MAA Text Mining
Spring 2019

Group 9

Susan Wang (M20170548)
Carolina Bellani (M20170098)
Yu Chun Liu (M20180220)
Bhupendra Roy (M20170361)

Hotel Reviews

Positive & Truthful

Well located and well staffed. The Amalfi was a clean and comfortable place to stay. It lacks a restaurant or bar but many are near by. Free continental breakfast and evening cocktail.

Negative & Truthful

no bell boys there when you need them, sully staff, no smiles at all in this place. had to wait 30 min for my luggage and was never greeted by any of the staff. boy, sure is a lousy fairmont compared to the one in SF. never here again!

Positive & Deceptive

I stayed at Swissotel Chicago when I was on business and it was very nice. The staff was very helpful and room was very clean. I would stay again in a heart beat!

Negative & Deceptive

The James Hotel in Chicago looks nice on the website, but looks can be deceiving! The employees are extremely rude and the rooms are disgusting. I would definitely advise against staying there.

Balanced Dataset

400 RECORDS
POSITIVE
DECEPTIVE
MECHANICAL TURK
20 HOTELS

400 RECORDS
POSITIVE
TRUTHFUL
TRIP ADVISOR
20 HOTELS

400 RECORDS
NEGATIVE
DECEPTIVE
MECHANICAL TURK
20 HOTELS

400 RECORDS
NEGATIVE
TRUTHFUL
WEB
20 HOTELS

WWW.KAGGLE.COM/RTATMAN/DECEPTIVE-OPINION-SPAM-CORPUS

Literature Review

FINDING DECEPTIVE OPINION SPAM BY ANY STRETCH OF THE IMAGINATION

Humans Review

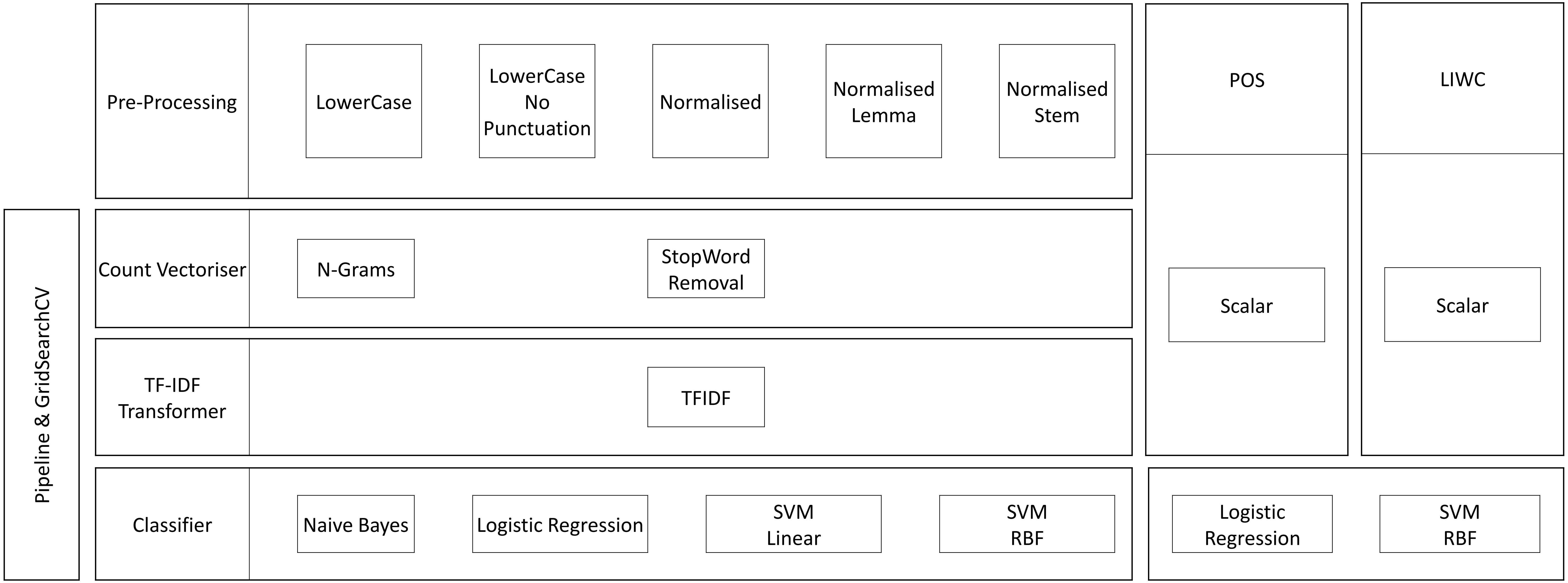
			TRUTHFUL			DECEPTIVE		
		Accuracy	**P**	**R**	**F**	**P**	**R**	**F**
HUMAN	JUDGE 1	**61.9%**	57.9	87.5	**69.7**	74.4	36.3	48.7
	JUDGE 2	56.9%	53.9	**95.0**	68.8	**78.9**	18.8	30.3
	JUDGE 3	53.1%	52.3	70.0	59.9	54.7	36.3	43.6
META	MAJORITY	58.1%	54.8	92.5	68.8	76.0	23.8	36.2
	SKEPTIC	60.6%	**60.8**	60.0	60.4	60.5	**61.3**	**60.9**

N-GRAMS POS LIWC

FEATURE ENGINEERING

https://www.aclweb.org/anthology/P11-1032

Methodology



Grid Search & Cross Validation

The red marks are the best performed parameters for deceptive opionion identification

Naive Bayes

```
param_grid = [
{'vect__stop_words': ['english', None],
vect__min_df': [1, 2, 5],
'vect__ngram_range': [(1, 1), (1, 2), (1, 3)],
clf__alpha': (1, 1e-2, 1e-3)}
]
```

Logistic Regression

```
param_grid = [
{'vect__stop_words': ['english', None],
vect__min_df': [1, 2, 5],
'vect__ngram_range': [(1, 1), (1, 2), (1, 3)],
'clf__C': [0.1, 1, 10, 100, 1000]}
]
```

SVM Rdf

```
param_grid = [
vect__min_df': [1, 2, 5],
'vect__ngram_range': [(1, 1), (1, 2), (1, 3)],
'clf__kernel': ['rbf'],
'clf__C': [0.1, 1, 10, 100],
'clf__gamma': [0.01, 0.1, 1, 10]}
]
```

SVM Linear

```
param_grid = [
{vect__min_df': [1, 2, 5],
'vect__ngram_range': [(1, 1), (1, 2), (1, 3)],
'clf__kernel': ['linear'],
'clf__C': [0.1, 1, 10, 100]}
]
```

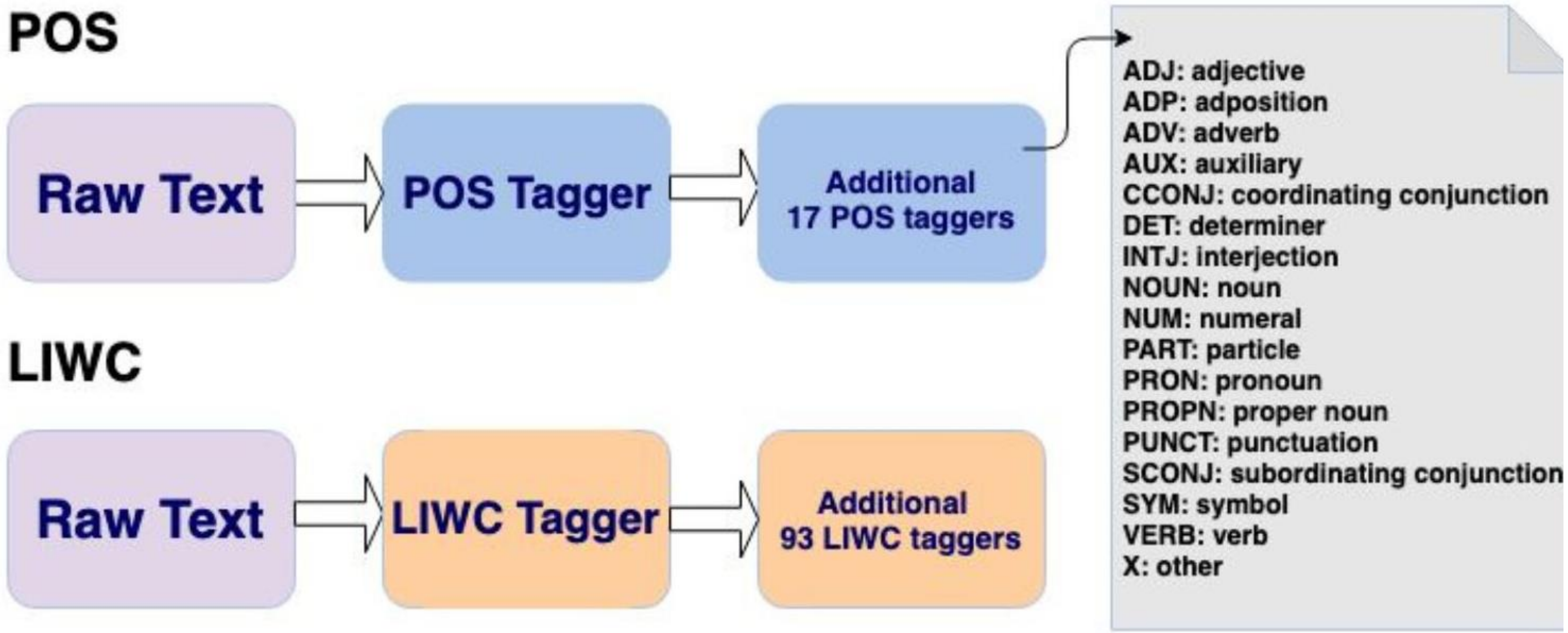
POS & LR

```
param_grid = [
{'vect__stop_words': ['english', None],
'vect__min_df': [1, 2, 5],
'vect__ngram_range': [(1, 1), (1, 2), (1, 3)],
'clf__C': [0.1, 1, 10, 100, 1000]}
]
```

LIWC & SVM

```
param_grid = [
{'vect__stop_words': ['english', None],
'vect__min_df': [1, 2],
'vect__ngram_range': [(1, 2), (1, 3)],
clf__kernel': ['rbf'],
'clf__gamma': [0.01, 0.1],
'clf__C': [1, 10, 100]}]
```

Feature Engineering

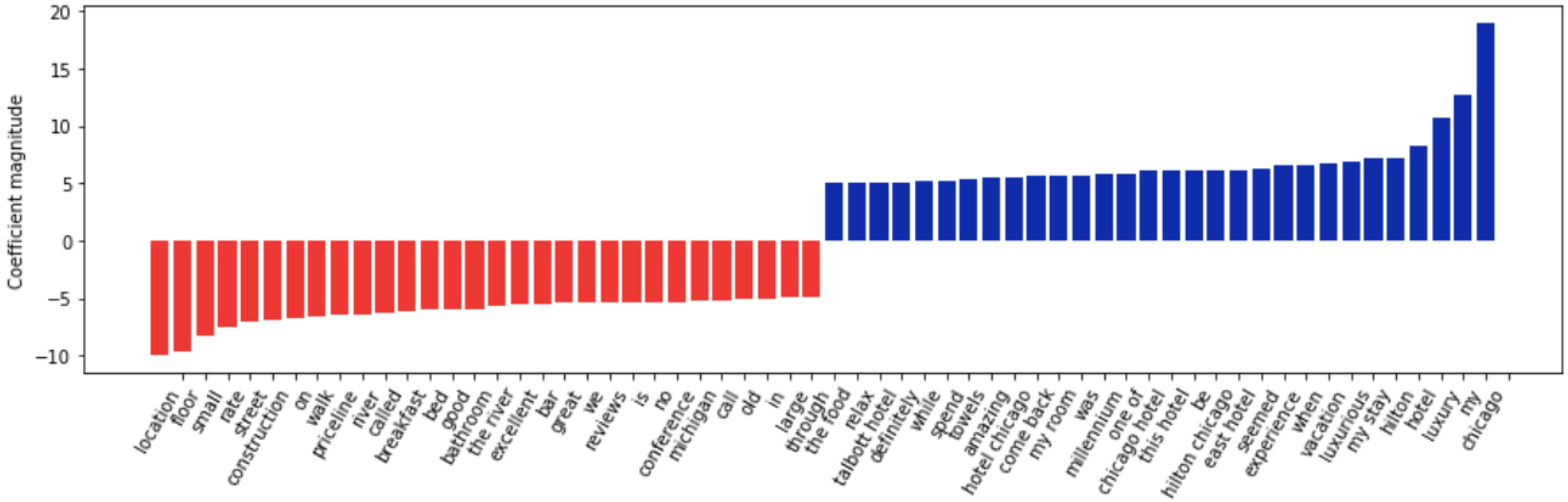


FINAL DATASET

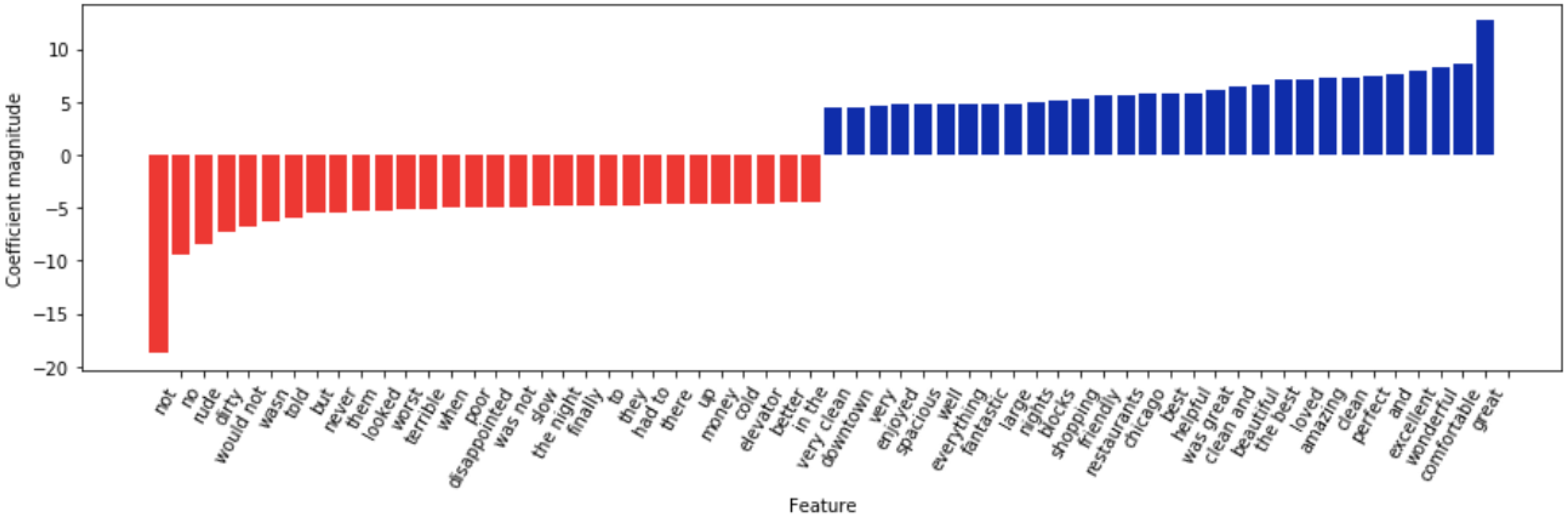
DATA			
Raw Text	Pre Processing	POS	LIWC
	Lowercase, Punctation Number, Stemming Lemmatization		

Identify important feature with text

Deceptive Opinion

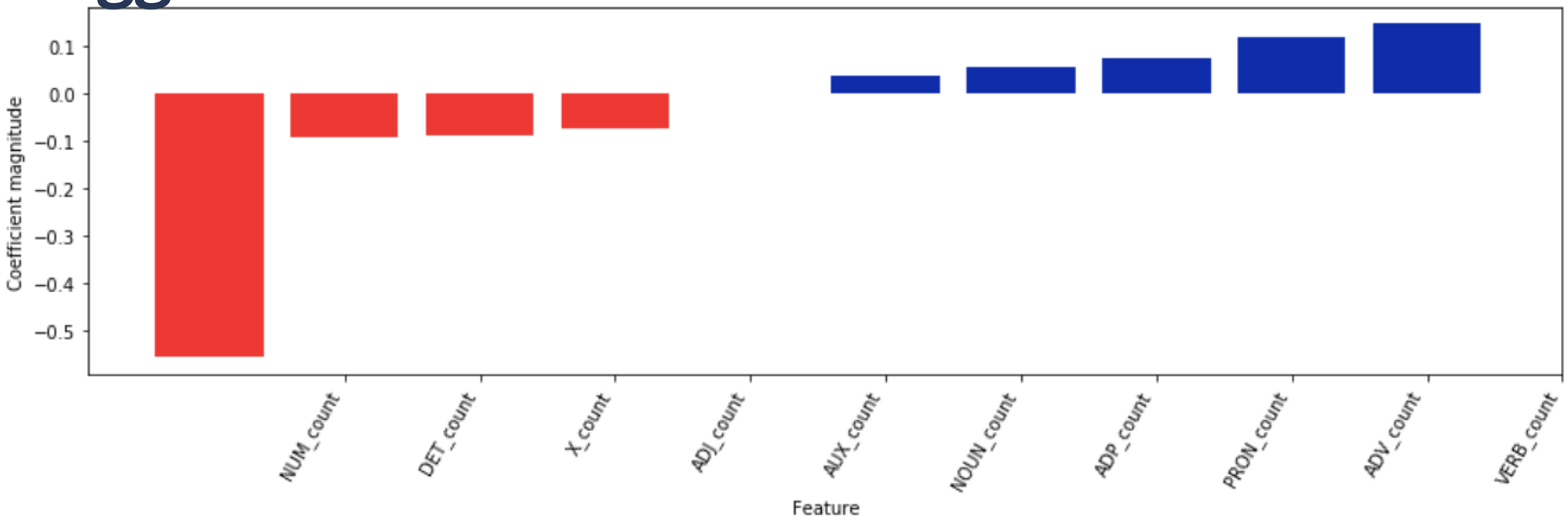


Sentiment Analysis

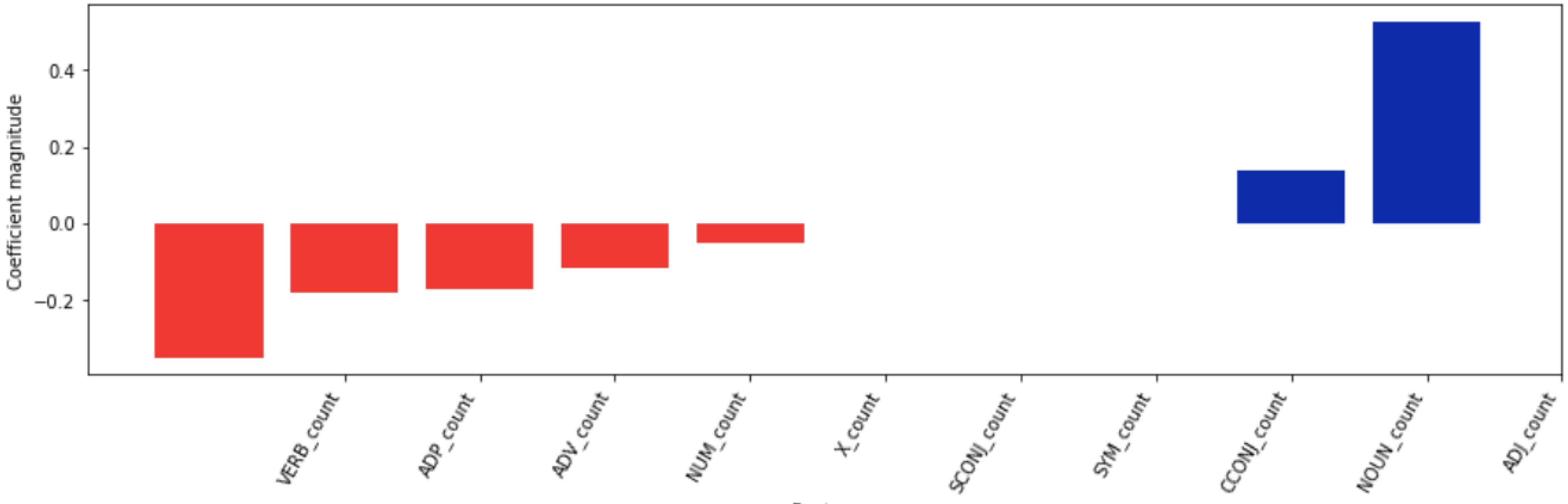


Identify important feature with POS Tagger

Deceptive Opinion

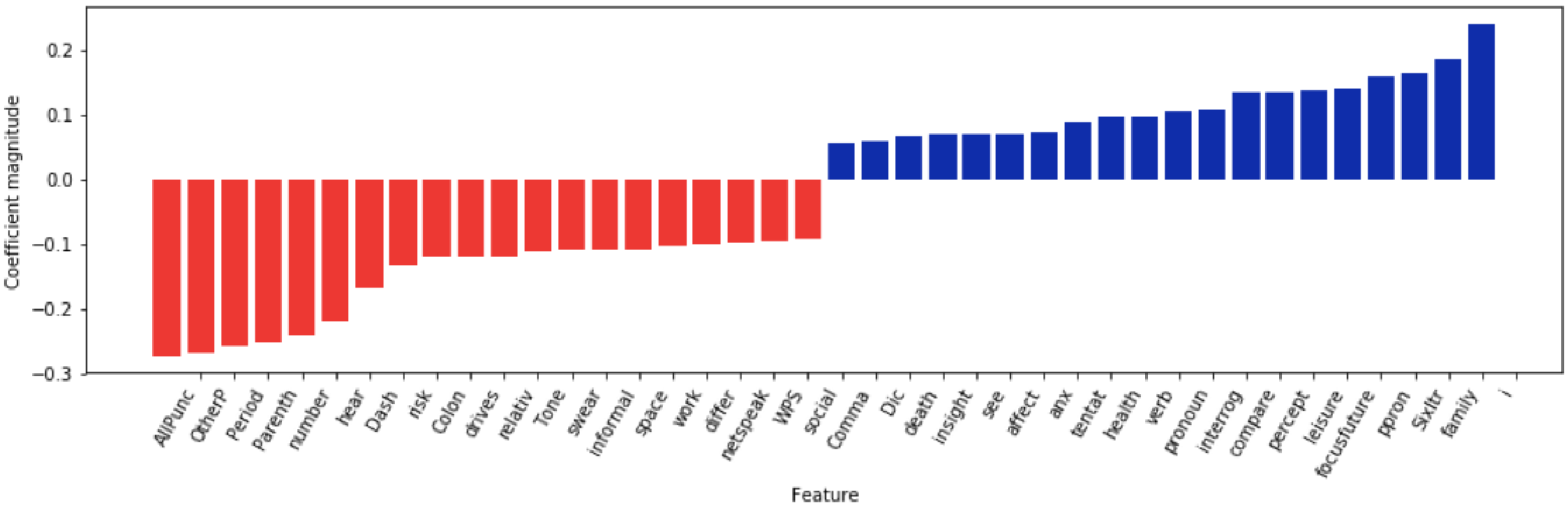


Sentiment Analysis

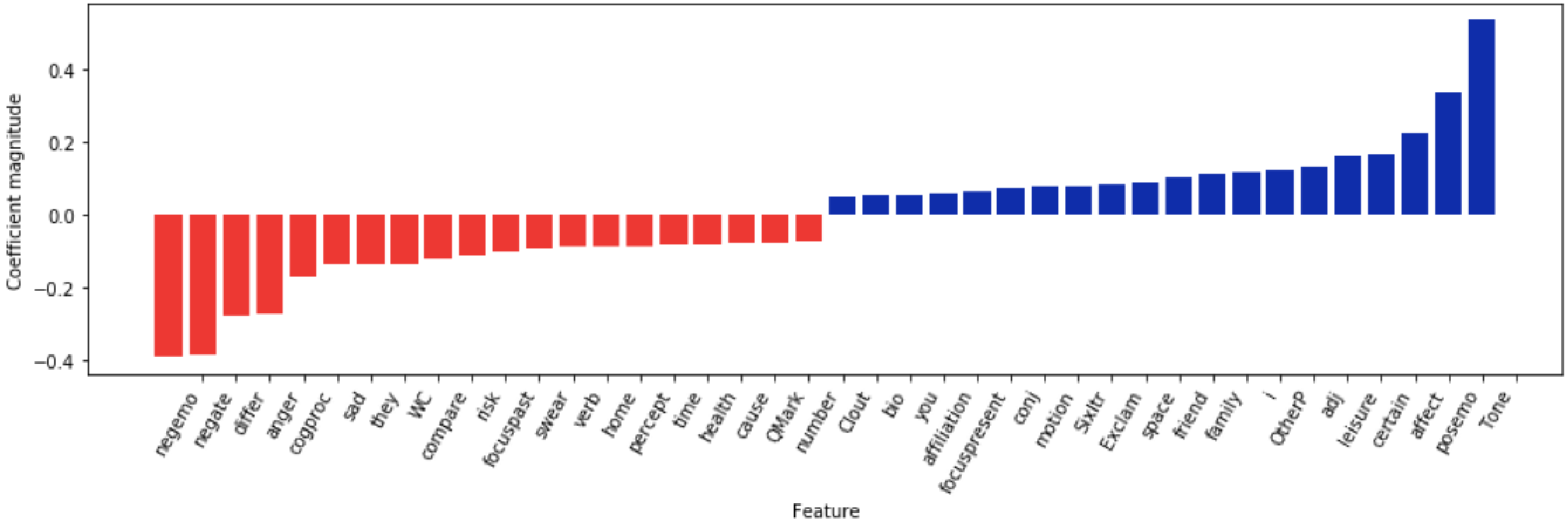


Identify important feature with LIWC Tagger

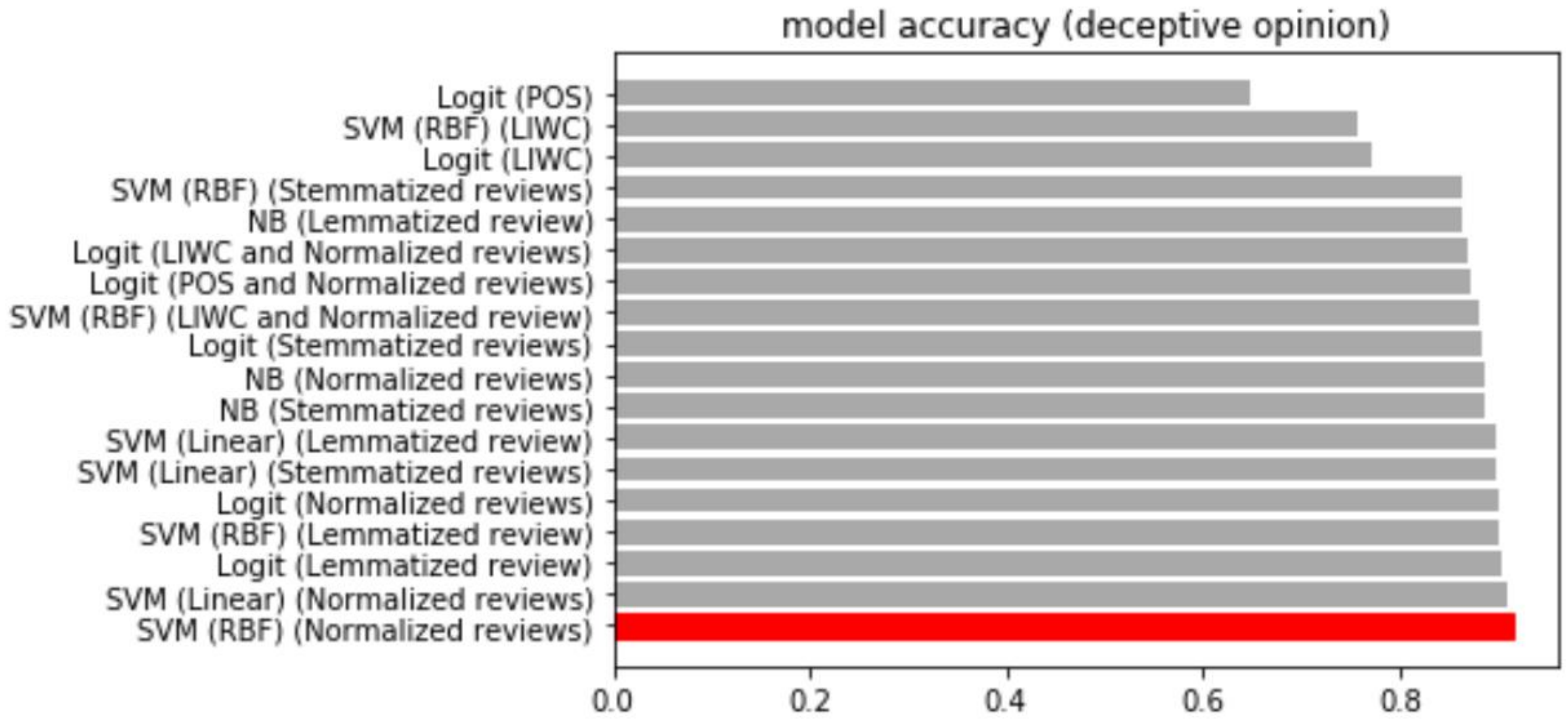
Deceptive Opinion



Sentiment Analysis

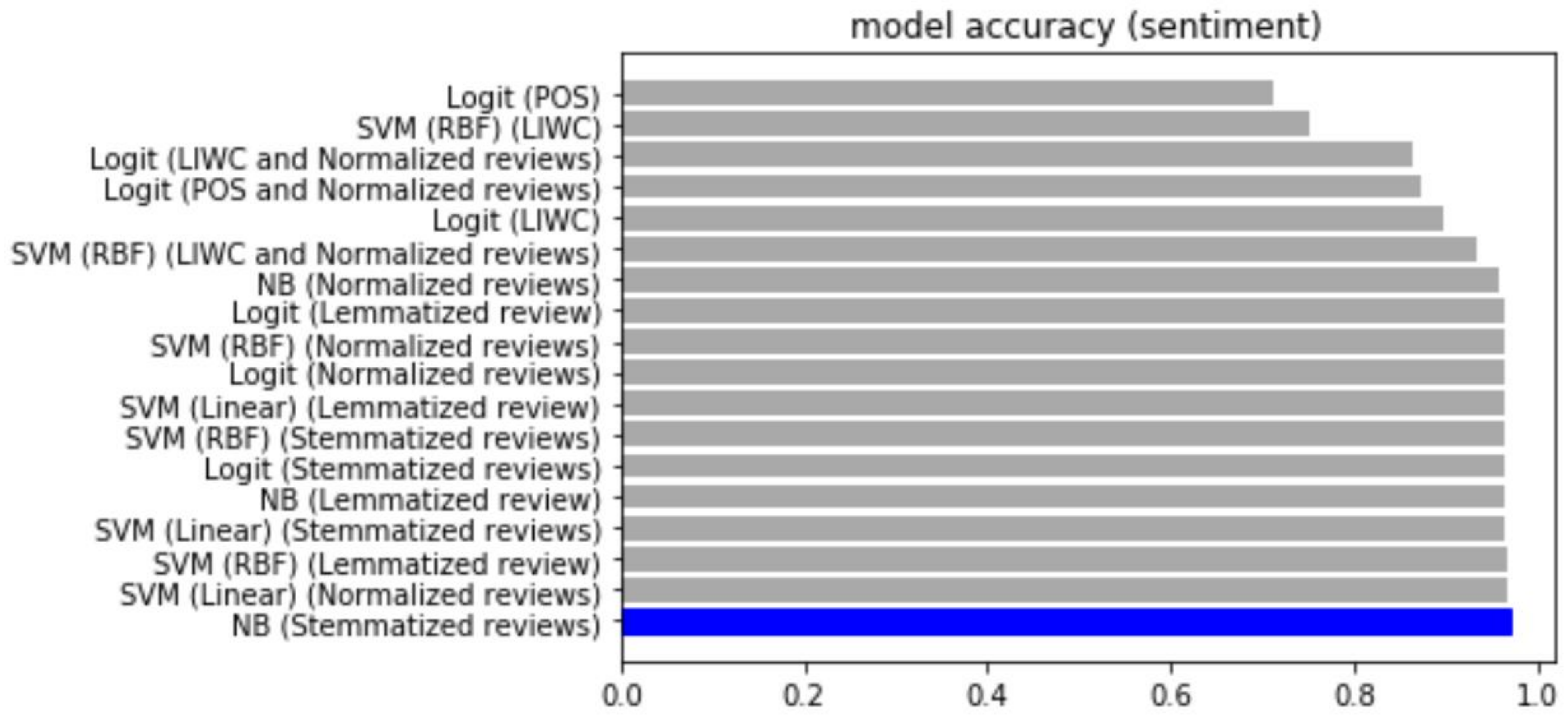


Deception		Test				Validation	
Input	Model	Accuracy	precision	recall	f1-score	f1 (crossvalid)	Training time
Normalized reviews	Logit	0.9	0.884	0.9208	0.902	0.878	70.462s
Stemmatized reviews	Logit	0.8833	0.8459	0.9375	0.8893	0.878	179.334s
Lemmatized review	Logit	0.9042	0.888	0.925	0.9061	0.88	306.583s
Normalized reviews	NB	0.8854	0.8715	0.9042	0.8875	0.883	42.409s
Stemmatized reviews	NB	0.8854	0.8627	0.9167	0.8889	0.886	125.702s
Lemmatized review	NB	0.8625	0.8372	0.9	0.8675	0.885	129.238s
Normalized reviews	**SVM (RBF)**	**0.9167**	**0.9032**	**0.9333**	**0.918**	**0.876**	**430.707s**
Stemmatized reviews	SVM (RBF)	0.8625	0.8508	0.8792	0.8648	0.873	1914.185s
Lemmatized review	SVM (RBF)	0.9021	0.8845	0.925	0.9043	0.882	1183.171s
Normalized reviews	SVM (Linear)	0.9083	0.9016	0.9167	0.9091	0.874	107.592s
Stemmatized reviews	SVM (Linear)	0.8979	0.8716	0.9333	0.9014	0.875	245.793s
Lemmatized review	SVM (Linear)	0.8979	0.8775	0.925	0.9006	0.879	340.068s
POS	Logit	0.6479	0.6282	0.725	0.6731	0.6932	3.133s
POS and Normalized reviews	Logit	0.8729	0.851	0.904	0.8768	0.867	78.691s
LIWC	Logit	0.7708	0.7462	0.8208	0.7817	0.7837	3.606s
LIWC and Normalized reviews	Logit	0.8688	0.8391	0.9125	0.8743	0.881	277.551s
LIWC	SVM (RBF)	0.7583	0.7583	0.7583	0.7583	0.7738	5.789s
LIWC and Normalized review	SVM (RBF)	0.8813	0.8617	0.9083	0.8844	0.8804	550.859s



Result - Deception Opinion

Sentimental		Test				Validation	
Input	Model	Accuracy	precision	recall	f1-score	f1 (crossvalid)	Training time
Normalized reviews	Logit	0.9625	0.9744	0.95	0.962	0.956	64.920s
Stemmatized reviews	Logit	0.9646	0.9705	0.9583	0.9644	0.96	332.299s
Lemmatized review	Logit	0.9625	0.9703	0.9542	0.9622	0.962	202.189s
Normalized reviews	NB	0.9583	0.9583	0.9583	0.9583	0.956	38.244s
Stemmatized reviews	**NB**	**0.9688**	**0.9747**	**0.9625**	**0.9686**	**0.959**	**125.014s**
Lemmatized review	NB	0.9646	0.9705	0.9583	0.9644	0.96	128.190s
Normalized reviews	SVM (RBF)	0.9625	0.9703	0.9542	0.9622	0.959	410.832s
Stemmatized reviews	SVM (RBF)	0.9646	0.9705	0.9583	0.9644	0.962	1101.111s
Lemmatized review	SVM (RBF)	0.9667	0.9706	0.9625	0.9665	0.963	1537.769s
Normalized reviews	SVM (Linear)	0.9667	0.9746	0.9583	0.9664	0.959	96.871s
Stemmatized reviews	SVM (Linear)	0.9646	0.9705	0.9583	0.9644	0.961	262.080s
Lemmatized review	SVM (Linear)	0.9625	0.9625	0.9625	0.9625	0.963	539.229s
POS	Logit	0.7125	0.7198	0.6958	0.7076	0.704	0.206s
POS and Normalized reviews	Logit	0.952	0.9697	0.9333	0.9512	0.944	81.215s
LIWC	SVM (RBF)	0.75	0.75	0.75	0.75	0.891	5.434s
LIWC and Normalized reviews	SVM (RBF)	0.933	0.913	0.958	0.935	0.935	406.423s
LIWC	Logit	0.8979	0.8996	0.8958	0.8977	0.908	0.645s
LIWC and Normalized reviews	Logit	0.9188	0.9102	0.9292	0.8648	0.867	108.592s



Result - Sentiment Analysis

Conclusions

1. Best result for deception was 0.9167 while sentiment analysis was 0.9688
2. Decent accuracy (0.89-0.91) were achieved for deceptive opinion spam with many based models using grid search cross validation
3. Feature Engineering (POS&LIWC) provided interesting insights when exploring significant features; but didn't improve results when combined with TFIDF n-gram
4. Deceptive opinion spams share some similar characteristics, like more descriptive and superfluous language, indications of imaginative writing

Future Works

1. Apply multi-class classification method
2. Different techniques and features selections could be explored; such as deep learning and embedding
3. Exploring more hyperparameters tuning, and changing weights of feature union
4. Refactor code to make it more streamlined

Reference

[1] M. Ott, Y. Choi, C. Cardie, and J.T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.

[2] M. Ott, C. Cardie, and J.T. Hancock. 2013. Negative Deceptive Opinion Spam. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.