US Airline sentiment analysis using Twitter data

Yaya Liu

Agenda

1. Background and project objectives
2. Data
    - Data description and visualizations
    - Data related risks
    - Data balancing
    - Data pre-processing
3. Sentiment analysis framework
4. Multi-classification model evaluation
5. Machine learning models
6. Findings and feature work

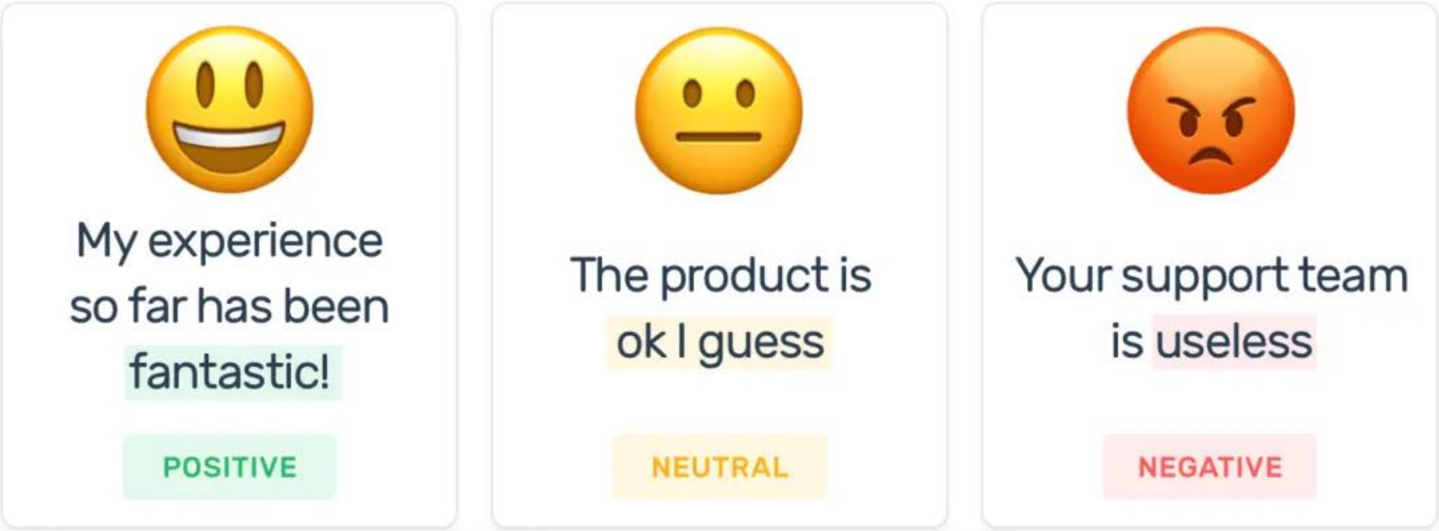
Background

- As of 2019, 3.5 billion people actively using social platforms. 500 million Tweets are tweeted every day. People share their genuine emotions, feelings, opinions and experiences on social media.
- More and more companies start utilizing online data to improve customer service, enhance business competitiveness, perform crisis management. One of the popular text analysis techniques is sentiment analysis.
- Sentiment analysis started in early 2000s. Multiple approaches have been developed and a lot of research has been done in various fields afterwards. But only a few studies directly focused on the area of airlines based on Twitter data

Project objectives

- The main goals of this project are to perform sentiment analysis in the area of US airline service using Twitter data and explore techniques that are related to sentiment analysis.

- Sentiment analysis: inspecting the given Tweet and determining a user's attitude as positive, negative, or neutral.



My experience so far has been fantastic!

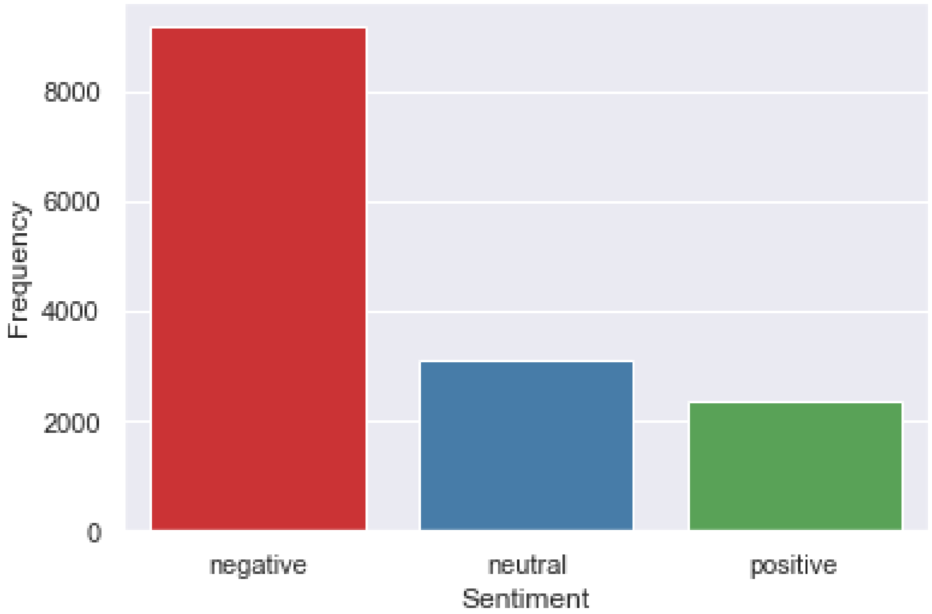
The product is ok I guess

Your support team is useless

Data

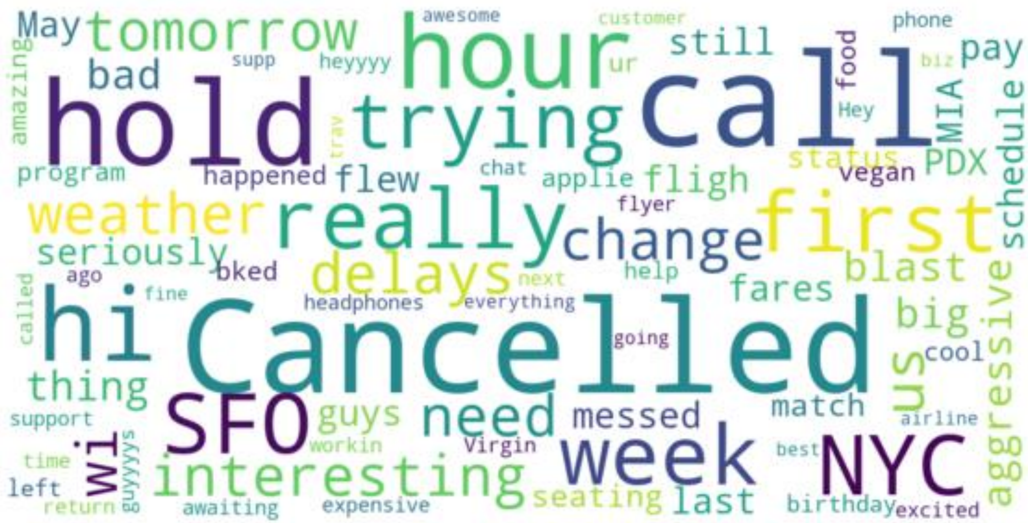
Data Source & Description

- Data is from Kaggle.
- Data includes 14, 640 Tweets covering six U.S. airline companies: American, Delta, Southwest, United, US Airways, Virgin America.
- Each Tweet has already been labeled as “negative”, “neutral” or “positive” class.
- Class distribution is skewed towards negative Tweets.

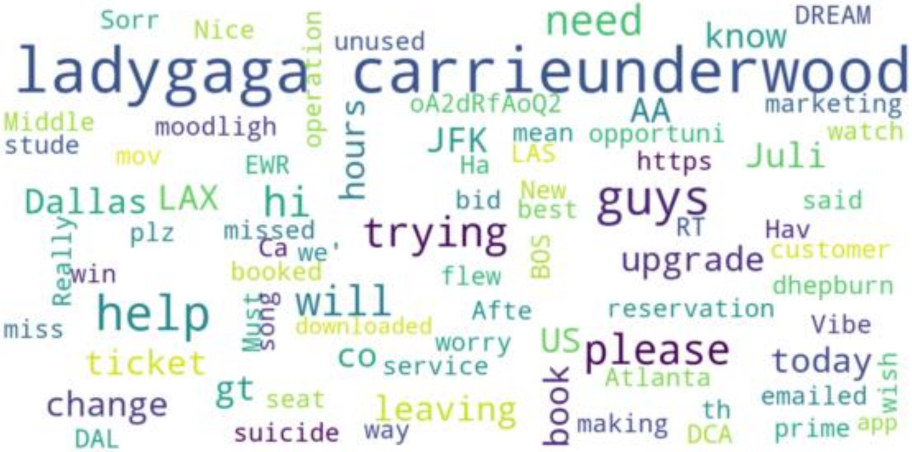


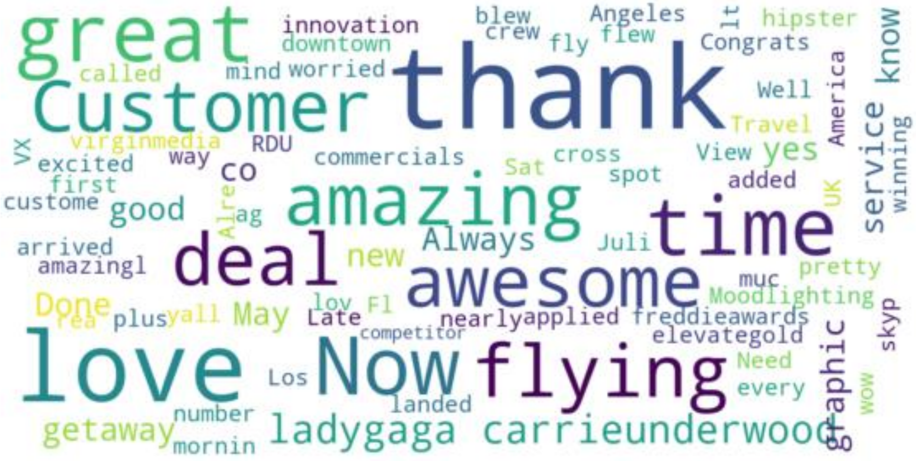
- Negative: 9, 178
- Neutral: 3, 099
- Positive: 2, 363

Data Visualization



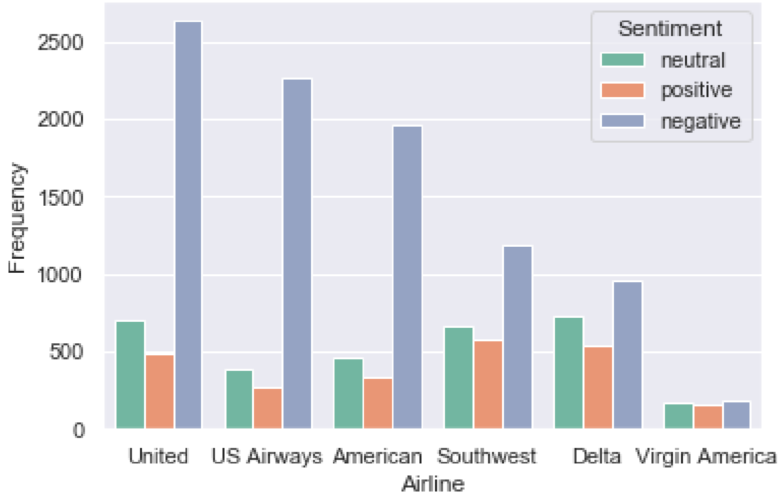
- Negative Tweets



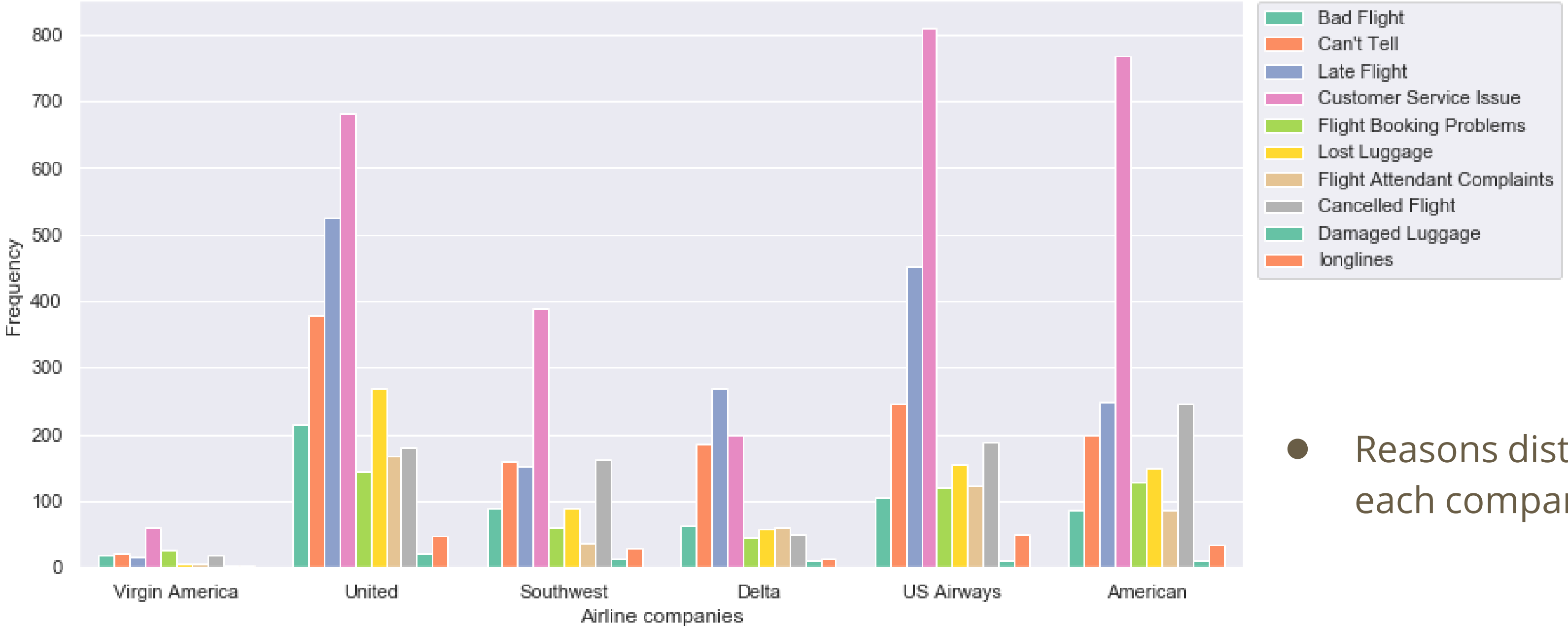


- Neutral Tweets
- Positive Tweets

Data Visualization

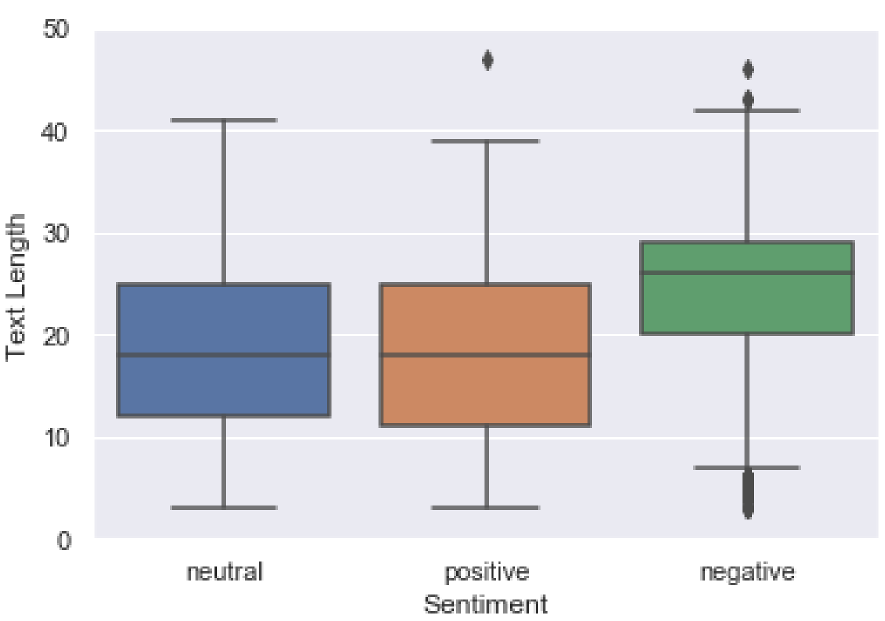


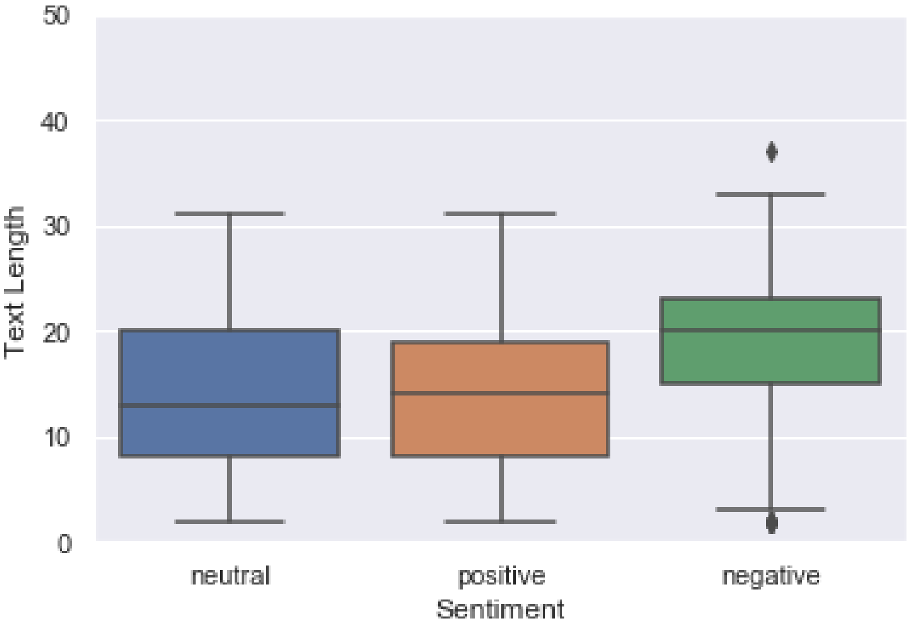
- Number of Tweets on each company



- Reasons distribution on each company

Data Visualization





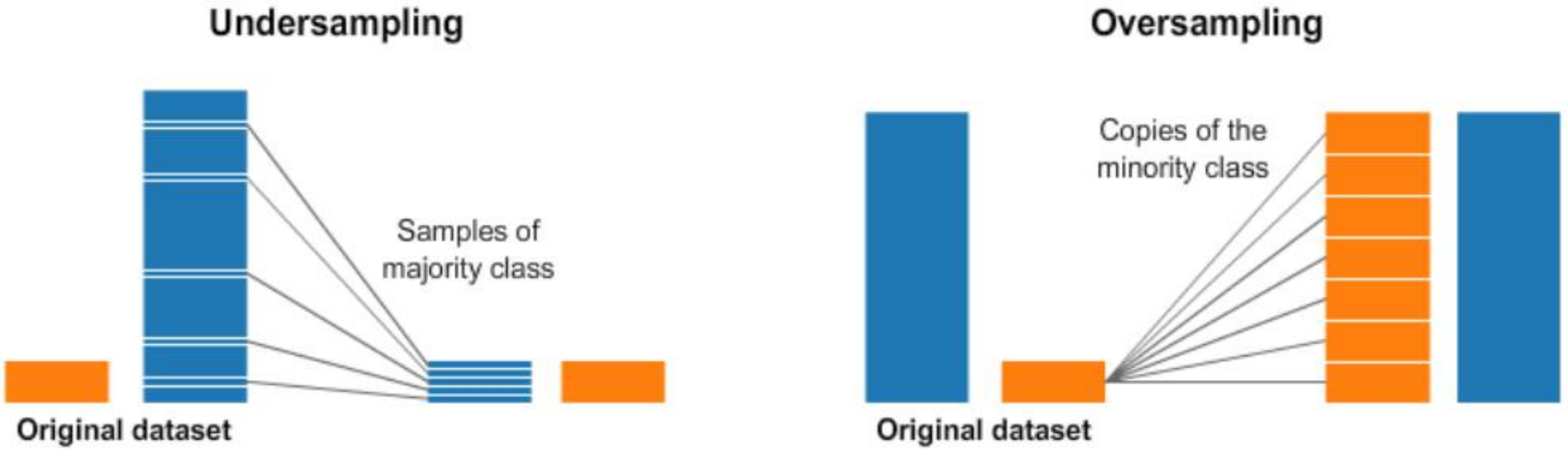
- Number of words and punctuations in each class (before text pre-processing)
- Number of words in each class (after text pre-processing)

Data related risks

- The service-related data from Twitter is often imbalanced. There are more negative Tweets than neutral and positive Tweets.
- Most of the Tweets are very short. They may not include enough context to identify the users’ attitudes.
- Tweets have a lot of “noise” comparing to published articles, such as emojis, external links, user mention, hashtags, etc. But since most of the Tweets are very short, do we need to remove all of them?

Data Balancing

- under-sampling: removing samples from the majority class.
- over-sampling: adding more examples from the minority class.



risk: information loss

risk: overfitting

Referenced from https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

Data pre-processing

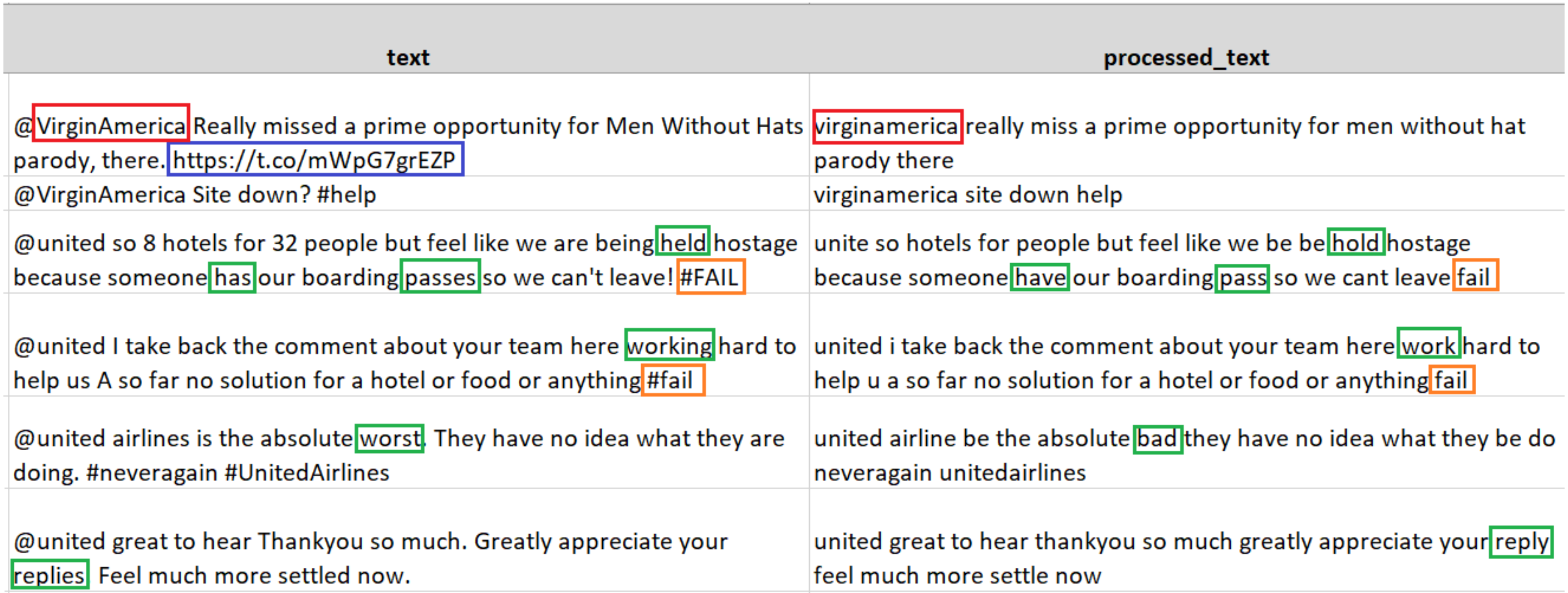
- Decapitalized all characters to reduce the number of word types.
- Removed special characters, emojis, numeric digits, web links and punctuations.
- Lemmatization with WordNet lexical database based on POS tags.

-> replace a word with its base form, which is known as lemma.

Why not stemming?

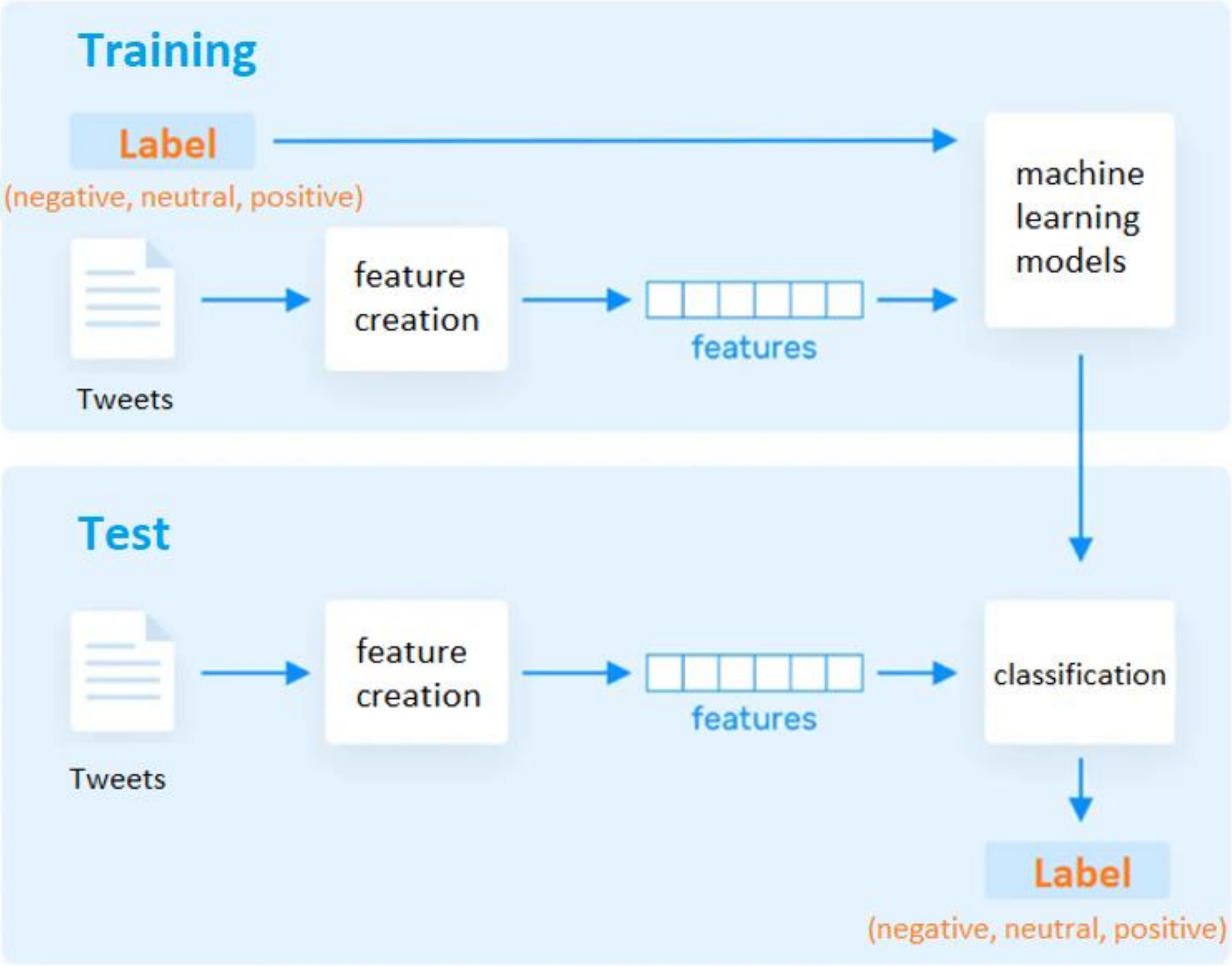
	really missed a prime opportunity for men without hats parody
Lemmatization	really miss a prime opportunity for men without hat parody
stemming	realli miss a prime opportun for men without hat parodi

Data pre-processing

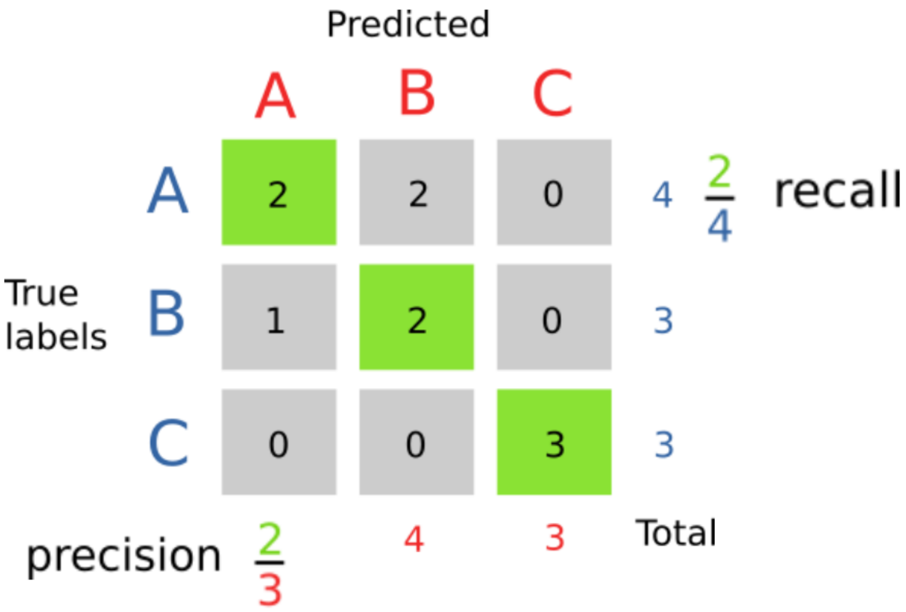


text	processed_text
@VirginAmerica Really missed a prime opportunity for Men Without Hats parody, there. https://t.co/mWpG7grEZP	virginamerica really miss a prime opportunity for men without hat parody there
@VirginAmerica Site down? #help	virginamerica site down help
@united so 8 hotels for 32 people but feel like we are being held hostage because someone has our boarding passes so we can't leave! #FAIL	unite so hotels for people but feel like we be be hold hostage because someone have our boarding pass so we cant leave fail
@united I take back the comment about your team here working hard to help us A so far no solution for a hotel or food or anything #fail	united i take back the comment about your team here work hard to help u a so far no solution for a hotel or food or anything fail
@united airlines is the absolute worst. They have no idea what they are doing. #neveragain #UnitedAirlines	united airline be the absolute bad they have no idea what they be do neveragain unitedairlines
@united great to hear Thankyou so much. Greatly appreciate your replies Feel much more settled now.	united great to hear thankyou so much greatly appreciate your reply feel much more settle now

Sentiment Analysis framework



Multi-classification model evaluation



Predicted

True labels

	A	B	C			
A	2	2	0	4	$\frac{2}{4}$	recall
B	1	2	0	3		
C	0	0	3	3		
precision	$\frac{2}{3}$	4	3	Total		

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

How to measure model's performance?

- Accuracy = Number of correct predictions / Total number of instances
- Macro F1-score

  = Average of $F1_{class1} + F1_{class2} + \cdots + F1_{classN}$

- Weighted F1-score

$$F1_{class1} * W_1 + F1_{class2} * W_2 + \cdots + F1_{classN} * W_N$$

Machine learning models

Naïve Bayes

- Naïve Bayes model finds the probabilities of classes assigned to texts by using the prior probability of the class and the likelihood, which is the probability of the text based on the class.
- In order to calculate the likelihood, Naïve Bayesian model treats each Tweet as a bag-of-words, and it assumes the position of the words doesn't matter and the features' probabilities are independent with each other given the class.

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(d \mid c) P(c)$$

$$= \underset{c \in C}{\operatorname{argmax}} \underbrace{P(x_1, x_2, \ldots, x_n \mid c)}_{\text{likelihood}} \underbrace{P(c)}_{\text{prior}}$$

Document d represented as features x1..xn

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

Naïve Bayes

- TF-IDF + Bag-of-words

	Precision (Macro Average)	Recall (Macro Average)	F1 Score (Macro Average)	Accuracy (Overall)
Original data	0.80	0.41	0.41	0.67
Oversampling data	0.71	0.71	0.71	0.78
Undersampling data	0.71	0.72	0.71	0.78

- TF-IDF + Combination of Bag-of-words, bigrams and trigrams

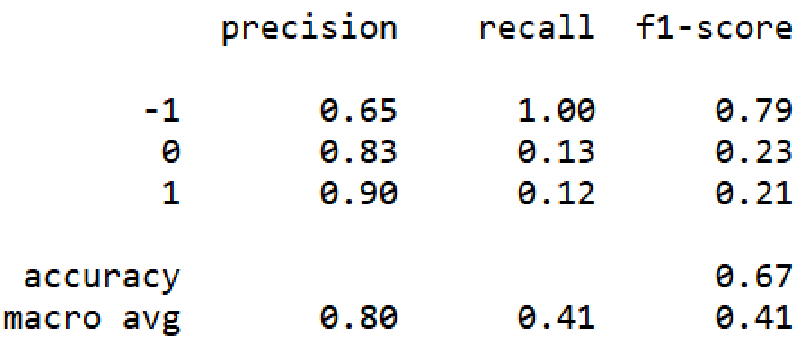
	Precision (Macro Average)	Recall (Macro Average)	F1 Score (Macro Average)	Accuracy (Overall)
Original data	0.76	0.39	0.36	0.65
Oversampling data	0.74	0.74	0.74	0.79
Undersampling data	0.72	0.71	0.71	0.78

Naïve Bayes

- Why the model has low recall and F1 Score when original data is used?

Predicted

True	-1	0	1	All
-1	1811	5	1	1817
0	541	82	5	628
1	415	12	56	483
All	2767	99	62	2928

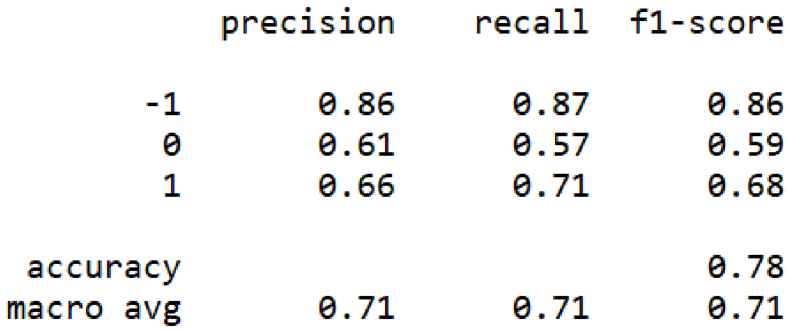


```
              precision    recall  f1-score

          -1       0.65      1.00      0.79
           0       0.83      0.13      0.23
           1       0.90      0.12      0.21

    accuracy                           0.67
   macro avg       0.80      0.41      0.41
```

Original data with TF-IDF + Bag-of-words

Predicted

True	-1	0	1	All
-1	1574	162	81	1817
0	177	357	94	628
1	77	65	341	483
All	1828	584	516	2928



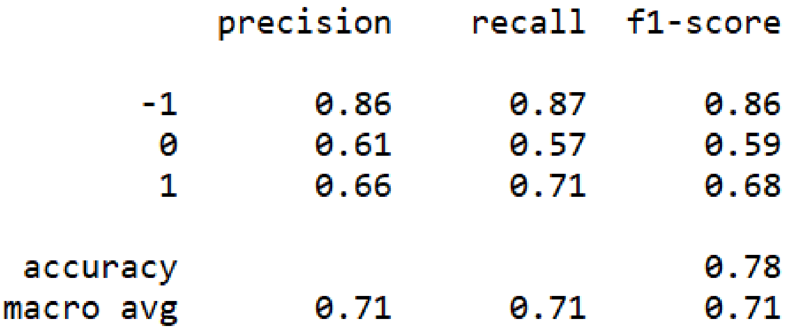
```
              precision    recall  f1-score

          -1       0.86      0.87      0.86
           0       0.61      0.57      0.59
           1       0.66      0.71      0.68

    accuracy                           0.78
   macro avg       0.71      0.71      0.71
```

Oversampling data with TF-IDF + Bag-of-words

Naïve Bayes

- Why over-sampling data with N-gram (Bag-of-words, bigrams and trigrams) outperforms oversampling data with bag-of-words?

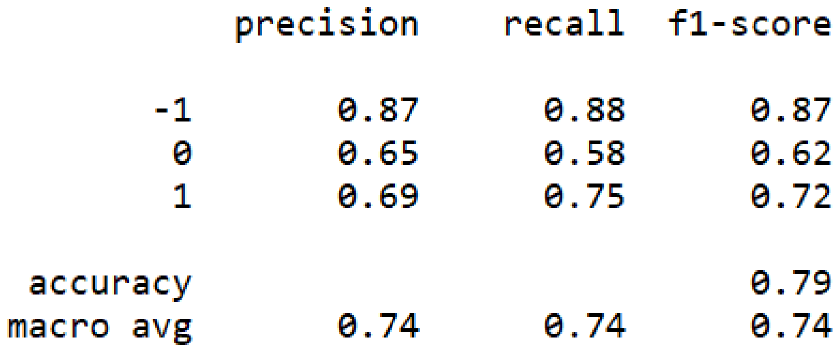
		Predicted			
		-1	0	1	All
True	-1	1574	162	81	1817
	0	177	357	94	628
	1	77	65	341	483
	All	1828	584	516	2928



```
              precision    recall  f1-score

          -1       0.86      0.87      0.86
           0       0.61      0.57      0.59
           1       0.66      0.71      0.68

    accuracy                           0.78
   macro avg       0.71      0.71      0.71
```

Over-sampling data with TF-IDF + Bag-of-words

		Predicted			
		-1	0	1	All
True	-1	1594	144	79	1817
	0	179	367	82	628
	1	65	54	364	483
	All	1838	565	525	2928

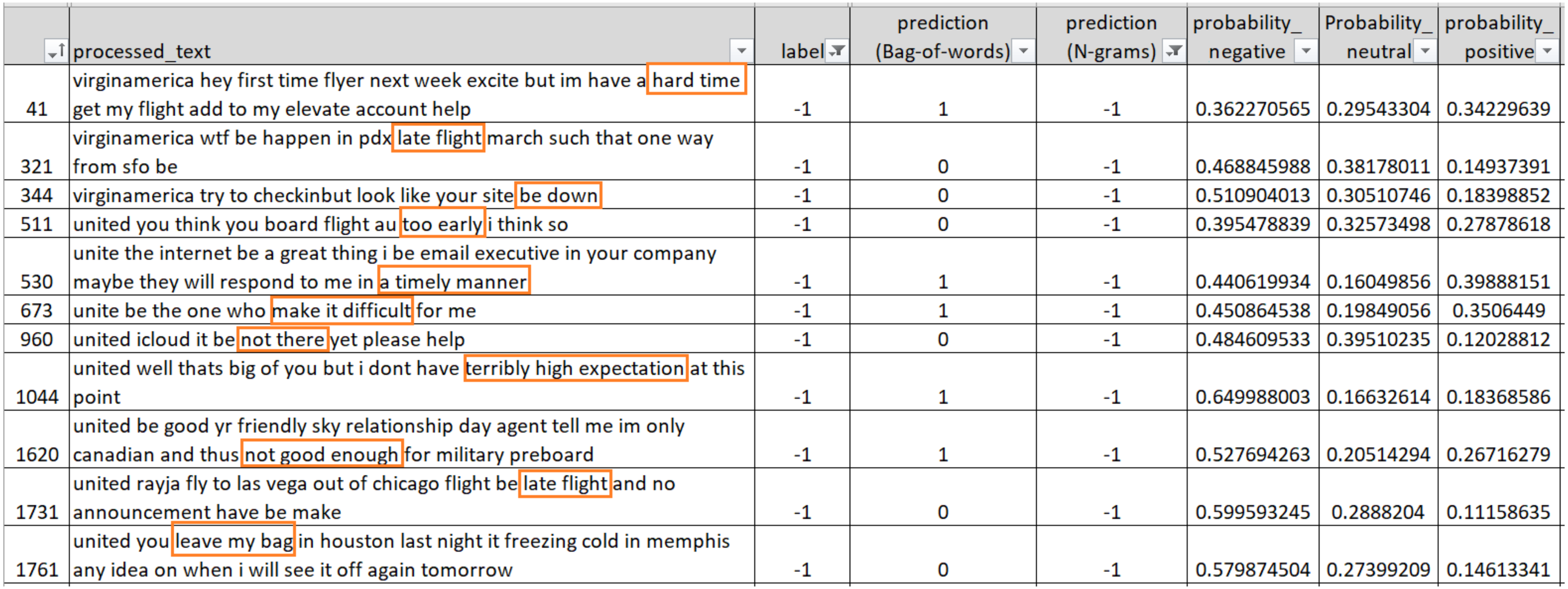


```
              precision    recall  f1-score

          -1       0.87      0.88      0.87
           0       0.65      0.58      0.62
           1       0.69      0.75      0.72

    accuracy                           0.79
   macro avg       0.74      0.74      0.74
```

Over-sampling data with TF-IDF + N-gram

Naïve Bayes

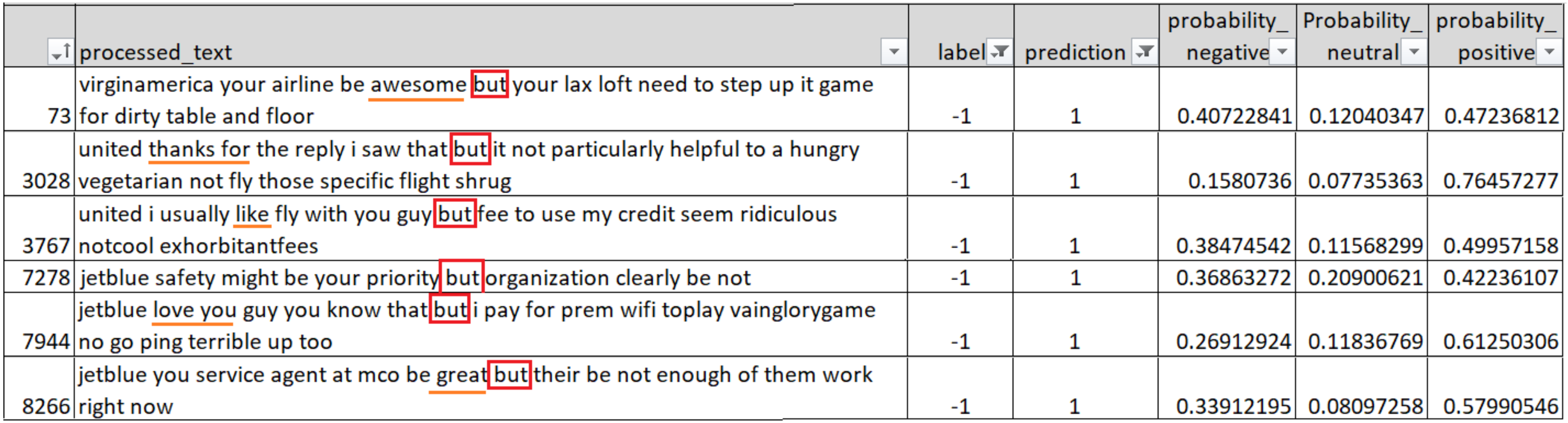
- Why over-sampling data with N-gram (Bag-of-words, bigrams and trigrams) outperforms oversampling data with bag-of-words?



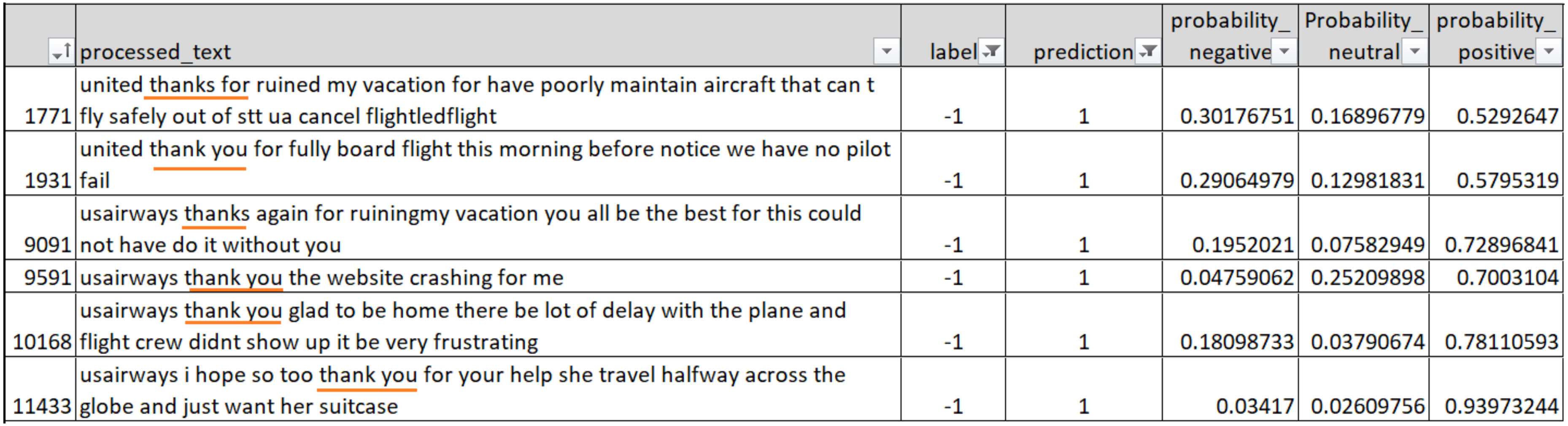
	processed_text	label	prediction (Bag-of-words)	prediction (N-grams)	probability_ negative	Probability_ neutral	probability_ positive
41	virginamerica hey first time flyer next week excite but im have a hard time get my flight add to my elevate account help	-1	1	-1	0.362270565	0.29543304	0.34229639
321	virginamerica wtf be happen in pdx late flight march such that one way from sfo be	-1	0	-1	0.468845988	0.38178011	0.14937391
344	virginamerica try to checkinbut look like your site be down	-1	0	-1	0.510904013	0.30510746	0.18398852
511	united you think you board flight au too early i think so	-1	0	-1	0.395478839	0.32573498	0.27878618
530	unite the internet be a great thing i be email executive in your company maybe they will respond to me in a timely manner	-1	1	-1	0.440619934	0.16049856	0.39888151
673	unite be the one who make it difficult for me	-1	1	-1	0.450864538	0.19849056	0.3506449
960	united icloud it be not there yet please help	-1	0	-1	0.484609533	0.39510235	0.12028812
1044	united well thats big of you but i dont have terribly high expectation at this point	-1	1	-1	0.649988003	0.16632614	0.18368586
1620	united be good yr friendly sky relationship day agent tell me im only canadian and thus not good enough for military preboard	-1	1	-1	0.527694263	0.20514294	0.26716279
1731	united rayja fly to las vega out of chicago flight be late flight and no announcement have be make	-1	0	-1	0.599593245	0.2888204	0.11158635
1761	united you leave my bag in houston last night it freezing cold in memphis any idea on when i will see it off again tomorrow	-1	0	-1	0.579874504	0.27399209	0.14613341

Naïve Bayes

- Error analysis based on NB + over-sampling + N-gram



	processed_text	label	prediction	probability_ negative	Probability_ neutral	probability_ positive
73	virginamerica your airline be awesome but your lax loft need to step up it game for dirty table and floor	-1	1	0.40722841	0.12040347	0.47236812
3028	united thanks for the reply i saw that but it not particularly helpful to a hungry vegetarian not fly those specific flight shrug	-1	1	0.1580736	0.07735363	0.76457277
3767	united i usually like fly with you guy but fee to use my credit seem ridiculous notcool exhorbitantfees	-1	1	0.38474542	0.11568299	0.49957158
7278	jetblue safety might be your priority but organization clearly be not	-1	1	0.36863272	0.20900621	0.42236107
7944	jetblue love you guy you know that but i pay for prem wifi toplay vainglorygame no go ping terrible up too	-1	1	0.26912924	0.11836769	0.61250306
8266	jetblue you service agent at mco be great but their be not enough of them work right now	-1	1	0.33912195	0.08097258	0.57990546

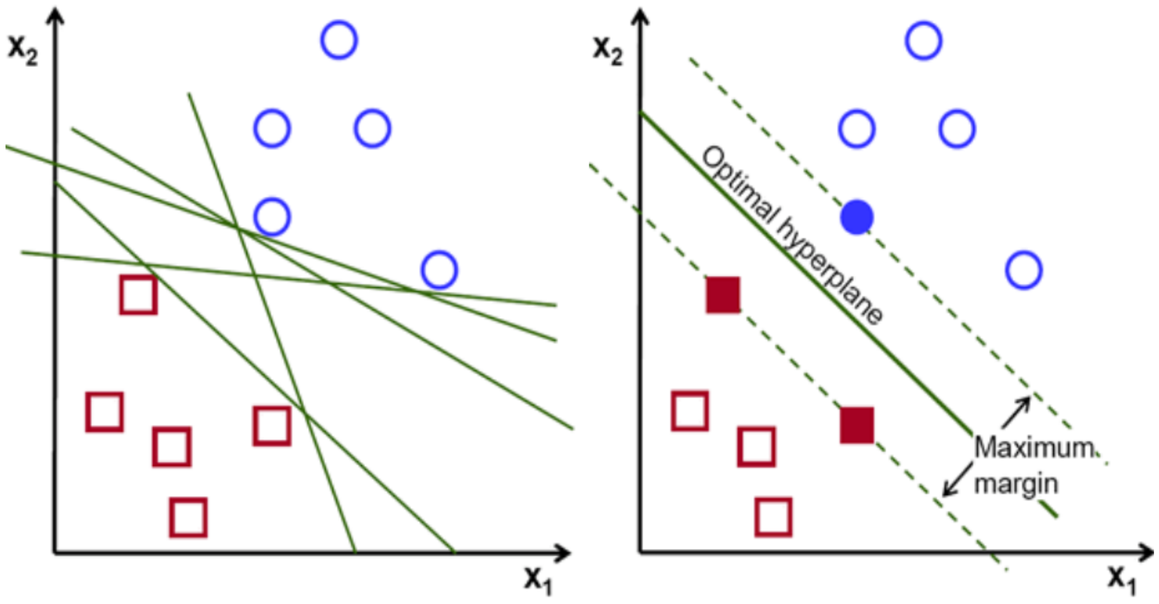


	processed_text	label	prediction	probability_ negative	Probability_ neutral	probability_ positive
1771	united thanks for ruined my vacation for have poorly maintain aircraft that can t fly safely out of stt ua cancel flightledflight	-1	1	0.30176751	0.16896779	0.5292647
1931	united thank you for fully board flight this morning before notice we have no pilot fail	-1	1	0.29064979	0.12981831	0.5795319
9091	usairways thanks again for ruiningmy vacation you all be the best for this could not have do it without you	-1	1	0.1952021	0.07582949	0.72896841
9591	usairways thank you the website crashing for me	-1	1	0.04759062	0.25209898	0.7003104
10168	usairways thank you glad to be home there be lot of delay with the plane and flight crew didnt show up it be very frustrating	-1	1	0.18098733	0.03790674	0.78110593
11433	usairways i hope so too thank you for your help she travel halfway across the globe and just want her suitcase	-1	1	0.03417	0.02609756	0.93973244

Support Vector Machine (SVM)

- SVM finds the hyperplane that maximizes the margin as well as minimizes the probability of misclassification.

- The instances on the edge of the margin are called support vectors, and the decision boundary is fully determined by the support vectors.



Referenced from https://towardsdatascience.com/svm-feature-selection-and-kernels-840781cc1a6c

Support Vector Machine (SVM) / TF-IDF

- TF-IDF + Bag-of-words

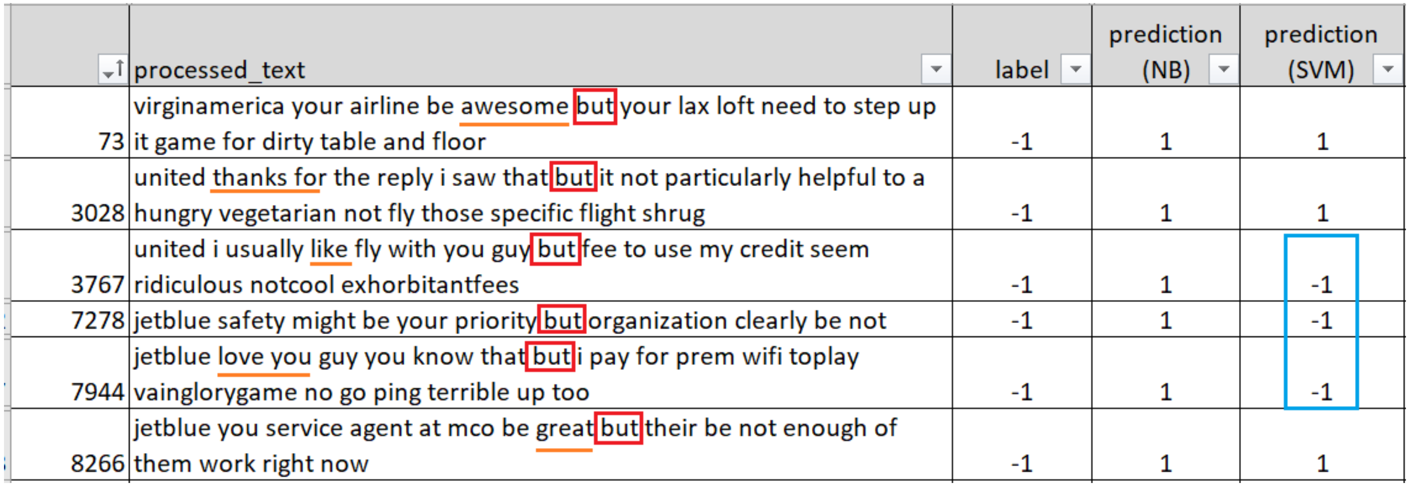
	Precision (Macro Average)	Recall (Macro Average)	F1 Score (Macro Average)	Accuracy (Overall)
Original data	0.77	0.70	0.73	0.80
Oversampling data	0.72	0.73	0.72	0.77
Undersampling data	0.69	0.73	0.70	0.75

- TF-IDF + Combination of Bag-of-words, bigrams and trigrams

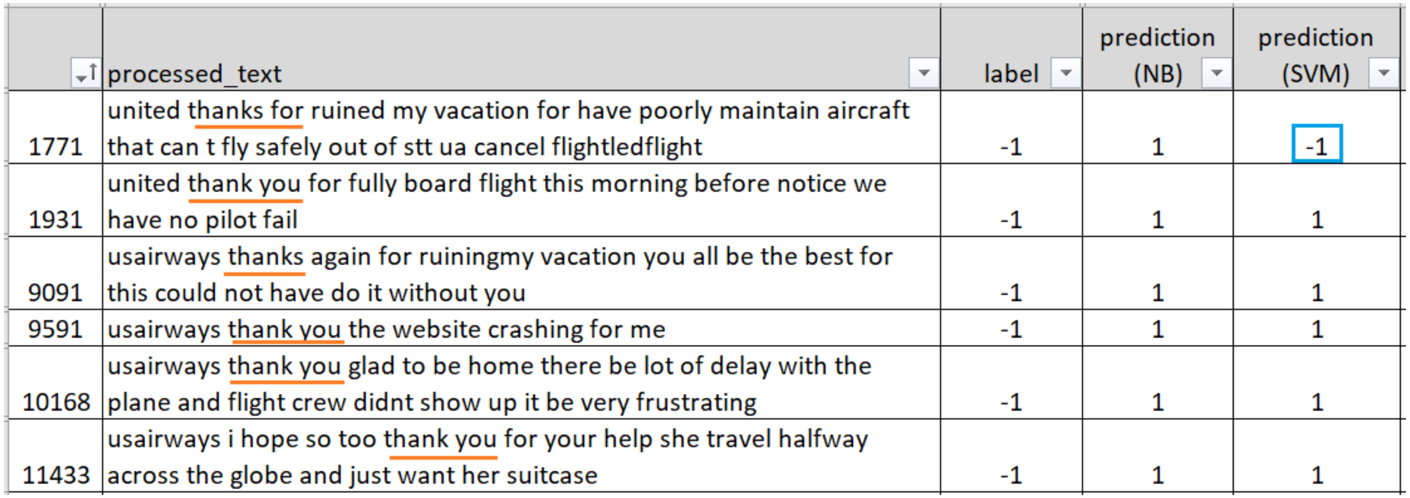
	Precision (Macro Average)	Recall (Macro Average)	F1 Score (Macro Average)	Accuracy (Overall)
Original data	0.79	0.70	0.73	0.80
Oversampling data	0.76	0.73	0.74	0.80
Undersampling data	0.71	0.73	0.72	0.78

Support Vector Machine (SVM) / IF-IDF

- Error analysis based on SVM + over-sampling data + N-gram



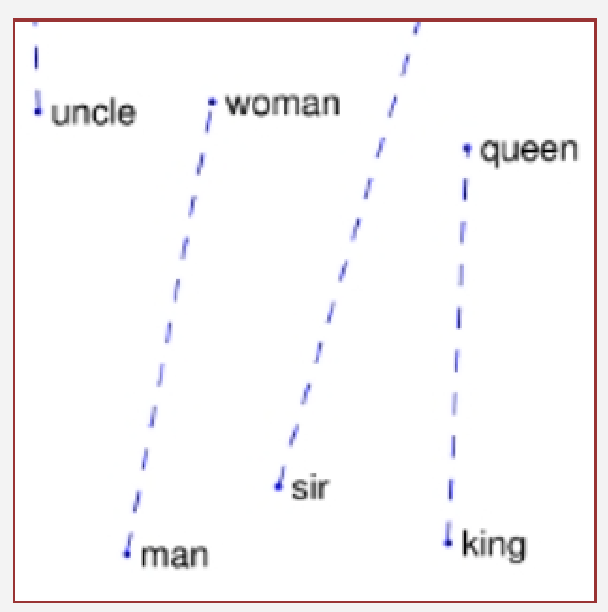
	processed_text	label	prediction (NB)	prediction (SVM)
73	virginamerica your airline be awesome but your lax loft need to step up it game for dirty table and floor	-1	1	1
3028	united thanks for the reply i saw that but it not particularly helpful to a hungry vegetarian not fly those specific flight shrug	-1	1	1
3767	united i usually like fly with you guy but fee to use my credit seem ridiculous notcool exhorbitantfees	-1	1	-1
7278	jetblue safety might be your priority but organization clearly be not	-1	1	-1
7944	jetblue love you guy you know that but i pay for prem wifi toplay vainglorygame no go ping terrible up too	-1	1	-1
8266	jetblue you service agent at mco be great but their be not enough of them work right now	-1	1	1



	processed_text	label	prediction (NB)	prediction (SVM)
1771	united thanks for ruined my vacation for have poorly maintain aircraft that can t fly safely out of stt ua cancel flightledflight	-1	1	-1
1931	united thank you for fully board flight this morning before notice we have no pilot fail	-1	1	1
9091	usairways thanks again for ruiningmy vacation you all be the best for this could not have do it without you	-1	1	1
9591	usairways thank you the website crashing for me	-1	1	1
10168	usairways thank you glad to be home there be lot of delay with the plane and flight crew didnt show up it be very frustrating	-1	1	1
11433	usairways i hope so too thank you for your help she travel halfway across the globe and just want her suitcase	-1	1	1

GloVe (global vectors for word representation)

- GloVe is an unsupervised machine learning algorithm for getting vector representations for words.
- The GloVe model is trained on a global word-word co-occurrence matrix, which records how frequently words co-occur with one another in a given corpus.
- If two words are co-exist many times, both words may have similar meaning, such as man and woman. GloVe uses a linear structure to capture the difference of frequently co-exist words.



$$King - Man + Woman = Queen$$

Referenced from https://nlp.stanford.edu/projects/glove/

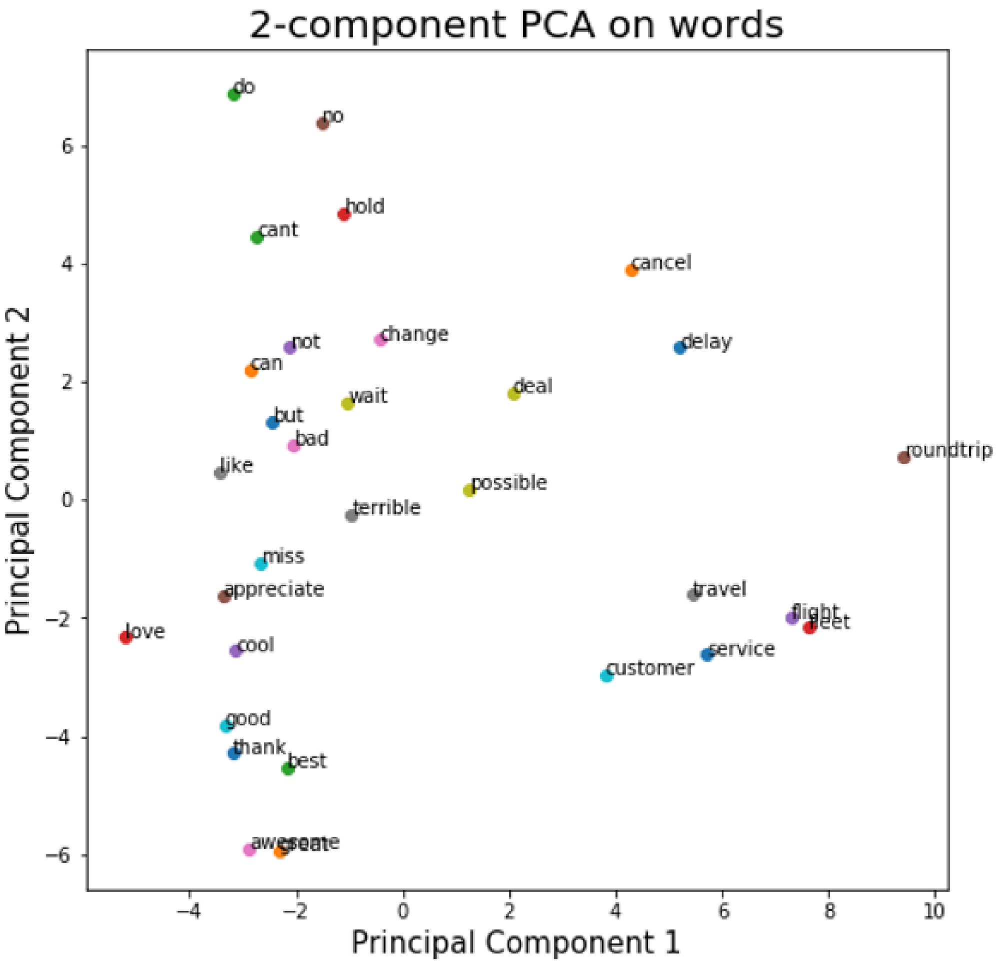
Support Vector Machine (SVM) / GloVe

- Over-sampling and under-sampling improve the recall.
- Results are much worse than Naïve Bayes and SVM with bag-of-words and N-gram.

	Precision (Macro Average)	Recall (Macro Average)	F1 Score (Macro Average)	Accuracy (Overall)
Original dataset	0.72	0.63	0.66	0.76
Oversampling data	0.65	0.68	0.66	0.71
Undersampling data	0.66	0.69	0.67	0.72

Support Vector Machine (SVM) / GloVe

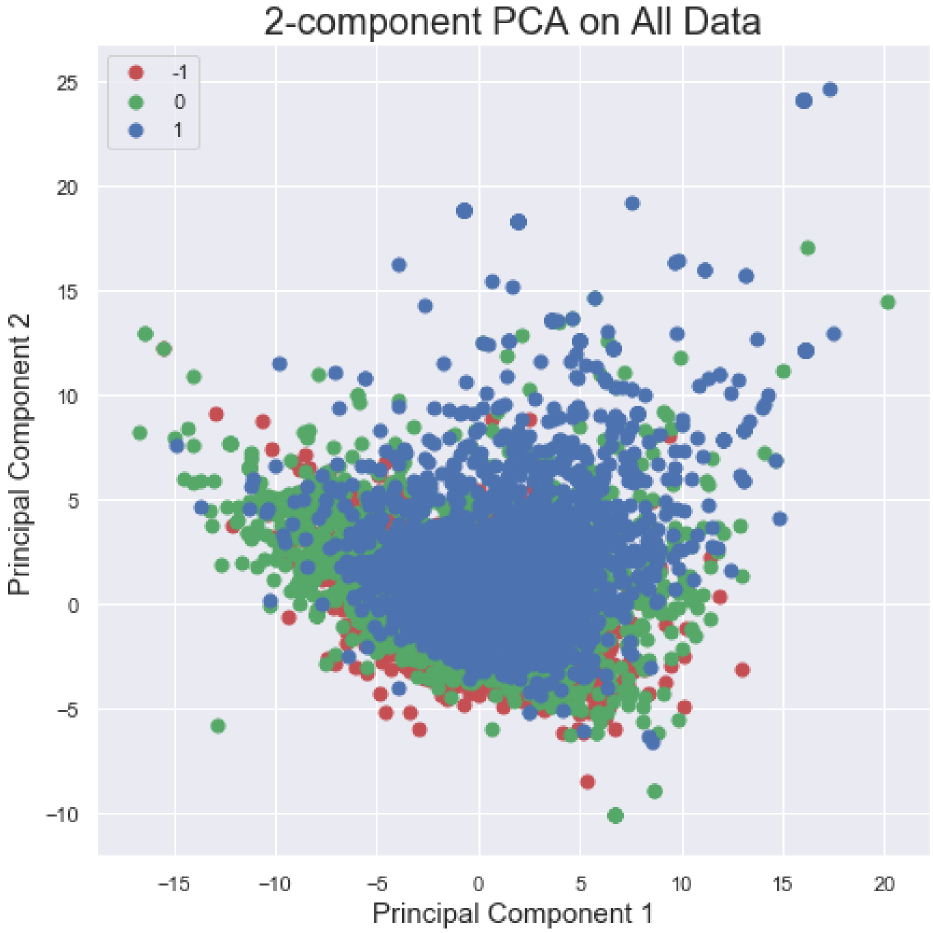
- Why SVM with GloVe is worse than SVM with TF-IDF and bag-of-words/N-gram?



- Pre-trained GloVe word vectors are from Stanford https://nlp.stanford.edu/projects/glove/
- Word vectors are trained on 2B tweets, 27B tokens, 1.2M vocabularies.
- A 100-dimension vector is used to represent a word.

Support Vector Machine (SVM) / GloVe

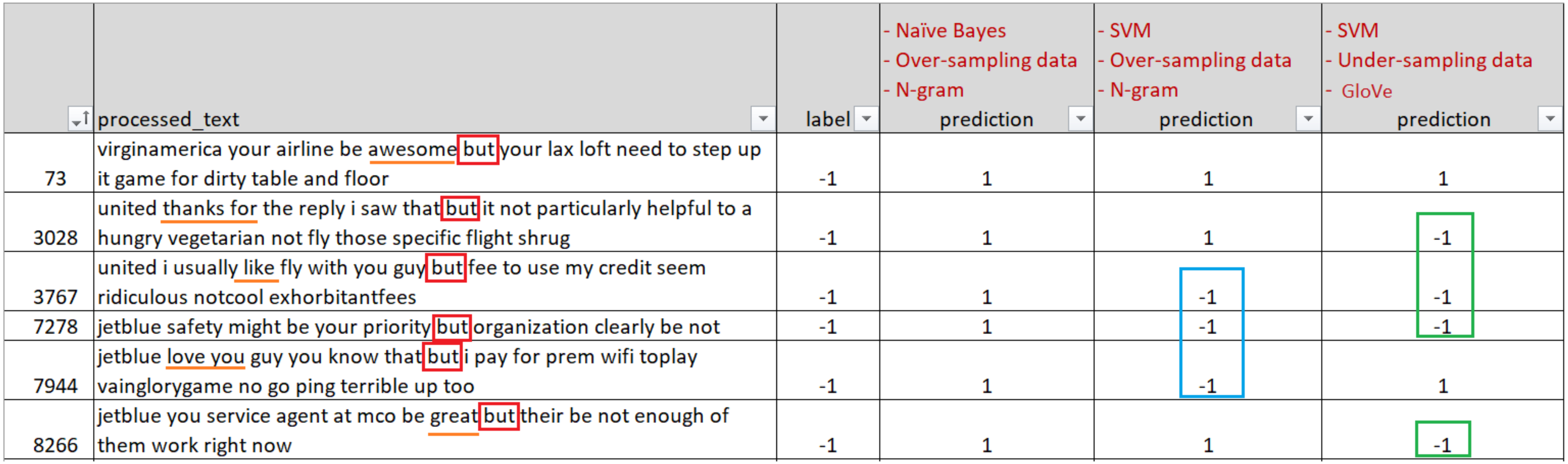
- Why SVM with GloVe is worse than SVM with TF-IDF and bag-of-words/N-gram?



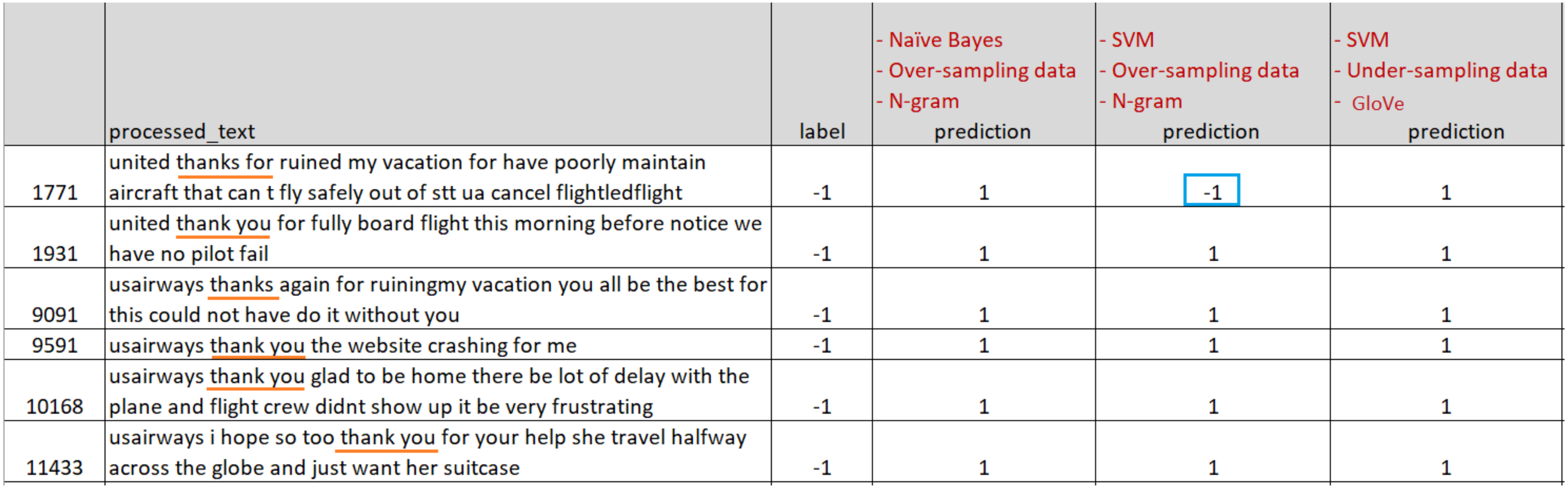
- Original data: 11672 out of 249986 (4.7 %) words are not in pre-trained model.
- Over-sampling data: 20492 out of 391057 (5.2 %) words are not in pre-trained model.
- Under-sampling data: 6928 out of 137055 (5 %) words are not in pre-trained model.

Support Vector Machine (SVM) / GloVe

- Error analysis based on SVM + under-sampling + GloVe



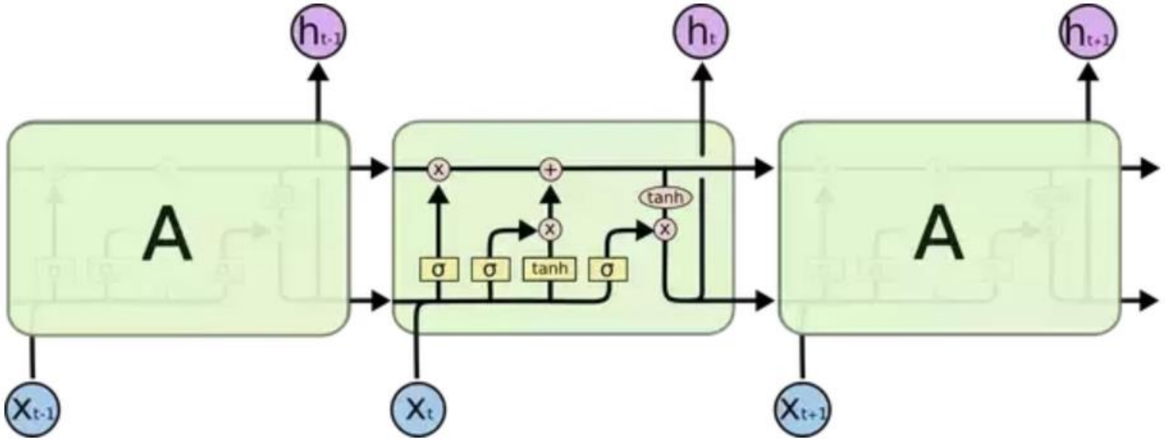
	processed_text	label	- Naïve Bayes - Over-sampling data - N-gram prediction	- SVM - Over-sampling data - N-gram prediction	- SVM - Under-sampling data - GloVe prediction
73	virginamerica your airline be awesome but your lax loft need to step up it game for dirty table and floor	-1	1	1	1
3028	united thanks for the reply i saw that but it not particularly helpful to a hungry vegetarian not fly those specific flight shrug	-1	1	1	-1
3767	united i usually like fly with you guy but fee to use my credit seem ridiculous notcool exhorbitantfees	-1	1	-1	-1
7278	jetblue safety might be your priority but organization clearly be not	-1	1	-1	-1
7944	jetblue love you guy you know that but i pay for prem wifi toplay vainglorygame no go ping terrible up too	-1	1	-1	1
8266	jetblue you service agent at mco be great but their be not enough of them work right now	-1	1	1	-1

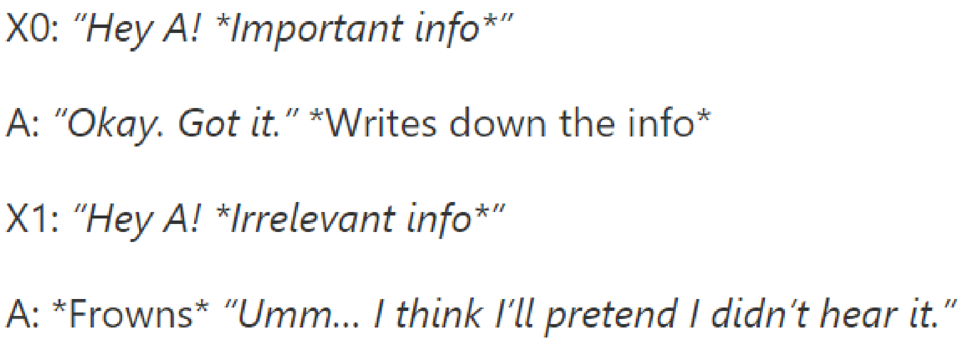


	processed_text	label	- Naïve Bayes - Over-sampling data - N-gram prediction	- SVM - Over-sampling data - N-gram prediction	- SVM - Under-sampling data - GloVe prediction
1771	united thanks for ruined my vacation for have poorly maintain aircraft that can t fly safely out of stt ua cancel flightledflight	-1	1	-1	1
1931	united thank you for fully board flight this morning before notice we have no pilot fail	-1	1	1	1
9091	usairways thanks again for ruiningmy vacation you all be the best for this could not have do it without you	-1	1	1	1
9591	usairways thank you the website crashing for me	-1	1	1	1
10168	usairways thank you glad to be home there be lot of delay with the plane and flight crew didnt show up it be very frustrating	-1	1	1	1
11433	usairways i hope so too thank you for your help she travel halfway across the globe and just want her suitcase	-1	1	1	1

Long Short-Term Memory Networks

- LSTM networks are a type of recurrent neural networks (RNNs).
- LSTM networks can “remember” or “forget” information in the cell state by using specialized neurons called “gates”.
- This gating mechanism fixes the “short-term memory” problem of RNNs.





X0: *“Hey A! *Important info*”*

A: *“Okay. Got it.”* *Writes down the info*

X1: *“Hey A! *Irrelevant info*”*

A: *Frowns* *“Umm... I think I’ll pretend I didn’t hear it.”*

Referenced from: https://www.quora.com/How-is-LSTM-different-from-RNN-In-a-layman-explanation

Long Short-Term Memory Networks

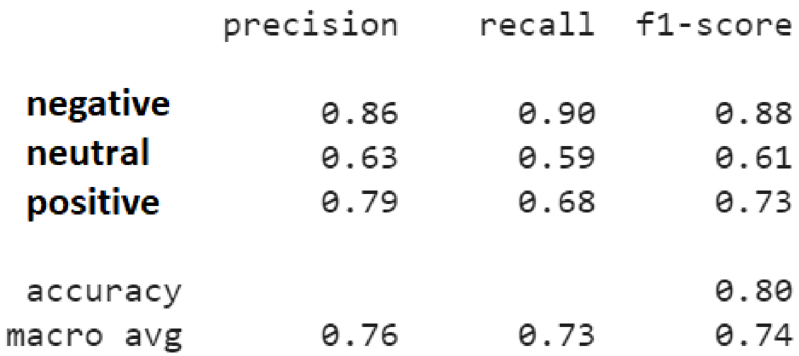
- Model configuration
  - 1 embedding layer.
    - Size of the vocabulary: 10347
    - Size of word embedding vector : 16
  - LSTM with 8 neurons.
  - 1 dense layer (activation: softmax).
  - Optimizer: adam, epochs: 8, batch size: 48.

	Precision (Macro Average)	Recall (Macro Average)	F1 Score (Macro Average)	Accuracy (Overall)
Original data	0.76	0.73	0.74	0.80

Long Short-Term Memory Networks

Predicted

True	negative	neutral	positive	All
negative	1641	146	30	1817
neutral	197	371	60	628
positive	80	74	329	483
All	1918	591	419	2928

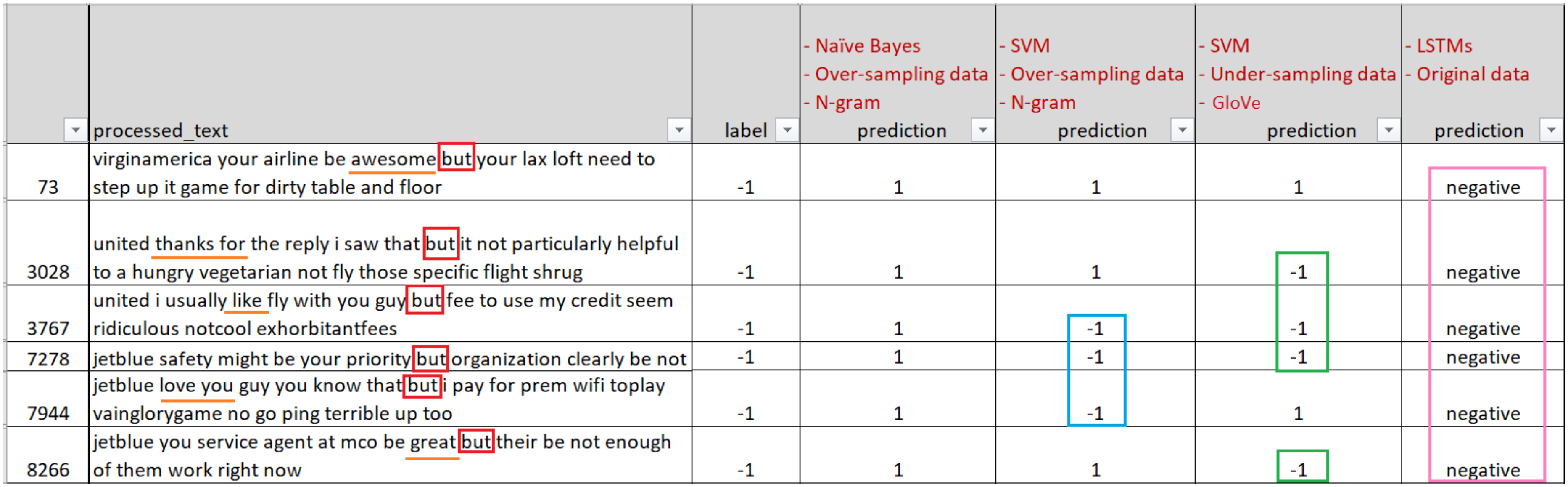


```
               precision    recall  f1-score

    negative        0.86      0.90      0.88
    neutral         0.63      0.59      0.61
    positive        0.79      0.68      0.73

    accuracy                            0.80
   macro avg        0.76      0.73      0.74
```

LSTM networks with original data

Long Short-Term Memory Networks

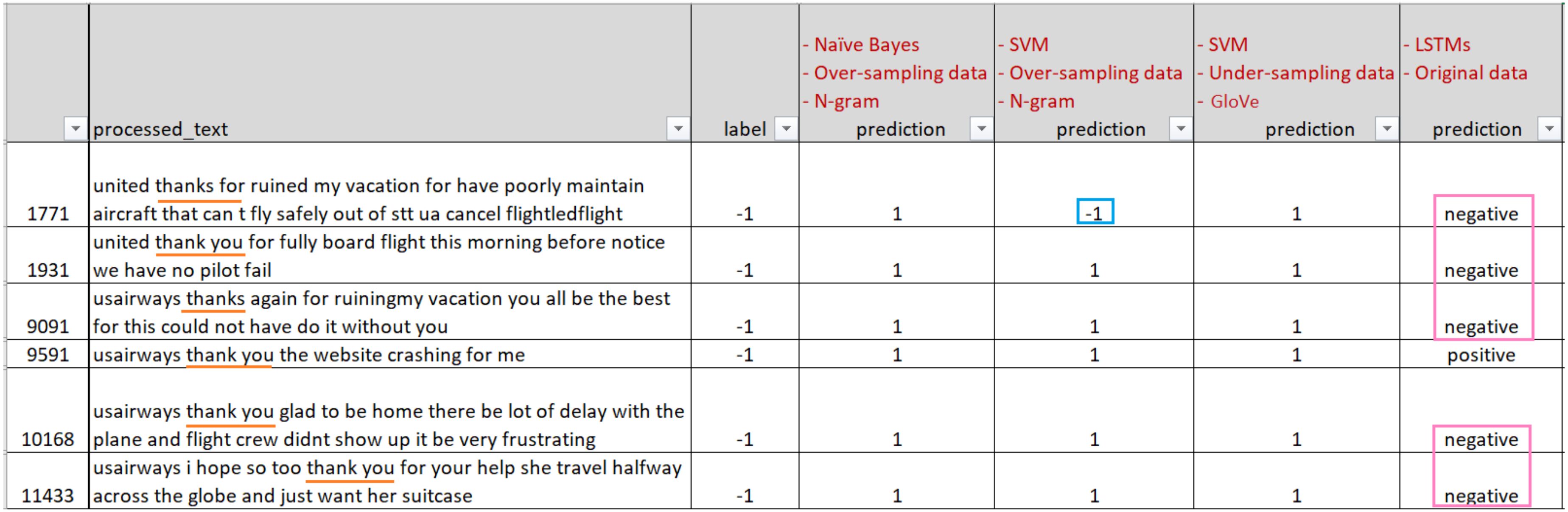
- Error analysis based on LSTM networks with original data.



			- Naïve Bayes - Over-sampling data - N-gram	- SVM - Over-sampling data - N-gram	- SVM - Under-sampling data - GloVe	- LSTMs - Original data
	processed_text	label	prediction	prediction	prediction	prediction
73	virginamerica your airline be awesome but your lax loft need to step up it game for dirty table and floor	-1	1	1	1	negative
3028	united thanks for the reply i saw that but it not particularly helpful to a hungry vegetarian not fly those specific flight shrug	-1	1	1	-1	negative
3767	united i usually like fly with you guy but fee to use my credit seem ridiculous notcool exhorbitantfees	-1	1	-1	-1	negative
7278	jetblue safety might be your priority but organization clearly be not	-1	1	-1	-1	negative
7944	jetblue love you guy you know that but i pay for prem wifi toplay vainglorygame no go ping terrible up too	-1	1	-1	1	negative
8266	jetblue you service agent at mco be great but their be not enough of them work right now	-1	1	1	-1	negative

Long Short-Term Memory Networks

- Error analysis based on LSTM networks with original data.



	processed_text	label	- Naïve Bayes - Over-sampling data - N-gram prediction	- SVM - Over-sampling data - N-gram prediction	- SVM - Under-sampling data - GloVe prediction	- LSTMs - Original data prediction
1771	united thanks for ruined my vacation for have poorly maintain aircraft that can t fly safely out of stt ua cancel flightledflight	-1	1	-1	1	negative
1931	united thank you for fully board flight this morning before notice we have no pilot fail	-1	1	1	1	negative
9091	usairways thanks again for ruiningmy vacation you all be the best for this could not have do it without you	-1	1	1	1	negative
9591	usairways thank you the website crashing for me	-1	1	1	1	positive
10168	usairways thank you glad to be home there be lot of delay with the plane and flight crew didnt show up it be very frustrating	-1	1	1	1	negative
11433	usairways i hope so too thank you for your help she travel halfway across the globe and just want her suitcase	-1	1	1	1	negative

Results

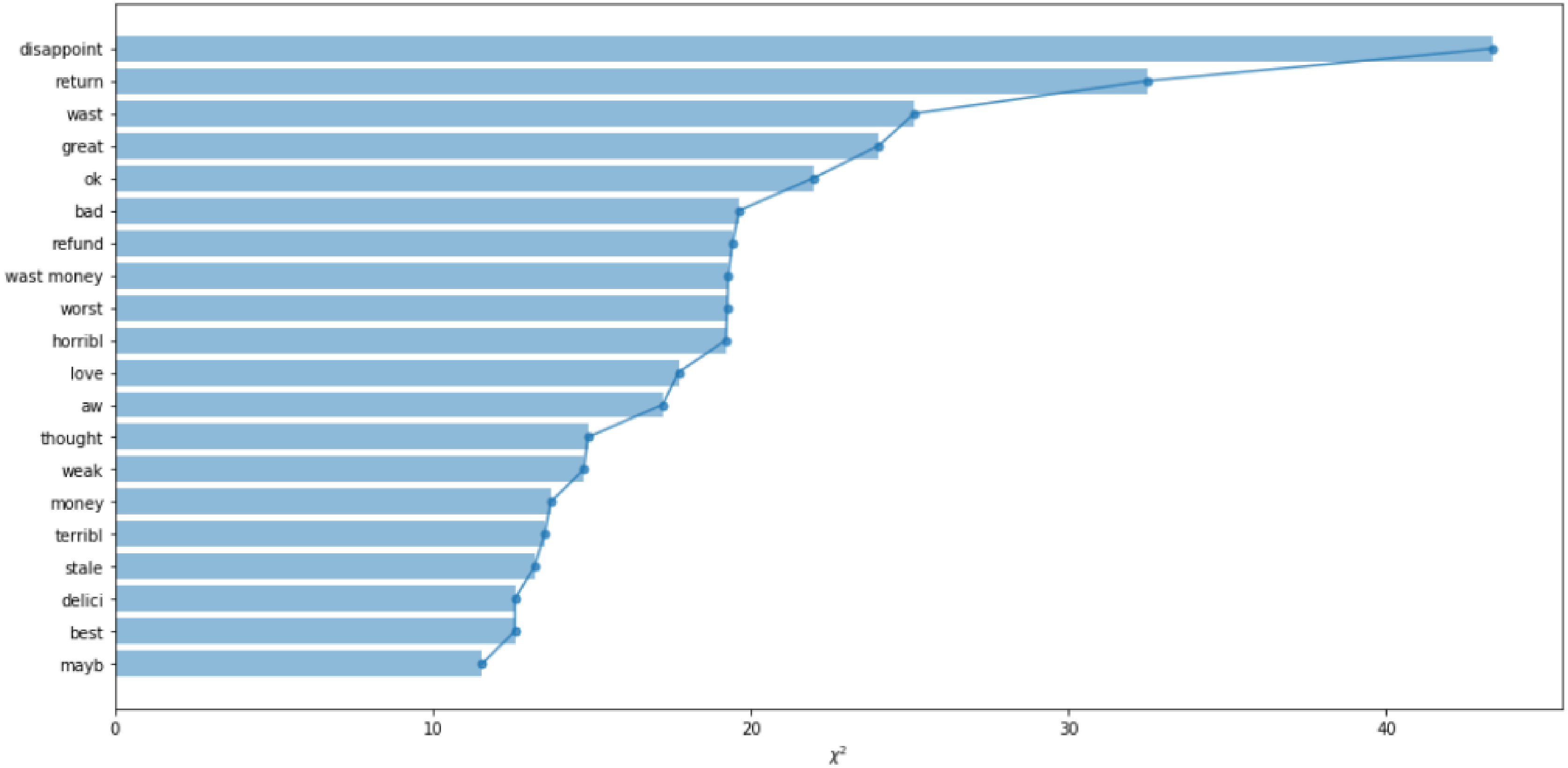
	Precision (Macro Average)	Recall (Macro Average)	F1 Score (Macro Average)	Accuracy (Overall)
Naïve Bayes - Over-sampling data - Bag-of-words - TF-IDF	0.71	0.71	0.71	0.78
Naïve Bayes - Over-sampling data - N-gram - TF-IDF	0.74	0.74	0.74	0.79
SVM - Over-sampling data - Bag-of-words - TF-IDF	0.72	0.73	0.72	0.77
SVM - Over-sampling data - N-gram - TF-IDF	0.76	0.73	0.74	0.80
SVM - Under-sampling data - GloVe pre-trained vectors	0.66	0.69	0.67	0.72
LSTM networks	0.76	0.73	0.74	0.80

Findings

- Data balancing techniques can offset the Naïve Bayes and SVM models' bias towards the majority class. Generally speaking, over-sampling has a better performance than under-sampling on short texts.
- Features
  - Naïve Bayes and SVM models has a better performance with N-gram than bag-of-words.
  - There is no guarantee that GloVe will yield better results than bag-of-words or N-gram on any data.
- Models
  - Naïve Bayes model is more sensitive to imbalanced data than SVM model. However, LSTM networks can handle imbalanced data well.
  - Although Naïve Bayes/SVM models with oversampling data and N-gram have similar performance with LSTM networks with original data, but LSTM networks has relatively higher recall on negative Tweets, and they give better predictions on Tweets that contain a twist and sarcasm.
  - LSTM networks is prone to overfitting. Therefore, it requires a lot of work on model tuning.

Future work

- Building ensemble models
- Feature selection



Questions?