Big Data Project
Twitter Sentiment & Topic Modelling

Group : Kevin & Junaid

Agenda

- Background
- Objective
- Data Description
- Workflow
- Sentiment Analysis (Classification)
- Topic Modelling (Clustering)
- Athena
- QuickSight Dashboard
- Challenges
- Conclusion
- Future considerations

Background

- Twitter sentiment analysis enables entities to track what the public opinion about a product, service and/or topic
- Helps detect things like angry customers or negative mentions before they become problematic to one's reputation.
- Can provide valuable insights that drive business decisions

Objectives

- Data: familiarization, cleanup, and helper functions.
- Sentimental Analysis: method and results review.
- Machine Learning: Logistic Regression and Random Forest.
- Topic Modelling.
- Collect all the data in S3 bucket and load it in Athena for further analysis.
- Create a dashboard containing visualizations in Quick Sight .
- Final Discussion.

Data Description

- #Inflation Tweets: 890K
- All topics tweets: 55 Million (reduced to 3.6 Million)

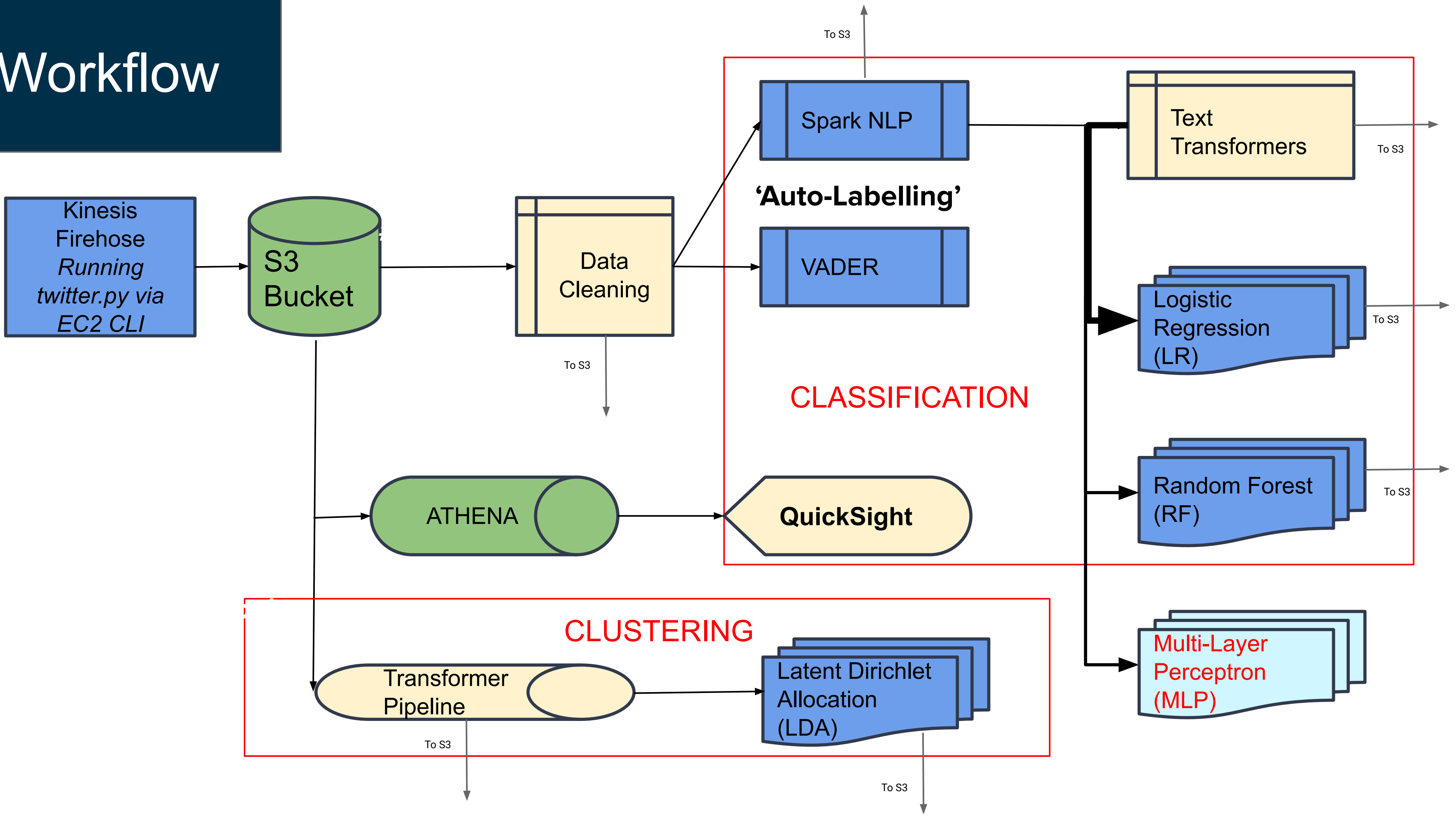
	folders	totalsize_mb
0	twitter/AI	2.45
1	twitter/BankofCanada	0.44
2	twitter/BlackFriday	476.65
3	twitter/CERB	0.0
4	twitter/CSIS	4.08
5	twitter/CanadaHousing	0.14
6	twitter/ElonMusk	73.72
7	twitter/Flames	0.07
8	twitter/Inflation	206.28
9	twitter/Interest_rate	1.26
10	twitter/Iran	159.17
11	twitter/MTA	121.38
12	twitter/StudentLoanRelief	0.71
13	twitter/WorldCup	4964.33
14	twitter/cancer	44.25
15	twitter/greenbelt	0.01
16	twitter/thanksgiving	947.83
17	twitter/twitter	5088.85
18	twitter/wecan	6.56

	id	name	username
1	1602659952749076480	Marc Burr	marcburr
2	1602659953285951493	STOCK TRAIN	stocktrain2
3	1602659953621204992	Caleb Kaplan	CalebKaplan
4	1602659955949043712	JD	JaedenJD
5	1602659956016418816	Crutcial News of Crypto's	CrusNewsCrypto
6	1602659957240868867	Fred Randall	FredRandall15

followers_count	location	geo	created_at
337	Billings, Montana	None	Tue Dec 13 13:41:23
34024	None	None	Tue Dec 13 13:41:23
2050	Georgia, USA	None	Tue Dec 13 13:41:23
182	Singapore	None	Tue Dec 13 13:41:24
51446	None	None	Tue Dec 13 13:41:24
23	None	None	Tue Dec 13 13:41:24

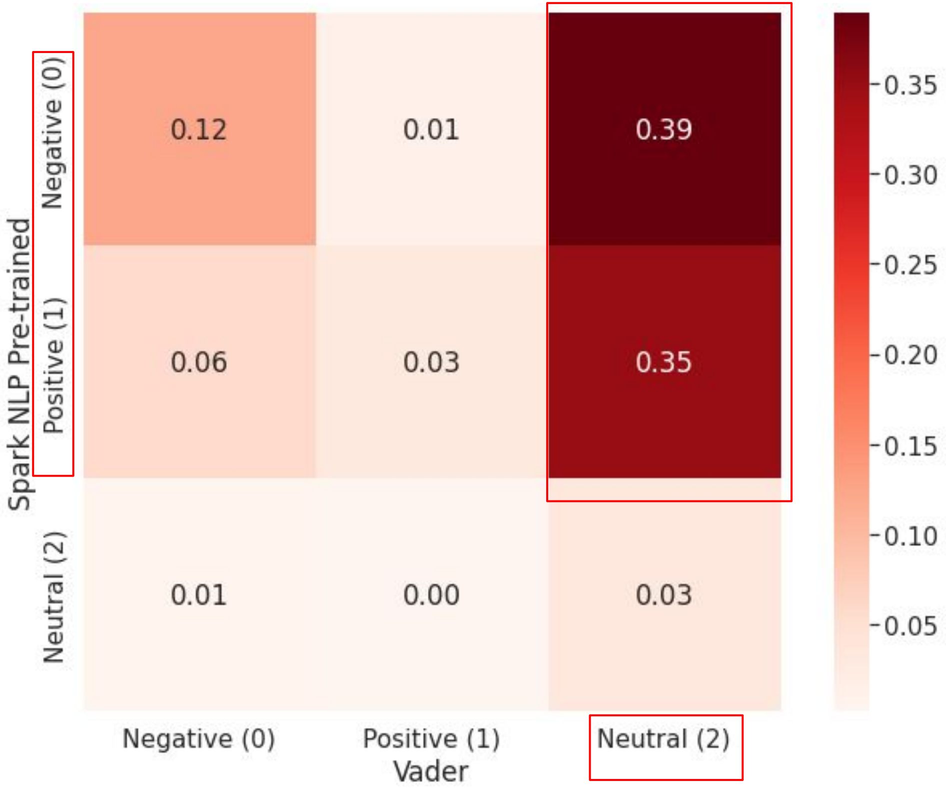
tweet
RT @MorningBrew: If there's one thing the market loves, it's better than expected inflation data. Happy Tuesday everyone. https://t.co/jjGc...
10-year Treasury yield drops below 3.5% after inflation reading @CNBC https://t.co/qwJgRZngMw
RT @RepJasonSmith: The $2 trillion American Rescue Plan sparked the worst inflation in 40 years, forcing every family to pay $8,600 more th...
RT @BusinessInsider: Inflation cooled again in November to the slowest pace in a year https://t.co/CBUM5wQpWb
BREAKING: U.S. inflation slowed again last month in the latest sign that price increases are slowly cooling despite... https://t.co/GsyGAFPHof
@StonedSportDude @disclosetv Ummm dude they just sent more money to Ukraine. Watch after Christmas. 2023 inflation will be worse.

Workflow

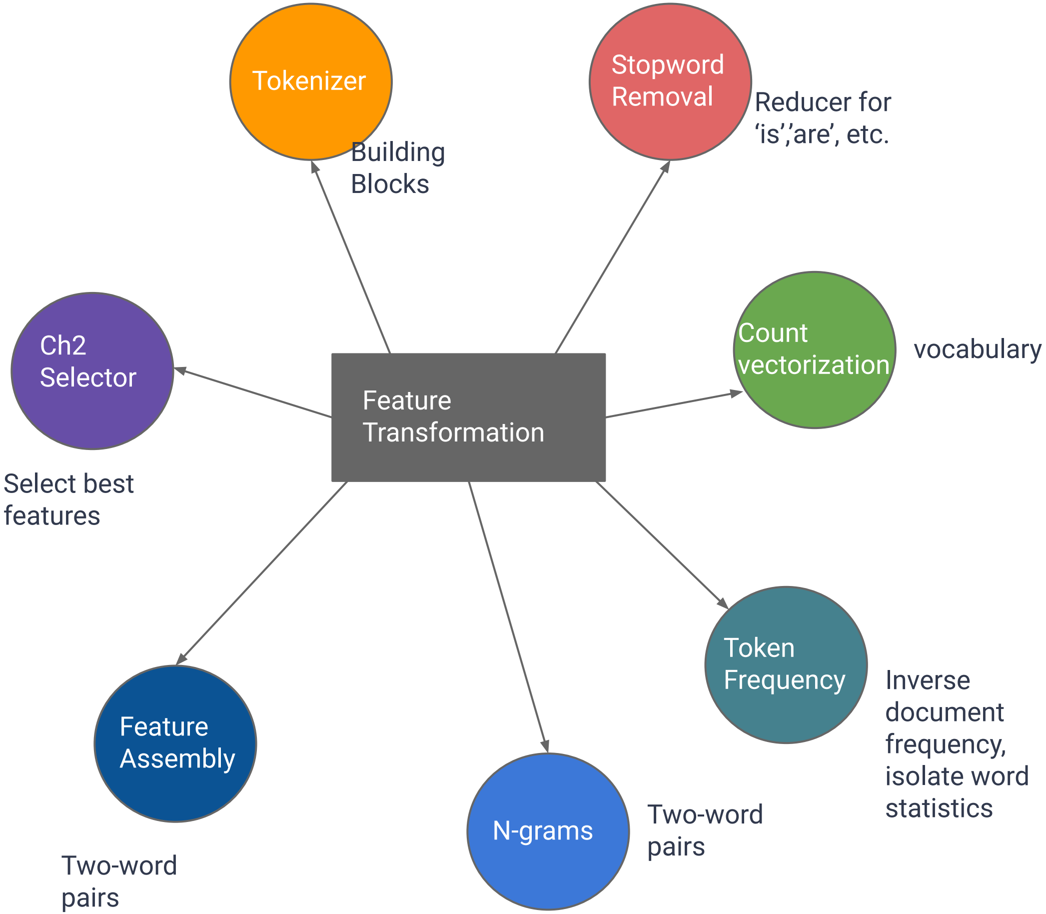

Kinesis Firehose *Running twitter.py via EC2 CLI*
S3 Bucket
Data Cleaning
To S3
Spark NLP
To S3
'Auto-Labelling'
VADER
Text Transformers
To S3
Logistic Regression (LR)
To S3
CLASSIFICATION
Random Forest (RF)
To S3
ATHENA
QuickSight
Multi-Layer Perceptron (MLP)
CLUSTERING
Transformer Pipeline
To S3
Latent Dirichlet Allocation (LDA)
To S3

Sentiment Analysis (Classification)

- Cleaning:
    - Regular expressions
    - No RTs = ~330k tweets
- In reality: fraction of data would need to be human-labelled (auto-TURK, self)
- Labellers: {negative, positive, neutral} = {0,1,2}
    - VADER - Lexicon/Rule-based;
        - Used distribution to make it more neutral
    - Pre-train spark-NLP pipeline model; specifically for twitter sentiment = generally polarized

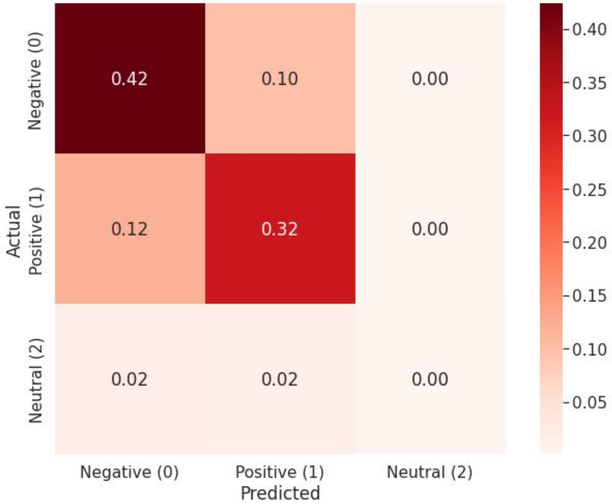


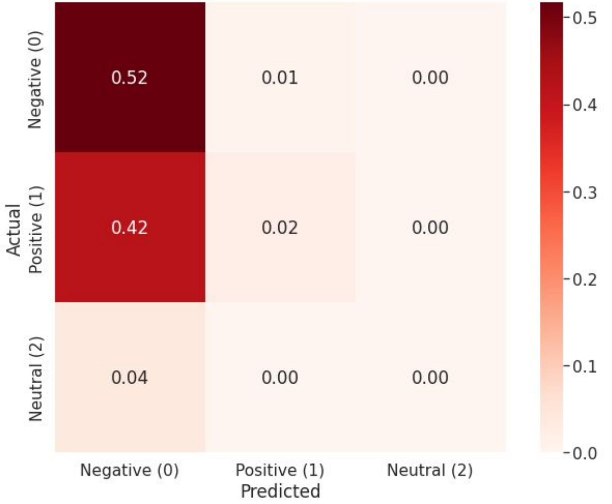
Analysis (Classification)



Sentiment Analysis (Classification)

- 70-30 Train-Test split
- Classifier GridSearchCV (Test Results)
  - Logistic Regression:
    - α = [0.1, 0.3, 0.5,0.7,0.9]
    - λ = [0.0,0.15, 0.3,0.5,0.75,0.9]
    - ~Lasso
    - **Accuracy: 74.0%**
    - ROC AUC: 73.4%
  - Random Forest (abandoned)
    - Accuracy: 54.0%
    - ROC AUC: 40.2%

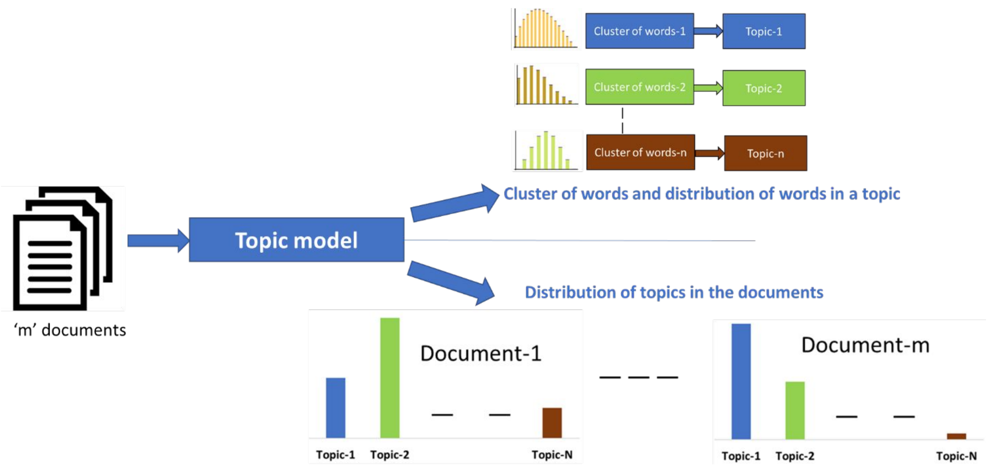


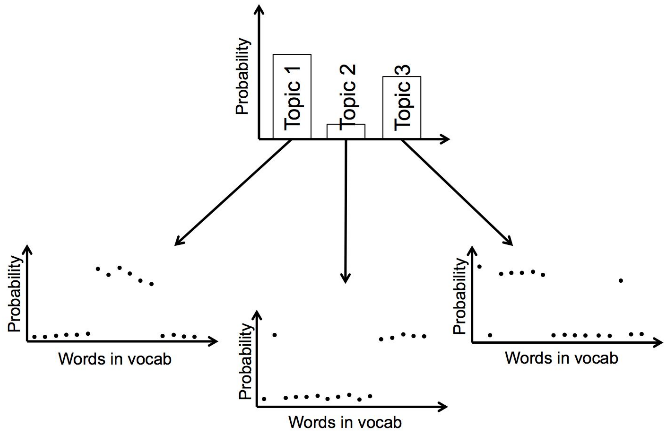


https://www.analyticsvidhya.com/blog/2021/06/part-18-step-by-step-guide-to-master-nlp-topic-modelling-using-lda-probabilistic-approach/

Topic Modelling (Clustering)

- More realistic exercise without labels (unsupervised)
- Explored inflation data (900k tweets) & all topics (55mil. To 3.6mil.)
- Transformer pipeline: process raw tweets
    - **Custom transformer class** for cleaning+ same classification transformers
- Latent Dirichlet Allocation (LDA) clustering algorithm:
    - word distribution of each topic(θ) & topic distribution over corpus (Z)

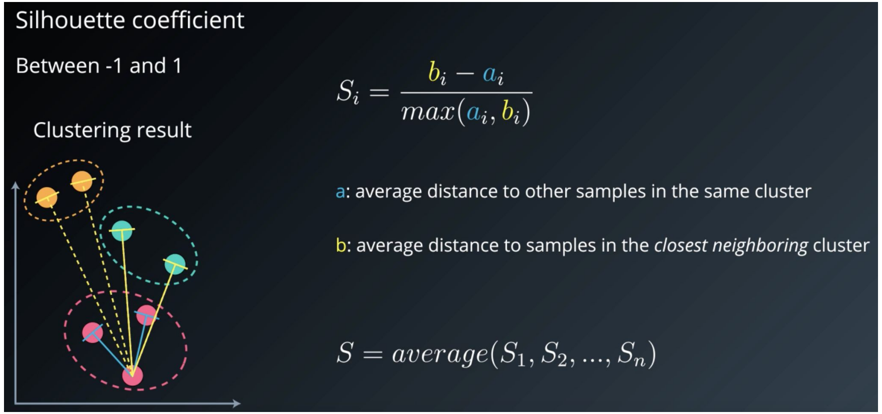




https://www.analyticsvidhya.com/blog/2021/06/part-18-step-by-step-guide-to-master-nlp-topic-modelling-using-lda-probabilistic-approach/

Topic Modelling (Clustering)

- #Inflation tweets: hard to distinguish clusters (tried more clusters with similar results)
- Clustering Evaluation: Silhouette Measurement = **-0.01**
    - In range of [-1,1], with being close to 1 = points in a clusters are close to other points in the cluster // far from points of other clusters



	topic	topicWords
1	0	["rate", "fed", "november", "us", "cpi", "data", "year", "even", "energy", "price", "go", "market", "really", "interest", "work", "high", "lower", "every", "joe", "consumer"]
2	1	["money", "people", "rates", "wages", "bank", "good", "like", "know", "get", "trav", "biden", "many", "us", "re", "global", "fed", "interest", "government", "high", "m"]
3	2	["prices", "gas", "biden", "pay", "high", "bitcoins", "still", "years", "going", "new", "like", "year", "world", "much", "&", "government", "need", "one", "m", "economy"]

Topic Modelling (Clustering)

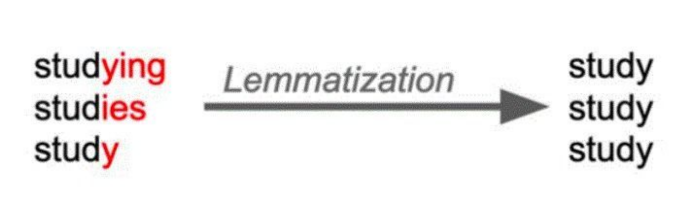
- 55mil. tweets
- World cup dominates most topics
- Poor topic distribution
- Hard to iterate on model parameters on dataset size

	topic	topicWords	count
1	0	["black", "friday", "black friday", "sale", "", "rt", "friday sale", "none", "&", "hours", "hear", "call", "big", "opportunity", "rt &"]	3755410
2	1	["pokemon", "day", "like", "like pokemon", "every", "drop", "world", "cup", "world cup", "real", "done", "m", "never", "united", "every day"]	2673166
3	2	["final", "world cup", "world", "cup", "cup final", "finals", "goals", "mbappe", "messi", "quarter", "elon", "kylian", "musk", "elon musk", "scored"]	2992402
4	3	["like", "dinner", "thanksgiving", "thanksgiving dinner", "right", "always", "meet", "world", "cup", "world cup", "team", "free", "turkey", "best", "first"]	2676399
5	4	["dreamers", "official", "fifa", "fifa world", "jung", "kook", "jung kook", "world", "iran", "world cup", "cup", "music", "qatar", "kook dreamers", "video"]	2339480
6	5	["world", "cup", "world cup", "morocco", "african", "semi", "next", "reach", "first", "ve", "portugal", "congratulations", "cup semi", "last", "team"]	2550828
7	6	["world", "world cup", "cup", "best", "best world", "post", "trump", "one", "take", "remember", "ever", "rica", "costa", "literally", "costa rica"]	2376501
8	7	["back", "x", "back back", "back world", "cup", "world", "world cup", "claim", "super", "rm", "cute", "wl", "eyes", "inflation", "special"]	3253272
9	8	["morning", "good", "world", "whole", "man", "good morning", "final", "world cup", "cup", "hope", "final world", "field", "cristiano", "cristiano ronaldo", "ronaldo"]	2342622
10	9	["m", "year", "much", "looking", "things", "come", "eat", "business", "money", "love", "think", "dear", "ll", "person", "thanksgiving"]	3429911
11	10	["&gt", "fuck", "let", "&gt &gt", "see", "twitter", "friend", "care", "rewards", "new", "nice", "inflation", "like", "go", "thing"]	3264477
12	11	["win", "win world", "world", "cup", "world cup", "old", "messi", "leo", "leo messi", "year old", "watching", "year", "times", "games", "goals"]	2326056
13	12	["oh", "god", "thanksgiving", "happy", "everyone", "world", "cup", "world cup", "thanksgiving everyone", "thank", "sb", "ger", "break", "damn", "today"]	2491105
14	13	["follow", "retweet", "giveaway", "good", "&", "rt", "like", "luck", "enter", "tweet", "good luck", "guys", "tag", "need", "thanksgiving"]	3632419
15	14	["winter", "war", "candy", "smtown", "kst", "smcu", "pm", "smtown smcu", "palace", "beautiful", "wahl", "congrats", "grant", "smcu palace", "world"]	1748113
16	15	["soon", "girl", "thanksgiving", "car", "m", "social", "palestine", "&", "want", "forever", "boys", "action", "edition", "mins", "media"]	2305587
17	16	["live", "vs", "stream", "link", "live stream", "fifa", "watch", "fifa world", "world cup", "cup", "world", "hd", "live link", "france", "qatar"]	2656341
18	17	["happy", "happy thanksgiving", "thanksgiving", "family", "cup", "world", "world cup", "asian", "thankful", "vote", "thanks", "acoty", "countries", "day", "celebrating"]	2616279
19	18	["goal", "winning", "world", "cup", "world cup", "first", "winning world", "first world", "cup goal", "wow", "omg", "winner", "qatar", "fifa", "sad"]	1990870
20	19	["cup", "world cup", "world", "argentina", "saudi", "won", "arabia", "saudi arabia", "messi", "won world", "history", "far", "match", "first", "football"]	3616690

Topic Modelling (Clustering)

- Removed retweets
- Downsampled WorldCup & twitter buckets (5%)
- = 3.6mil. Tweets of all topics in the bucket
- **Add: Custom Transformer class for lemmatization**
- **Grid Search:** n topics & topic distribution (beta)

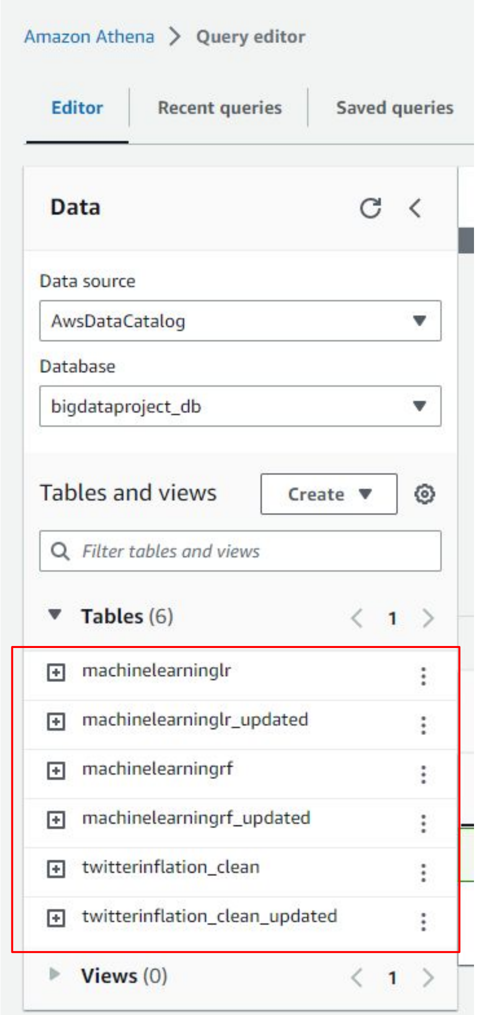
topicWords_k10_beta0_1	count_k10_beta0_
['might' 'moment' 'cancer' 'task' 'discovering' 'aptitude' 'aptitude task' 'might moment' 'moment discovering' 'discovering aptitude']	176688
['thanksgiving' 'happy' 'family' 'happy thanksgiving' 'hope' 'day' 'inflation' 'good' 'great' 'dinner']	542810
['black' 'friday' 'black friday' 'happy' 'happy thanksgiving' 'sale' 'deal' 'thanksgiving' 'friday sale' 'friday deal']	588362
['ukraine' 'well' 'thanksgiving' 'inflation' 'get' 'yes' 'people' 'got' 'money' 'russia']	485641
['world' 'cup' 'world cup' 'live' 'war' 'fifa' '2022' 'fifa world' 'stream' 'iran']	300879
['know' 'believe' 'probably' 'bring' 'cancer' 'pursuit' 'believe know' 'bring cancer' 'probably believe' 'know pursuit']	275196
['inflation' 'leftover' 'fed' 'thanksgiving leftover' 'trump' 'make' 'need' 'little' 'market' 'hear']	355182
['elon' 'musk' 'elon musk' 'twitter' 'world' 'thanksgiving' 'feel' 'love' 'nft' 'messi']	229746
['thanksgiving' 'year' 'one' 'inflation' 'look' 'american' 'guy' 'time' 'like' 'thanksgiving weekend']	342573
['thank' 'god' 'thanksgiving' 'black' 'record' 'friday' 'black friday' 'beautiful' 'via' 'holiday']	297353



https://www.turing.com/kb/stemming-vs-lemmatization-in-python

Athena

- Data from S3 was imported into athena
  - Inflation Tweet Data
  - Logistic Regression ML
  - Random Forest ML
- CTAS method was used to created another updated table
  - Data cleaning performed
  - Additional column created
  - Removal of unnecessary columns
  - Cleaning location data
  - Date extracted from string column
- Clustering Data wasn't included due to vector import limitations

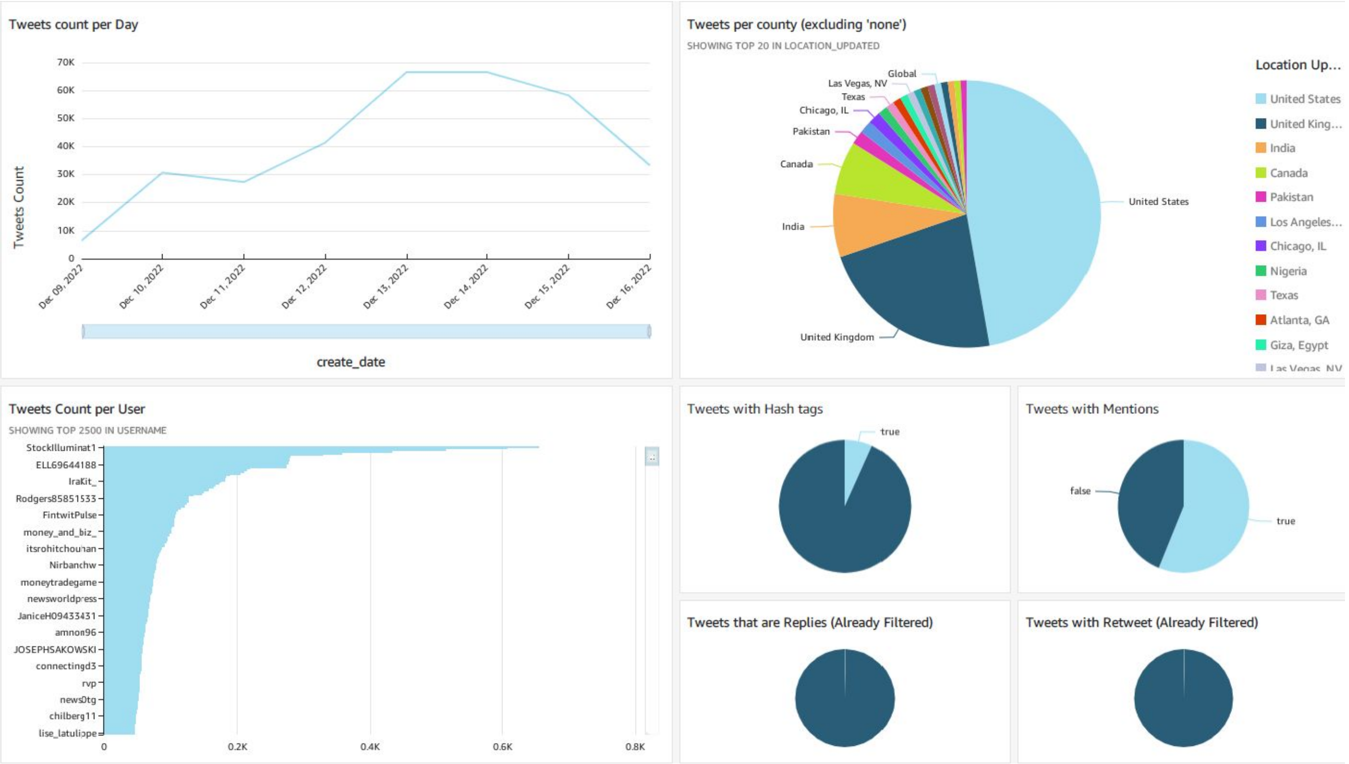



```sql
CREATE EXTERNAL TABLE IF NOT EXISTS `bigdataproject_db`.`machinelearninglr` (
  `id` string,
  `name` string,
  `username` string,
  `tweet` string,
  `followers_count` int,
  `location` string,
  `created_at` string,
  `text` string,
  `f_retweet` string,
  `f_reply` string,
  `f_mentions` string,
  `f_hashtag` string,
  `vader_score` int,
  `vader_label` int,
  `snlp_sentiment` string,
  `label` int,
  `predictions` int
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'
WITH SERDEPROPERTIES ('field.delim' = '|')
STORED AS INPUTFORMAT 'org.apache.hadoop.mapred.TextInputFormat' OUTPUTFORMAT '
LOCATION 's3://ptb2-junaid/twitterInflation_lr_pred_all_delimited.csv/'
TBLPROPERTIES ('classification' = 'csv');
```

```sql
CREATE TABLE machinelearninglr_updated AS
Select
 id
,username
,tweet
,followers_count
,SUBSTR(created_at,5,6) || ', ' || SUBSTR(created_at,27,4) as create_date --Step #2
,f_reply as tweet_as_reply
,f_mentions as tweet_has_mention
,f_hashtag as tweet_has_hashtag
,snlp_sentiment as sentiment
,CASE
WHEN Predictions = 0 THEN 'Negative'
WHEN Predictions = 1 THEN 'Neutral'
WHEN Predictions = 2 THEN 'Positive'
END AS prediction_sentiment -- Step #6
,CASE
WHEN predictions = label then 'Accurate'
ELSE 'Inaccurate'
END as Prediction_Accuracy -- Step #5
,CASE
When location like '%USA' Then 'United States'
When location like '%UK' Then 'United Kingdom'
When location like '%London%' Then 'United Kingdom'
When location like '%, NY' Then 'United States'
When location like '%India' Then 'India'
When location like '%Pakistan' Then 'Pakistan'
When location like '%England%' Then 'United Kingdom'
When location like '%United Kingdom%' Then 'United Kingdom'
When location like '%Canada%' Then 'Canada'
When location like '%New York%' Then 'United States'
When location like '%Washington%' Then 'United States'
When location like '%Toronto%' Then 'Canada'
When location like 'None' Then 'Earth'
ELSE location
END AS location_updated -- Step #3
from "bigdataproject_db"."machinelearninglr"
where 1=1
and f_retweet LIKE 'false' -- Step #1
```

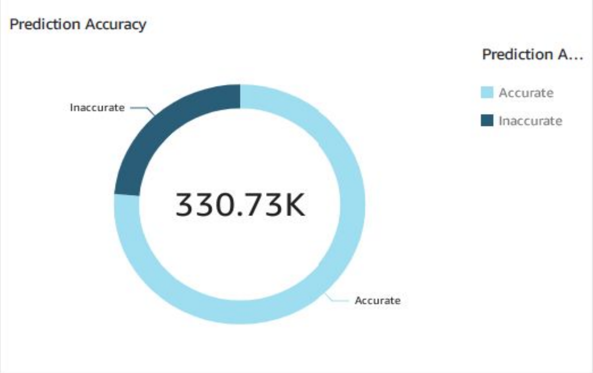
QuickSight

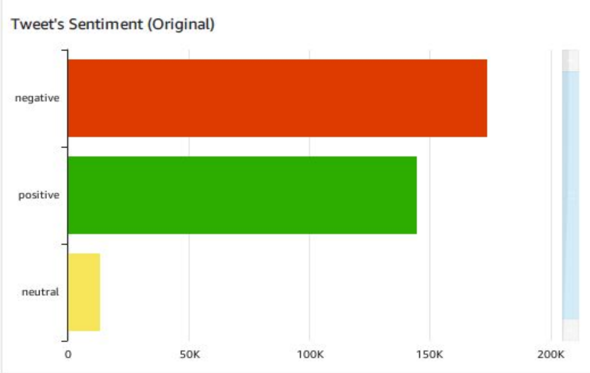
- Pre ML Tweets data
- Count per day
- by country
  - Mostly US and UK
- Count per user
- Count with hashtags
- Count with mentions
- Count with replies
- Count for retweets

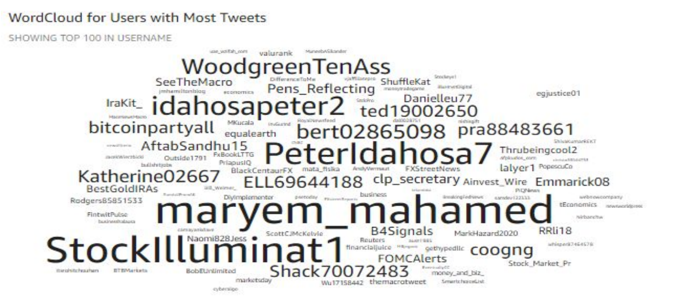


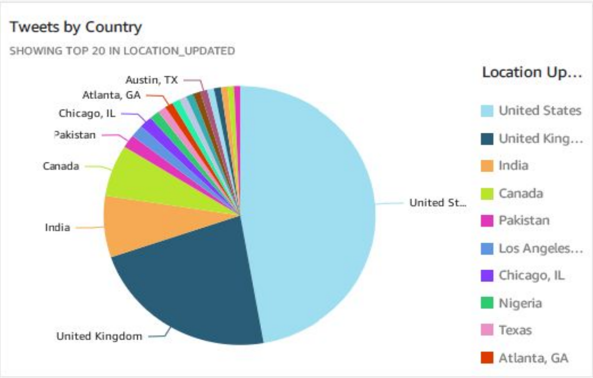
QuickSight

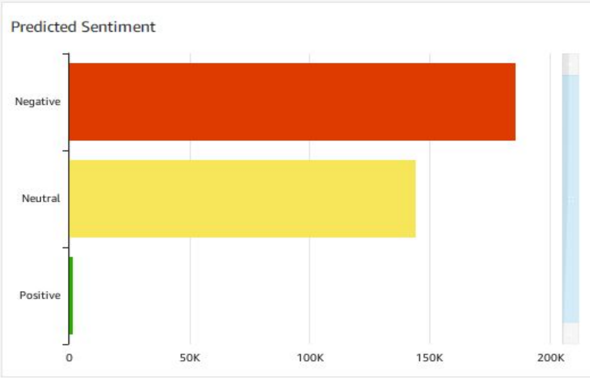
- Post Logistic Regression ML model
- Prediction Accuracy
- Tweets sentiments snlp
- Predicted sentiments
- Tweets by country











Challenges

- Twitter streaming - rate limit
- DataBricks Community version had computing limitations
- DataBricks “free-trial”:
  - Free Databricks usage but still charged for EC2 clusters
  - Experimentation and errors were costly
- Spark
  - Limited resource availability
  - Advanced statistical/ML methods are not natively available in PySpark
- Spark-nlp setup
  - Difference between standard ML packages vs. PySpark ML vs. Spark NLP
- Athena
  - Import issues with delimiter selection
  - Certain datasets do not get imported well
  - Limitations to update, insert or delete due to data source being S3

Conclusion

- Classification model produced results that were 76% accurate
  - Decent; suggests that the model is well-suited to the task it was trained on.
  - LR = supervised; so the better the model selection and feature engineering the better the accuracy

- Unsupervised learning clustering was also used to analyze a much bigger data set
  - The clustering was fine-tuned in a custom grid-search procedure
  - Resulted in fairly distinct clusters/topics, given the tools available in PySpark.

Future Considerations

- Stream data (if not tweets another type of text data, ex. News APIs).
- Build a more stable ETL method and automate various manual workflows.
- Explore more advanced NLP:
  - More transformers
  - pre-train models, embeddings
- Deep learning
  - Classification: Fix issue with MLP & try others
  - Clustering: Variational Autoencoders, Deep Adaptive Clustering.
- Manual labelling to check and tweak labels.
  - Building custom lexicon rather than relying on general linguistic packages.
- Clean location data = geographic plots.
- Multi-language tweet processing