Data Science Project 2. Airline Tweets

Team 4
George Hill, Koti Jaddu, Lauren Alie, Rory Tallon, David Marples, Oliver Little

The Problem

Analysis of tweets about US airlines

- Gain insights from exploratory analysis
- Compare non neural network and neural network models
- Understand public sentiment towards the airlines and reasons
- Consider future technical and business steps in this investigation

The Approach

- Discuss potential insights
- Migrate tweet data into a database
- Gather data from other datasets to explore
- Research different approaches, then assign to individuals
- Finalise model results & develop presentation

Working as a Team

- Discord
  - Discuss the task
  - Share understanding
  - Hold weekly meetings
- GitHub
  - Collaborate on code
  - Kanban board to assign tasks and keep track of progress in an agile way
- Google Slides
  - Collaborate on presentation

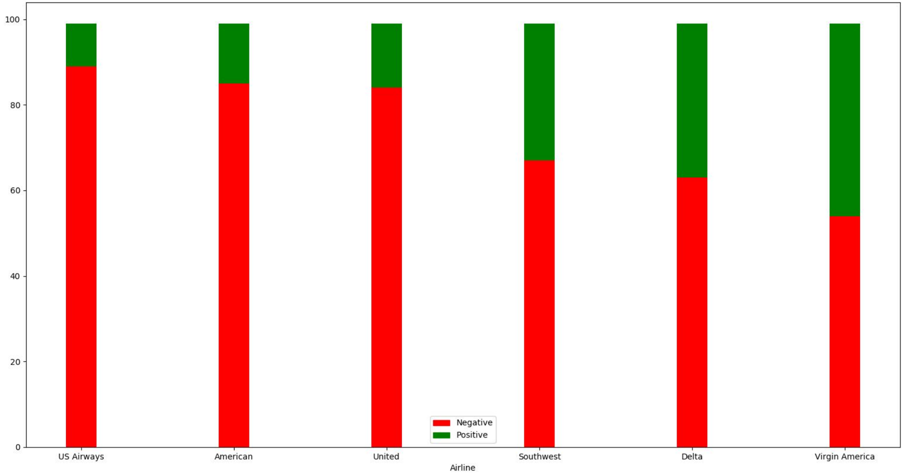
Preliminary Analysis

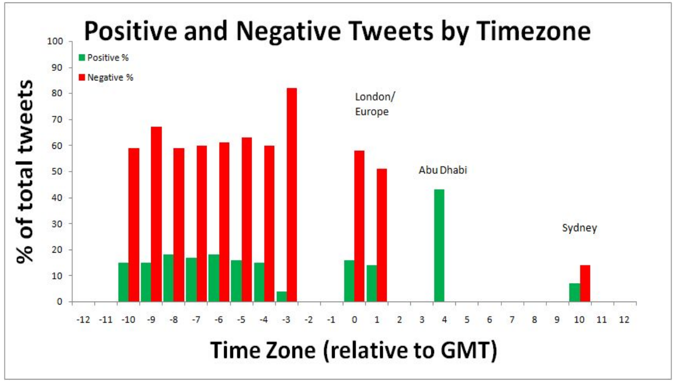
First look at the data

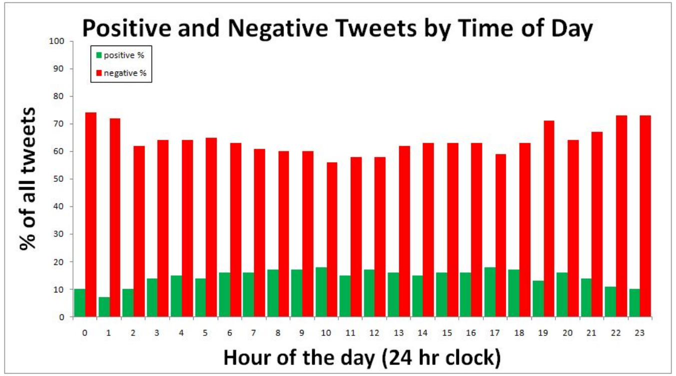
- JetBlue is labelled wrong
- Not all columns filled for all tweets
    - Tweet coordinates few and far between
- Tweet Location appears to be user-reported
    - e.g. Tweets from “1/1 Loner Squad” and “somewhere celebrating life”
- Sarcasm consistently classified wrong
    - e.g. “plus you’ve added commercials to the experience… tacky.” → Positive!?

Preliminary Analysis

Worst Airlines by proportion of negative tweets

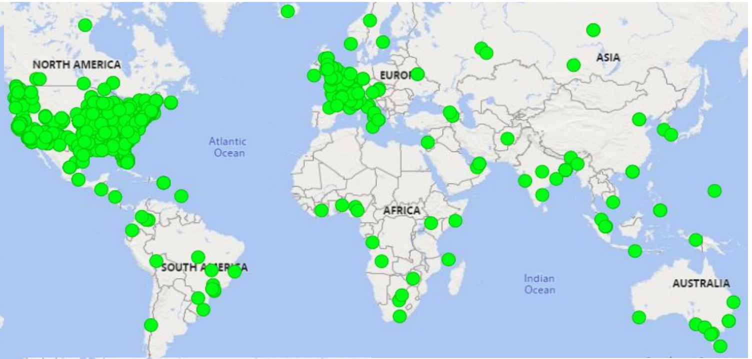






Anything you notice about data?

Negative



Positive

Neutral

Custom Location Analysis

Attempted to infer location from the text:

- Specific keywords that typically had a location mentioned before or after.
- Used lists of airport codes and city names to work out if a word was actually a location

Relatively successful:

- Over 2000 tweets in the dataset had a match in one of the location lists.
- Some locations were particularly negatively mentioned, others were positive.

Issues:

- Limited transferability to other datasets.
- Could be improved by combining with latitude/longitude location.

Latent Semantic Analysis

Process of automatically categorising the dataset into topics. Two methods:

- Singular Value Decomposition
- Latent Dirichlet Allocation

Findings:

- Same topics generally appear across the whole dataset, even if the dataset is categorised by airline.
- Some patterns visible, but issues with tweets being in the wrong category.
- Manual method produce better results.

Preprocessing

- Remove non-alphabetical characters
- Convert to lowercase
- Stopwords and length limit
- N-grams
- Stemming/Lemmatization
- TF-IDF
- Emoji Translation
- Converting #HashtagsWithTitleCase to words
- Removing some abbreviations such as w/, &, and -->
- Removing punctuation and contractions

Modelling

Baseline Modelling Approaches

All models used pre-processed text from tweets.csv as data for classification, correlating the words with the "airline_sentiment".

1) Dictionary-based:
    - Dictionary of all words constructed (~8200 words)
    - "Key" words ("high" frequency; strong positive or negative associations) identified (~1800 words)
    - Test tweets analysed by identifying key words to assign a classification. Accuracy 73%.
    - Vector created from the included key words for...

2) Simple neural net: 1 minute to train; Accuracy 80%

3) Naive Bayes: tweets clustered by similar word content. Training time 0.2 seconds; Accuracy 78%

4) Logistic Regression: Training time 3.4 seconds; Accuracy 67%

Recurrent Neural Network

Accuracy: ~90.2%
Training time: ~34 minutes
Prediction time: ~1.342s

Embedding Layer -> LSTM Layer -> LSTM Layer -> Dense Layer

Extension Data

Extension Data: Analysis

Two new sets of tweet data (Jet2, 393 tweets; Royal Caribbean, 244 tweets), graded by 3 humans.

Jet2 data:

- 103 disagreements out of 1179 ratings (~9%): so human grading is quite robust
- 18% positive, and 30% negative
- Dictionary classifier scored ~53% (with no retraining)
- Basic NN classifier scored ~57% (training on ¾, testing on ¼)

Royal Caribbean data:

- 68 disagreements out of 732 ratings (~9%)
- 36% positive, and 22% negative. Many tweets anticipated the cruise.
- Dictionary classifier scored ~52% (with no retraining)
- Basic NN classifier scored ~49% (training on ¾, testing on ¼)

Extension Data: Testing RNN

Two new sets of tweet data (Jet2, 393 tweets; Royal Caribbean, 244 tweets), graded by 3 humans.

Jet2 data:

- LSTM RNN classifier scored ~ 87.7%

Royal Caribbean data:

- LSTM RNN classifier scored ~ 86.5%

Next Steps

Choosing a Model: Business Factors

- Accuracy
- Training/Testing time
- Cost (including staff)
- Adaptability
- Relevant to business goals?
- Technical Constraints - can we store/process the data fast enough?
- Scalability/Reusability
- Governance
- Stakeholder Satisfaction
- Easy interpretation by non-technical users

Business Next Steps

- Next steps for airline(s)
  - Data from beyond Feb ‘15 - do opinions change over time?
  - Look at competitors in more detail - are we better at something than our competitors (and can we market that?)
- Improve Topic Detection (LSA) to provide better feedback on what causes the most customer complaints.
-

Technical Next Steps

- Data from beyond Feb ‘15 - do opinions change over time?
- NN vs Non-NN
  - NN has better performance in general
  - NN requires lots of training data
  - Training required for NN but not non-NN
  - 2 sides
- Branch out into different industries (cruise lines, train companies, etc.) to improve accuracy and applicability to more problems
- Broaden range of topics considered in sentiment decision-making

Lessons Learned

- More loosely-defined problem than last time requires more out-of-the-box thinking
- Importance of using other datasets to validate findings
- NLP is technically challenging
- Improved teamwork
- Made use of shared code and data

Any Questions?

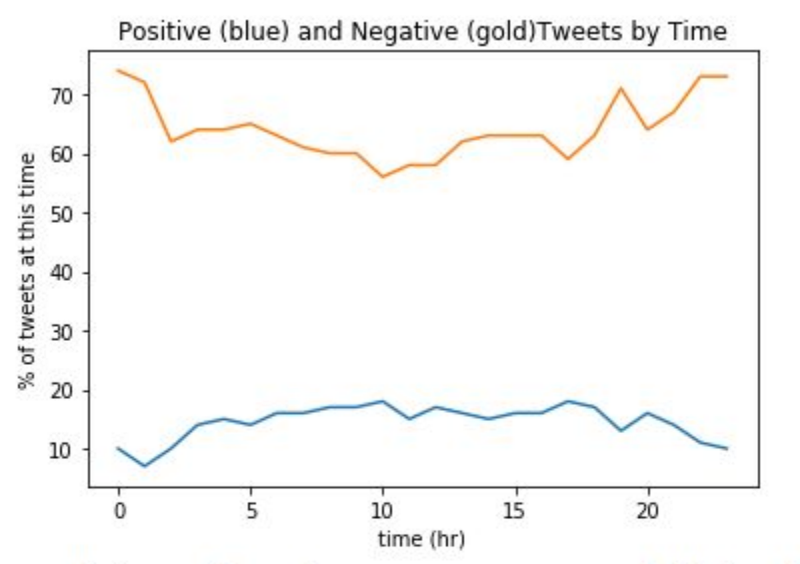
Convolutional Neural Network with Data Augmentation

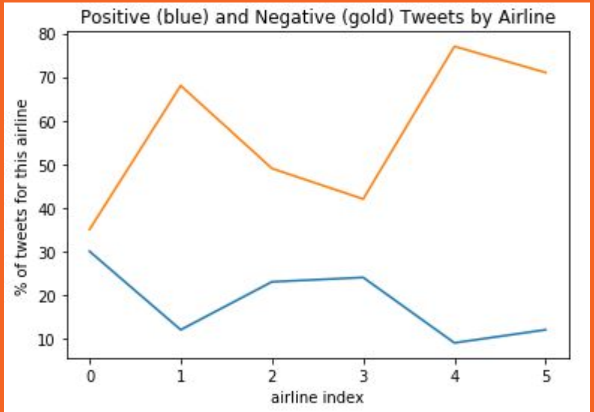
Accuracy: ~99.7%
Training Time: 6120s
Prediction Time: 1.61s

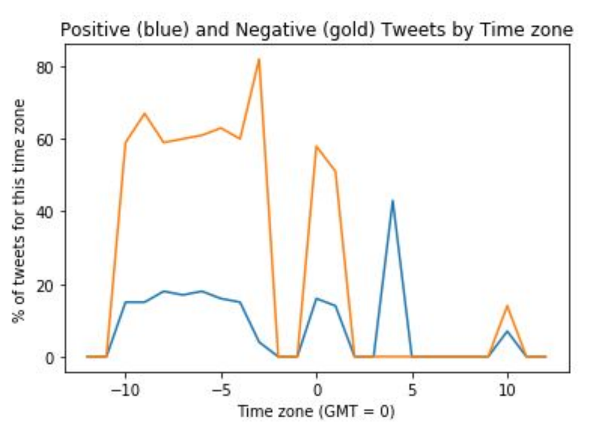
		Actual									
		0	1	2	3	4	5	6	7	8	9
Predicted	0	100	0	0	0	0	0	0	0	0	0
	1	0	100	0	0	0	0	0	0	0	0
	2	0	0	100	0	0	0	0	0	0	0
	3	0	0	0	100	0	0	0	0	0	0
	4	0	0	0	0	100	0	0	0	0	1
	5	0	0	0	0	0	100	0	0	0	0
	6	0	0	0	0	0	0	100	0	0	0
	7	0	0	0	0	0	0	0	100	0	0
	8	0	0	0	0	0	0	0	0	100	0
	9	0	0	0	0	0	0	0	0	0	99

General Structure

- Exploratory analysis
- Initial Insights
- Basic Model
- Improving our model
  - Preprocessing
- Applying model to other data (our USP!)
- Next Steps
  - Technical
  - Business
- Lessons Learned 🙄


Positive (blue) and Negative (gold)Tweets by Time
% of tweets at this time
time (hr)
0
5
10
15
20
10
20
30
40
50
60
70


Positive (blue) and Negative (gold) Tweets by Airline
% of tweets for this airline
airline index
0
1
2
3
4
5
10
20
30
40
50
60
70
80


Positive (blue) and Negative (gold) Tweets by Time zone
% of tweets for this time zone
Time zone (GMT = 0)
−10
−5
0
5
10
0
20
40
60
80

Basic Analysis of Airline Tweets

1) Many more tweets during the day (~ 1000 at 9am, vs ~100 at 1am)

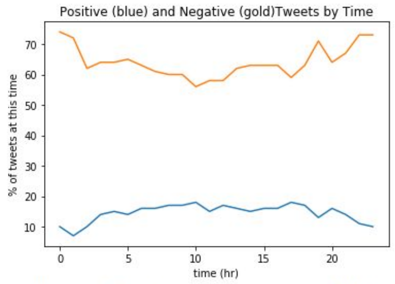
2) Most tweets to airlines are negative (~2/3 vs 1/6 positive)

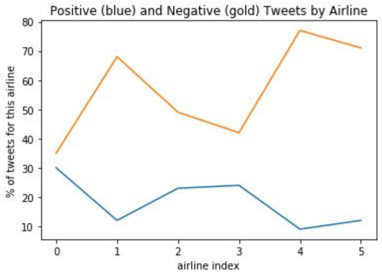
3) They are proportionally more negative at night.

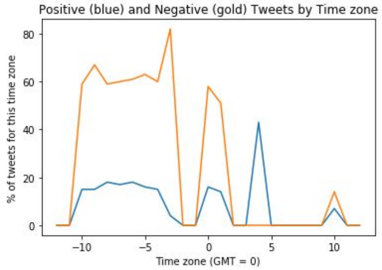
4) Some airlines cop it worse than others:

- Virgin America few tweets, but most positive (30%) and fewest negative (35%):
- Southwest and Delta ~25% positive and ~50% negative,
- United, US Airways and American ~10% positive and ~70% negative.

5) Time zone analysis: some surprising blips (artifacts?). Europeans seem a little less negative, and Australians are so laid back they don't give a 4X.





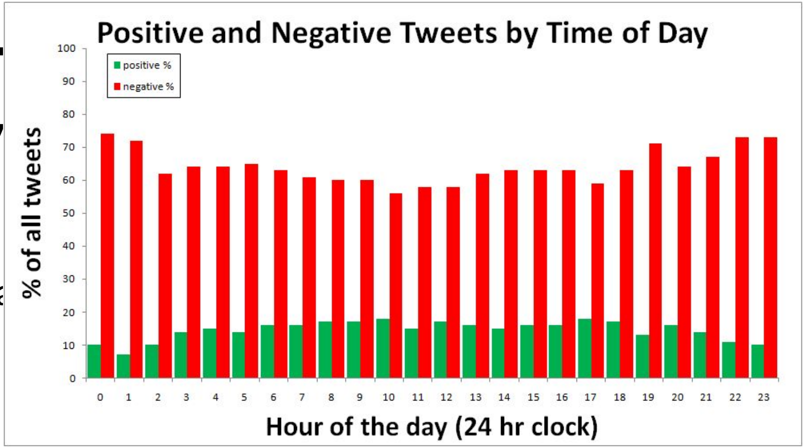


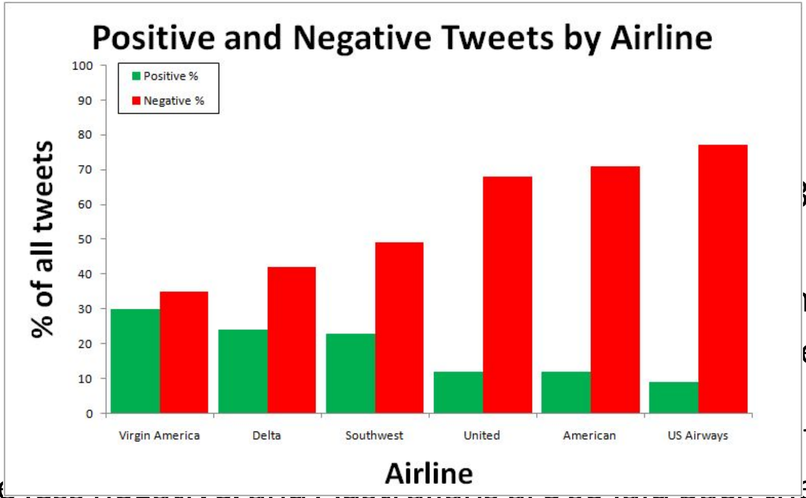
Basic Analysis of Airline Tw

1) Many more tweets during the day (~ 1000 at 9am, vs ~100 a

2) Most tweets to airlines are negative (~2/3 vs 1/6 positive)

3)



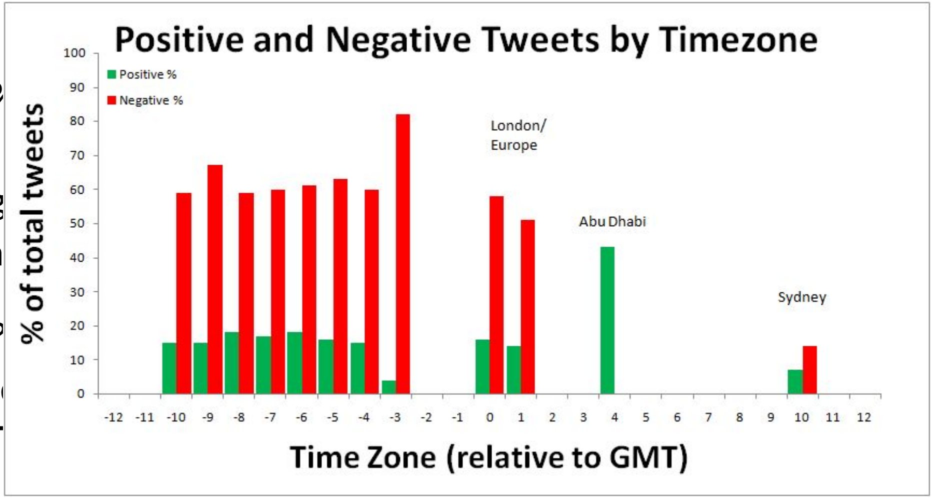


4)

- 30%
- neg
- e a

5) ts?

littl ey (



David's Dictionary and Keys

2) Gave a dictionary of ~8200 words from the given tweets, with their positive and negative associations.

3) Filtered to get a "key" set which

    (a) had at least 4 uses and

    (b) had negative or positive associations "significantly" different from the overall set.

    Current set has ~1800 members.

David’s Dictionary Classifier

1) Looks at text of tweet:

    - Find “key” words, and
    - Calculate a weighted sentiment score, and
    - Use that to classify the tweet.

    Gives about 73% accuracy on the main set.

2) Generates:

    - Vectors (~1800 dimensions, one for each key word), and
    - “true” sentiment labels,

    to allow training of a NN

David’s Simple Neural Net Classifier

1) Feed the vectors and labels constructed above to:

2) A simple 3 layer neural net constructed in Tensorflow:

- Input layer determined dynamically by vector size
- Hidden layer of 512 neurons
- Output layer of 3 neurons, corresponding to positive, neutral and negative sentiment

3) Train with ¾ of the data for 10 epochs, and test on remaining ¼.

Took ~ 1 minute to train. Achieved 80% accuracy.

Naive Model: Naive Bayes

Accuracy: ~78.2%

Training time: ~0.188s

Prediction time: ~0.043s

Naive Model: Logistic Regression

Accuracy: ~67.2%

Training time: ~3.4s

Prediction time: ~0.12s