Emoji Predictions-Twitter 🐦

ISE 599 - Fall, 2019

Kyle Wang, Hongyu Li, Lu Wang

USC Viterbi
School of Engineering

University of Southern California

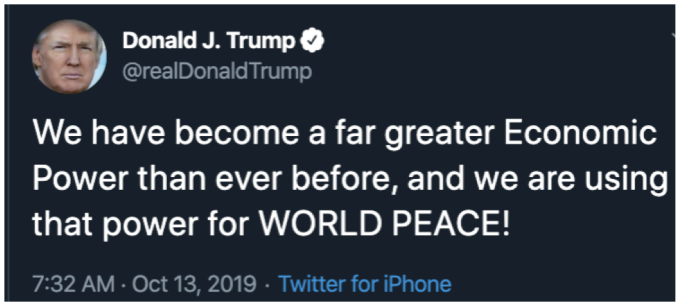
I. ⚠️ Problem Statement / Goals

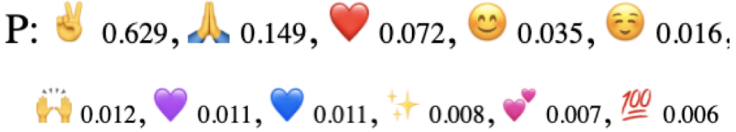
- **Emoji Suggestion when posting a tweet**

  - Currently, emoji can be recommended only based on a single word meaning

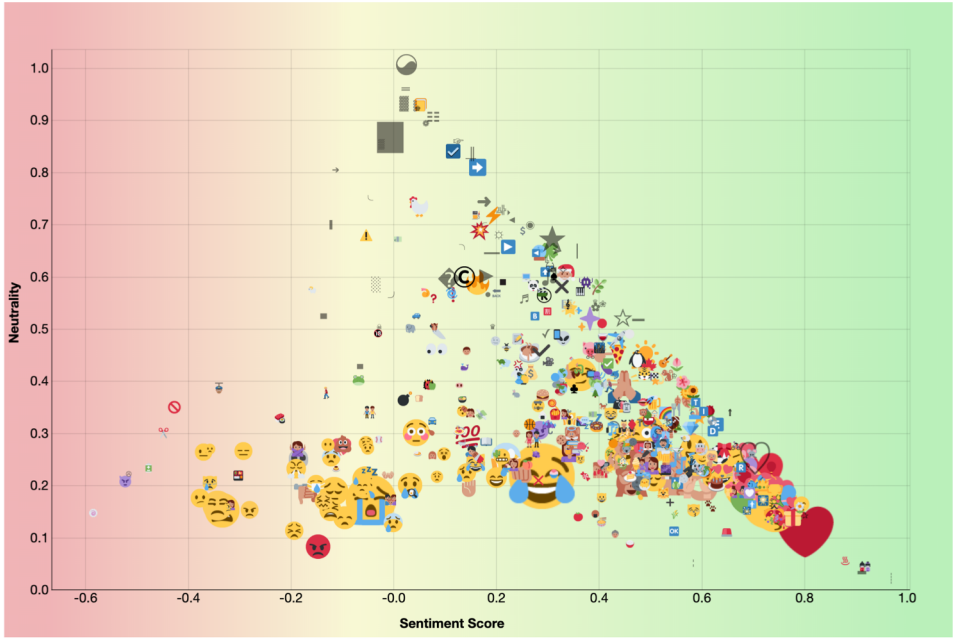
- **Information Retrieval via emoji**

  - Currently, searching emoji will only return a tweet that contain that emoji





II. ⛏️ Dataset - Emoji Selection



emojitracker: realtime emoji use on twitter

2582480656	1242463662	965050334	935384848	776346305
383539634	367082159	354531910	353836027	330163828
230228547	228828752	227902143	224609951	213631430
181421601	177174700	176840906	165906122	161482063
142460161	136608412	133868090	131739599	129630974
107029281	106136583	105328953	104935013	104722335
95069819	92180862	92030822	89237343	89156181
77535599	77321210	76780814	76178219	74582200
62631657	62257124	62002966	61809162	61247868
47945459	46477292	44784865	43875713	43727524
38572361	38480313	37665808	37661824	37510664
35692355	35193362	34830129	33779043	31777873
29776203	29117446	29090402	28756815	28243790
26897483	26189538	26070278	25887408	25859591
25334531	24384746	24189270	24158045	24029032
22570323	22494747	22451442	22240966	22224571
20152549	19886370	19741478	19710561	19464743
18086502	17971658	17953474	17554752	17486203
15802863	15739886	15655387	15599836	15482391
14779039	14356809	13865985	13691088	13586911
13053675	13037101	12940683	12940020	12824872
12216917	12197072	12013044	11951454	11872026
11051383	10955929	10943882	10911363	10848423

II. ⛏️ Dataset - Emoji Selection

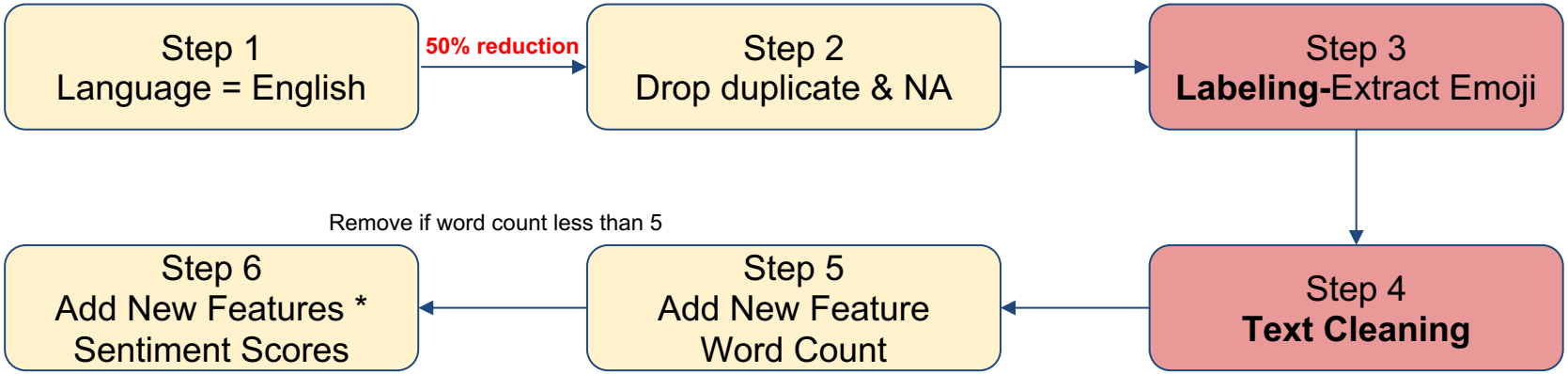
	Emoji	Name	tag	Unicode		Emoji	Name	tag	Unicode
1	😂	joy	tears	\ud83d\ude02	11	🐶	dog	pet	\ud83d\udc36
2	🙃	upside_down_face		\ud83d\ude43	12	🙈	see_no_evil	monkey; ignore	\ud83d\ude48
3	😘	kissing_heart	flirt	\ud83d\ude18	13	♻️	recycle	environment;green	\u267b\ufe0f
4	😭	sob	sad; cry;bawling	\ud83d\ude2d	14	😡	rage; pout	angry	\ud83d\ude21
5	❤️	heart	love	\u2764\ufe0f	15	😋	yum	tongue; lick	\ud83d\ude0b
6	💯	100	score; perfect	\ud83d\udcaf	16	😱	scream	horror; shocked	\ud83d\ude31
7	🌚	New_moon_with_face		\ud83c\udf1a	17	🙏	pray	please; hope; wish	\ud83d\ude4f
8	🔥	fire	burn	\ud83d\udd25	18	👍	+1; thumbsup	approve; ok	\ud83d\udc4d
9	🙌	Raised_hands	hooray	\ud83d\ude4c	19	👫	couple	date	\ud83d\udc6b
10	🔞	underage		\ud83d\udd1e	20	🇺🇸	us	flag; united; america	\ud83c\uddfa\ud83c\uddf8

III. 💽 Data Collection - Real-time Scraping / tweepy

	created_at	id	id_str	text	source	truncated	in_reply_to_status_id	in_reply_to_statu
0	2019-10-15 19:50:13+00:00	1.184195e+18	1.184195e+18	RT @SidemenClothing: 🚨 COMPETITION TIME!! 🚨 \n\...	<a href="http://twitter.com/download/iphone" r...	0.0	NaN	
1	2019-10-15 19:50:13+00:00	1.184195e+18	1.184195e+18	So Ghost wanted to get out the game but Tasha ...	<a href="http://twitter.com/download/iphone" r...	0.0	NaN	
2	2019-10-15 19:50:13+00:00	1.184195e+18	1.184195e+18	@_niallssmile THANK YOU SO MUCH ❤️	<a href="http://twitter.com/download/iphone" r...	0.0	1.184194e+18	1.184
3	2019-10-15 19:50:13+00:00	1.184195e+18	1.184195e+18	@ourrachblogs My dad's birthday was Christmas...	<a href="http://twitter.com/download/iphone" r...	0.0	1.184082e+18	1.184
4	2019-10-15 19:50:13+00:00	1.184195e+18	1.184195e+18	@karelyruiz001 Hermosa 😘	<a href="http://twitter.com/download/android" ...	0.0	1.183956e+18	1.183
...	...	...	...	...	...	...	...	
56342	2019-10-17 21:30:29+00:00	1.184945e+18	1.184945e+18	🎁 20€ GRATIS\n*. Disfruta de la experiencia\n🇪🇸 ...	<a href="https://www.hootsuite.com" rel="nofol...	1.0	NaN	
56343	2019-10-17 21:30:29+00:00	1.184945e+18	1.184945e+18	RT @roannapadilla: Concert for a cause! ❤️ htt...	<a href="http://twitter.com/download/iphone" r...	0.0	NaN	
56344	2019-10-17 21:30:29+00:00	1.184945e+18	1.184945e+18	RT @fannydeschamps3: Cette vidéo je m'en lasse...	<a href="http://twitter.com/download/iphone" r...	0.0	NaN	
56345	2019-10-17 21:30:29+00:00	1.184945e+18	1.184945e+18	RT @Grenda_Bastian: Quiero ir al #TakeOverAcap...	<a href="http://twitter.com/download/android" ...	0.0	NaN	
56346	2019-10-17 21:30:29+00:00	1.184945e+18	1.184945e+18	اتمنى بس ما توقفون عند 😹\n\n بيبان بيوتنا وقف...	<a href="http://twitter.com/download/iphone" r...	0.0	NaN	

- 10 days, different time
- 10 JSON files, total 10.15 GB
- 1,106,130 rows
- 38 columns
- via tweet API
- time consuming

IV. 🧹 Data Cleaning - Overview



*using textBlob.sentiment()

Polarity [-1,1] - where 1 means positive statement and -1 means negative statement

Subjectivity [0,1] - personal opinion, emotion or judgment

IV. 🧹 Data Cleaning - Labeling

Step 3
Labeling-Extract Emoji

- Input: dataframe after initial preprocessing
- Generate mapping dictionary for selected emojis

{'😂': 0, '🙃': 1, '😘': 2, '😭': 3, '❤️': 4, '💯': 5, '🌚': 6, '🔥': 7, '🙌': 8, '🔞': 9, '🐶':10, '🙈':11, '♻️': 12. '😡':13, '😋':14, '😱':15, '🙏':16, '👍':17, '👫':18, '🇺🇸': 19}

- Select only tweet that contain **one** targeted emoji (80% reduction)
- Output: dataframe with label column

IV. 🧹 Data Cleaning - Text Cleaning

Step 4
Text Cleaning

- Basic Preprocessing : preprocessor.clean()
- Extra Preprocessing - Remove
  - Emoji patterns (\U000000A0-\U0001FA90)
  - Emoticons (':-)', ':)', ';)', ':o)', :-b', ':b', '>:)', '>;)', '>:-)', '<3', etc)
  - Punctuation, Hashtags, Links, @, stopwords
- Lower case
- Word tokenizing
- Time Permitting- Spell Checking and Auto Correction using pyspellchecker

  ({'showin', 'feelin'}, {'everday'}, {'youre'}, {'aint'},{'idk'},{'wtf', 'lmao'})

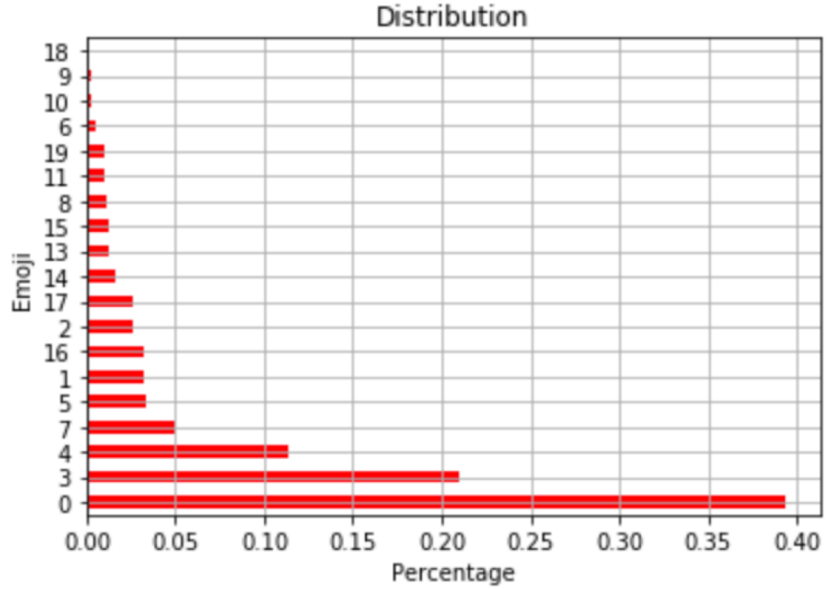
IV. 🧹 Data Cleaning - Constructed

	time	filtered_tweet	polarity	subjectivity	word_count	label
0	2019-11-17 20:55:38	today got greatest phone call time	1.000000	1.0000	11	0
1	2019-11-17 20:55:38	cute enough onlyfans	0.250000	0.7500	8	0
2	2019-11-17 20:55:38	stewards conclude investigations go barca seco...	0.000000	0.0000	19	0
3	2019-11-17 20:55:38	im trying find bubl response theres golden gif...	-0.050000	0.7500	20	0
4	2019-11-17 20:55:38	sight making emotional crowd waving lightstick...	0.200000	0.6750	17	3
...	...	...	...	...	...	...
149679	2019-11-17 03:58:36	unintentional night one talk nothing yay	0.000000	0.0000	14	1
149680	2019-11-17 03:58:36	thenailsecret cant nails wont get done	0.000000	0.0000	12	3
149681	2019-11-17 03:58:36	every time im liana people assume couple like	0.000000	0.0000	13	0
149682	2019-11-17 03:58:36	ni yeah smiled feels	0.600000	0.2000	8	3
149683	2019-11-17 03:58:38	hottie tori black fucking hard cock crowded co...	-0.229167	0.4875	10	0

148212 rows × 6 columns

- 148,212 rows
- 6 columns
- CSV file, 14 MB
- Highly Imbalanced

IV. 🧹 Data Cleaning - Remove Least 5 Frequent



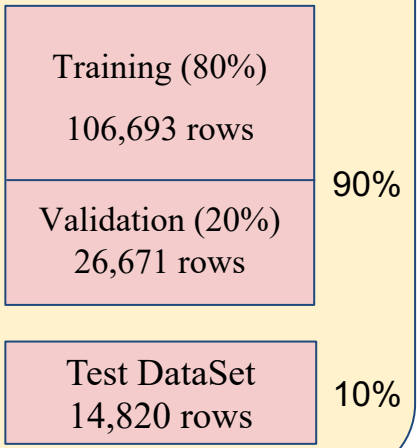
Note: Emoji 12: ♻️ has zero record

{'😂': 0, '🙃': 1, '😘': 2, '😭': 3,
'❤️': 4, '💯': 5, '~~🌑~~': 6, '🔥': 7,
'~~🙌~~': 8, '~~🤬~~': 9, '~~🐶~~':10, '~~🙈~~':11,
'~~♻️~~': 12, '😡':13, '😋':14, '😱':15,
'🙏':16, '👍':17, '~~🤦‍♀️~~':18, '🇺🇸': 19}

- 148,184 rows left

⏭ Next step...and a few more

Split Data (Stratified)



Models

- ❑ **Random**
- ❑ **Using tf-idf**
- ❑ **tf-idf w/others**
- ❑ **Using FastText**
- ❑ **bi-LSTM**

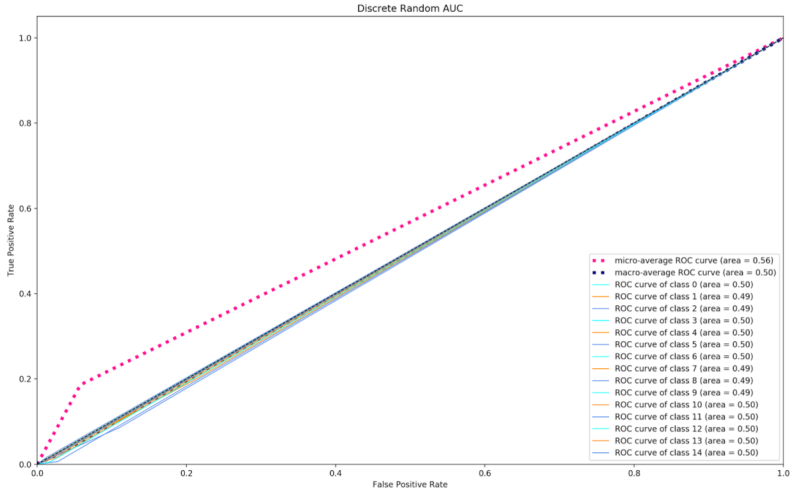
Validation & Metrics

- **Micro F1 score**, Macro F1 score
- AUC scores
- Recall at K=2, K=3
- NDCG at 3
- Computing time

V. 🌐 Model - #1 Discrete Random Model

- Using probability distribution in training set to randomly predict

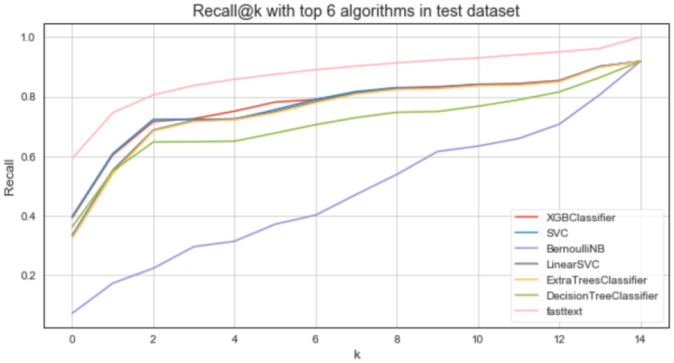
		Evaluation on Test Set					
	Training Time (second)	**Micro F1**	Macro F1	Micro ROC	**Recall at 2**	Recall at 3	NDCG at 3
#1	0.007	0.186	0.051	0.56	0.332	0.442	0.305



V. 🌐 Model - #2 Using tf-idf only

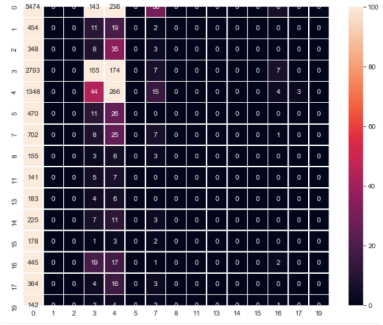
	MLA Name	MLA Parameters	MLA F1 Macro	MLA F1 Micro	MLA Time
16	XGBClassifier	{'base_score': 0.5, 'booster': 'gbtree', 'cols...	0.128311	0.4412	12.8865
12	SVC	{'C': 1.0, 'cache_size': 200, 'class_weight': ...	0.0380544	0.39945	15.5961
10	BernoulliNB	{'alpha': 1.0, 'binarize': 0.0, 'class_prior':...	0.0385706	0.39935	0.00430427
13	LinearSVC	{'C': 1.0, 'class_weight': None, 'dual': True,...	0.0405007	0.39935	0.239394
2	ExtraTreesClassifier	{'bootstrap': False, 'class_weight': None, 'cr...	0.0405007	0.39935	0.135596
15	ExtraTreeClassifier	{'class_weight': None, 'criterion': 'gini', 'm...	0.0405007	0.39935	0.0149227
14	DecisionTreeClassifier	{'class_weight': None, 'criterion': 'gini', 'm...	0.0405007	0.39935	0.0256711
7	RidgeClassifierCV	{'alphas': array([0.1, 1. , 10.]), 'class_w...	0.0383922	0.3993	3.9057
4	RandomForestClassifier	{'bootstrap': True, 'class_weight': None, 'cri...	0.0405121	0.39925	0.0692048
1	BaggingClassifier	{'base_estimator': None, 'bootstrap': True, 'b...	0.0407816	0.3992	0.174288
5	LogisticRegressionCV	{'Cs': 10, 'class_weight': None, 'cv': 'warn',...	0.0381103	0.3991	5.24862
3	GradientBoostingClassifier	{'criterion': 'friedman_mse', 'init': None, 'l...	0.0407733	0.399	5.55937
6	PassiveAggressiveClassifier	{'C': 1.0, 'average': False, 'class_weight': N...	0.0403966	0.39895	0.0508148
0	AdaBoostClassifier	{'algorithm': 'SAMME.R', 'base_estimator': Non...	0.0390009	0.3988	0.301531
11	KNeighborsClassifier	{'algorithm': 'auto', 'leaf_size': 30, 'metric...	0.033684	0.3374	0.00201316
8	SGDClassifier	{'alpha': 0.0001, 'average': False, 'class_wei...	0.0256376	0.22695	0.0627114
9	Perceptron	{'alpha': 0.0001, 'class_weight': None, 'early...	0.0242824	0.21805	0.0472102

- Tf-idf, max_features = 2000
- Training set sample 10,000 w/ Cross Validation
- Computationally intensive



V. 🌐 Model - #2 Using tf-idf only

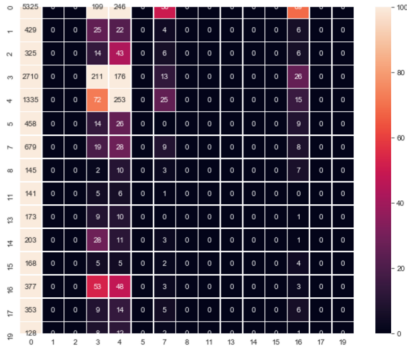
XGB Classifier



	precision	recall	f1-score	support
0	0.41	0.93	0.57	5889
1	0.00	0.00	0.00	486
2	0.00	0.00	0.00	394
3	0.36	0.05	0.09	3136
4	0.33	0.17	0.22	1700
5	0.00	0.00	0.00	507
7	0.09	0.01	0.02	743
8	0.00	0.00	0.00	167
11	0.00	0.00	0.00	153
13	0.00	0.00	0.00	193
14	0.00	0.00	0.00	246
15	0.00	0.00	0.00	184
16	0.10	0.00	0.01	484
17	0.00	0.00	0.00	387
19	0.00	0.00	0.00	151
accuracy			0.40	14820
macro avg	0.09	0.08	0.06	14820
weighted avg	0.28	0.40	0.27	14820

[tf-idf max features = 100]

Linear SVC

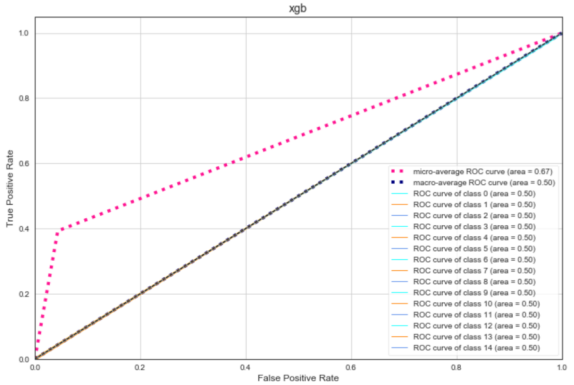


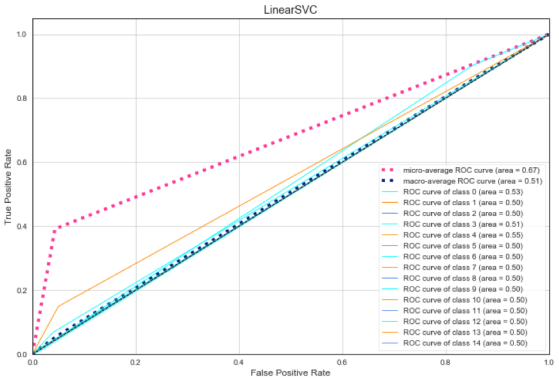
	precision	recall	f1-score	support
0	0.41	0.90	0.57	5889
1	0.00	0.00	0.00	486
2	0.00	0.00	0.00	394
3	0.31	0.07	0.11	3136
4	0.28	0.15	0.19	1700
5	0.00	0.00	0.00	507
7	0.07	0.01	0.02	743
8	0.00	0.00	0.00	167
11	0.00	0.00	0.00	153
13	0.00	0.00	0.00	193
14	0.00	0.00	0.00	246
15	0.00	0.00	0.00	184
16	0.02	0.01	0.01	484
17	0.00	0.00	0.00	387
19	0.00	0.00	0.00	151
accuracy			0.39	14820
macro avg	0.07	0.08	0.06	14820
weighted avg	0.27	0.39	0.27	14820

V. 🌐 Model - #2 Using tf-idf only

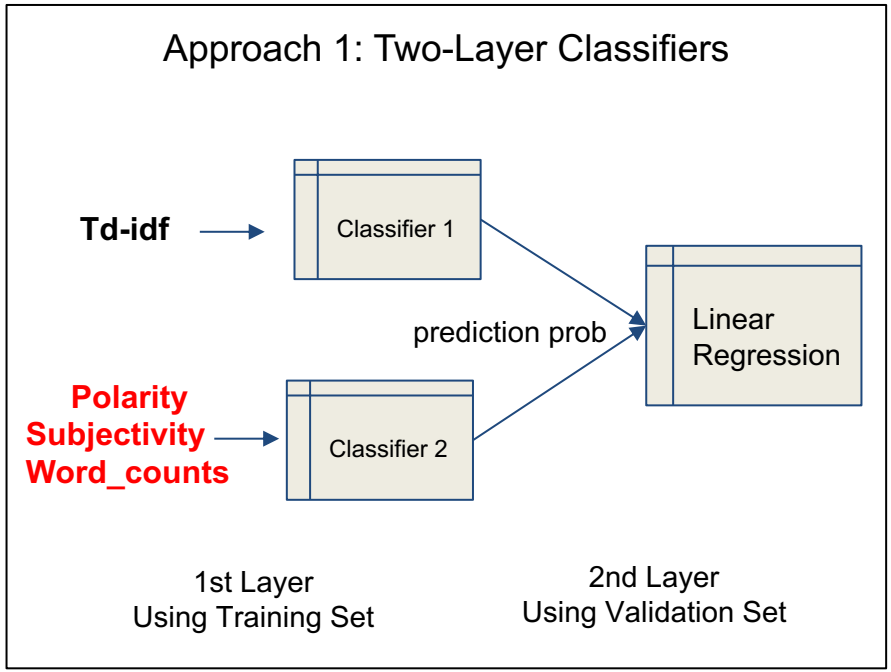
- Using best parameters from sub-training set,
- Tf-idf max features = 100

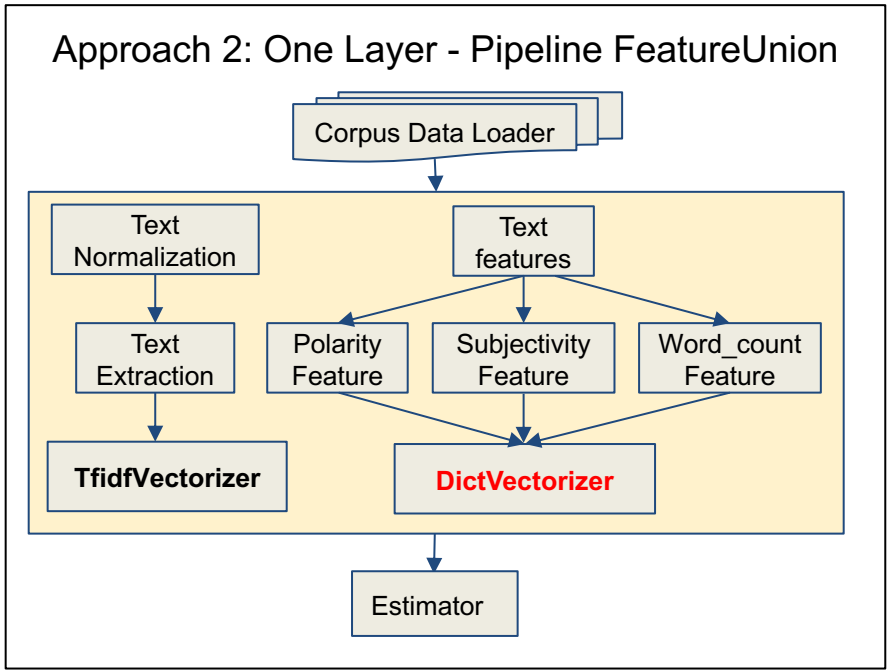
		Evaluation on Test Set					
	Training Time (second)	**Micro F1**	Macro F1	Micro ROC	**Recall at 2**	Recall at 3	NDCG at 3
#1 Random	0.007	0.186	0.051	0.56	0.332	0.442	0.305
#2 tf-idf XGB	89.169	0.393	0.039	0.67	0.605	0.716	0.585
#3 tf-idf Linear SVC	8.280	0.402	0.058	0.67	0.610	0.688	0.584





V. 🌐 Model - #3 Using tf-idf w/ other features





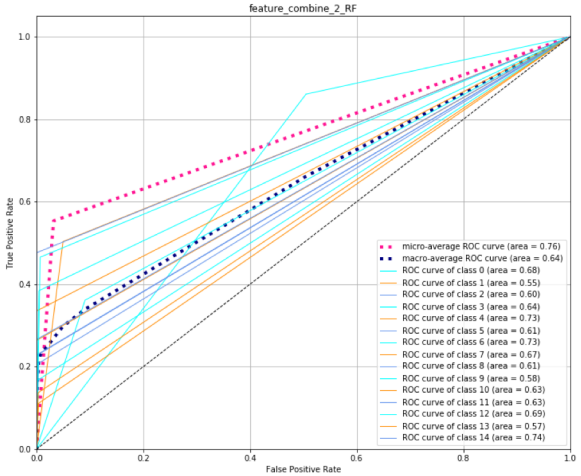
V. 🌐 Model - #3 Using tf-idf w/ other features

	Models		Evaluation on Test Set					
		Training Time (second)	**Micro F1**	Macro F1	Micro ROC	**Recall at 2**	Recall at 3	NDCG at 3
*Approach 1 Two Layers classifiers	XGB	228.405	0.397	0.038	0.68	-	-	-
	LinearSVC	238.390	0.397	0.038	0.68	-	-	-
	RandomForest	201.164	0.397	0.038	0.68	-	-	-
Approach 2 One Layer - Pipeline	XGB	598.546	0.451	0.157	0.71	0.433	0.546	0.607
	LinearSVC	394.712	0.543	0.355	0.76	0.408	0.501	0.637
	RandomForest	104.597	0.554	0.422	0.76	0.425	0.497	0.599

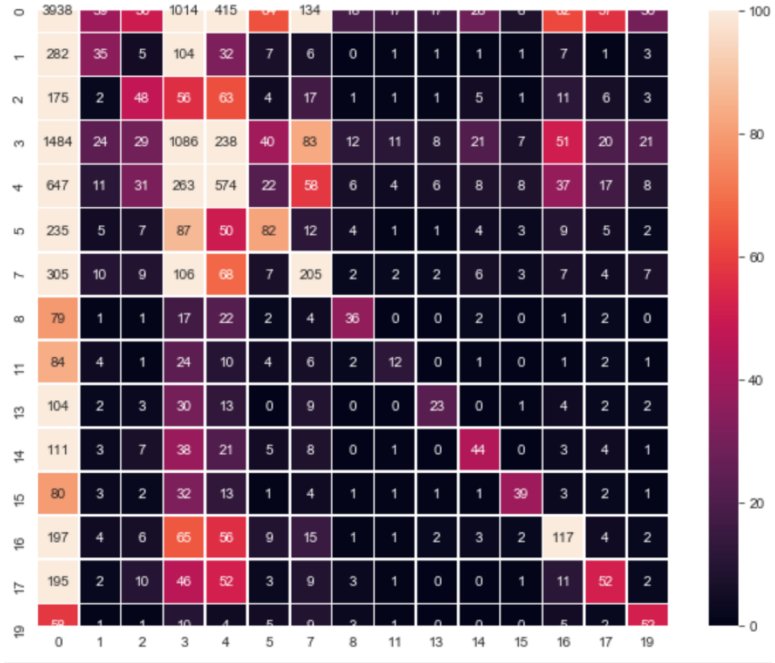
***Note: Approchah 1 has no probability ranking list. (second layer classifier is linear regression)**

V. 🌐 Model - #3 Using tf-idf w/ other features

		Evaluation on Test Set					
	Training Time (second)	**Micro F1**	Macro F1	Micro ROC	**Recall at 2**	Recall at 3	NDCG at 3
#1 Random	0.007	0.186	0.051	0.56	0.332	0.442	0.305
#2 tf-idf XGB	89.169	0.393	0.039	0.67	0.605	0.716	0.585
#3 tf-idf Linear SVC	8.280	0.402	0.058	0.67	0.610	0.688	0.584
#4 Feature Union RandomForest	104.597	0.554	0.422	0.76	0.425	0.497	0.599



V. 🌐 Model - #4 Using FastText only: features = 100



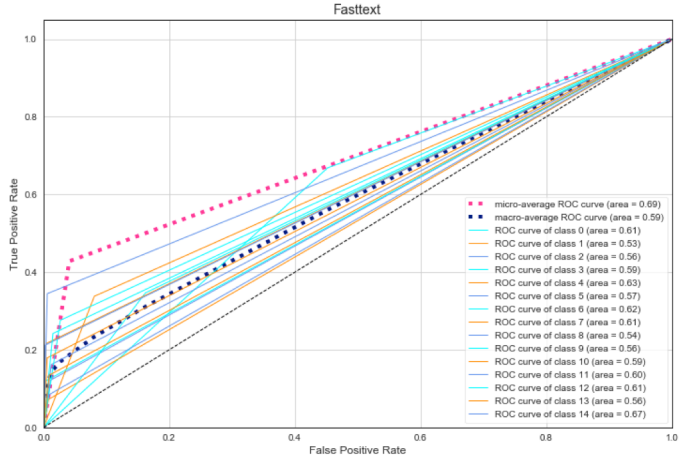
	precision	recall	f1-score	support
0	0.49	0.67	0.57	5891
1	0.24	0.07	0.11	486
2	0.23	0.12	0.16	394
3	0.36	0.35	0.36	3135
4	0.35	0.34	0.34	1700
5	0.32	0.16	0.22	507
7	0.35	0.28	0.31	743
8	0.40	0.22	0.28	167
11	0.22	0.08	0.12	152
13	0.37	0.12	0.18	193
14	0.35	0.18	0.24	246
15	0.53	0.21	0.30	184
16	0.36	0.24	0.29	484
17	0.29	0.13	0.18	387
19	0.39	0.34	0.36	151
accuracy			0.43	14820
macro avg	0.35	0.23	0.27	14820
weighted avg	0.40	0.43	0.40	14820

{'😂': 0, '🙃': 1, '😘': 2, '😭': 3, '❤️': 4, '💯': 5, '🔥': 7, '🙌': 8, '😱':11, '😡':13, '😋':14, '😱':15, '🙏':16, '👍':17, '🇺🇸': 19}

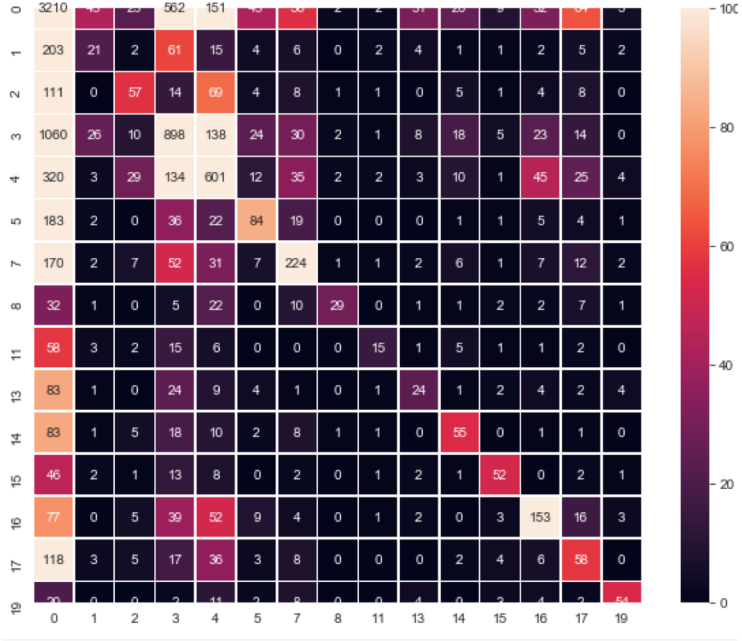
V. 🌐 Model - #4 Using FastText only

- Epoch: 14 & Learning rate: 0.1
- Max feature = 100

		Evaluation on Test Set					
	Training Time (second)	**Micro F1**	Macro F1	Micro ROC	**Recall at 2**	Recall at 3	NDCG at 3
#1 Random	0.007	0.186	0.051	0.56	0.332	0.442	0.305
#2 XGB	89.169	0.393	0.039	0.67	0.605	0.716	0.585
#3 Linear SVC	8.280	0.402	0.058	0.67	0.610	0.688	0.584
#4 Feature Union RandomForest	104.597	0.554	0.422	0.76	0.425	0.497	0.599
#5 FastText	5.628	0.428	0.268	0.69	0.744	0.804	0.398



V. 🌐 Model - #5 Bidirectional-LSTM - Heatmap



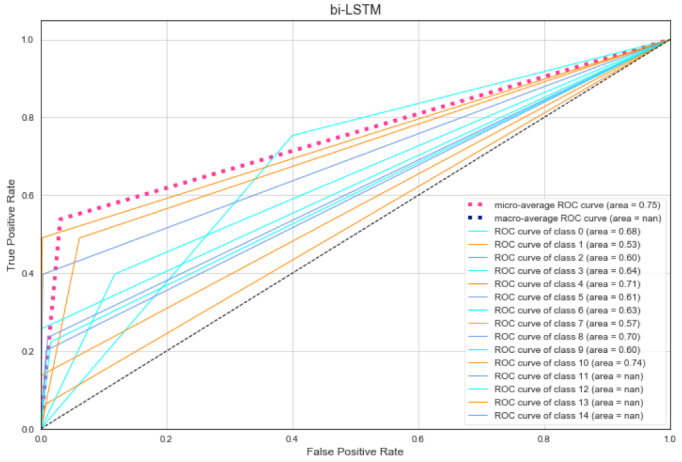
	precision	recall	f1-score	support
0	0.56	0.75	0.64	4259
1	0.19	0.06	0.10	329
2	0.39	0.20	0.27	283
3	0.48	0.40	0.43	2257
4	0.51	0.49	0.50	1226
5	0.42	0.23	0.30	358
7	0.53	0.43	0.47	525
8	0.76	0.26	0.38	113
11	0.54	0.14	0.22	109
13	0.29	0.15	0.20	160
14	0.42	0.30	0.35	186
15	0.60	0.40	0.48	131
16	0.53	0.42	0.47	364
17	0.26	0.22	0.24	260
19	0.70	0.49	0.58	110
accuracy			0.52	10670
macro avg	0.48	0.33	0.37	10670
weighted avg	0.50	0.52	0.50	10670

{'😂': 0, '🙃': 1, '😘': 2, '😭': 3, '❤️': 4, '💯': 5, '🔥': 7, '🙌': 8, '🐶':11, '😡':13, '😋':14, '😱':15, '🙏':16, '👍':17, '🇺🇸': 19}

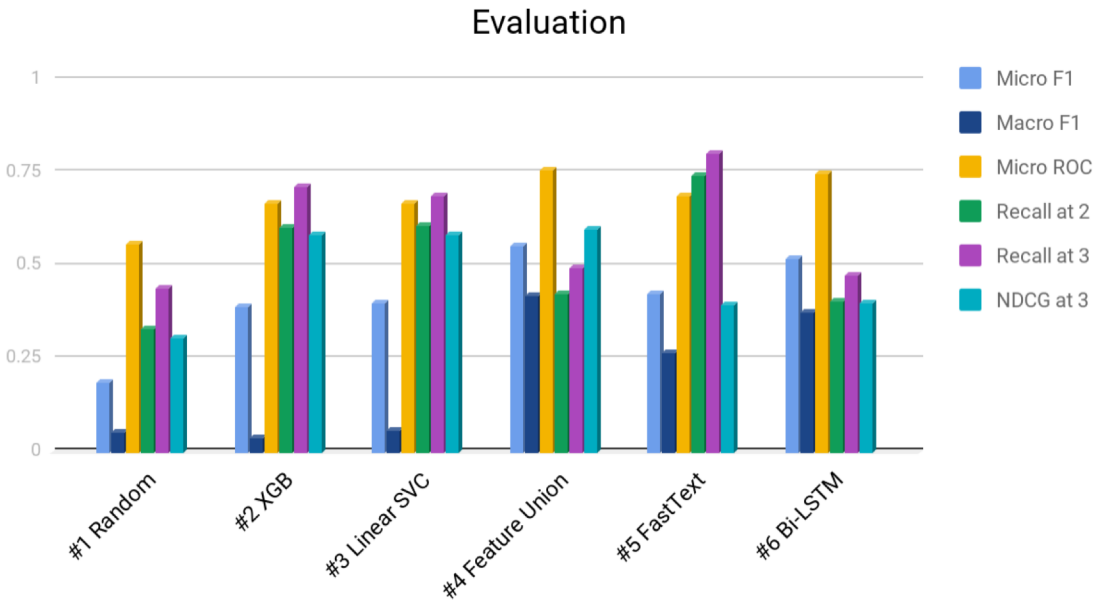
V. 🌐 Model - #5 Bidirectional-LSTM

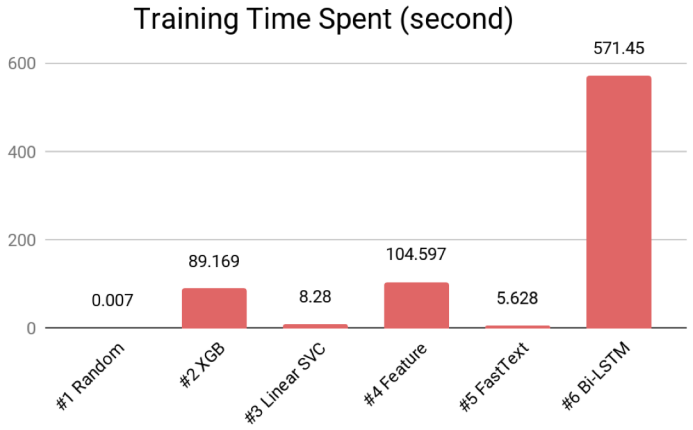
- Epoch: 5 & Batch size: 64
- First hidden layers with 100 neurons, dropout = 0.2
- Second hidden layers with 50 neurons, dropout = 0.2
- Activation function: “softmax”

		Evaluation on Test Set					
	Training Time (second)	**Micro F1**	Macro F1	Micro ROC	**Recall at 2**	Recall at 3	NDCG at 3
#1 Random	0.007	0.186	0.051	0.56	0.332	0.442	0.305
#2 XGB	89.169	0.393	0.039	0.67	0.605	0.716	0.585
#3 Linear SVC	8.280	0.402	0.058	0.67	0.610	0.688	0.584
#4 Feature Union RandomForest	104.597	0.554	0.422	0.76	0.425	0.497	0.599
#5 FastText	5.628	0.428	0.268	0.69	0.744	0.804	0.398
#6 Bi-LSTM	571.45	0.519	0.375	0.75	0.405	0.478	0.399

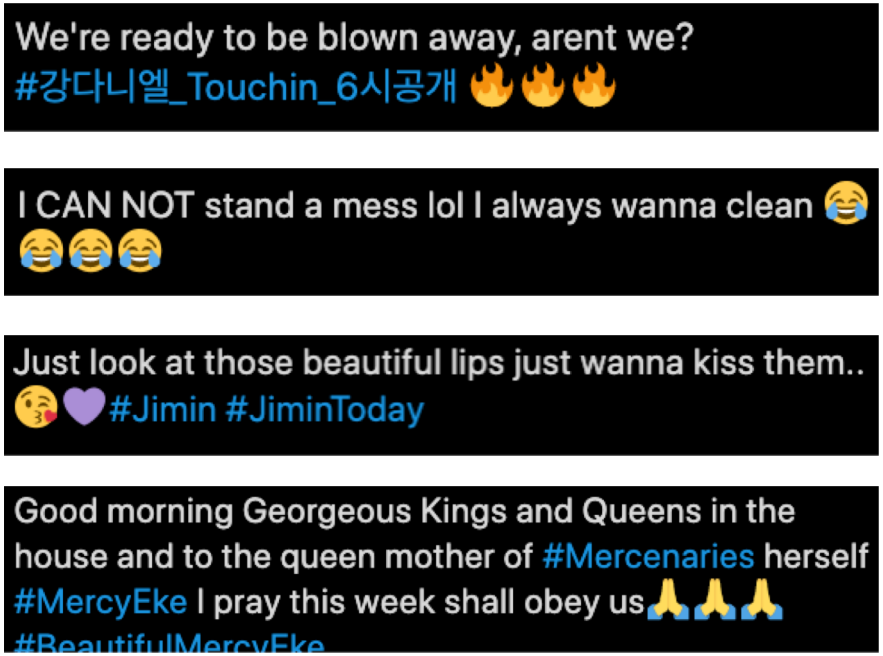


VI. 📈 Summary & Comparison





VII. 🐦 Potentials



FastText
{😂: 0.484, 🔥: 0.014, 🙂:0.004}
{😂: 0.672, 😭: 0.014, 🙂:0.014}
{😘: 0.335, ❤️: 0.043, 🙌::0.024}
{🙏:0.492 , ❤️:0.245, 🇺🇸: 0.144}

Challenges

- Useful data is hard to get, large file reduction after cleaning
- Data is highly imbalanced
- Computation Intense except for FastText model

VIII. 🙋 Future Works

- Scrap more data
- Spell Checking and Auto Correction using pyspellchecker

  ({'showin', 'feelin'}, {'everday'}, {'youre'}, {'aint'},{'idk'},{'wtf', 'lmao'})
- Try different Word Embedding: Random Indexing Vectors
- Try different models s.t. Combine FastText with other features
- Study the learning curve for each method
- Publish Paper

Github: https://github.com/kyleearth/emojiprediction.git

Reference

Barbieri, Francesco, Ballesteros, Miguel, Ronzano, Francesco, Saggion, Horacio. Multimodal Emoji Prediction. arXiv.org. April 2018. http://search.proquest.com/docview/2072033947/.

Li X., Yan R., Zhang M. (2017) Joint Emoji Classification and Embedding Learning. In: Chen L., Jensen C., Shahabi C., Yang X., Lian X. (eds) Web and Big Data. APWeb-WAIM 2017. Lecture Notes in Computer Science, vol 10367. Springer, Cham

Francesco Barbieri, Luis Espinnosa-Anke, Jose Camacho-Collados, Steven Schockaert, adn Horacio Saggion. 2018. Interpretable Emoji Prediction via Label-Wise Attention LSTMs. *Association for Computational Linguistics.* P.4766--4771

Petra Kralj Novak, Jasmina Smailovic´, Borut Sluban, and Igor Mozetic˘. 2015. Sentiment of emojis. *PloS one*, 10(12):e0144296.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Emoticons vs. emojis on Twitter: A causal inference approach. *arXiv preprint arXiv:1510.08480.*