CSML1010 – Milestone 1

Group 11 – Alex Fung and Patrick Osborne
Instructor – Dr. En-Shiun Annie Lee

Twitter US Airline – Sentiment Analysis

- **Milestone 1**
  - Feature Engineering
  - Feature Selection
  - Benchmarking

- **Feature Engineering**
  - Idea is to pull as much information out of the text as we can
  - Various ways of parsing and interpreting the text

- **Feature Selection**
  - Too many features – will overwhelm our model. Need to keep the best and discard the rest

Feature Engineering – Basic Methods

– Bag of Words

	00	000	000114	000lb	00a	00am	00p	00pm	01	01pm	...	zambia	zcc82u	zero	zig	zip	zipper	zone	zoom	zuke	zurich
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
14635	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
14636	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
14637	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
14638	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
14639	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

14640 rows × 8669 columns

– Bag of N-Grams

	00 27	00 bag	00 check	00 don	00 flight	00 goodwill	00 happy	00 phone	00 pm	00 say	...	zone precious	zone space	zone thank	zoom sauce	zoom scroll	zuke non	zurich bc	zurich credit	zurich jfk	zurich new
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
14635	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
14636	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
14637	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
14638	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
14639	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

14640 rows × 62070 columns

YORK U UNIVERSITÉ UNIVERSITY

Feature Engineering – Basic Methods

– TF-IDF

	00	000	000114	000lb	00a	00am	00p	00pm	01	01pm	...	zambia	zcc82u	zero	zig	zip	zipper	zone	zoom	zuke	zurich
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
14635	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14636	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14637	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14638	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14639	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

14640 rows × 8669 columns

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

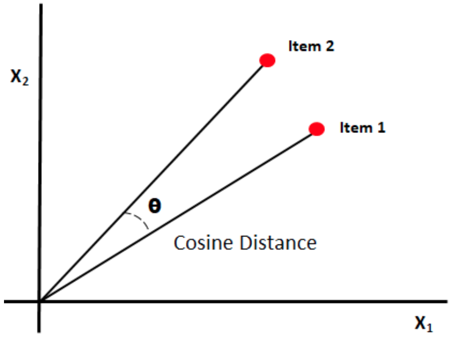
$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

– Cosine Similarity

	0	1	2	3	4	5	6	7	8	9	...	14630	14631	14632	14633	14634	14635	14636	14637	14638	14639
0	1.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.00000	0.000000	0.0	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.000000
1	0.0	1.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.00000	0.000000	0.0	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.000000
2	0.0	0.0	1.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.00000	0.000000	0.0	0.084757	0.0	0.000000	0.000000	0.0	0.000000	0.114465
3	0.0	0.0	0.000000	1.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.00000	0.000000	0.0	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.000000
4	0.0	0.0	0.000000	0.0	1.0	0.307344	0.0	0.0	0.0	0.0	...	0.00000	0.000000	0.0	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
14635	0.0	0.0	0.000000	0.0	0.0	0.028866	0.0	0.0	0.0	0.0	...	0.30427	0.188921	0.0	0.052285	0.0	1.000000	0.046479	0.0	0.033641	0.070611
14636	0.0	0.0	0.000000	0.0	0.0	0.026716	0.0	0.0	0.0	0.0	...	0.00000	0.000000	0.0	0.098288	0.0	0.046479	1.000000	0.0	0.031135	0.065351
14637	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.00000	0.000000	0.0	0.000000	0.0	0.000000	0.000000	1.0	0.000000	0.000000
14638	0.0	0.0	0.000000	0.0	0.0	0.019337	0.0	0.0	0.0	0.0	...	0.00000	0.000000	0.0	0.035024	0.0	0.033641	0.031135	0.0	1.000000	0.047300
14639	0.0	0.0	0.114465	0.0	0.0	0.116877	0.0	0.0	0.0	0.0	...	0.00000	0.000000	0.0	0.130844	0.0	0.070611	0.065351	0.0	0.047300	1.000000

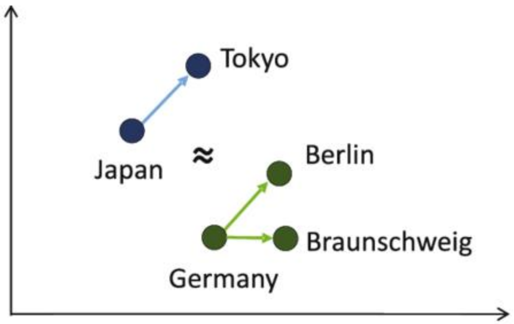
14640 rows × 14640 columns

Cosine Distance/Similarity

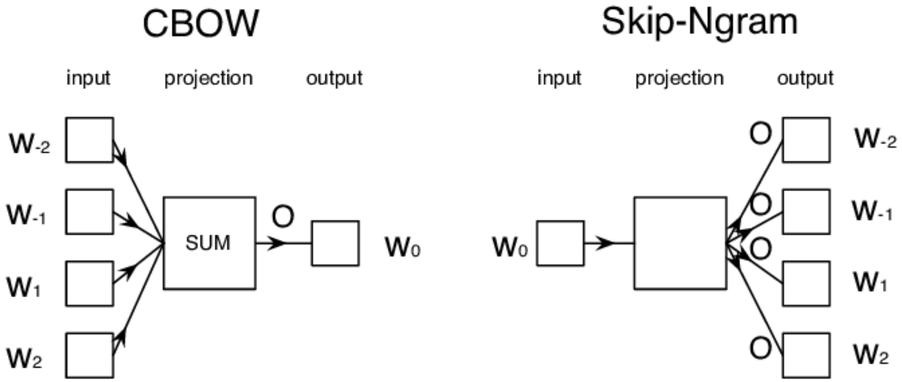


Feature Engineering – Word2Vec

- **Word2Vec**
  - A form of word embedding
  - Represent individual words in a way that similar words are represented in similar ways
  - Generates numerical vectors



- **CBOW vs Skip-gram**
  - Frequency/speed/accuracy trade-offs



Feature Engineering – Word2Vec - Models

- **Tensorflow / Keras Trained**
  - Manual process
  - Slow even when GPU accelerated

- **Gensim Trained**
  - Very fast
  - Possibly less fine-grained control

- **Google Pre-Trained**
  - Fastest
  - Advantage of training on existing corpus
  - Cannot custom train to our dataset

Feature Engineering – GloVe

- **GloVe Word Embedding**
  - Global Vectors for Word Representation
  - An alternate model

	0	1	2	3	4	5	6	7	8	9	...	290	291	292
kickin	0.075017	-0.023001	0.051686	-0.367610	0.43434	0.433910	-0.093183	-0.159130	-0.027413	-0.11179	...	0.33969	-0.911680	-0.408490
lite	0.217430	-0.281690	0.400900	0.065049	0.12261	0.286080	0.427170	-0.229840	0.510550	-0.42872	...	-0.43788	-0.061199	0.033553
click	0.124620	-0.000954	-0.340720	0.278290	0.12420	-0.186700	-0.084583	-0.385110	0.006960	0.92982	...	-0.13069	0.272240	0.197760
electronic	-0.117840	0.131870	0.450020	0.249900	0.15860	-0.024728	-0.442390	-1.149400	-0.503550	0.98614	...	-0.87140	-0.218680	0.429410
7:50pm	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.00000	...	0.00000	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
undelayed	0.307060	-0.434910	0.004983	-0.154580	-0.24499	0.635660	-0.212920	0.082008	-0.116720	-1.34320	...	0.13149	0.336960	-0.095300
range	-0.138620	0.629800	0.253930	-0.157670	0.77029	-0.378440	0.177030	0.343710	0.051385	1.14790	...	-0.30940	0.392230	-0.102660
20hrs	0.650210	0.443270	-0.015427	-0.123740	-0.10975	-0.141080	-0.289020	0.069521	0.173960	-1.17750	...	0.29694	-0.153780	-0.458250
2/23/15	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.00000	...	0.00000	0.000000	0.000000
ticket(s	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.00000	...	0.00000	0.000000	0.000000

9541 rows × 300 columns

Benchmarking – Logistic Regression

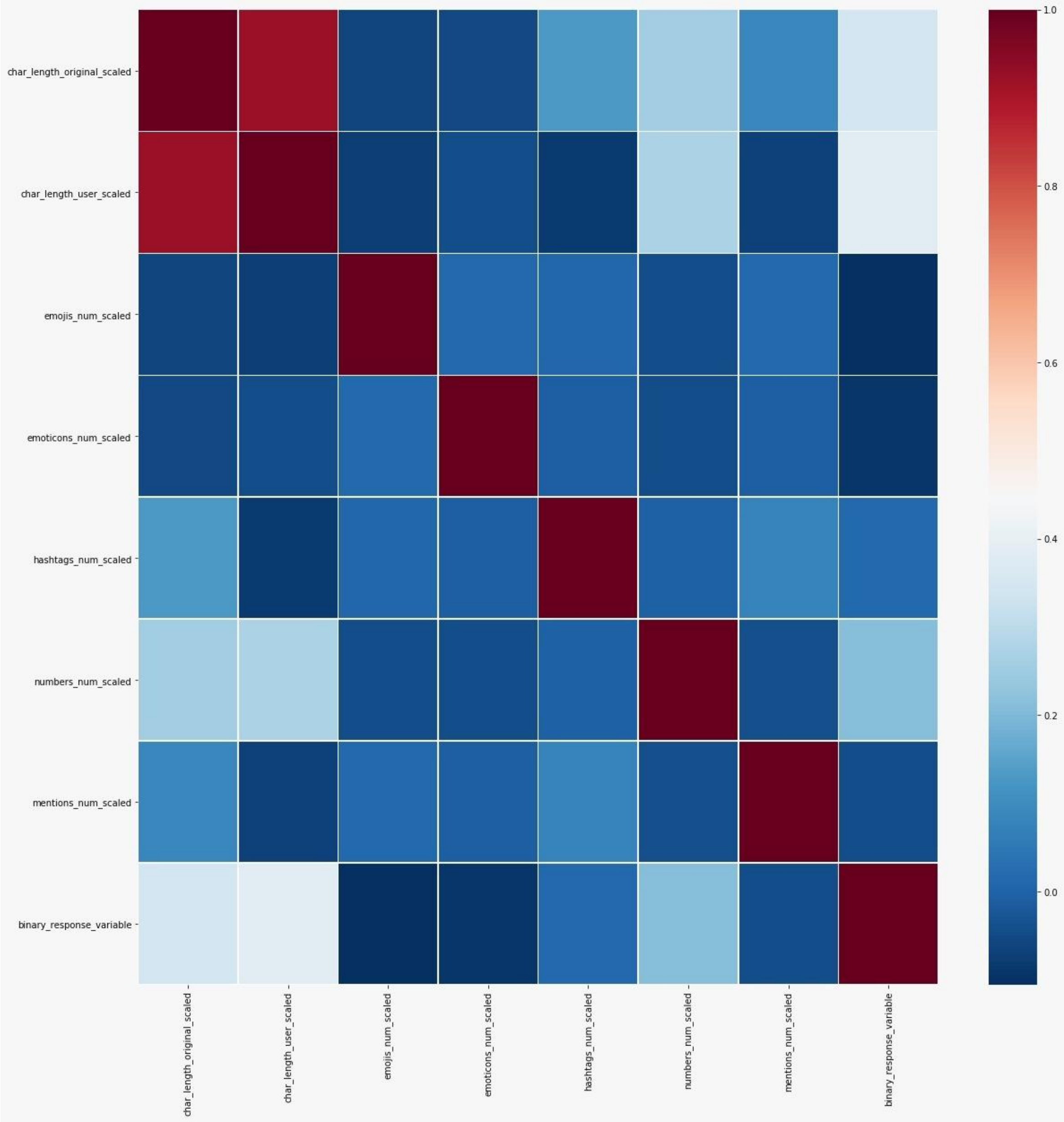
- **Gensim Word2Vec Word Embeddings**

	precision	recall	f1-score	support
0	0.79	0.71	0.75	1639
1	0.84	0.89	0.86	2753
micro avg	0.82	0.82	0.82	4392
macro avg	0.81	0.80	0.81	4392
weighted avg	0.82	0.82	0.82	4392

- **Google Word2Vec Pre-Trained Word Embeddings**

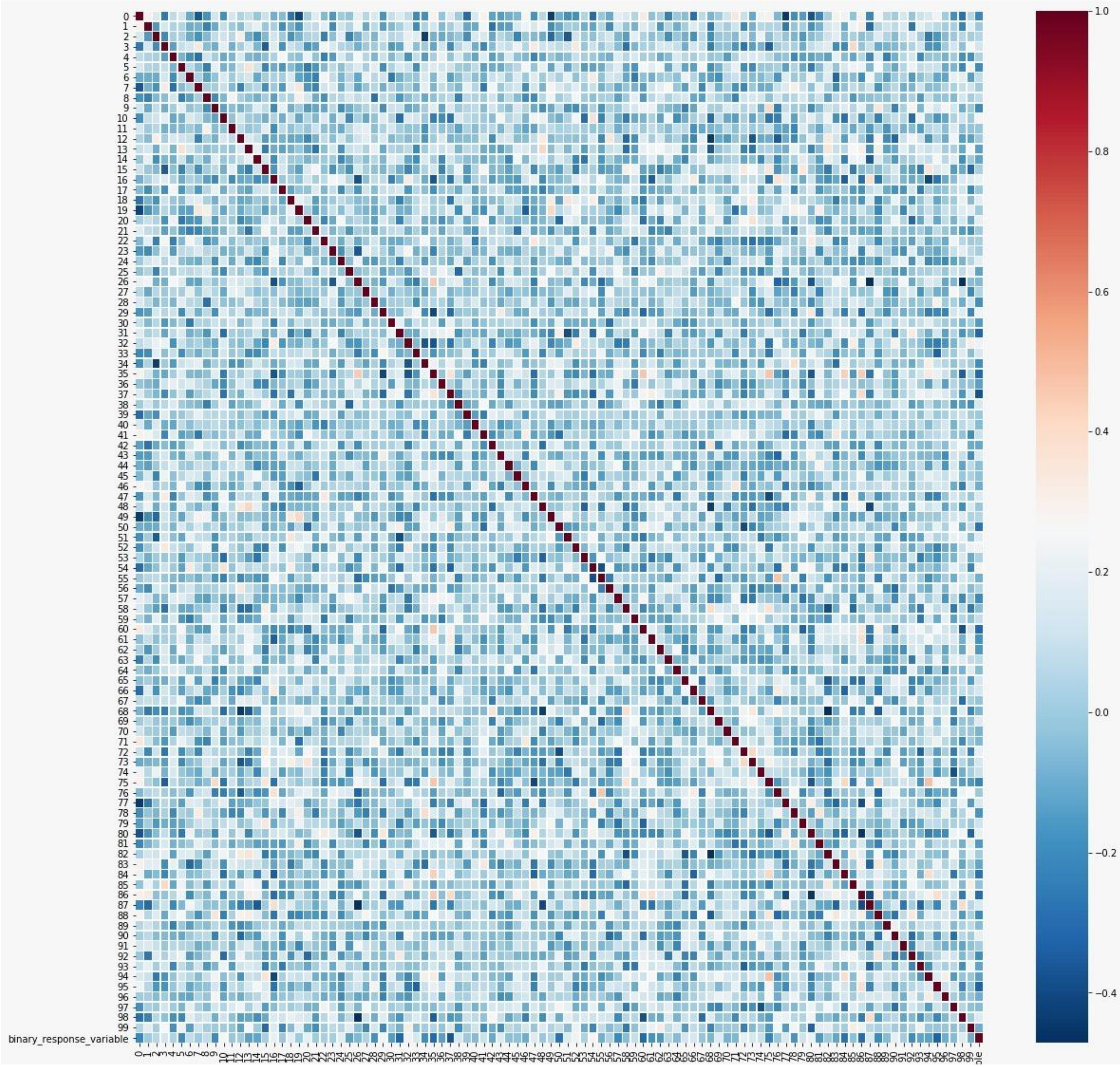
	precision	recall	f1-score	support
0	0.79	0.70	0.74	1639
1	0.83	0.89	0.86	2753
micro avg	0.82	0.82	0.82	4392
macro avg	0.81	0.80	0.80	4392
weighted avg	0.82	0.82	0.82	4392

Feature Selection

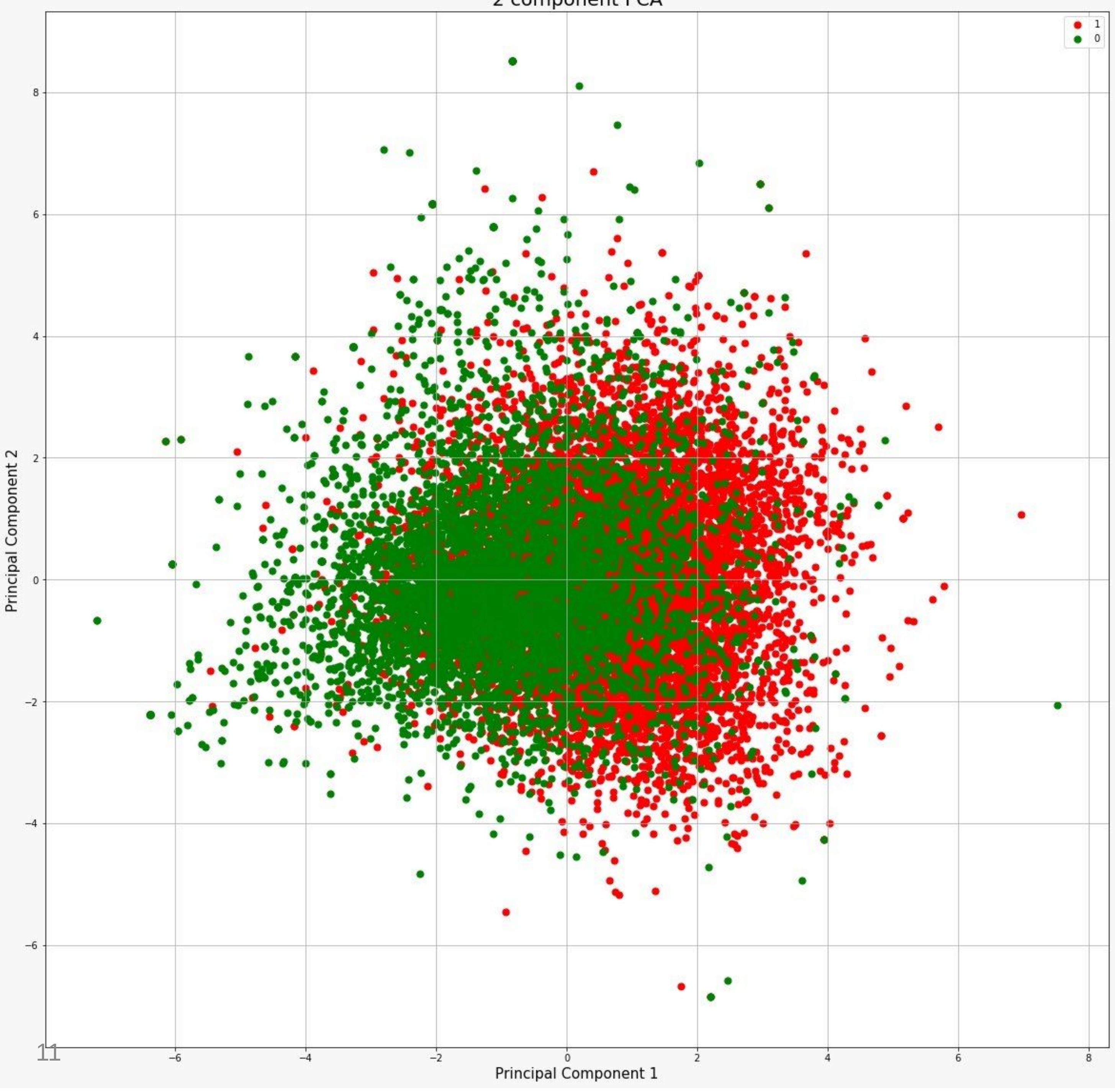


Correlation Matrix of Engineered Features

Feature Selection



Correlation Matrix of Gensim word2vec



Principal Component Analysis

(Dimensionality Reduction)

Feature Selection

- Top 10 Features Ranked

```
[(1, '28'),
 (1, '31'),
 (1, '32'),
 (1, '37'),
 (1, '66'),
 (1, '81'),
 (1, '95'),
 (1, '99'),
 (1, 'char_length_user_scaled'),
 (1, 'emoticons_flag'),
 (2, '44'),
 (3, '10'),
 (4, '73'),
 (5, '9'),
 (6, '2'),
 (7, '63'),
 (8, 'mentions_num_scaled'),
 (9, '69'),
 (10, '74'),
```

- Benchmark Results

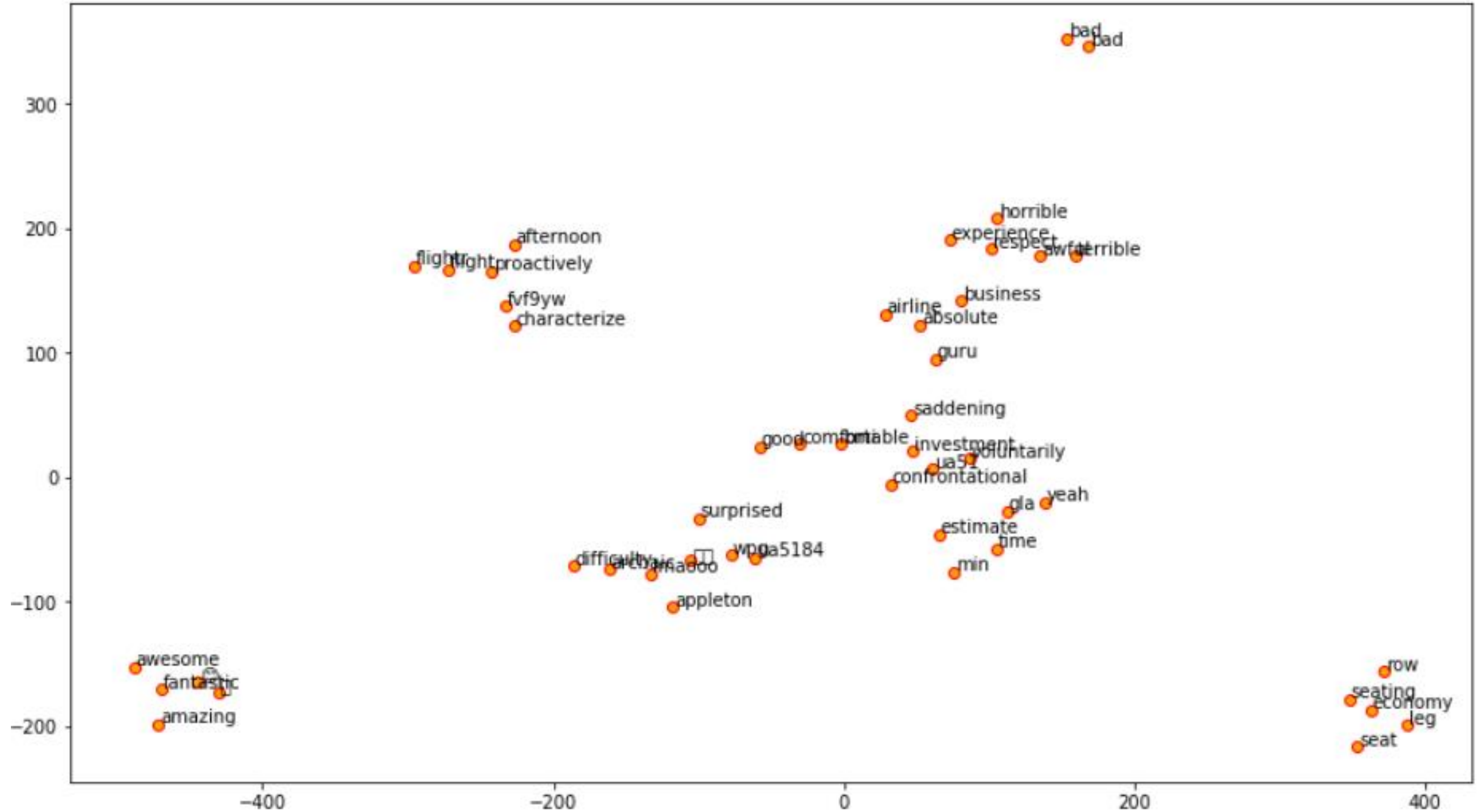
	precision	recall	f1-score	support
0.0	0.76	0.63	0.69	5462
1.0	0.80	0.88	0.84	9178
micro avg	0.79	0.79	0.79	14640
macro avg	0.78	0.76	0.76	14640
weighted avg	0.79	0.79	0.78	14640

Recursive Feature Elimination

YORK UNIVERSITÉ UNIVERSITY

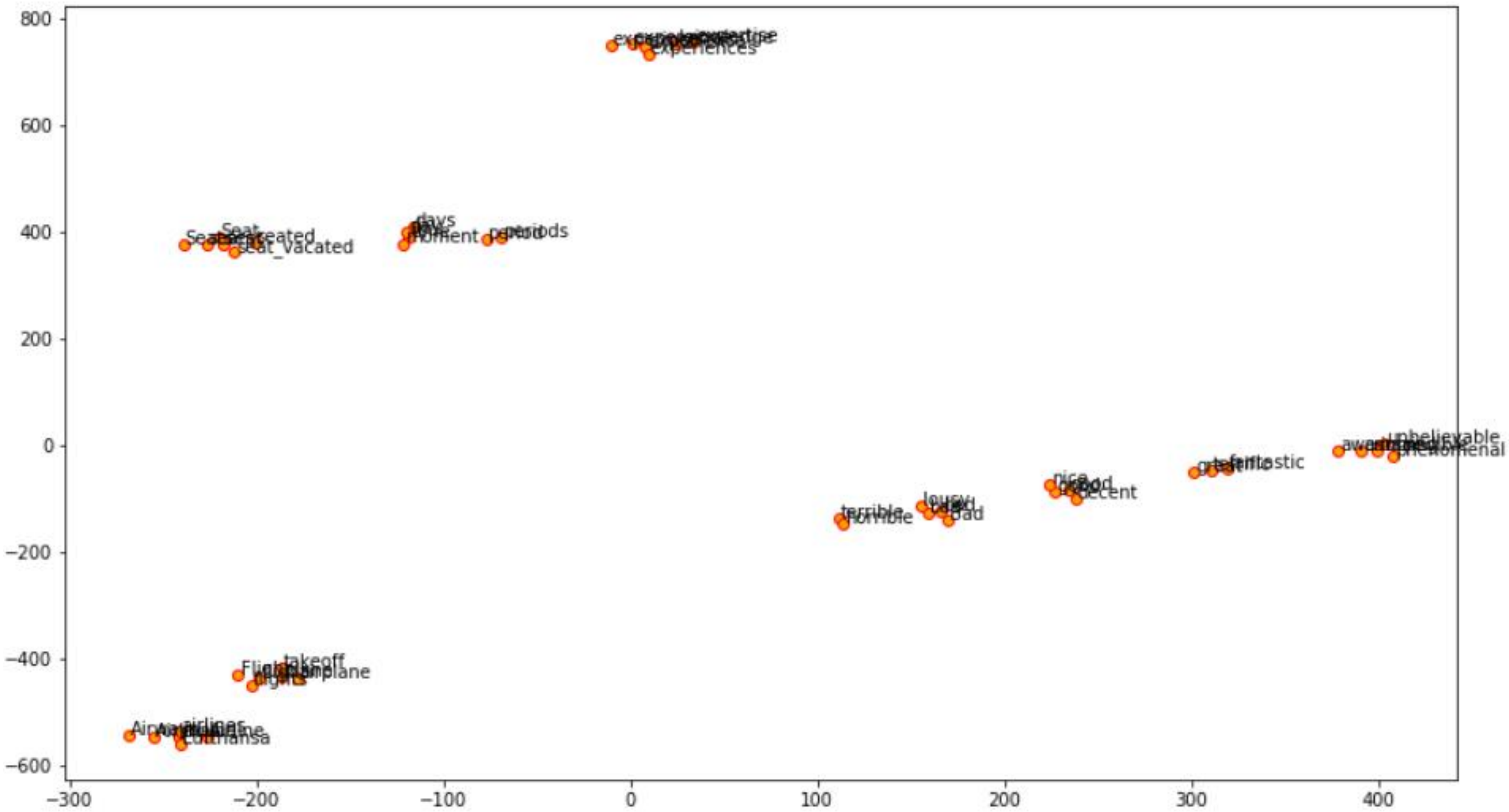
Feature Exploration

– **Gensim Word2Vec**



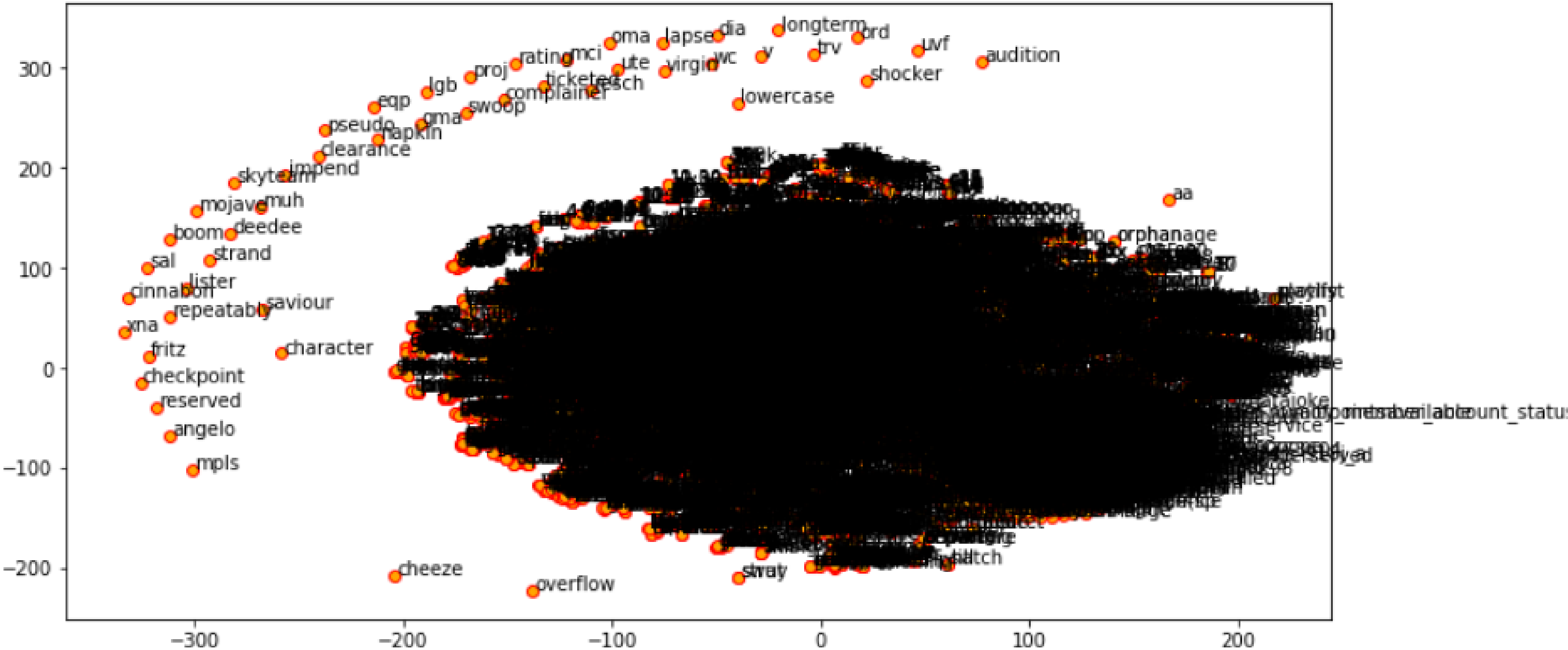
Feature Exploration

- **Google Word2Vec**



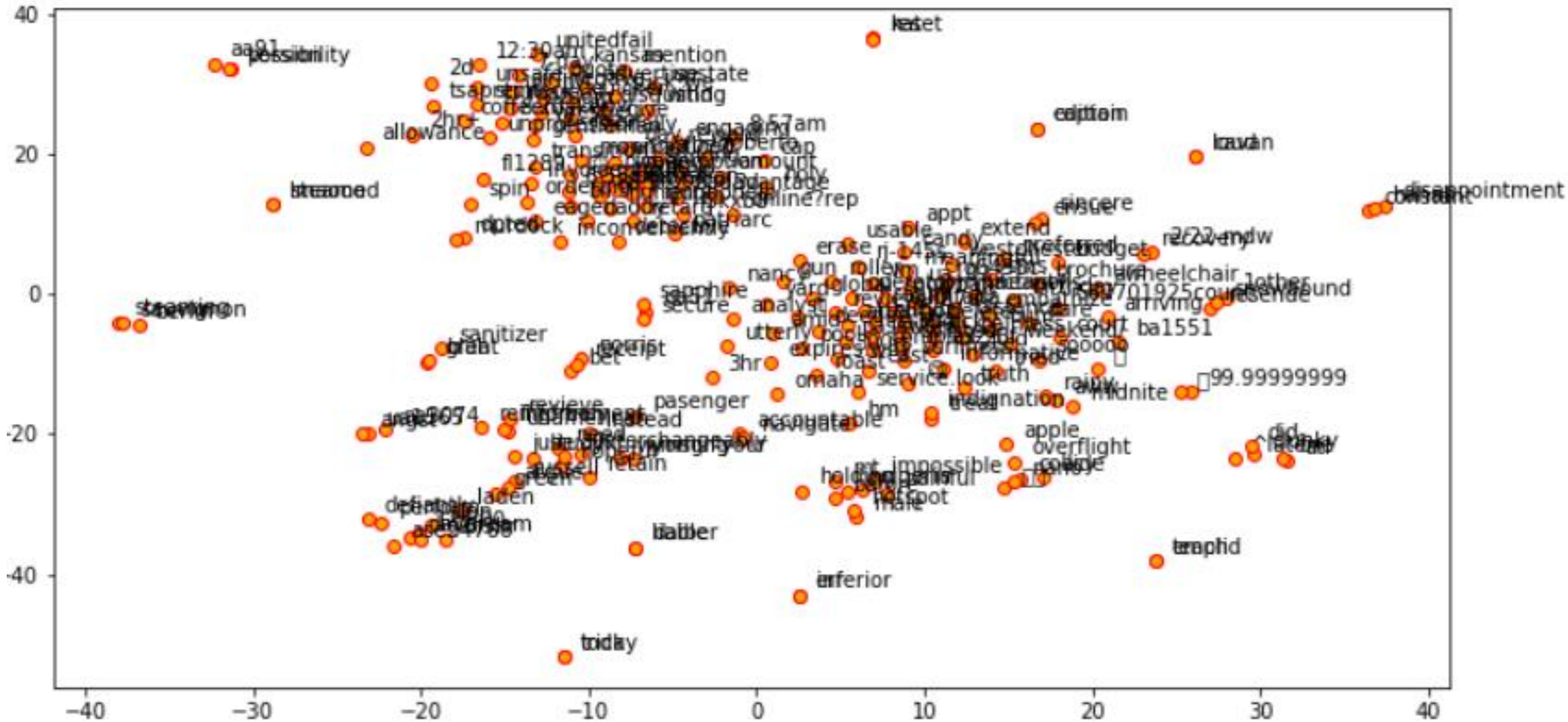
Feature Exploration

– **GloVe**
  ▪ All word embeddings



Feature Exploration

- **GloVe**
  - 250 item sample



Thank You!