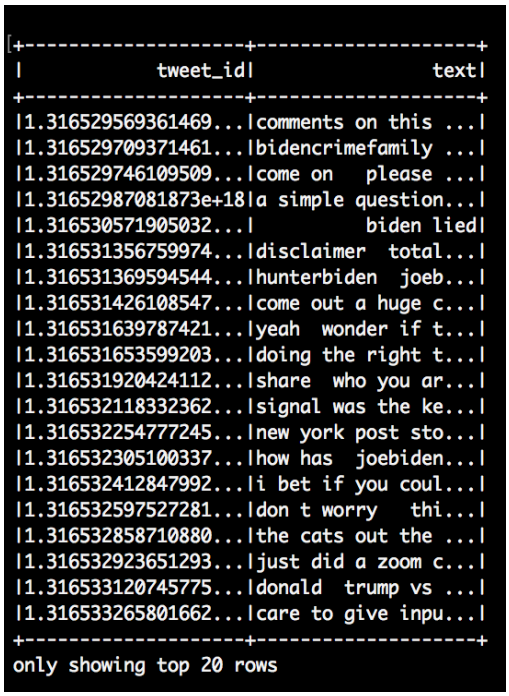
CS 657 Massive Mining Datasets Assignment -3

Sai Sahana Bhargavi - G01330358

Introduction: The task is to find out if using the Topic Modeling information helps improve the classification accuracy in this problem.The dataset contains 2020 US Election Tweets Analysis descriptions and the file hashtag_biden.csv contains the tweets.

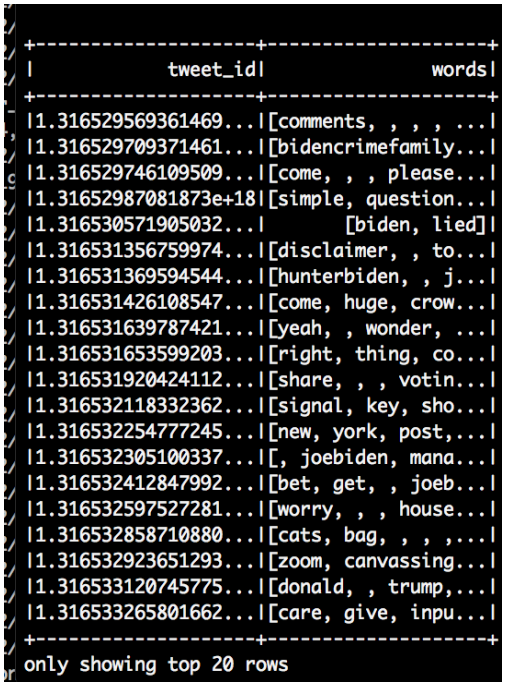
Preprocessing Data :

Cleaned the data present in 'tweet' columns by removing punctuation marks, alpha-numeric characters, spaces and lowering the cases using regexp_replace() included with parameters for respective functionalities.



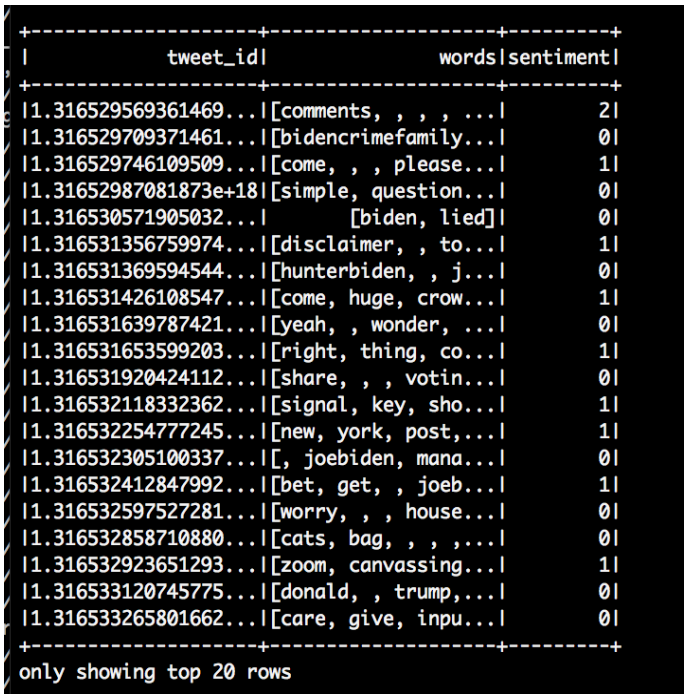
```
+--------------------+--------------------+
|            tweet_id|                text|
+--------------------+--------------------+
|1.316529569361469...|comments on this ...|
|1.316529709371461...|bidencrimefamily ...|
|1.316529746109509...|come on   please ...|
|1.31652987081873e+18|a simple question...|
|1.316530571905032...|          biden lied|
|1.316531356759974...|disclaimer  total...|
|1.316531369594544...|hunterbiden  joeb...|
|1.316531426108547...|come out a huge c...|
|1.316531639787421...|yeah  wonder if t...|
|1.316531653599203...|doing the right t...|
|1.316531920424112...|share  who you ar...|
|1.316532118332362...|signal was the ke...|
|1.316532254777245...|new york post sto...|
|1.316532305100337...|how has  joebiden...|
|1.316532412847992...|i bet if you coul...|
|1.316532597527281...|don t worry   thi...|
|1.316532858710880...|the cats out the ...|
|1.316532923651293...|just did a zoom c...|
|1.316533120745775...|donald  trump vs ...|
|1.316533265801662...|care to give inpu...|
+--------------------+--------------------+
only showing top 20 rows
```

- Converted the processed text in 'tweet', by breaking it down into words using feature extractor Tokenizer() , removed stop words using StopWordsRemover() and converted into vector format using feature TF_IDF feature and applied transformers for term frequencyCountVectorizer() and inverse document frequency IDF().



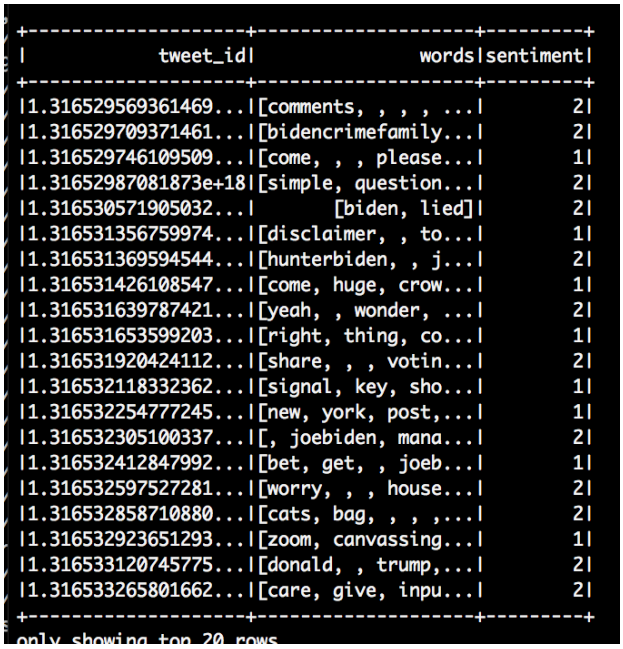
```
+-------------------+-------------------+
|           tweet_id|              words|
+-------------------+-------------------+
|1.316529569361469...|[comments, , , , ...|
|1.316529709371461...|[bidencrimefamily...|
|1.316529746109509...|[come, , , please...|
|1.31652987081873e+18|[simple, question...|
|1.316530571905032...|       [biden, lied]|
|1.316531356759974...|[disclaimer, , to...|
|1.316531369594544...|[hunterbiden, , j...|
|1.316531426108547...|[come, huge, crow...|
|1.316531639787421...|[yeah, , wonder, ...|
|1.316531653599203...|[right, thing, co...|
|1.316531920424112...|[share, , , votin...|
|1.316532118332362...|[signal, key, sho...|
|1.316532254777245...|[new, york, post,...|
|1.316532305100337...|[, joebiden, mana...|
|1.316532412847992...|[bet, get, , joeb...|
|1.316532597527281...|[worry, , , house...|
|1.316532858710880...|[cats, bag, , , ,...|
|1.316532923651293...|[zoom, canvassing...|
|1.316533120745775...|[donald, , trump,...|
|1.316533265801662...|[care, give, inpu...|
+-------------------+-------------------+
only showing top 20 rows
```

- assigned one of three classes (positive, neutral, negative) to every tweet in the dataset using TextBlob.sentiment.polarity feature.



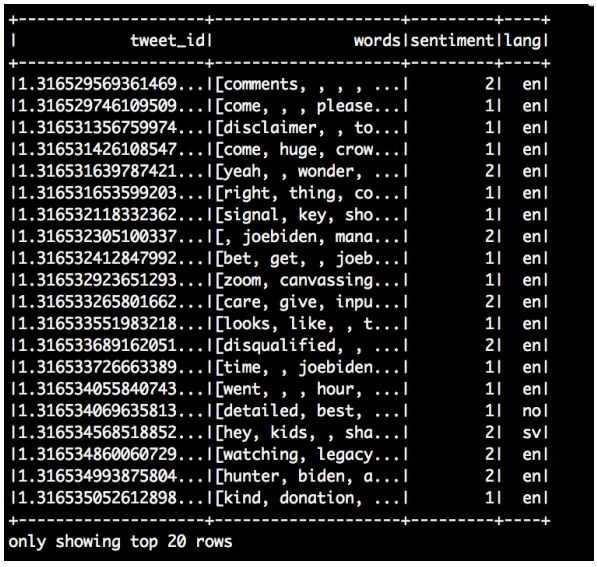
```
+-------------------+-------------------+---------+
|           tweet_id|              words|sentiment|
+-------------------+-------------------+---------+
|1.316529569361469...|[comments, , , , ...|        2|
|1.316529709371461...|[bidencrimefamily...|        0|
|1.316529746109509...|[come, , , please...|        1|
|1.31652987081873e+18|[simple, question...|        0|
|1.316530571905032...|       [biden, lied]|        0|
|1.316531356759974...|[disclaimer, , to...|        1|
|1.316531369594544...|[hunterbiden, , j...|        0|
|1.316531426108547...|[come, huge, crow...|        1|
|1.316531639787421...|[yeah, , wonder, ...|        0|
|1.316531653599203...|[right, thing, co...|        1|
|1.316531920424112...|[share, , , votin...|        0|
|1.316532118332362...|[signal, key, sho...|        1|
|1.316532254777245...|[new, york, post,...|        1|
|1.316532305100337...|[, joebiden, mana...|        0|
|1.316532412847992...|[bet, get, , joeb...|        1|
|1.316532597527281...|[worry, , , house...|        0|
|1.316532858710880...|[cats, bag, , , ,...|        0|
|1.316532923651293...|[zoom, canvassing...|        1|
|1.316533120745775...|[donald, , trump,...|        0|
|1.316533265801662...|[care, give, inpu...|        0|
+-------------------+-------------------+---------+
only showing top 20 rows
```

- Converted neutral polarity values to negative labels.



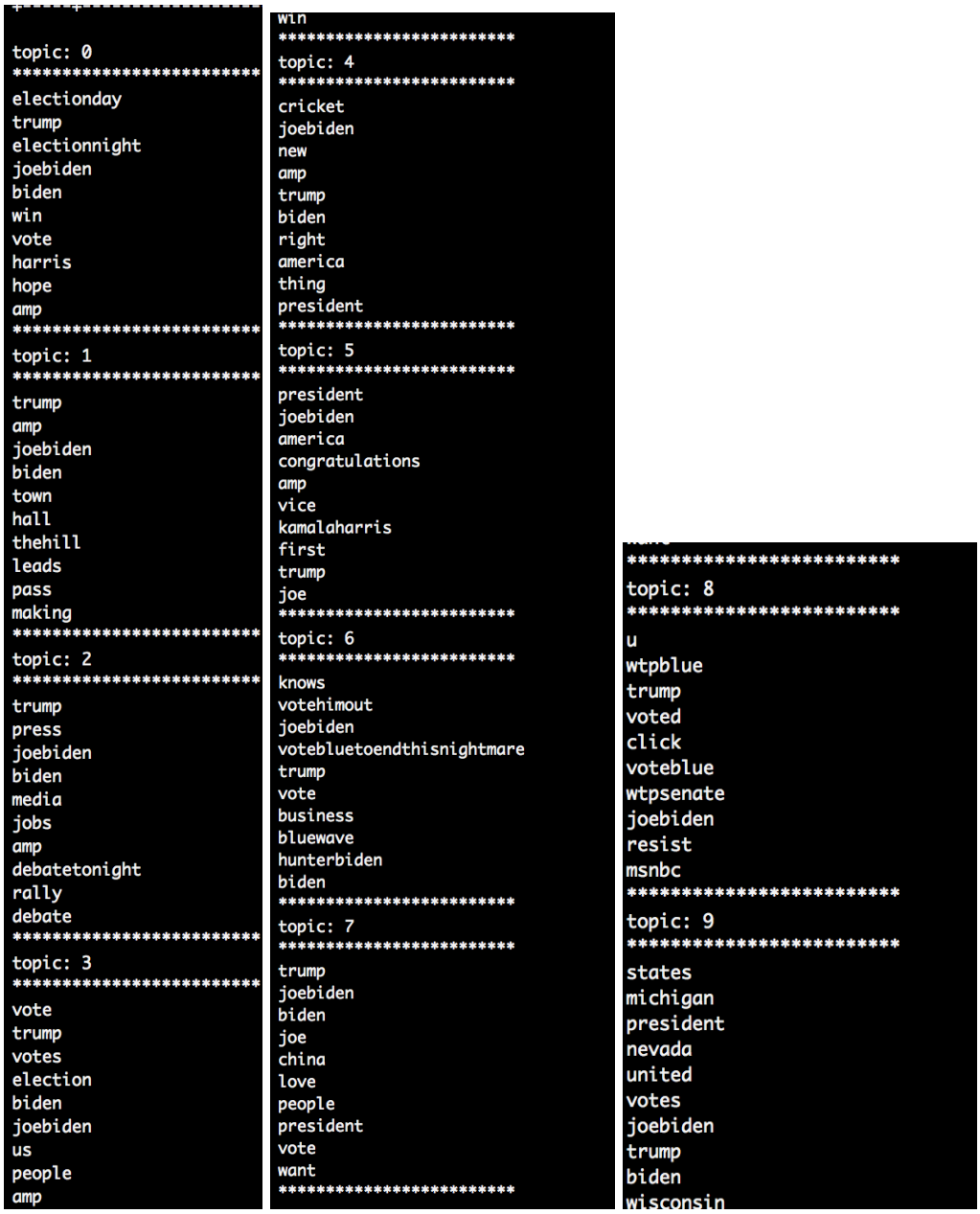
```
+--------------------+--------------------+---------+
|            tweet_id|               words|sentiment|
+--------------------+--------------------+---------+
|1.316529569361469...|[comments, , , , ...|        2|
|1.316529709371461...|[bidencrimefamily...|        2|
|1.316529746109509...|[come, , , please...|        1|
|1.31652987081873e+18|[simple, question...|        2|
|1.316530571905032...|       [biden, lied]|        2|
|1.316531356759974...|[disclaimer, , to...|        1|
|1.316531369594544...|[hunterbiden, , j...|        2|
|1.316531426108547...|[come, huge, crow...|        1|
|1.316531639787421...|[yeah, , wonder, ...|        2|
|1.316531653599203...|[right, thing, co...|        1|
|1.316531920424112...|[share, , , votin...|        2|
|1.316532118332362...|[signal, key, sho...|        1|
|1.316532254777245...|[new, york, post,...|        1|
|1.316532305100337...|[, joebiden, mana...|        2|
|1.316532412847992...|[bet, get, , joeb...|        1|
|1.316532597527281...|[worry, , , house...|        2|
|1.316532858710880...|[cats, bag, , , ,...|        2|
|1.316532923651293...|[zoom, canvassing...|        1|
|1.316533120745775...|[donald, , trump,...|        2|
|1.316533265801662...|[care, give, inpu...|        2|
+--------------------+--------------------+---------+
only showing top 20 rows
```

- Detected language of each tweet using detect(string) function of langdetect package and discarded the english tweets.

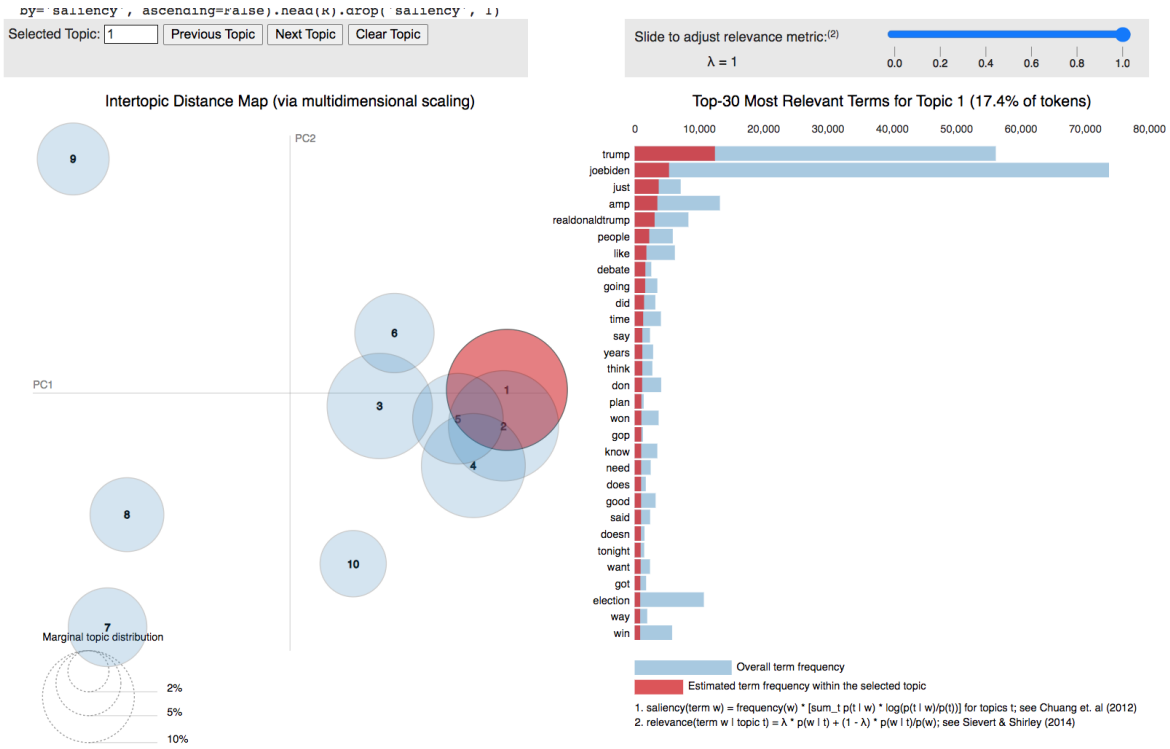


```
+--------------------+--------------------+---------+----+
|            tweet_id|               words|sentiment|lang|
+--------------------+--------------------+---------+----+
|1.316529569361469...|[comments, , , , ...|        2|  en|
|1.316529746109509...|[come, , , please...|        1|  en|
|1.316531356759974...|[disclaimer, , to...|        1|  en|
|1.316531426108547...|[come, huge, crow...|        1|  en|
|1.316531639787421...|[yeah, , wonder, ...|        2|  en|
|1.316531653599203...|[right, thing, co...|        1|  en|
|1.316532118332362...|[signal, key, sho...|        1|  en|
|1.316532305100337...|[, joebiden, mana...|        2|  en|
|1.316532412847992...|[bet, get, , joeb...|        1|  en|
|1.316532923651293...|[zoom, canvassing...|        1|  en|
|1.316533265801662...|[care, give, inpu...|        2|  en|
|1.316533551983218...|[looks, like, , t...|        1|  en|
|1.316533689162051...|[disqualified, , ...|        2|  en|
|1.316533726663389...|[time, , joebiden...|        1|  en|
|1.316534055840743...|[went, , , hour, ...|        1|  en|
|1.316534069635813...|[detailed, best, ...|        1|  no|
|1.316534568518852...|[hey, kids, , sha...|        2|  sv|
|1.316534860060729...|[watching, legacy...|        2|  en|
|1.316534993875804...|[hunter, biden, a...|        2|  en|
|1.316535052612898...|[kind, donation, ...|        1|  en|
+--------------------+--------------------+---------+----+
only showing top 20 rows
```

- Implemented LDA-LatentDirichletAllocation using pyspark.ml.clustering in Spark to find topic composition and membership for the data provided.Tuned the model for number of topics=10 and displayed the most important distribution of words for each topic .


topic: 0
****
electionday
trump
electionnight
joebiden
biden
win
vote
harris
hope
amp
****
topic: 1
****
trump
amp
joebiden
biden
town
hall
thehill
leads
pass
making
****
topic: 2
****
trump
press
joebiden
biden
media
jobs
amp
debatetonight
rally
debate
****
topic: 3
****
vote
trump
votes
election
biden
joebiden
us
people
amp
win
****
topic: 4
****
cricket
joebiden
new
amp
trump
biden
right
america
thing
president
****
topic: 5
****
president
joebiden
america
congratulations
amp
vice
kamalaharris
first
trump
joe
****
topic: 6
****
knows
votehimout
joebiden
votebluetoendthisnightmare
trump
vote
business
bluewave
hunterbiden
biden
****
topic: 7
****
trump
joebiden
biden
joe
china
love
people
president
vote
want
****
****
topic: 8
****
u
wtpblue
trump
voted
click
voteblue
wtpsenate
joebiden
resist
msnbc
****
topic: 9
****
states
michigan
president
nevada
united
votes
joebiden
trump
biden
wisconsin

- Used pyLDAvis visualization tools for presentation of the results.



- Used logistic regression in pySpark, by using the TF-IDF representation of their vectors. classified the tweets after performing 10-fold cross validation and ROC is calculatedand Enhance the tweet representation by adding the topic probability distribution for the tweet for the previous step and we obtained the ROC value as 0.85 without Topic Modelling and when applied Topic Modelling we obtained the ROC value as 0.72 using Logistic Regression, which is less when compared with without Topic Modelling.