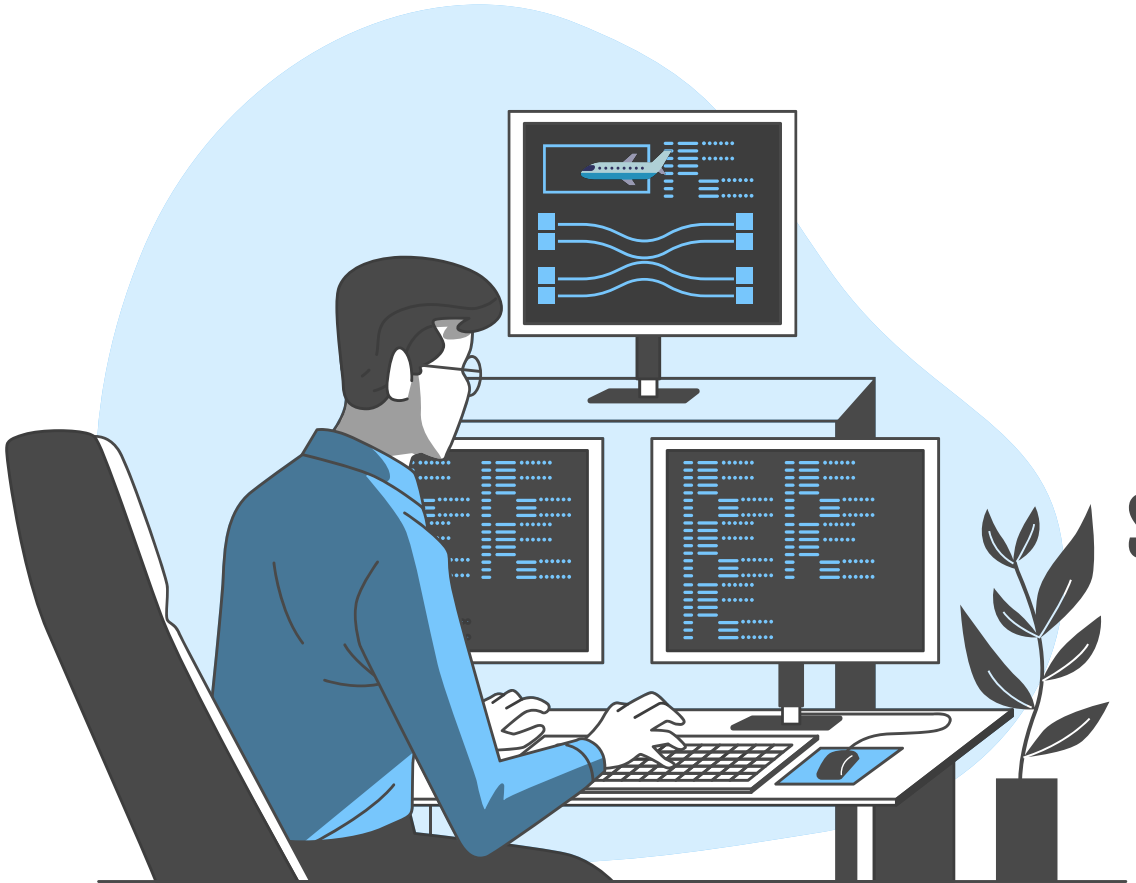


Airline tweets sentiment analysis

Wirach (Joe) Leelakiatiwong

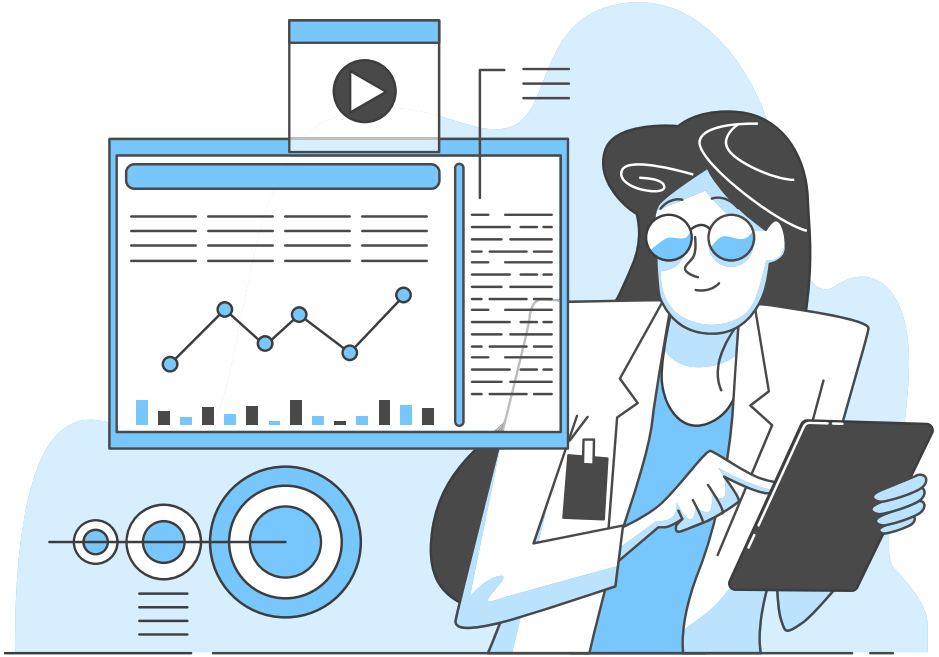
Table of Contents

01 **The Problem**
Why twitter? Why sentiment analysis?

02 **The Data**
What data do we have?

03 **The approach**
How can we flag negative sentiment?

04 **The Conclusion**
Model performance and further improvements



A little bit of background

Twitter

Twitter is a **'microblogging**' system that allows you to **send** and **receive** short posts called **tweets**. Tweets can be up to **280 characters long** and can include links to relevant websites and resources.

https://esrc.ukri.org/research/impact-toolkit/social-media/twitter/what-is-twitter/

BBC Breaking News
@BBCBreaking

France and Germany join the US in advising their nationals in Libya immediately bbc.in/1rVmrDJ

RETWEETS 596 FAVORITES 223

US Airways
@USAirways

@ellerafter We welcome feedback, Elle. If travel is complete, you can detail it here fo review and follow-up: pic.twitter.com/vbeYgCuG25

Reply Retweet Favorite More

Donald J. Trump
@realDonaldTrump

Follow

Sorry losers and haters, but my I.Q. is one of the highest -and you all know it! Please don't feel so

Jim B.
@thetruefailure

Follow

Hey @Delta, you suck a lot and I hate having to fly your terrible airline.

7:37 AM - 13 Feb 2019

2 Likes

2 2

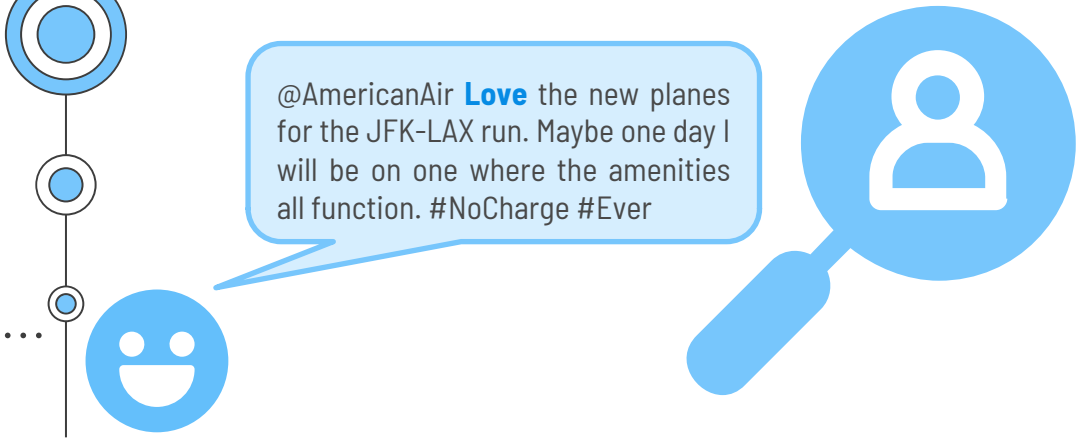
Delta @Delta · Feb 13

Replying to @thetruefailure

That's not good to hear, Jim. Pls follow/DM if I may be of assistance. *AAB

Send a private message

1



Sentiment Analysis

Sentiment analysis is a natural language processing technique used to **determine whether data is positive, negative or neutral** to help businesses monitor brand and product sentiment in customer feedback, and **understand customer needs.**

https://monkeylearn.com/sentiment-analysis/

01

The Problem

"Although airline industry — one of the world most competitive market — have been employing multiple feedback channels to collect customer feedback, they are all arduous. Thus making a social media analysis a more compelling approach. This project will identify **negative sentiment tweets** about the airline to categorize **them into subcategories and direct them to related departments to solve problems** and develop improvement strategies in the future."

02

The Data

Data: Twitter US Airline Sentiment dataset (Kaggle)

Coverage

Data was collected from 16th to 24th February of 2015

...

Airlines

Covered 6 major airlines in the USA (United, US Airways, American Airlines, Southwest, Delta, Virgin)

...

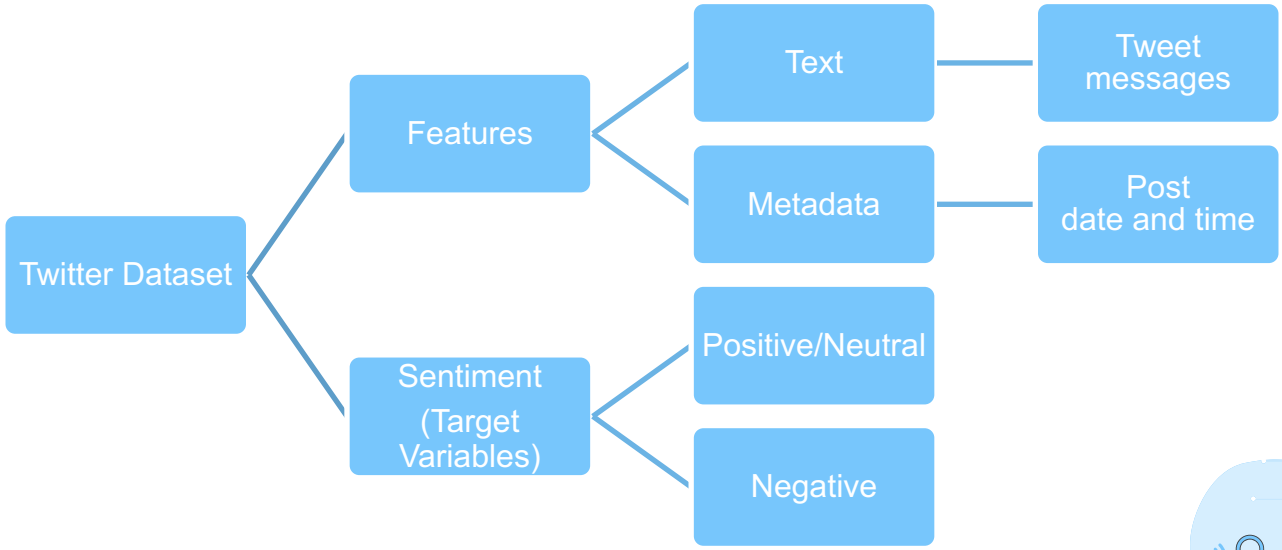
Size

Rows : 14640
Columns : 15

...

	tweet_id	airline_sentiment	airline	text	tweet_created
0	570306133677760513	neutral	Virgin America	@VirginAmerica What @dhepburn said.	2015-02-24 11:35:52-08:00
1	570301130888122368	positive	Virgin America	@VirginAmerica plus you've added commercials t...	2015-02-24 11:15:59-08:00
2	570301083672813571	neutral	Virgin America	@VirginAmerica I didn't today... Must mean I n...	2015-02-24 11:15:48-08:00
3	570301031407624196	negative	Virgin America	@VirginAmerica it's really aggressive to blast...	2015-02-24 11:15:36-08:00
4	570300817074462722	negative	Virgin America	@VirginAmerica and it's a really big bad thing...	2015-02-24 11:14:45-08:00
...	...	...	...	...	...
14635	569587686496825344	positive	American	@AmericanAir thank you we got on a different f...	2015-02-22 12:01:01-08:00
14636	569587371693355008	negative	American	@AmericanAir leaving over 20 minutes Late Flig...	2015-02-22 11:59:46-08:00
14637	569587242672398336	neutral	American	@AmericanAir Please bring American Airlines to...	2015-02-22 11:59:15-08:00
14638	569587188687634433	negative	American	@AmericanAir you have my money, you change my ...	2015-02-22 11:59:02-08:00
14639	569587140490866689	neutral	American	@AmericanAir we have 8 ppl so we need 2 know h...	2015-02-22 11:58:51-08:00

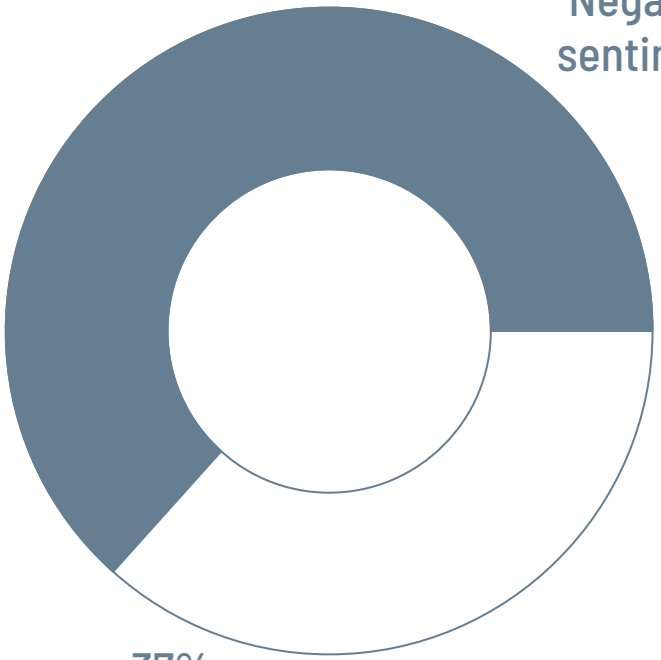
Data Dictionary (in the scope of project)



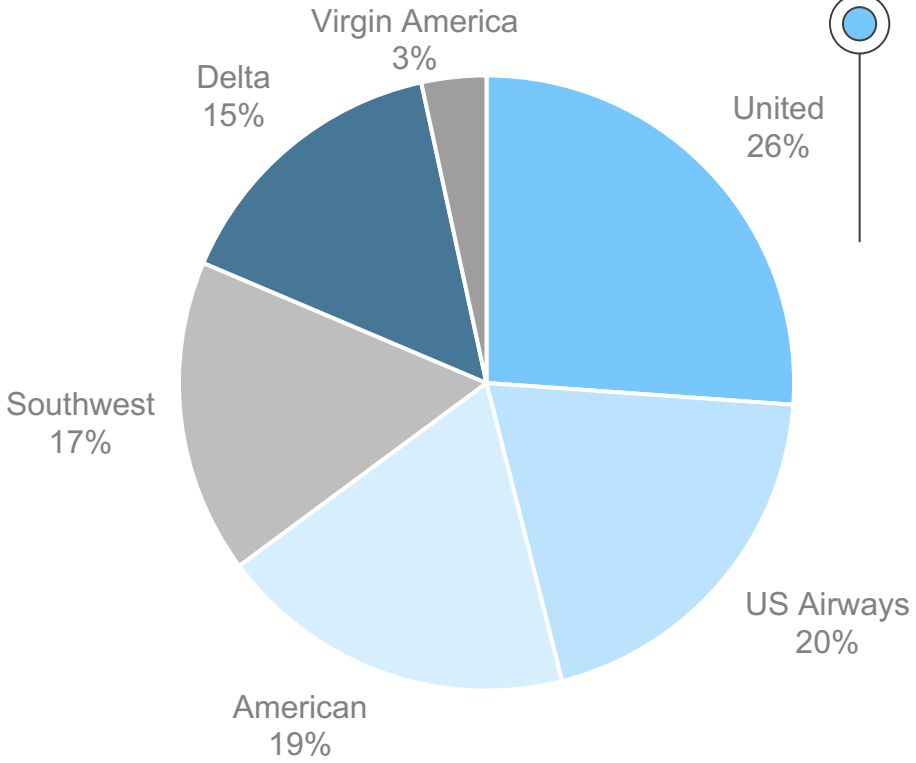
Exploratory Data Analysis

Data Overview

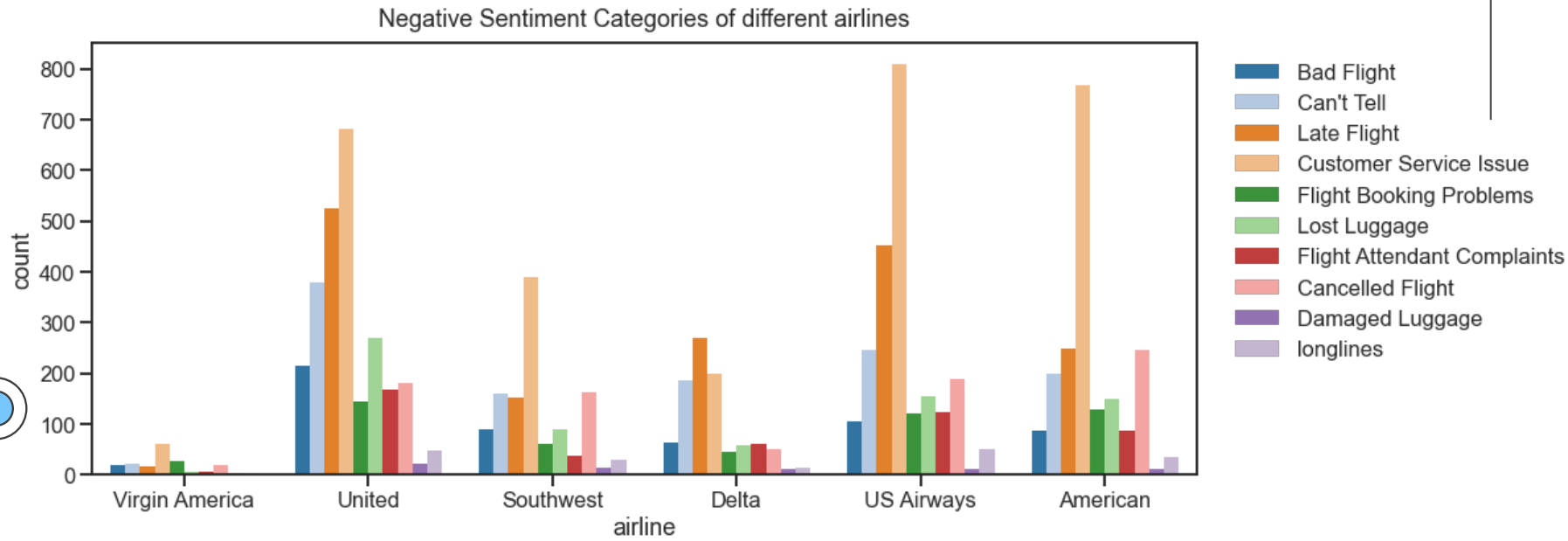
63 %
Negative
sentiment



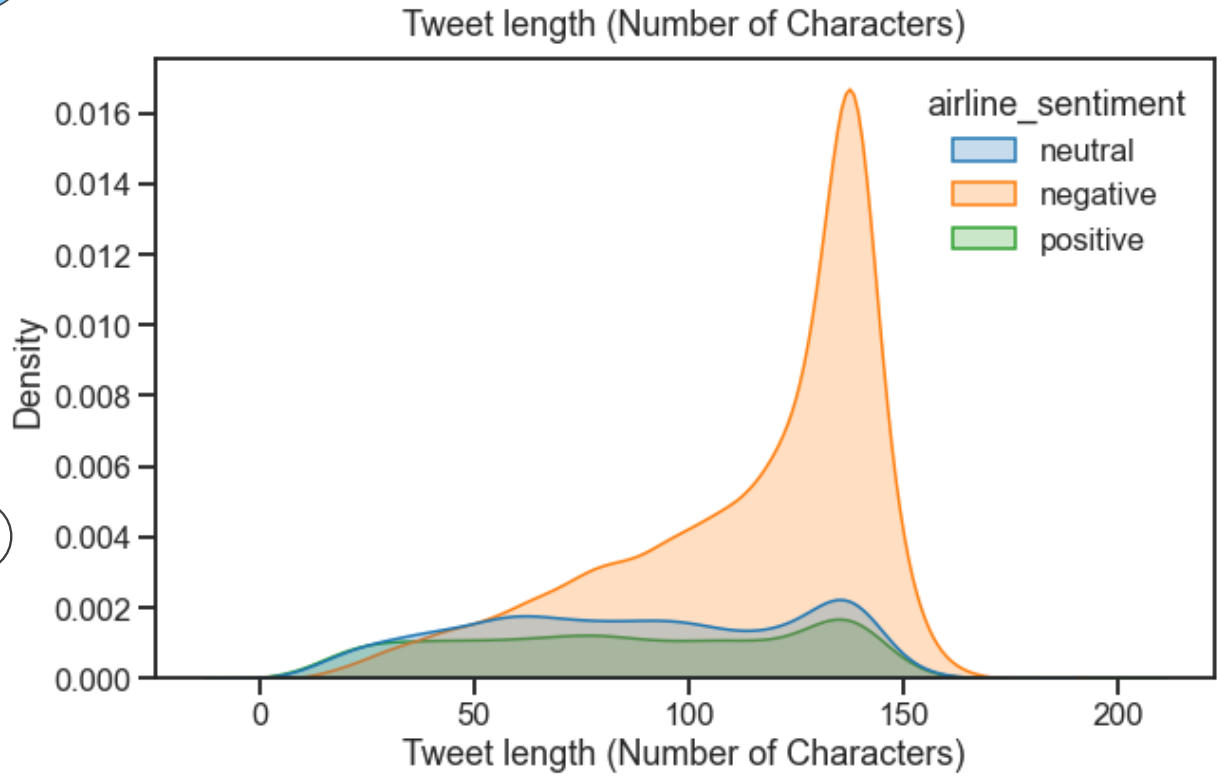
37%
Positive / Neutral
Sentiment

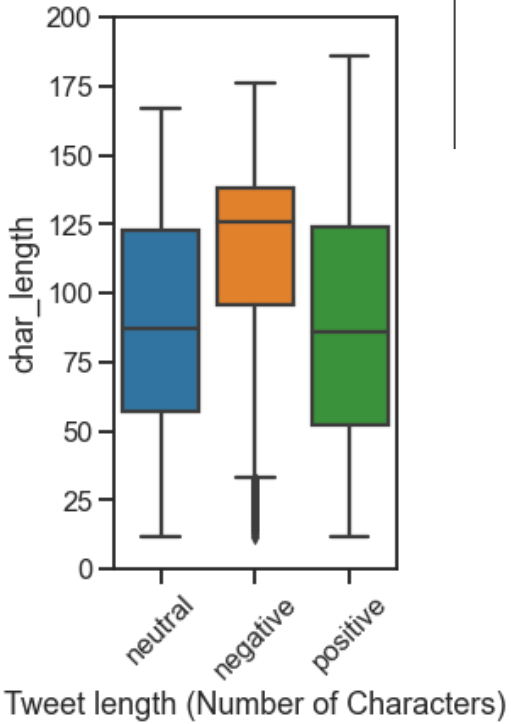


Airlines Overview

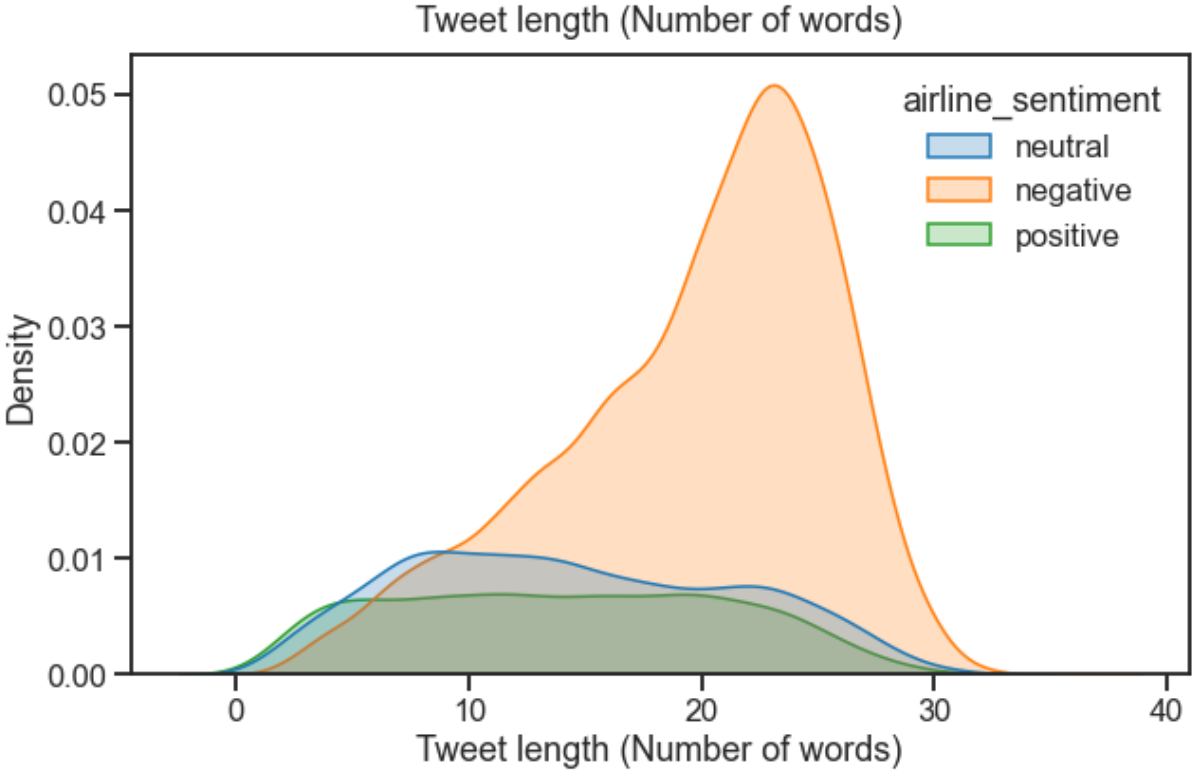


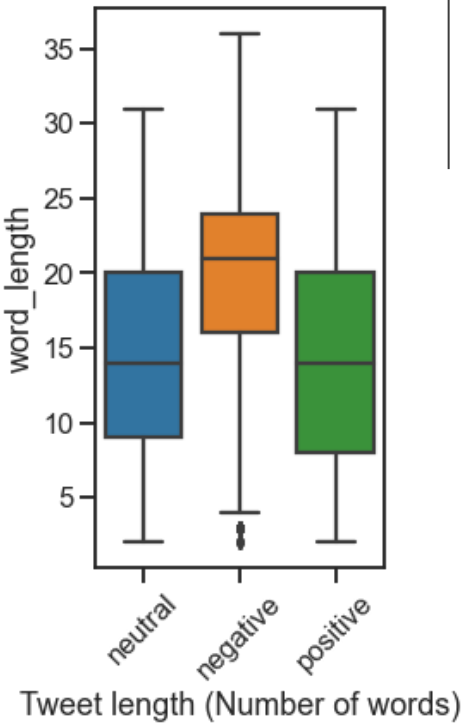
Tweet length (#characters)



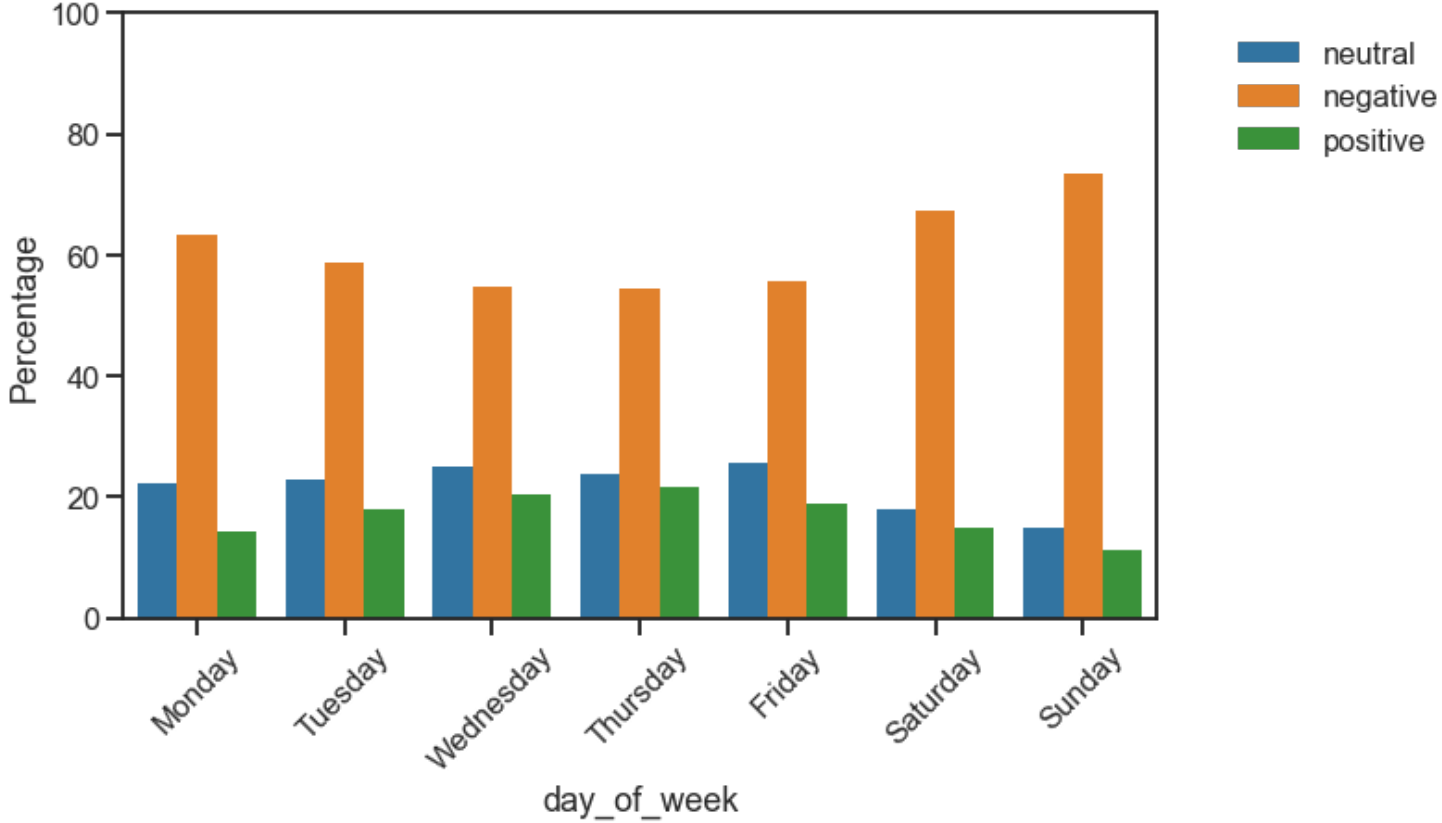


Tweet length (#words)

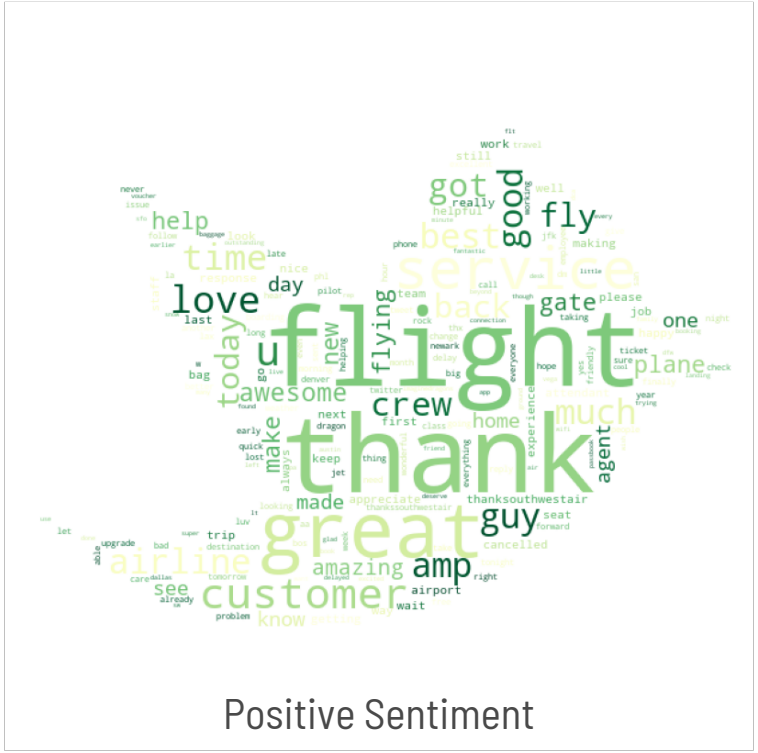




Days of week


100
80
60
40
20
0
Percentage
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
day_of_week
neutral
negative
positive

Negative vs Positive Sentiment words



Positive Sentiment

Negative Sentiment

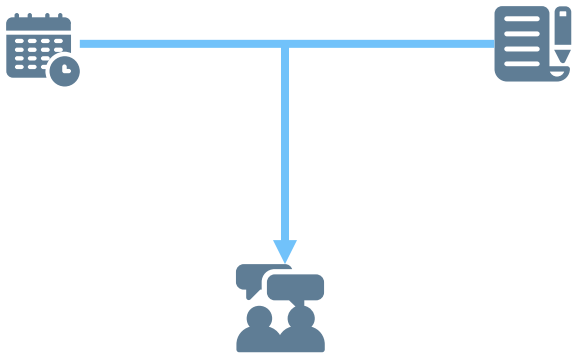
03

The Approach

Algorithm outline

Metamodel

"Take metadata such as day of week, tweet length as input"



Textual model

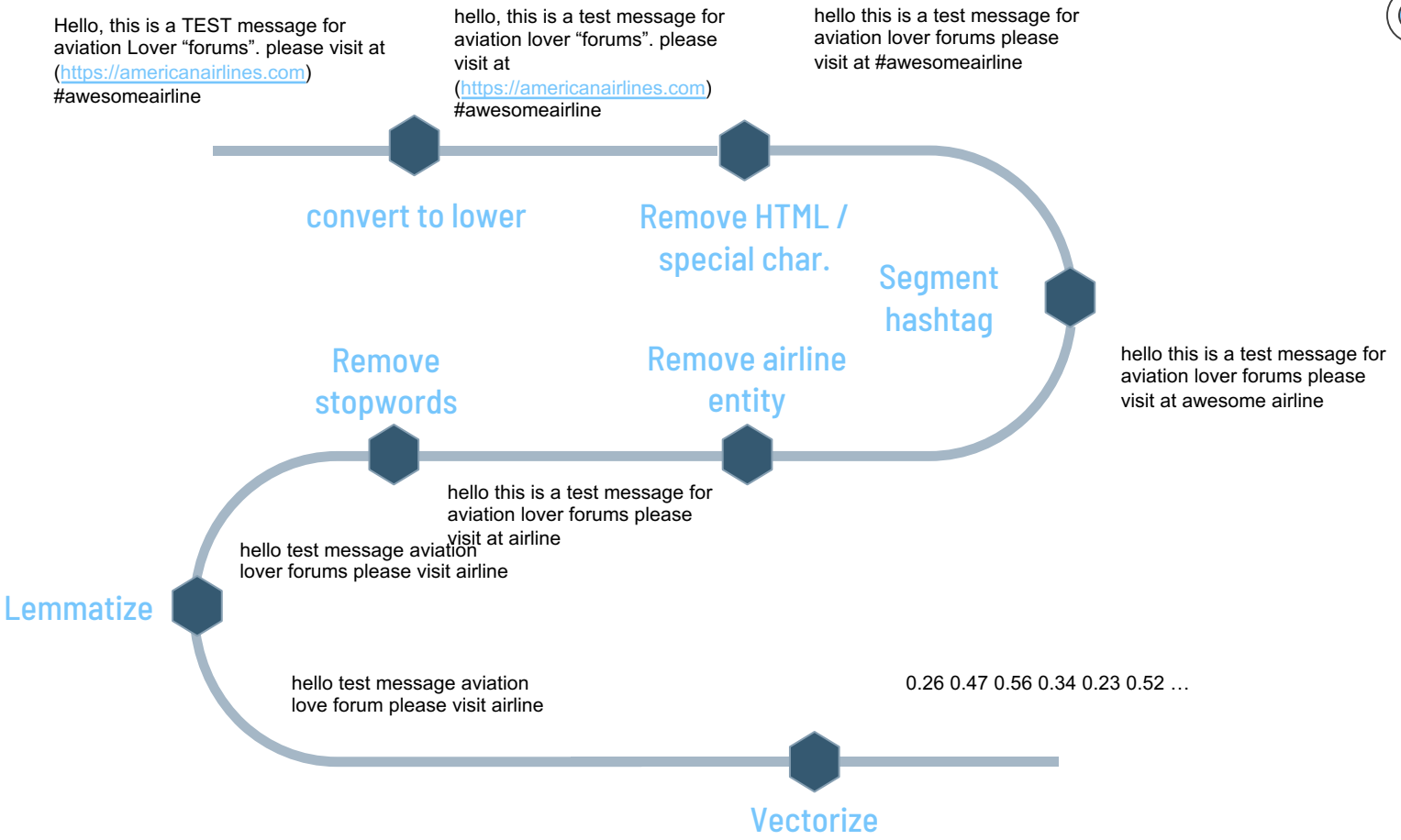
"Take processed tweet as input"

Combined model

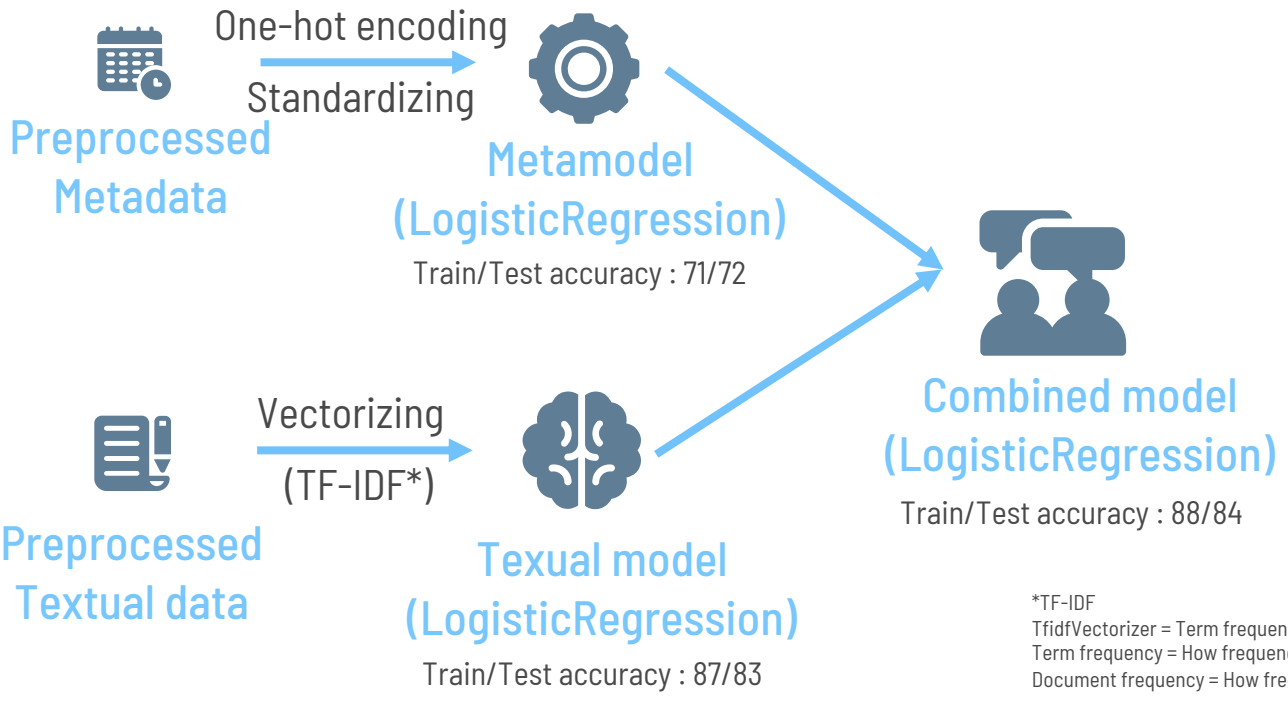
"Combine prediction from both model and predict final probability"

*Baseline Accuracy = 62.7%

Text Data Preprocessing



The Result



*TF-IDF
TfidfVectorizer = Term frequency (TF) x Inverse of Document Frequency (IDF)
Term frequency = How frequence of the word occured in that document
Document frequency = How frequence of the document containing that word

Words that occur often in one document but don't occur in many documents contain more predictive power (will get high vectorized value)

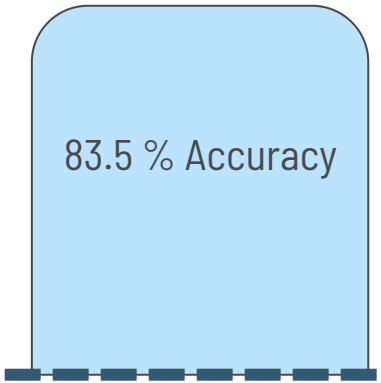
04

The Conclusion

Performance Summary

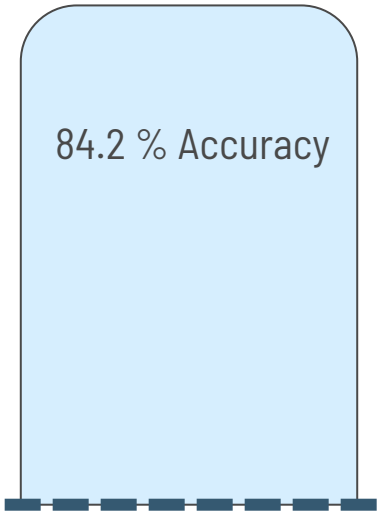
Textual model

Learned from tweet messages



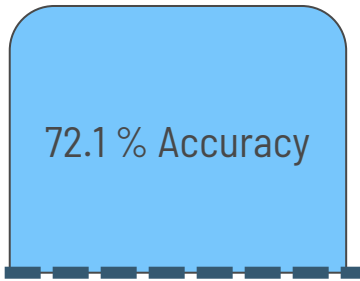
Combined model

Combine prediction from both models



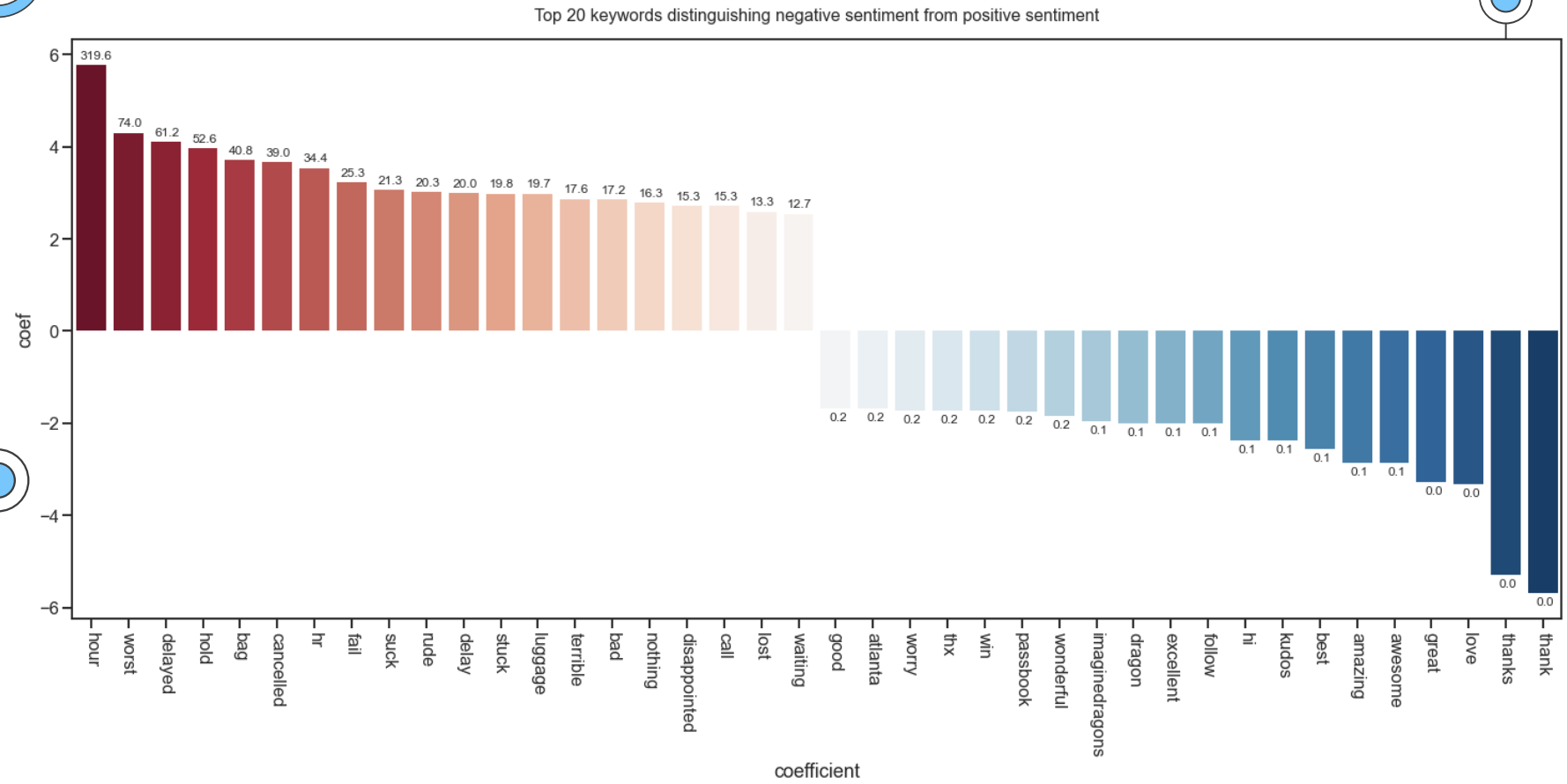
Metamodel

Learned from metadata



Baseline accuracy
62.7%

What can be infered from the model?



Prediction and Limitation

Customer services was horrible

88.2% probability of being negative sentiment

The landing was perfect!

20.3% probability of being negative sentiment

Although the departure was delayed, at least we are able to arrive on time

77.0% probability of being negative sentiment

Conclusion

01 How can it be used?

Hooked with the data pipeline to filter the negative sentiment

02 Limitation

A bag of words model may not effectively predict complex sentences

03 Further improvement (The data)

- Gathering data from extended time span
- Acquiring more data to train deep learning model

04 Further improvement (The model)

- Train more complex model (require massive dataset)
- Build classifier to subcategorize negative feedbacks

Extra

Topic modeling

Without topic modeling

Customer services

Too many things flow to my desk!

Lost and found

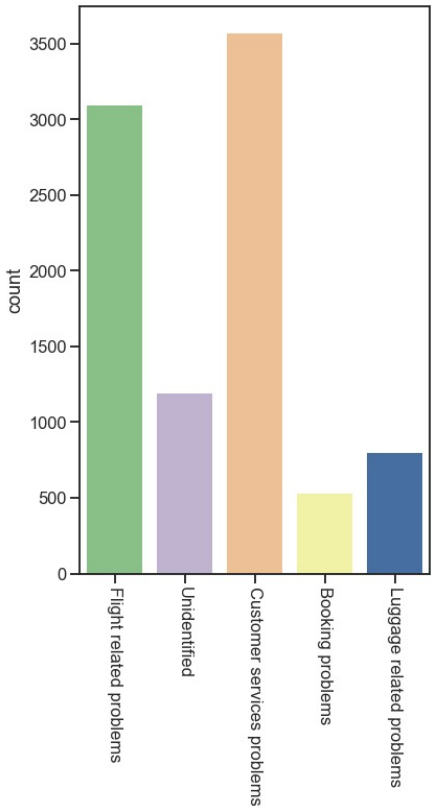
Customer complaint never reach me!

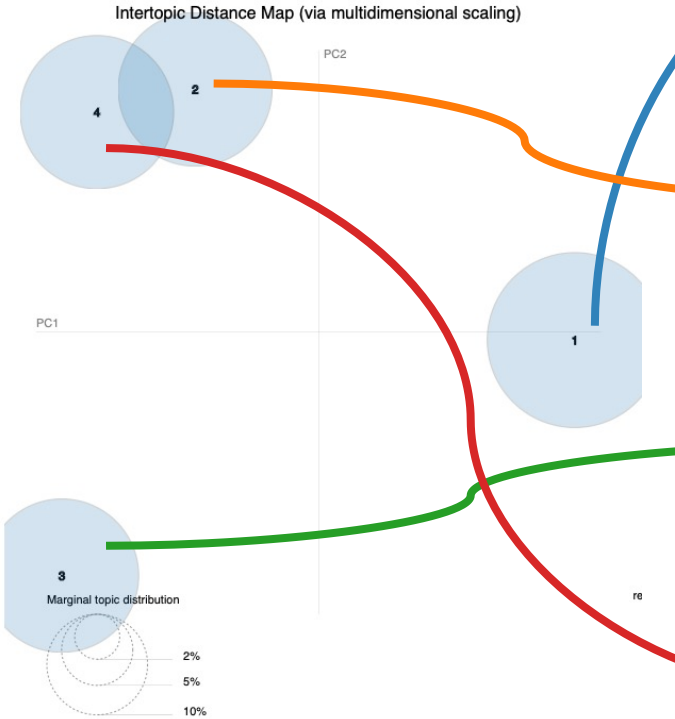
Ticketing

What can I do with the lost and found problem?

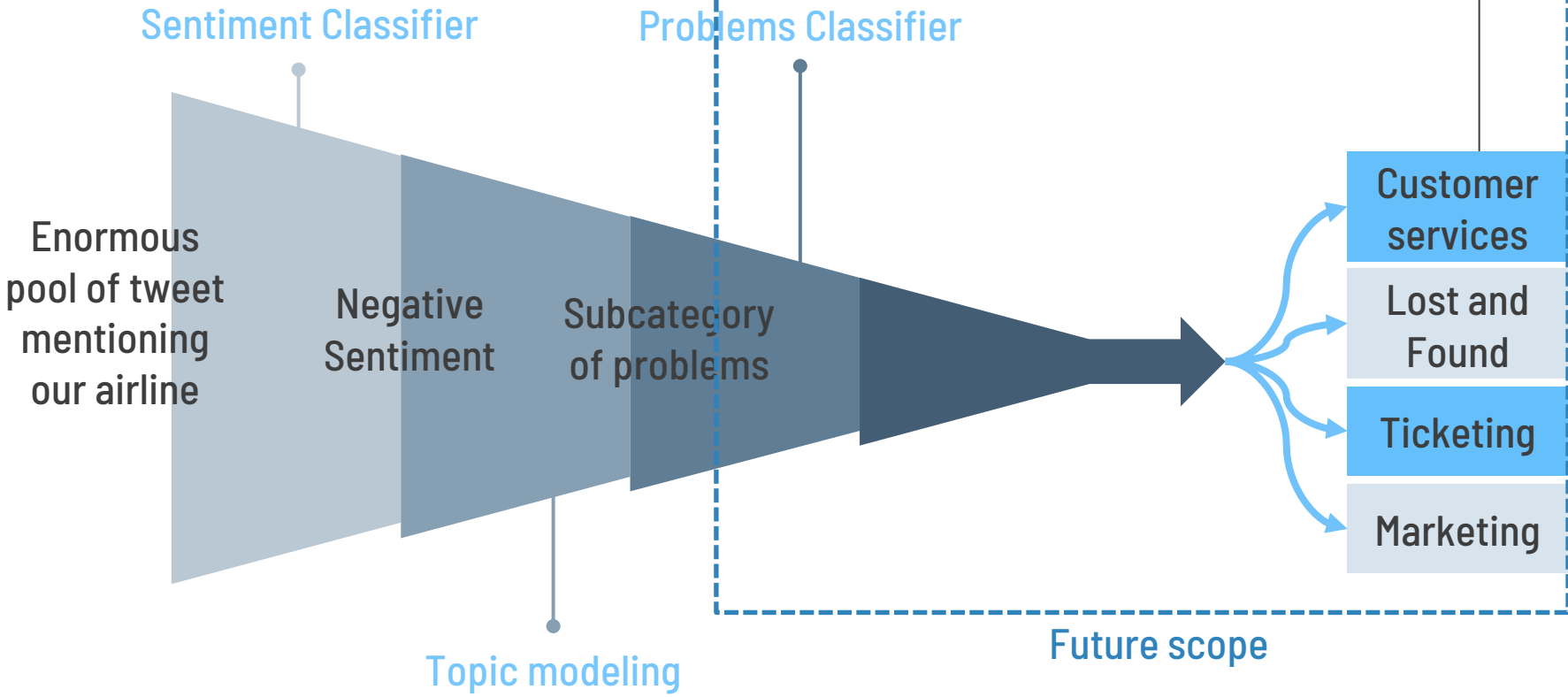
Ground truth and topic modeling result

Ground truth





What can topic modeling do?



Thanks!

Do you have any questions?

Wirach.lee@gmail.com
+66 81 453 6326