Mining Public Opinion: Sentiment Analysis of Tweets to Discover Regional & Demographic Trends

Team: Data Smugglers - Vishnu Alla, Vineeth Karjala, Surya Vakkalagadda

Project idea

- Sentiment analysis helps understand public emotions, reactions, and opinions
- Build sentiment classification models using Machine Learning and deep learning techniques on twitter dataset
- Use the results on airline customer satisfaction and Public sentiment of Squid Game

Objectives

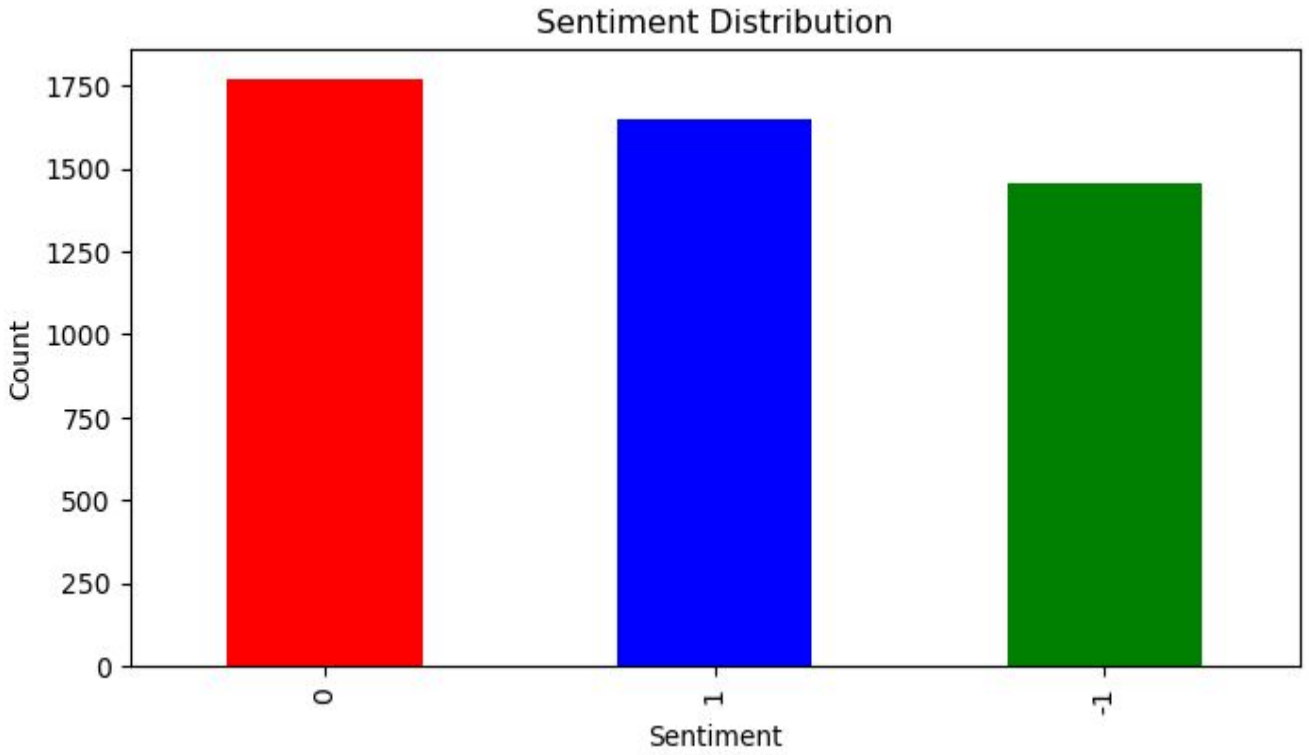
- Train and compare various models (SVM, Random Forest, Naive Bayes, Feedforward NN, LSTM) on labeled Twitter data
- Use BoW and TF-IDF for feature extraction
- Best model based on F1-score to balance precision and recall
- Apply the trained model to unlabeled datasets to infer sentiments

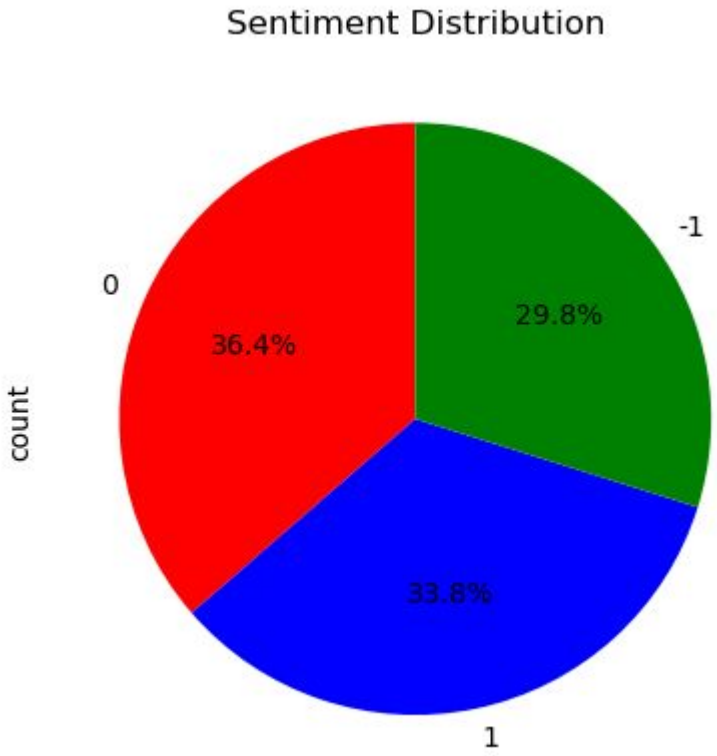
Data

Dataset	Usage	Size	Input Feature	Target Variable
Dataset 1	Training & Testing & Validation	4869	Tweet text	Sentiment Label
Dataset 2	Airline Sentiment Analysis	14640	Tweet text	Inferred sentiment
Dataset 3	Squid Game Sentiment by Location	56149	Tweet text	Inferred sentiment

How I feel today #legday #jelly #aching #gym
@ArrivaTW absolute disgrace two carriages from Bangor half way there standing room only #disgraced
This is my Valentine's from 1 of my nephews. I am elated; sometimes the little things are the biggest & best things!
betterfeelingfilms: RT via Instagram: First day of filming #powerless back in 2011. Can't ¡
Zoe's first love #Rattled @JohnnyHarper15

Sentiment distribution



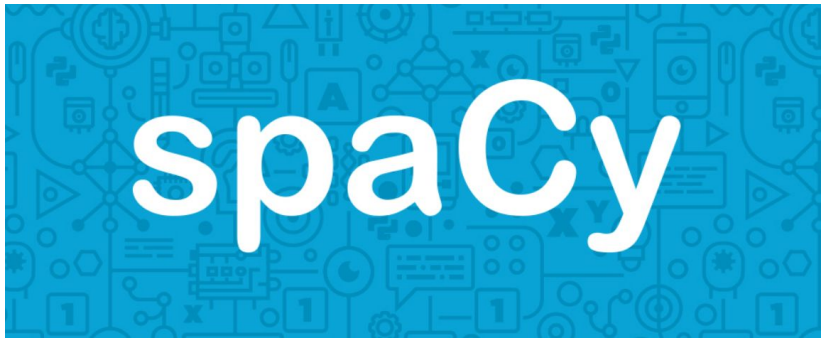


Methodology: Data Preprocessing

- Cleaning - removed URLs, mentions, hashtags
- Lowercasing, stopword removal
- Tokenization and lemmatization
- Label encoding negative, neutral and positive sentiments (-1, 0, 1)

Data Cleaning - continued

- Removed invalid/non-geographic user locations



Feature Extraction

- Bag of Words (BoW) - frequency of each word in the corpus vocabulary
- Term Frequency–Inverse Document Frequency (TF-IDF) - more weight to rare but important words

Model Development

Model	Input features	Hyperparameters tuned on
SVM	TF-IDF vectors	Cost [0.1, 1, 10] Kernel [linear, RBF]
Random Forest	TF-IDF vectors	n_estimators [100, 200] max_depth [None, 10, 20]
Naïve Bayes	TF-IDF vectors	Alpha [0.1, 0.5, 1.0]
SVM	BOW vectors	Cost [0.1, 1, 10] Kernel [linear, RBF]
Random Forest	BOW vectors	n_estimators [100, 200] max_depth [None, 10, 20]
Naïve Bayes	BOW vectors	Alpha [0.1, 0.5, 1.0]
Feedforward Neural Network	TF-IDF vectors	learning_rates = [0.001, 0.0005, 0.01] hidden_dims = [64, 128, 256] batch_sizes = [32, 64]
LSTM	TF-IDF vectors	learning_rates = [0.001, 0.0005, 0.01] hidden_dims = [64, 128, 256] batch_sizes = [32, 64]

Model Summary

Model	Feature used	F1 Score	Model Complexity	% Difference from Best Model
SVM	TF-IDF	0.632	Comparatively less	-2.17
SVM	BoW	0.616	Comparatively less	-4.64
Random Forest	TF-IDF	0.615	Comparatively less	–4.8
Random Forest	BoW	0.622	Comparatively less	-3.72
Naive Bayes	TF-IDF	0.606	Fast, simple	-6.2
Naive Bayes	BoW	0.62	Fast, simple	-4.02
Feed Forward NN	TF-IDF	0.646	High	-
LSTM Network	TF-IDF	0.62	High	-4.02

Hyperparameter Tuning Strategy

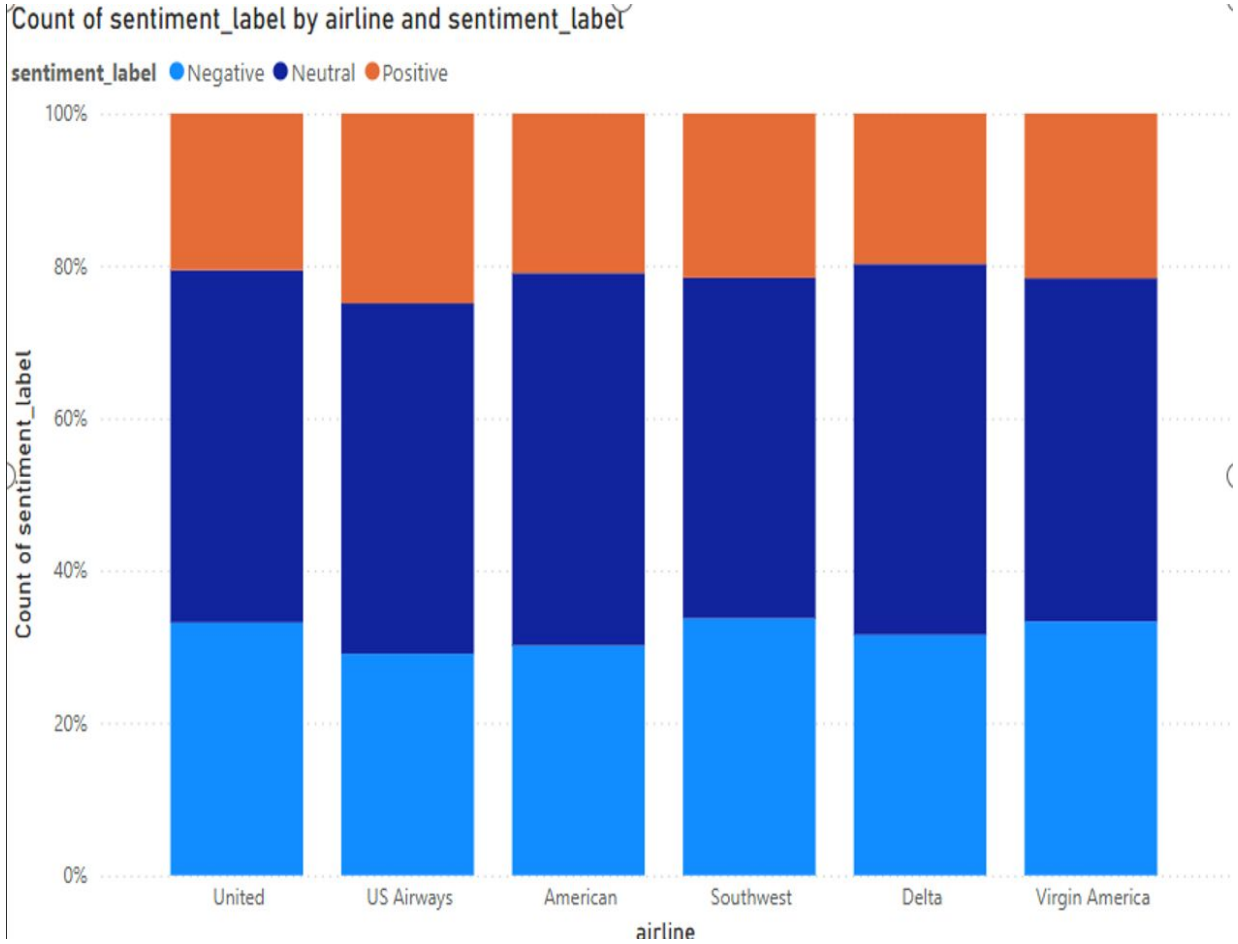
- Key params: SVM: C, kernel
  RF: n_estimators, max_depth
  FFNN/LSTM: LR, hidden layers, epochs
- Validation via 80:20 train-test split

Key Improvements

- Added geographic filtering for location-based sentiment
- Used Power BI for visual insights
- Balanced accuracy vs complexity

Airline Sentiment Analysis Results

- Model used: SVM (TF-IDF)
- United Airlines: High volume, mostly negative/neutral
- Virgin America: Lower volume, higher positive ratio
- Other carriers: Mixed but mostly negative



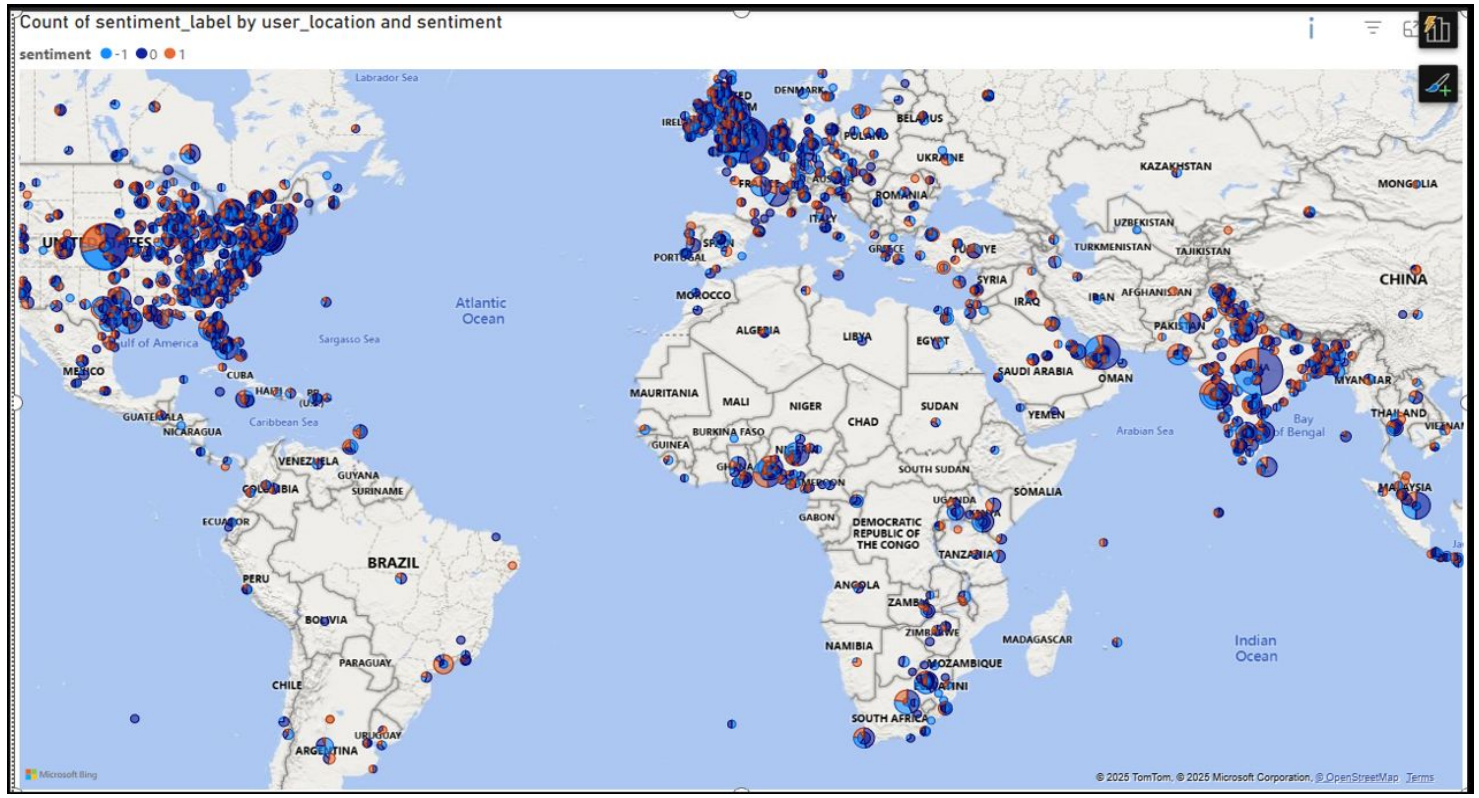
>United Airlines had the highest tweet volume with a large share of neutral and negative sentiments, indicating higher customer dissatisfaction or issues.

> American, US Airways, and Southwest also showed high proportions of negative tweets, though slightly better than United.

> Virgin America, despite receiving fewer tweets overall, had a notably higher proportion of positive tweets, suggesting a relatively better perception among users.

Geographic Sentiment for ‘Squid Game’

- Applied location filters to map sentiments globally
- North America & Europe: Mostly positive
- Other regions: Mixed/neutral/negative (possible cultural variation)



> A majority of tweets from North America and parts of Europe showed positive sentiment, indicating strong viewer appreciation in those regions.

> Some neutral and negative sentiments were observed in other parts of the world, possibly due to cultural differences, content preference, or expectations not being met.

Conclusion

- Built and evaluated ML and DL models for sentiment analysis
- SVM with TF-IDF was chosen as the best model though Feed Forward NN achieved highest accuracy because of it's simple architecture
- Used SVM with TF-IDF for inference on the other datasets - Airline and Squid Game sentiment mapping

Future work

- To train on more complex models like transformer-based models to check for significant gain in F1
- Incorporating pre trained word embeddings like GloVe and Word2Vec