num
ber neighbour school employer chemist set online shopping account poss adequate supply regular med but not
over order

TWEET 2 LEMMAS : vagabonds coronavirus australia woolworth give elderly disabled dedicated shopping hour am
id covid 19 outbreak

TWEET 3 LEMMAS : food stock not the only one which empty please don panic there will enough food for everyo
ne you not take more than you need stay calm stay safe

TWEET 4 LEMMAS : ready supermarket during the outbreak not because paranoid but because food stock litteral
y empty the serious thing but please don panic cause shortage
```

DATA VISUALIZATION: SENTIMENT

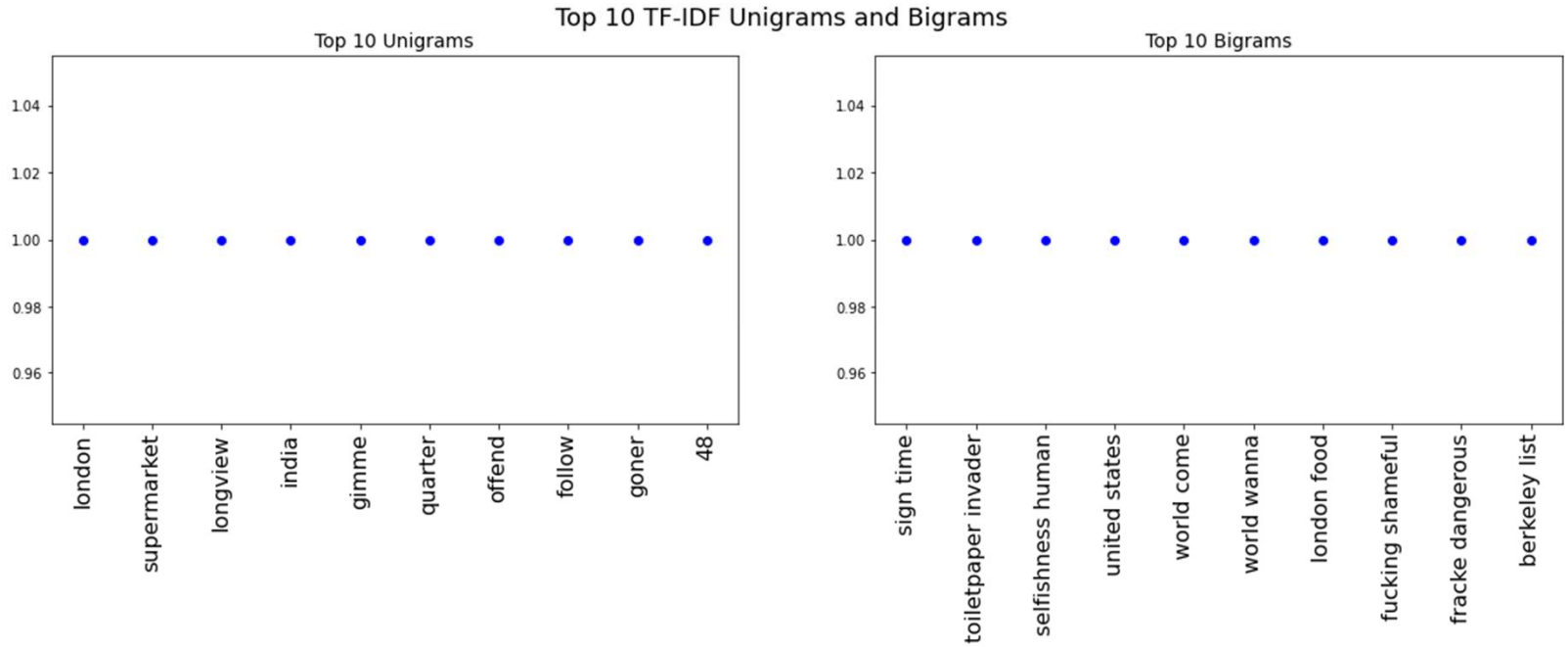
- First, combine “Extremely Negative” and “Negative” sentiments and “Positive” and “Extremely Positive” sentiments.
- Note that “Neutral” is underrepresented in the dataset compared to both “Negative” and “Positive”.



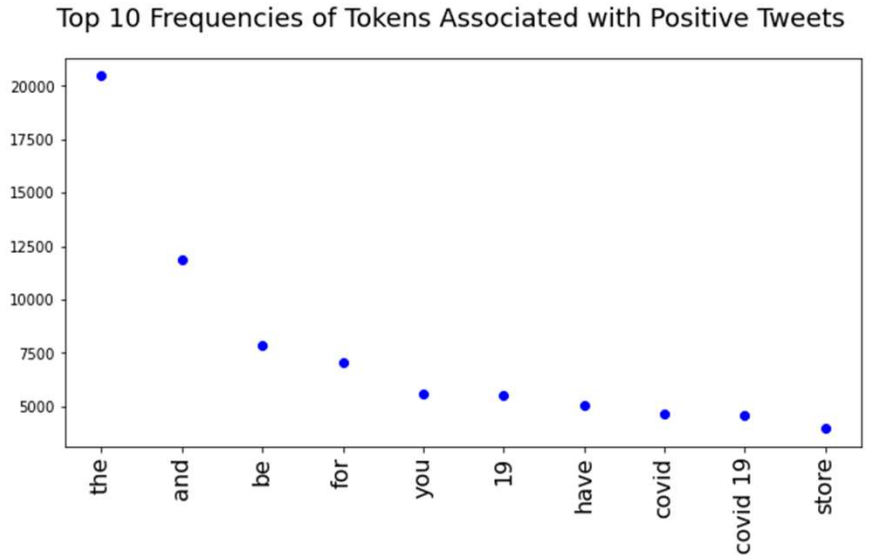
DATA VISUALIZATION: TWEET LENGTH AND WORD COUNT

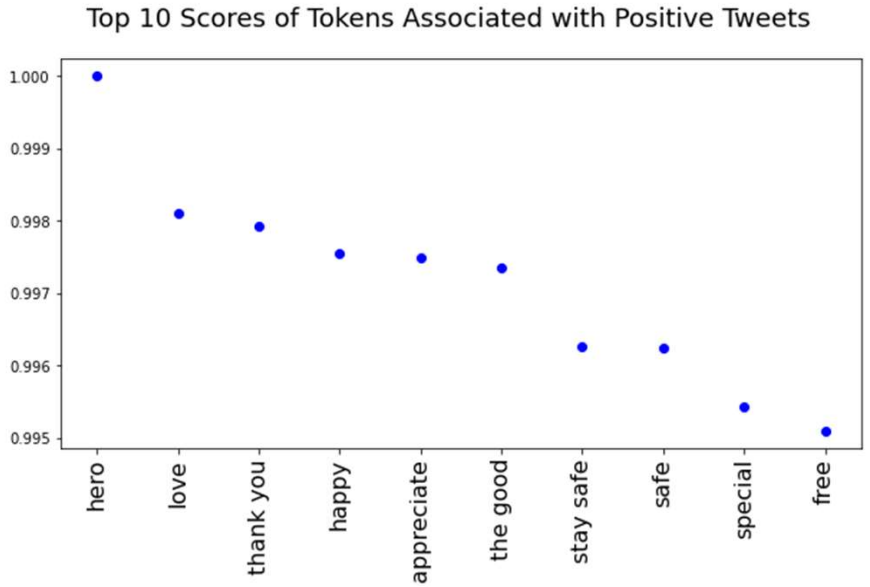


DATA VISUALIZATION: TOP TF-IDF UNIGRAMS AND BIGRAMS



DATA VISUALIZATION: TOP POSITIVE TOKENS





MODELING AND PREDICTION: VECTORIZING

Count Vectorizer:

- ngram_range = (1, 2)

```
len(tf_vectorizer.vocabulary_) # size of the vocabulary
```

```
397695
```

```
tf_matrix.shape # shape of the document word matrix
```

```
(41142, 397695)
```

Tf-idf Vectorizer:

```
len(tf_idf_vectorizer.vocabulary_) # size of the vocabulary
```

```
32392
```

```
tf_idf_matrix.shape # shape of the document word matrix
```

```
(41142, 32392)
```

MODELING AND PREDICTION: MODEL 1

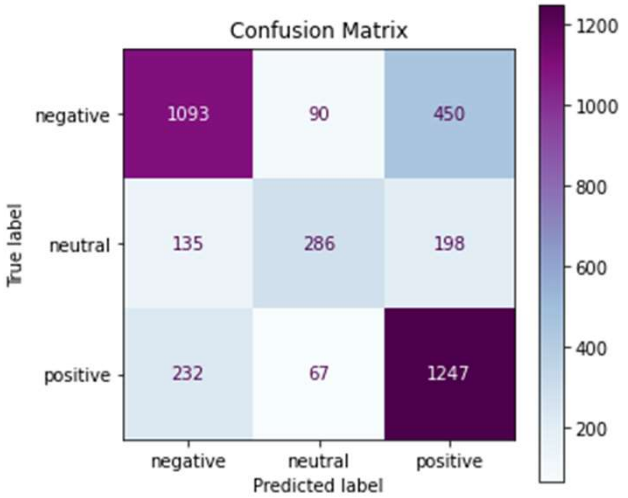
Before Pipeline:

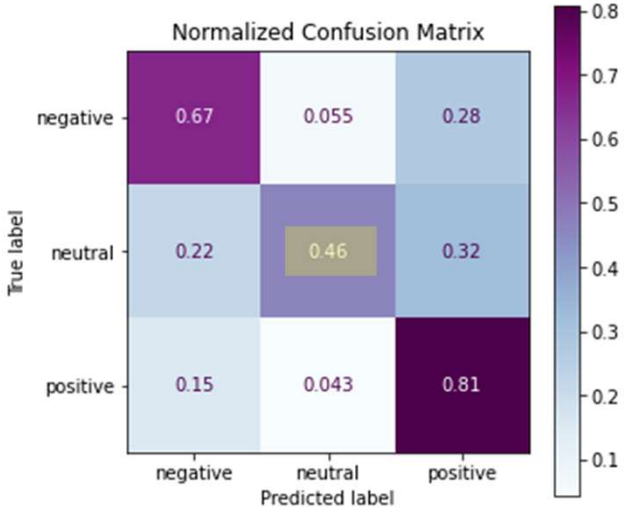
- Clean the texts.
- Replace all tokens with their lemmas.

Model 1 Pipeline:

- vectorizer = TfidfVectorizer()
- classifier = RandomForestClassifier(random_state = 100)

MODELING AND PREDICTION: MODEL 1 EVALUATION






```
accuracy = 0.69142
```

- The model only correctly classifies **46% of neutrals** while it correctly classifies **67% of negatives** and **81% of positives**.
- The misclassified negatives are largely classified as positives, and the misclassified positives are largely classified as negatives.

MODELING AND PREDICTION: MODEL 2

Model 2 Pipeline:

- vectorizer = TfidfVectorizer()
- classifier = RandomForestClassifier(random_state = 100)

Possible Parameters:

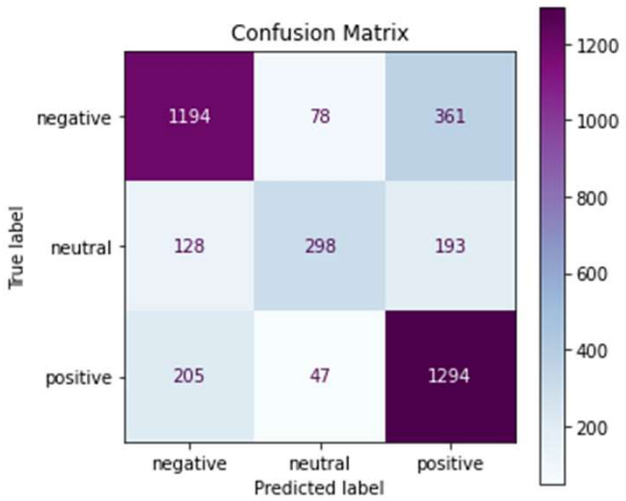
- **TfidfVectorizer():**
  - stop_words: (None, “english”)
  - max_df: (0.33, 0.66, 0.1)
- **RandomForestClassifier()**
  - min_samples_split: (2, 4)
  - min_samples_leaf: (1, 2, 4)

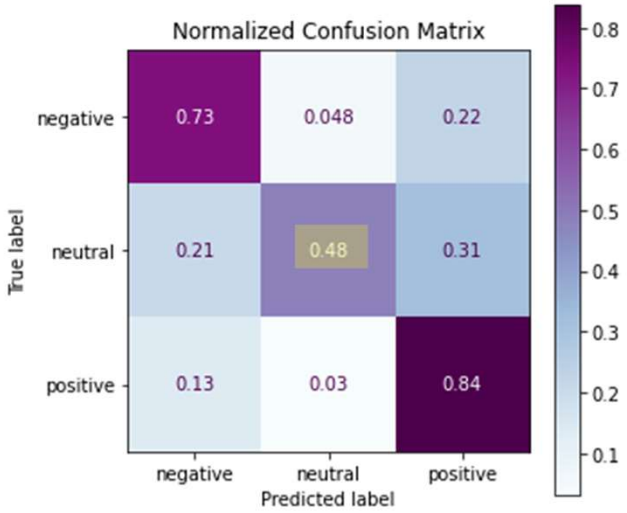
GridSearchCV(cv = 4)

Best Parameters:

- vectorizer__stop_words: “english”
- vectorizer__max_df: 0.33
- classifier__min_samples_split: 2
- classifier__min_samples_leaf: 2

MODELING AND PREDICTION: MODEL 2 EVALUATION






```
accuracy = 0.73354
```

- The accuracy for this model is a slight improvement compared to the first model, but this model struggles with classifying neutrals like the first model.
- The model only correctly classifies **48% of neutrals** while it correctly classifies **73% of negatives** and **84% of positives**.

MODELING AND PREDICTION: MODEL 3

Model 3 Pipeline:

- vectorizer = TfidfVectorizer(stop_words = “english”, max_df = 0.33)
- classifier = ClfSwitcher(random_state = 100)

Possible Estimators and Parameters:

- **LogisticRegression():**
  - penalty: [“l2”, “none”]
- **SVC():**
  - C: (0.5, 1, 2)
  - kernel: (“linear”, “rbf”)
- **RandomForestClassifier():**
  - min_samples_split: (2, 4)
  - min_samples_leaf: (1, 2, 4)

GridSearchCV(cv = 4)

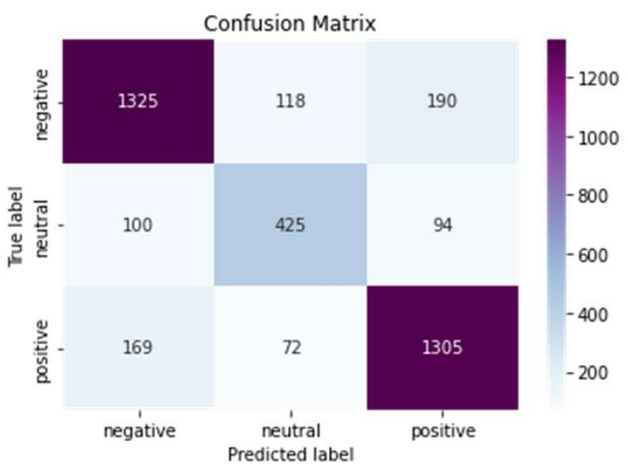
Best Model:

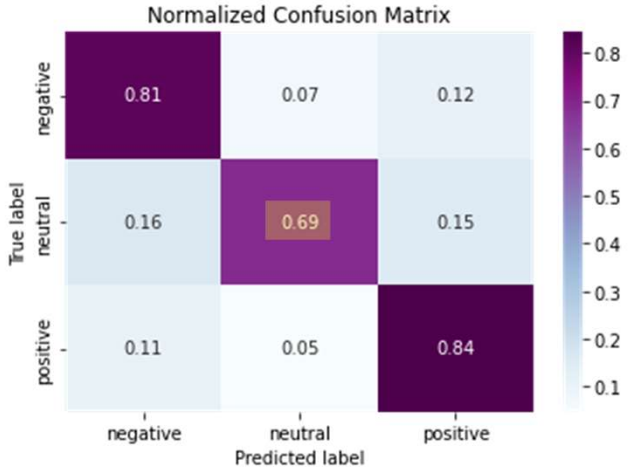
- Support Vector Classification

Best Parameters:

- C = 2
- kernel = “linear”

MODELING AND PREDICTION: MODEL 3 EVALUATION






```
accuracy = 0.80437
```

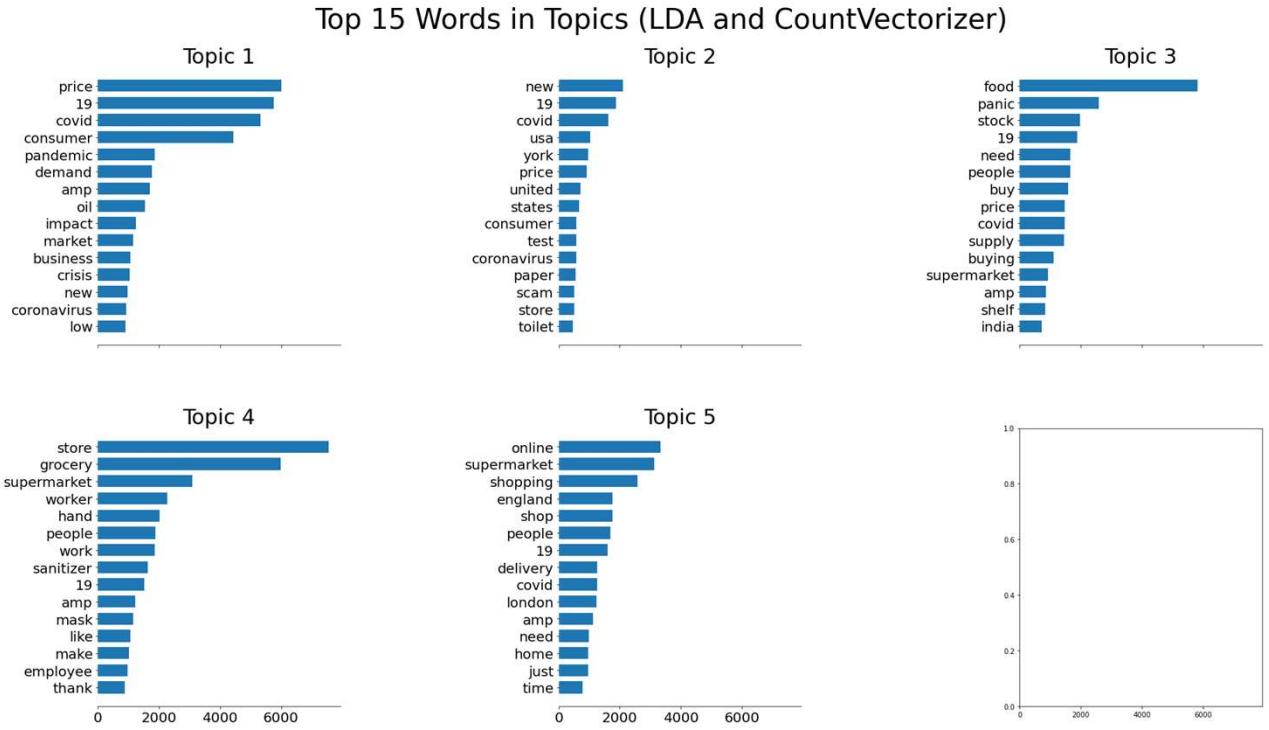
- Not only is this model's overall accuracy an improvement over the first two models, but its accuracy for neutrals is a dramatic improvement compared to these models.
- In this model, both negatives and positives are misclassified as neutrals more often than in the first two models.

TOPIC MODELING: STEPS

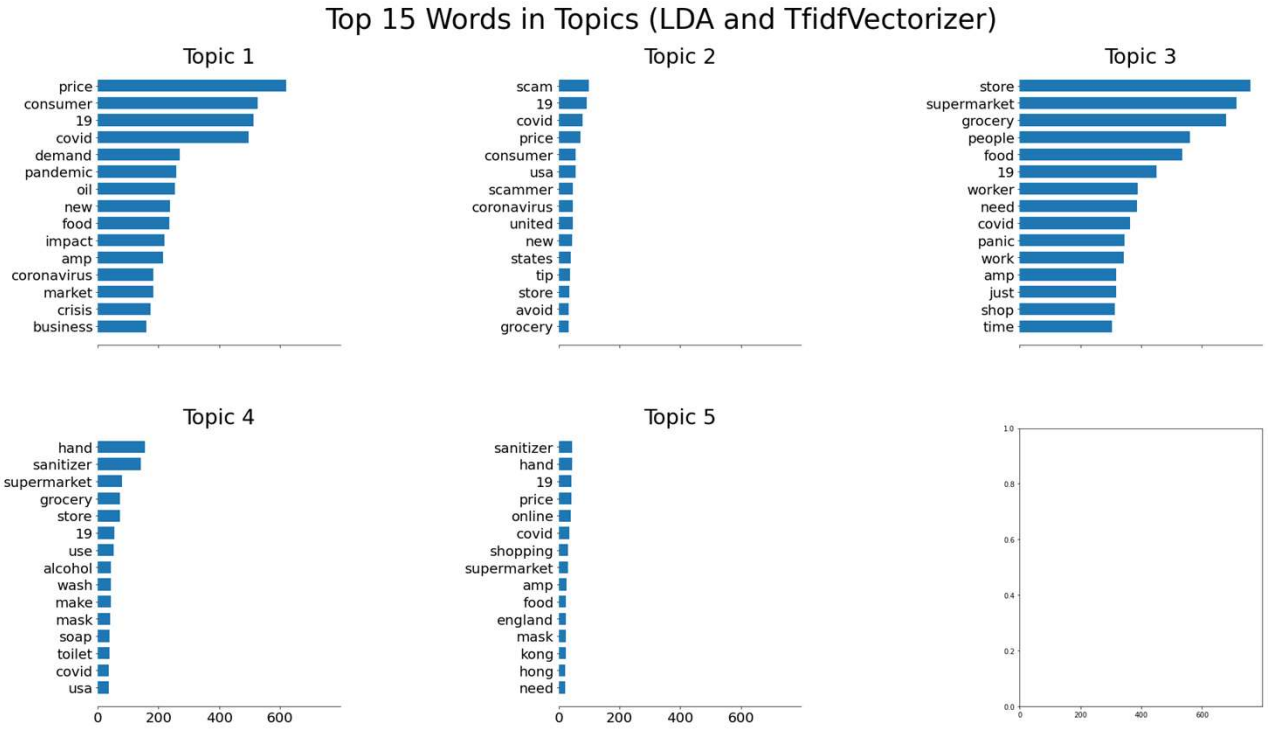
- Create a Vectorizer object using **CountVectorizer()** or **TfidfVectorizer()**.
  - Set **stop_words = "english"**.
- Fit the corpus using the Vectorizer object.
- Generate the **document word matrix** by transforming the corpus using the fitted Vectorizer object.
- Create a **topic model** using **LatentDirichletAllocation(random_state = 100)**.
  - Set **n_components = 5** to separate the observations into 5 topics.
- Fit the LDA model using the **document word matrix**.

```
LatentDirichletAllocation
LatentDirichletAllocation(n_components=5, random_state=100)
```

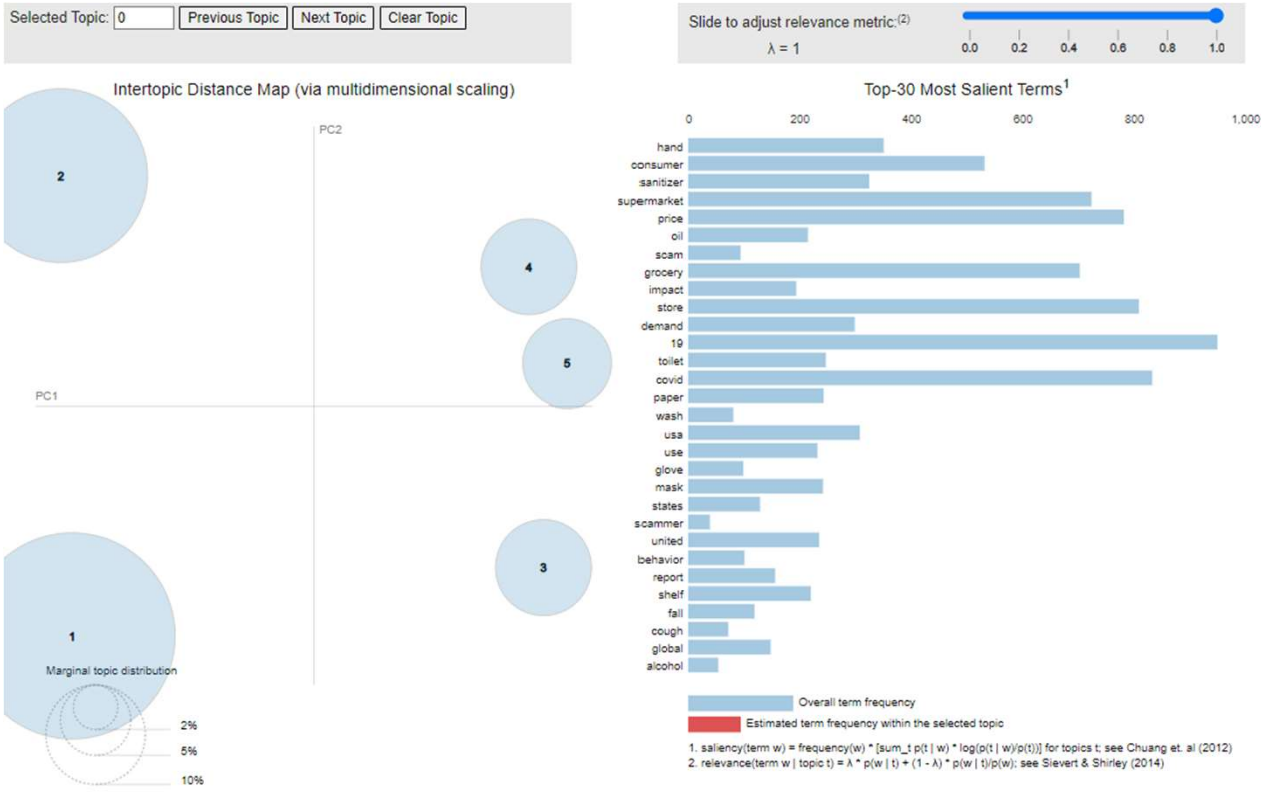
TOPIC MODELING: TOP 15 WORDS IN EACH TOPIC USING LDA AND COUNT VECTORIZER



TOPIC MODELING: TOP 15 WORDS IN EACH TOPIC USING LDA AND TF-IDF VECTORIZER



TOPIC MODELING: DIMENSION REDUCTION



TOPIC MODELING: DIMENSION REDUCTION (TOPIC 1)