Project: Conversation Dialogue Identification

CSML1010 – Milestone 1 Group 3: Jerry Khidaroo, Paul Doucet

Instructor: Dr. Annie En-Shiun Lee

CSML1010 – Milestone 1

- Dataset: Taskmaster-1 from Google
- NLP Multi-Class Text Classification Problem
- Data Preparation
- Data Clean Up
- Exploratory Data Analysis
- **Feature Extraction & Engineering**
- **Feature Scaling & Selection**
- Modeling
- Model Evaluation & Tuning

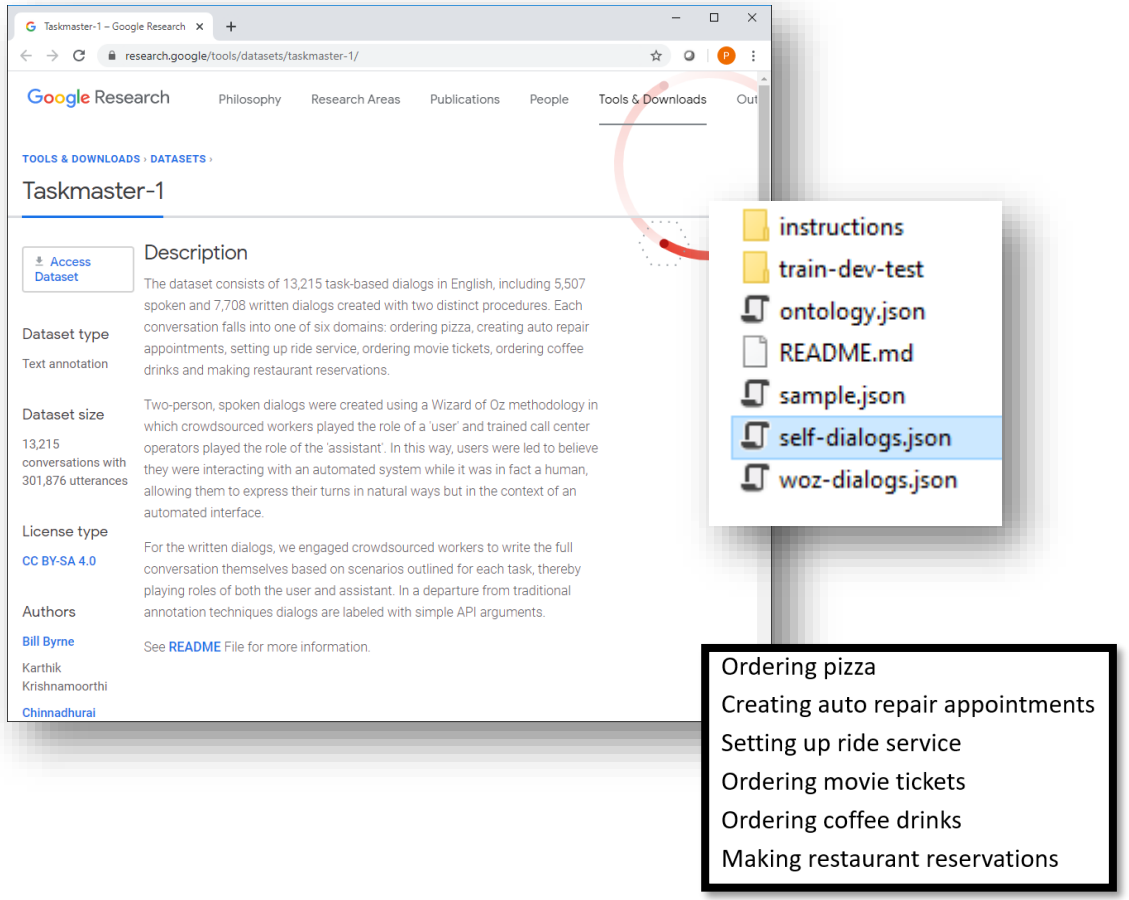
Dataset: Taskmaster-1 from Google

https://research.google/tools/datasets/taskmaster-1/

The dataset selected is the Taskmaster-1 from Google

The dataset consists of task-based dialogs falling into one of six domains divided into 14 categories

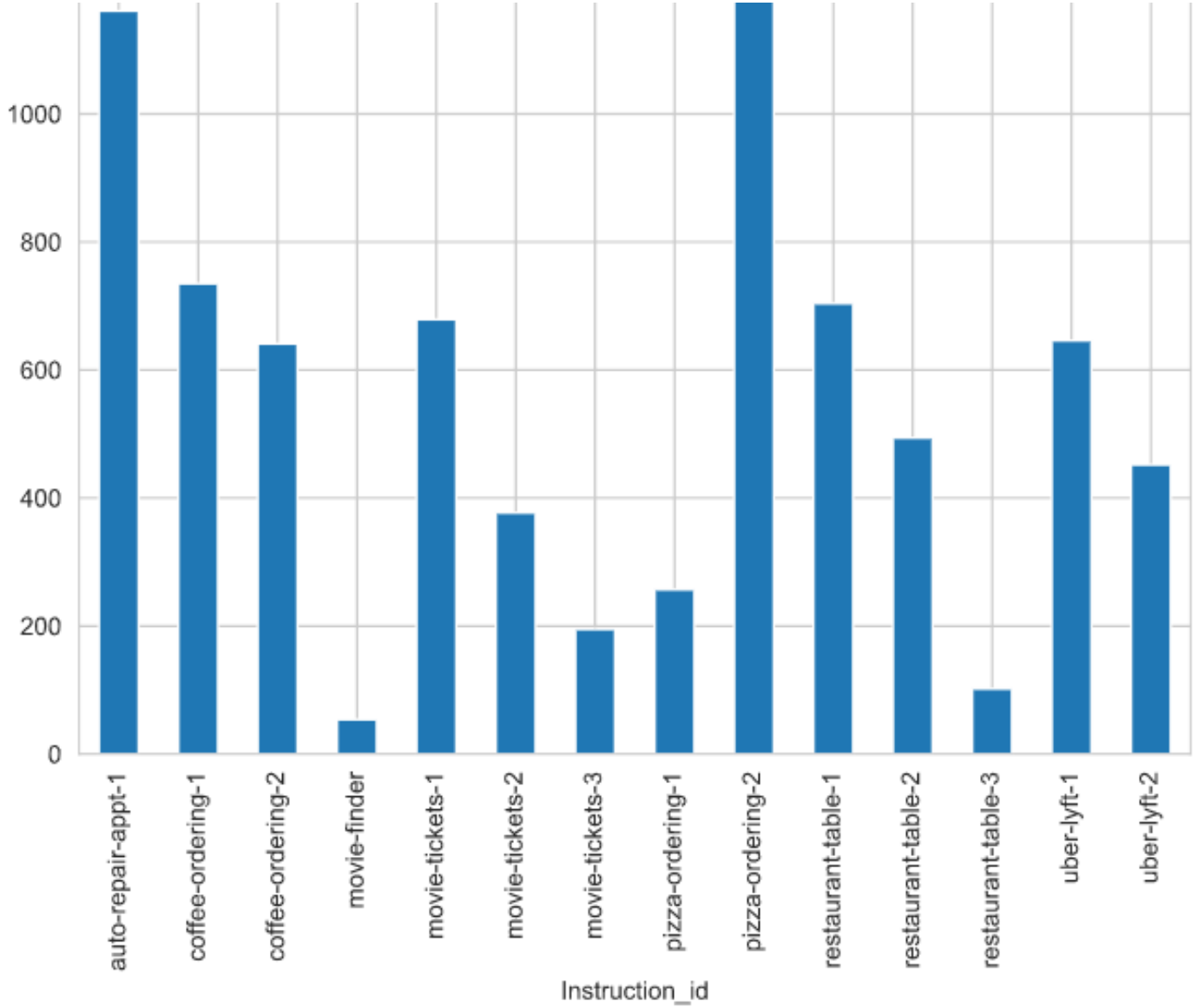
We will be using the self-dialogs file which contains 7,708 conversations: **‘self-dialogs.json’**



Description of Categorical Variable (**Instruction_id**)

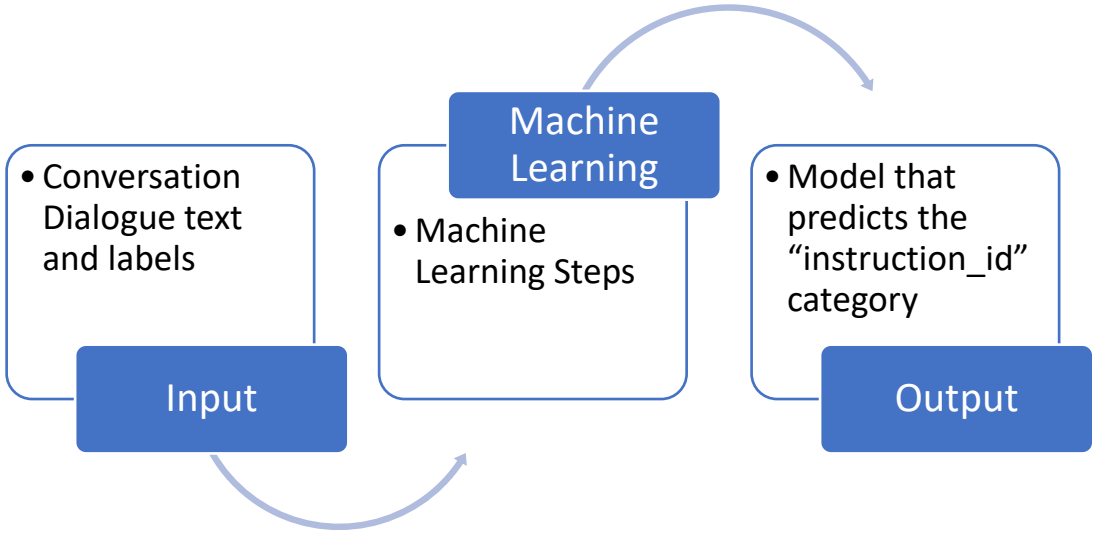
Instruction_id	Conversation Dialogue Description
Auto-repair-appt-1	Users will pretend they need to take their car to the mechanic, so they need to get an appointment scheduled
Coffee-ordering-1	Users will pretend they've decided to order a coffee drink from a coffee shop
Coffee-ordering-2	Users will pretend they've decided to order a coffee drink, and makes changes to the drink after the initial options have been requested
movie-finder	User is looking for a movie to see at home
movie-tickets-1	User wants to see a movie playing now
movie-tickets-2	User wants to see a movie playing now, settling for a second choice
movie-tickets-3	User wants to see one of two movies
Pizza-Ordering-1	User orders one pizza, and ask all relevant details
Pizza-Ordering-2	User orders one pizza with two toppings, and ask all relevant details
restaurant-table-1	Users will pretend they are searching for a restaurant and book a table
restaurant-table-2	Users will pretend they are searching for a restaurant and book a table, and will need to find an alternative when choice is not available
restaurant-table-3	Users will pretend they are searching for a restaurant and book a table, and will look at options at two restaurants
uber-lyft-1	Users will pretend they need to order a car for a ride inside a city
uber-lyft-2	Users will pretend they need to order a car for a ride inside a city, and looking for an alternative when choice is not available

Distribution of Categorical Variable (Instruction_id)



Project: Conversation Identification

- The problem we will examine is a supervised multi-class text classification problem.
- **Goal:** *Build a model that identifies the category for dialogue conversations*
  - **Input:** conversation dialogue text and labels
  - **Output:** model that predicts the “instruction_id” category



Feature Extraction and Selection Summary

Vector Types	Feature Extracted	Feature Selection Method
Count Vectors	Bag-of-words	Univariate Chi^2
	Bag of n-grams	Univariate Chi^2
	Bag-of-words	PCA
Combination Vector	Bag-of-words + Bag of n-grams	Univariate Chi^2
Word Level	TF-IDF	Univariate Chi^2
Word Embeddings	Word2Vec from Word2Vec model	Univariate Chi^2
	Word2Vec from FastText model	Univariate Chi^2
	GloVe Embeddings with Flair	Univariate Chi^2

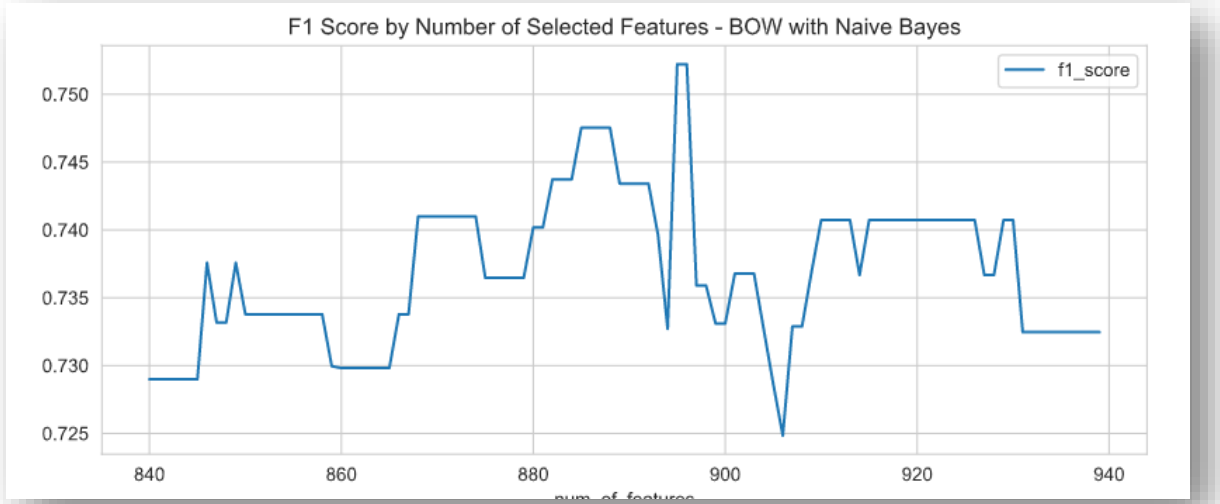
Bag of Words

Feature Extraction

	PAD	like	would	ok	yes	okay	pm	want	order	thank	tickets	please	one	time
0	0	0	0	10	6	0	1	3	0	2	0	0	0	0
1	0	1	0	2	4	0	1	1	0	7	0	3	0	2
2	0	2	3	0	1	1	4	0	0	1	2	3	1	0
3	0	0	0	0	0	4	0	0	0	2	0	0	0	1
4	0	0	0	0	1	1	0	2	2	0	0	2	3	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	0	4	1	2	1	5	0	0	2	3	0	5	0	0
996	0	1	0	0	1	1	5	0	0	1	0	0	0	0
997	0	0	0	0	2	0	0	1	1	2	0	1	0	0
998	0	3	1	4	3	3	2	5	0	1	1	2	4	2
999	0	6	5	0	2	4	8	0	1	0	0	1	5	2

1000 rows × 5805 columns

Feature Selection – Univariate with Chi^2



Benchmarking

	Features_Benchedmarked	Precision	Recall	f1_score	accuracy
0	BOW Naive Bayes Baseline	0.68	0.62	0.60	0.62
1	BOW Naive Bayes Optimal Features Selected: 896	0.79	0.75	0.75	0.75

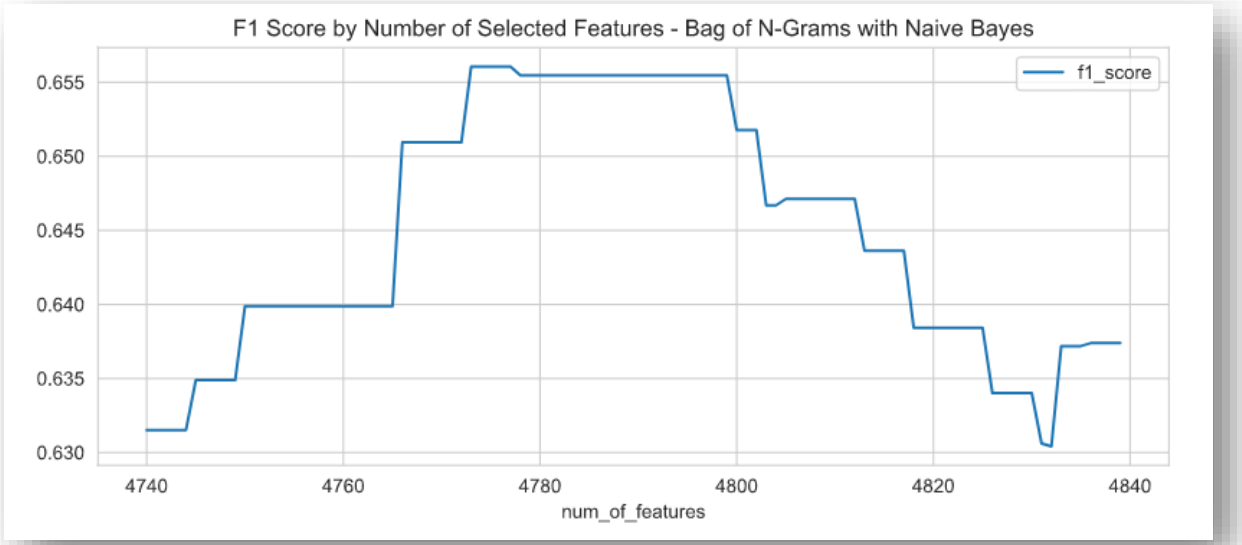
Bag of n-Grams

Feature Extraction

	abby normal	abe louies	abgout second	able accomadate	able accommodate	able arrange	able book	able call	able choose	able confirm	able get	able go
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
2	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
4	0	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
745	0	0	0	0	0	0	0	0	0	0	0	0
746	0	0	0	0	0	0	0	0	0	0	0	0
747	0	0	0	0	0	0	0	0	0	0	0	0
748	0	0	0	0	0	0	0	0	0	0	0	0
749	0	0	0	0	0	0	0	0	0	0	0	0

750 rows × 36335 columns

Feature Selection – Univariate with Chi^2



Benchmarking

	Features_Benchedmarked	Precision	Recall	f1_score	accuracy
0	Bag of N-Gram Naive Bayes baseline	0.70	0.64	0.62	0.64
1	Bag of N-Gram Naive Bayes Optimal Features Selected: 4774	0.69	0.66	0.66	0.66

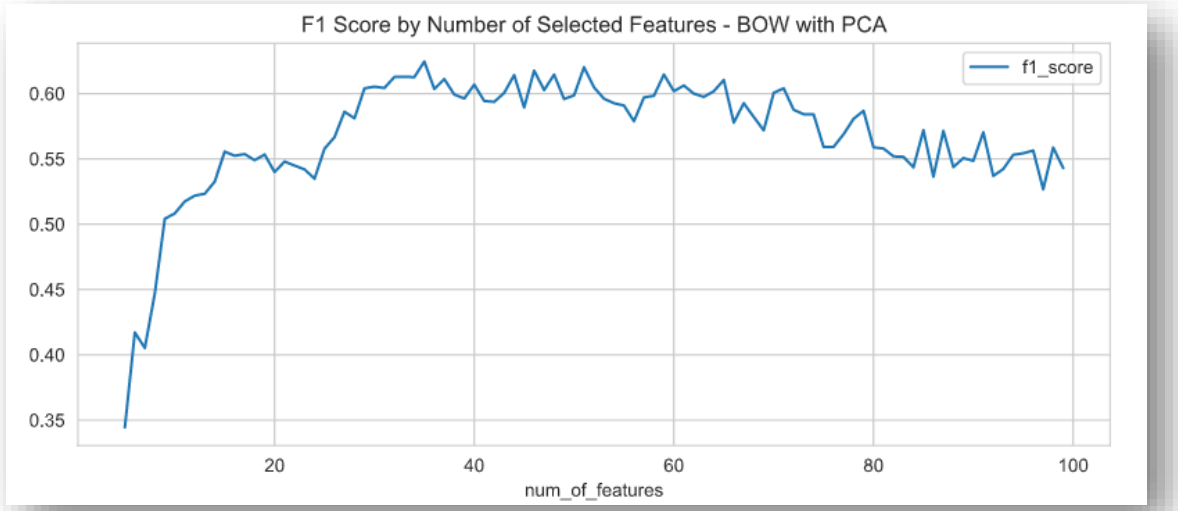
PCA on Bag of Words

Feature Extraction

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	1.37	-1.61	-0.65	0.40	-2.47	-1.83	0.66	0.93	0.35	0.45	1.39	1.58	-0.88
1	-2.02	-2.62	0.33	-1.64	-0.99	-0.85	-2.02	0.87	2.24	-2.80	-0.42	0.22	-0.51
2	-3.09	3.23	1.07	-2.81	0.11	-0.45	-0.94	-0.00	-0.34	0.01	-1.27	0.09	-2.37
3	2.66	0.69	-2.27	2.91	-3.35	3.07	-1.85	-1.83	2.40	2.49	1.14	-0.56	-0.33
4	-2.38	-4.20	-1.04	0.02	-0.68	0.32	-0.51	-0.28	-0.03	-1.49	0.60	-0.13	0.46
...	...	...	...	...	...	...	...	...	...	...	...	...	...
745	-3.07	-0.57	0.11	-3.86	-0.67	-1.92	-1.83	0.87	1.02	0.05	-0.16	-0.69	0.69
746	-1.78	-1.48	2.56	1.75	-0.34	3.39	0.92	3.44	0.29	-1.81	-0.19	-0.40	0.18
747	-3.72	1.31	-4.59	-3.17	2.58	0.59	4.80	-1.17	0.02	2.34	0.15	0.58	-0.36
748	-1.90	-4.91	-2.57	-0.83	0.13	1.19	2.24	-2.61	0.83	0.23	0.04	-2.68	1.55
749	0.58	5.87	-5.46	-4.87	0.16	0.72	3.34	-0.25	1.99	3.47	0.84	-0.67	1.92

750 rows × 38 columns

Feature Selection – PCA



Benchmarking

	Features_Benchedmarked	Precision	Recall	f1_score	accuracy
0	BOW and Bag of N-Grams Combined Baseline	0.70	0.62	0.59	0.62
1	BOW + Bag of NGrams Top: 1000 Features with Naive Bayes	0.77	0.74	0.74	0.74
2	BOW With Top: 35 PCA Components Seleted	0.63	0.61	0.61	0.61

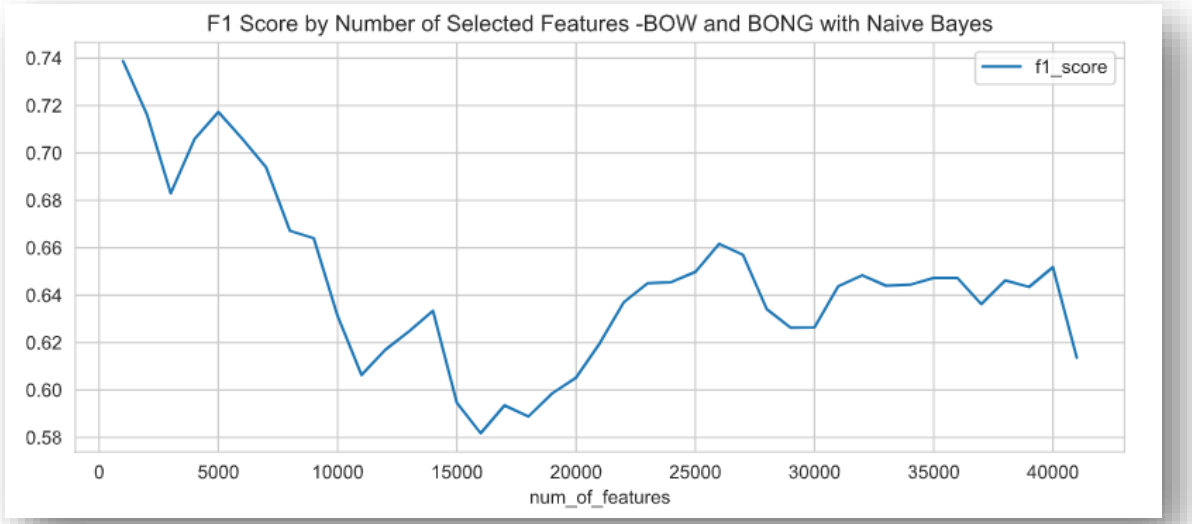
Combination – BOW and Bag of nGrams

Feature Extraction

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0.00	1.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	2.00	1.00	0.00	0.00
1	0.00	1.00	3.00	0.00	3.00	0.00	0.00	4.00	0.00	1.00	0.00	0.00	2.00
2	0.00	2.00	1.00	3.00	1.00	0.00	0.00	3.00	1.00	0.00	0.00	1.00	0.00
3	0.00	5.00	4.00	4.00	3.00	0.00	6.00	0.00	0.00	1.00	0.00	2.00	0.00
4	0.00	2.00	1.00	4.00	2.00	0.00	0.00	3.00	0.00	0.00	0.00	0.00	1.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...
245	0.00	2.00	2.00	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	2.00	0.00
246	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
247	0.00	2.00	3.00	0.00	1.00	1.00	4.00	0.00	0.00	1.00	2.00	3.00	1.00
248	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
249	0.00	4.00	4.00	0.00	2.00	5.00	0.00	0.00	5.00	2.00	0.00	0.00	2.00

250 rows × 42140 columns

Feature Selection – Univariate with Chi^2



Benchmarking

	Features_Benchedmarked	Precision	Recall	f1_score	accuracy
0	BOW and Bag of N-Grams Combined Baseline	0.70	0.62	0.59	0.62
1	BOW + Bag of NGrams Top: 1000 Features with Naive Bayes	0.77	0.74	0.74	0.74

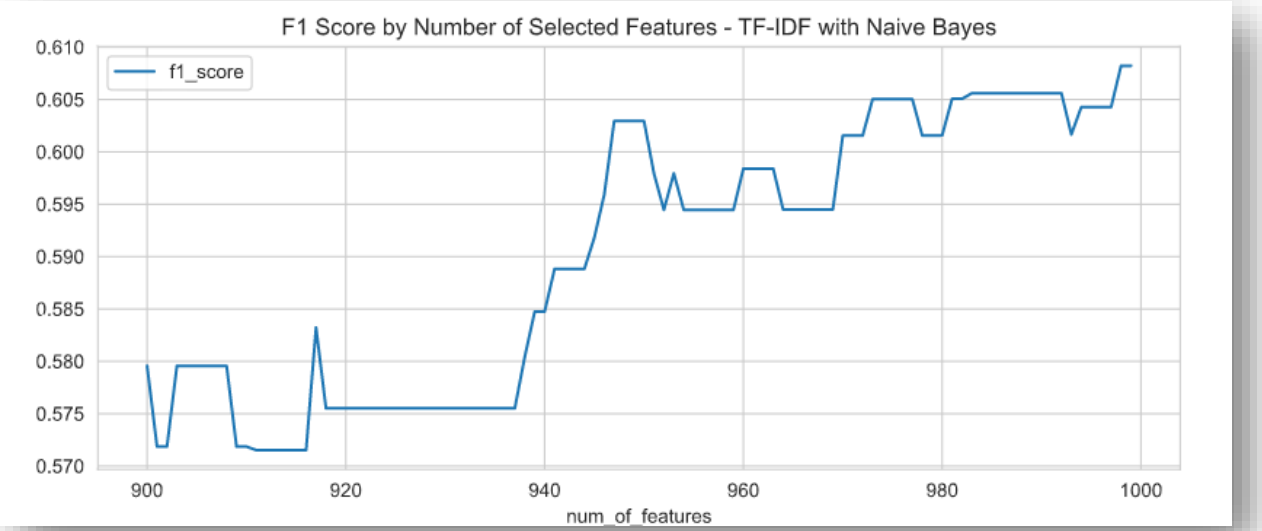
TF-IDF

Feature Extraction

	abbott	able	abosuolutely	abotu	abrams	absolutely	abyss	ac	academy	accelerate	accent
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
...	...	...	...	...	...	...	...	...	...	...	...
745	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
746	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
747	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
748	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
749	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

750 rows × 5018 columns

Feature Selection – Univariate with Chi^2



Benchmarking

	Features_Benchedmarked	Precision	Recall	f1_score	accuracy
0	TF-IDF Naive Bayes Baseline	0.54	0.56	0.50	0.56
1	TF-IDF Naive Bayes Optimal Features Selected: 999	0.74	0.64	0.61	0.64

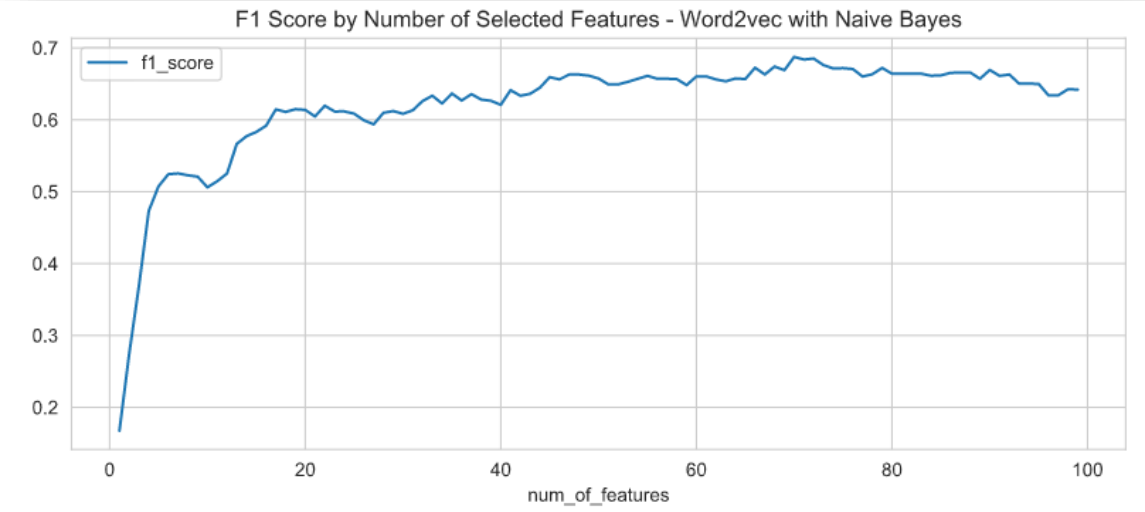
Word2Vec

Feature Extraction

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0.28	-0.62	-0.28	0.36	0.78	0.37	-0.23	-0.67	0.20	-0.53	0.01	-0.26	-0.77
1	0.31	-0.59	0.30	-0.64	-1.04	0.68	0.25	-0.07	-0.88	-0.27	-0.52	0.09	0.04
2	0.89	-1.03	1.05	-0.15	0.91	-0.25	-0.15	-0.16	-0.54	-0.61	-0.16	0.87	-0.07
3	0.46	-0.32	-0.53	-0.28	0.05	0.26	0.38	-0.02	0.21	0.17	0.11	1.07	-0.18
4	0.12	0.12	-0.46	-0.21	-0.09	-0.17	0.05	-0.07	0.16	0.23	0.26	1.25	-0.11
...	...	...	...	...	...	...	...	...	...	...	...	...	...
745	0.61	-1.87	1.25	-0.97	0.40	0.04	0.09	0.24	-1.06	-0.71	-0.20	0.29	-0.28
746	-0.11	-0.49	0.05	-0.31	0.51	-0.33	0.02	-0.05	-0.71	-0.04	0.38	0.47	0.29
747	1.09	-1.52	1.21	-0.56	0.49	0.35	-0.05	-0.34	-0.65	-0.63	-1.07	0.60	0.12
748	0.89	-0.80	-0.56	0.67	0.20	0.57	0.71	-0.27	-0.04	0.23	-0.32	1.22	-0.79
749	0.43	-1.15	0.94	-0.15	0.74	0.10	-0.35	0.11	-0.53	-1.06	-0.39	0.30	-0.09

750 rows × 100 columns

Feature Selection – Univariate with Chi^2



Benchmarking

	Features_Benchedmarked	Precision	Recall	f1_score	accuracy
0	Word2Vec Naive Bayes Baseline	0.68	0.65	0.64	0.65
1	Word2Vec Naive Bayes Optimal Features Selected: 70	0.71	0.69	0.69	0.69

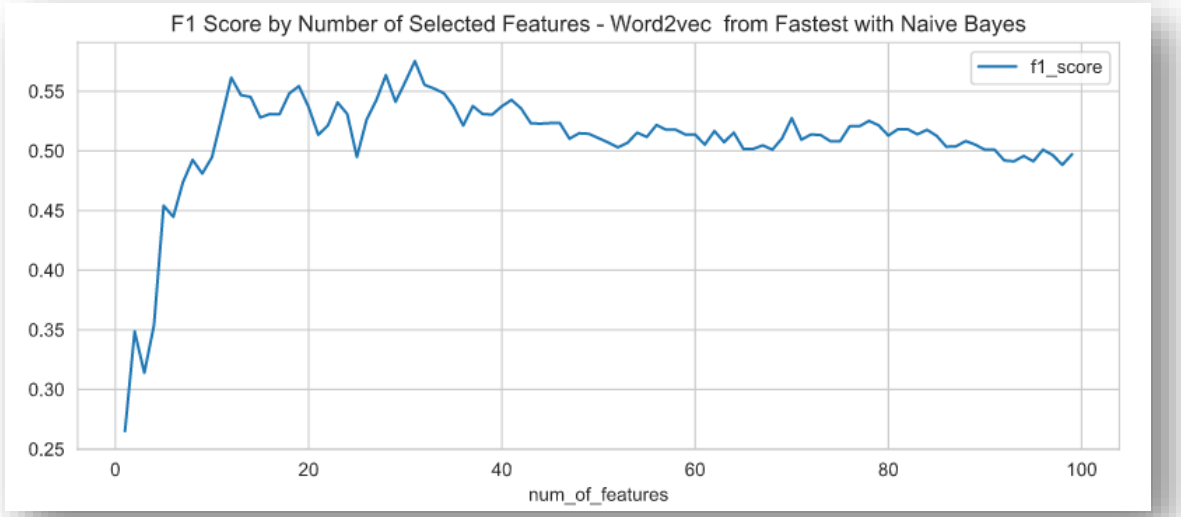
Word2Vec from FastText

Feature Extraction

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0.28	-0.62	-0.28	0.36	0.78	0.37	-0.23	-0.67	0.20	-0.53	0.01	-0.26	-0.77
1	0.31	-0.59	0.30	-0.64	-1.04	0.68	0.25	-0.07	-0.88	-0.27	-0.52	0.09	0.04
2	0.89	-1.03	1.05	-0.15	0.91	-0.25	-0.15	-0.16	-0.54	-0.61	-0.16	0.87	-0.07
3	0.46	-0.32	-0.53	-0.28	0.05	0.26	0.38	-0.02	0.21	0.17	0.11	1.07	-0.18
4	0.12	0.12	-0.46	-0.21	-0.09	-0.17	0.05	-0.07	0.16	0.23	0.26	1.25	-0.11
...	...	...	...	...	...	...	...	...	...	...	...	...	...
745	0.61	-1.87	1.25	-0.97	0.40	0.04	0.09	0.24	-1.06	-0.71	-0.20	0.29	-0.28
746	-0.11	-0.49	0.05	-0.31	0.51	-0.33	0.02	-0.05	-0.71	-0.04	0.38	0.47	0.29
747	1.09	-1.52	1.21	-0.56	0.49	0.35	-0.05	-0.34	-0.65	-0.63	-1.07	0.60	0.12
748	0.89	-0.80	-0.56	0.67	0.20	0.57	0.71	-0.27	-0.04	0.23	-0.32	1.22	-0.79
749	0.43	-1.15	0.94	-0.15	0.74	0.10	-0.35	0.11	-0.53	-1.06	-0.39	0.30	-0.09

750 rows × 100 columns

Feature Selection – Univariate with Chi^2



Benchmarking

	Features_Benchedmarked	Precision	Recall	f1_score	accuracy
0	Word2Vec Fastext Naive Bayes Baseline	0.56	0.56	0.50	0.56
1	Word2Vec from Fastest Naive Bayes Optimal Features Selected: 31	0.67	0.61	0.58	0.61

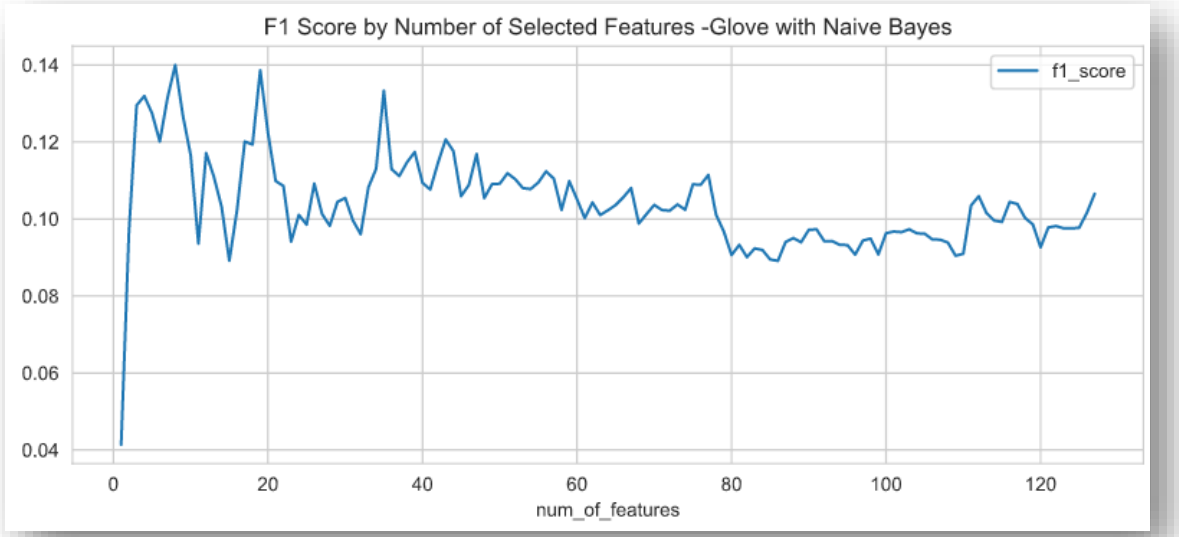
Glove from FLAIR

Feature Extraction

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	-0.20	0.25	0.13	0.17	0.48	-0.22	-0.23	0.05	0.09	0.01	-0.04	0.06	-0.16
1	-0.19	0.44	0.09	0.03	0.20	0.12	-0.28	0.02	-0.07	-0.06	-0.20	-0.34	-0.11
2	0.07	0.48	0.03	0.22	0.40	0.07	-0.23	0.17	0.13	0.29	-0.03	-0.15	-0.12
3	-0.17	0.22	-0.12	-0.08	0.23	-0.05	-0.23	0.09	0.12	0.11	-0.36	-0.13	-0.10
4	-0.07	0.55	-0.09	0.13	0.18	0.07	-0.28	0.08	0.24	0.19	-0.07	-0.03	0.04
...	...	...	...	...	...	...	...	...	...	...	...	...	...
245	0.10	0.28	0.13	0.20	0.54	-0.08	-0.44	-0.05	0.09	0.08	0.10	0.10	0.15
246	-0.22	0.32	-0.06	0.07	0.21	0.15	-0.13	-0.09	0.27	0.11	-0.20	-0.29	-0.22
247	-0.12	0.38	-0.04	0.05	0.10	0.07	-0.23	0.00	0.06	-0.08	-0.08	-0.01	-0.14
248	-0.20	0.39	0.07	0.15	0.19	-0.04	-0.35	0.02	0.14	-0.14	-0.05	-0.07	0.08
249	0.14	0.20	-0.06	0.23	0.35	0.08	-0.14	0.21	0.08	0.14	-0.17	0.02	0.13

250 rows × 128 columns

Feature Selection – Univariate with Chi^2



Benchmarking

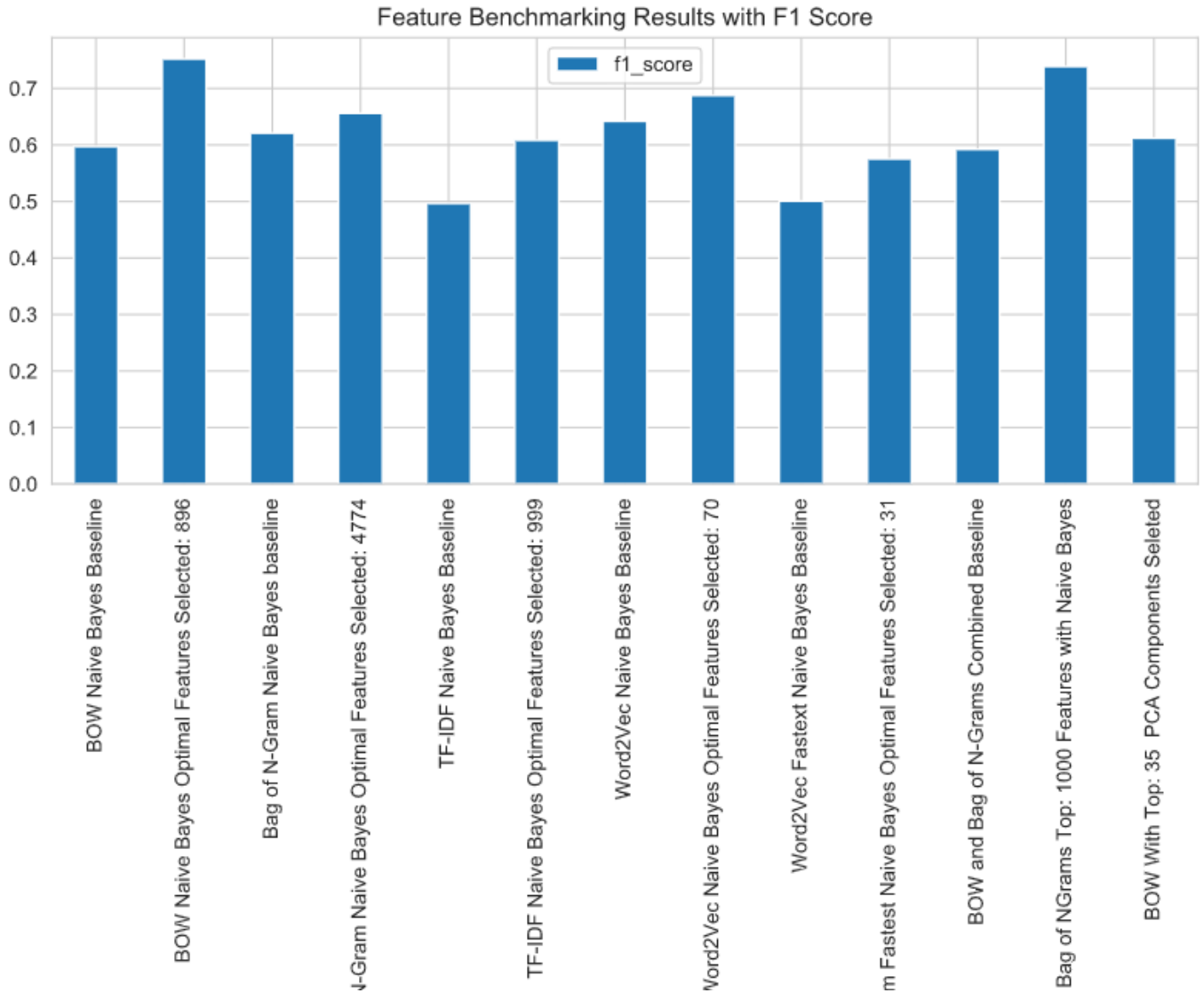
	Features_Benchedmarked	f1_score	accuracy
0	Glove with Naive Bayes All Features	0.13	0.14

Feature Engineering, Extraction and Selection Final Results

	Features_Benchedmarked	Precision	Recall	f1_score	accuracy
0	BOW Naive Bayes Baseline	0.68	0.62	0.60	0.62
1	BOW Naive Bayes Optimal Features Selected: 896	0.79	0.75	0.75	0.75
2	Bag of N-Gram Naive Bayes baseline	0.70	0.64	0.62	0.64
3	Bag of N-Gram Naive Bayes Optimal Features Selected: 4774	0.69	0.66	0.66	0.66
4	TF-IDF Naive Bayes Baseline	0.54	0.56	0.50	0.56
5	TF-IDF Naive Bayes Optimal Features Selected: 999	0.74	0.64	0.61	0.64
6	Word2Vec Naive Bayes Baseline	0.68	0.65	0.64	0.65
7	Word2Vec Naive Bayes Optimal Features Selected: 70	0.71	0.69	0.69	0.69
8	Word2Vec Fastext Naive Bayes Baseline	0.56	0.56	0.50	0.56
9	Word2Vec from Fastest Naive Bayes Optimal Features Selected: 31	0.67	0.61	0.58	0.61
10	BOW and Bag of N-Grams Combined Baseline	0.70	0.62	0.59	0.62
11	BOW + Bag of NGrams Top: 1000 Features with Naive Bayes	0.77	0.74	0.74	0.74
12	BOW With Top: 35 PCA Components Seleted	0.63	0.61	0.61	0.61

Feature Engineering, Extraction and Selection Final Results

Visualization of Benchmark Results



Feature Engineering, Extraction and Selection Final Results

Best results were produced from the following:	
Extracted Feature	Bag of Words
Feature Selection Method	Univariate chi^2
Reference Model	Naive Bayes Multinomial Variant

Label	precision	recall	f1-score	support
auto-repair-appt-1	1.00	1.00	1.00	18
coffee-ordering-1	0.79	0.58	0.67	19
coffee-ordering-2	0.62	0.81	0.70	16
movie-finder	1.00	0.91	0.95	11
movie-tickets-1	0.78	0.90	0.84	20
movie-tickets-2	0.79	0.68	0.73	22
movie-tickets-3	0.92	0.96	0.94	24
pizza-ordering-1	0.62	0.71	0.67	14
pizza-ordering-2	0.81	0.74	0.77	23
restaurant-table-1	1.00	0.41	0.58	22
restaurant-table-2	0.44	0.71	0.55	17
restaurant-table-3	0.53	0.62	0.57	16
uber-lyft-1	0.85	0.73	0.79	15
uber-lyft-2	0.73	0.85	0.79	13
accuracy			0.75	250
macro avg	0.78	0.76	0.75	250
weighted avg	0.79	0.75	0.75	250

Confusion Matrix Heatmap