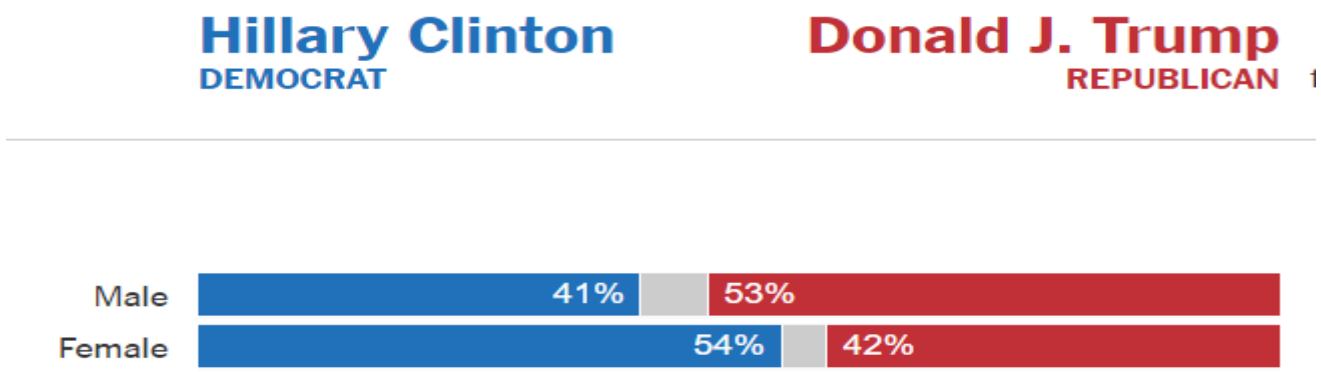
Analysis in Twitter Gender Classification

Chuong Trinh

Motivation

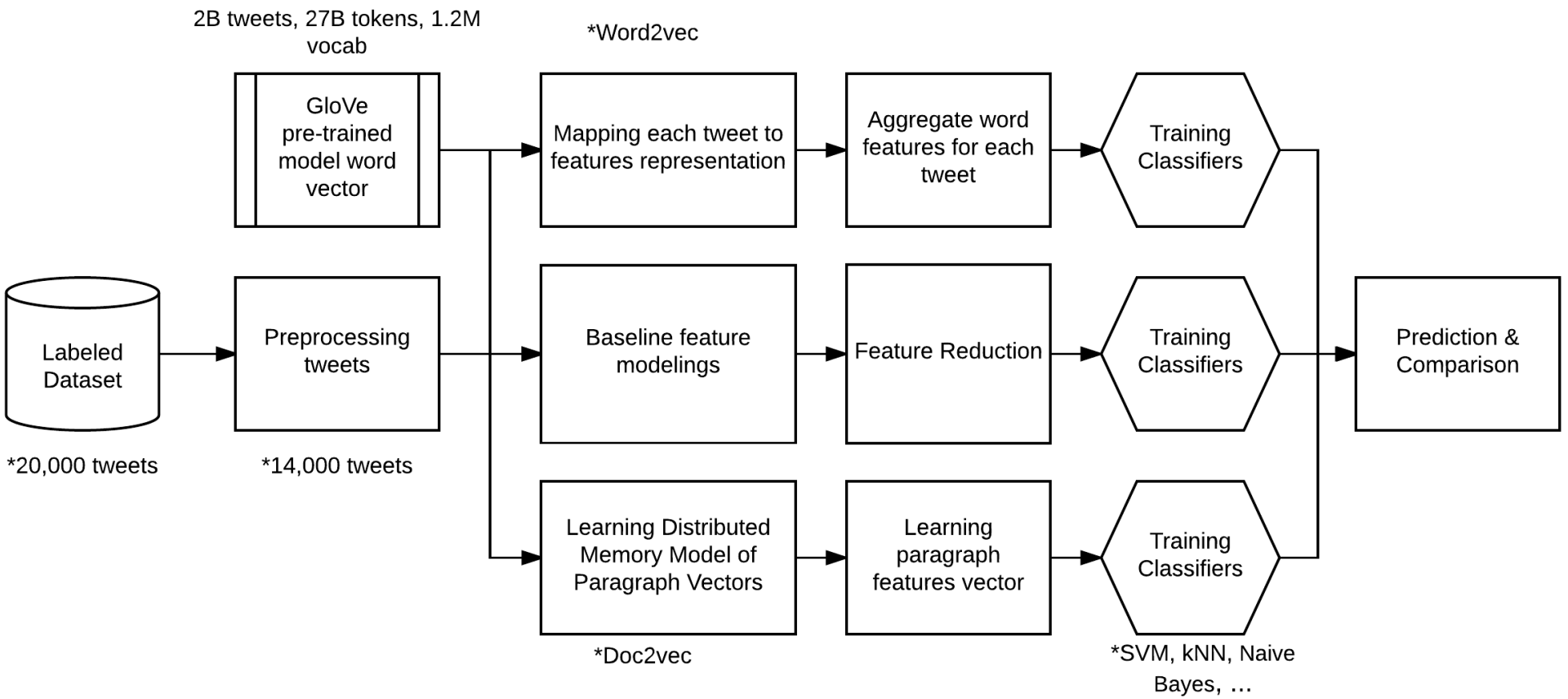
- Growing interest in automatically predicting the gender of authors from texts:
  - Opinions, political stances, styles, and preferences may be unique to each gender
  - Useful to individuals, companies, and governments for personal recommendation, customization, targeted advertising, political analysis, and policy formulation.



Why Gender Classification from Tweets is Hard!

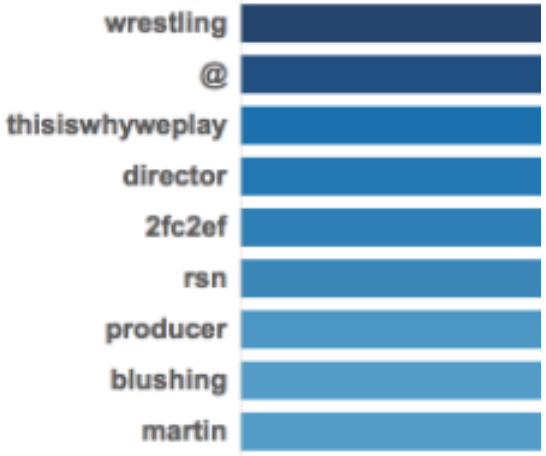
- Limited characters (140) per tweet
- Lots of spamming, advertising accounts, media sources, bots, etc.
- User's profile privacy
- Users construct their identity through interacting with other users! (Marwick and boyd, 2011) – all depend on the context
- For example
  - Tweet 1: I'm walking on sunshine <3 #and don't you feel good
  - Tweet 2: lalaloveya <3
  - Tweet 3: @USER loveyou ;D

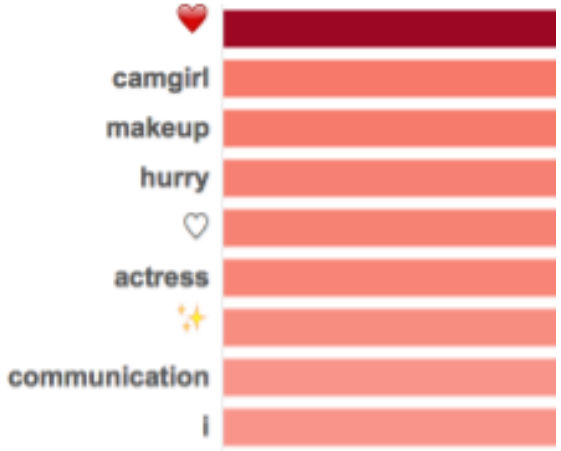
Pipeline


2B tweets, 27B tokens, 1.2M vocab
*Word2vec
GloVe pre-trained model word vector
Mapping each tweet to features representation
Aggregate word features for each tweet
Training Classifiers
Labeled Dataset
Preprocessing tweets
Baseline feature modelings
Feature Reduction
Training Classifiers
Prediction & Comparison
*20,000 tweets
*14,000 tweets
Learning Distributed Memory Model of Paragraph Vectors
Learning paragraph features vector
Training Classifiers
*Doc2vec
*SVM, kNN, Naive Bayes, …

Dataset & Baseline

- CrowdFlower (kaggle – data challenge site)
  - 20,000 tweets – collected in 2015
    - Human Amazon Turker labeling + CrowdFlower's labeling system
    - ~ 14,000 tweets can be used (non-English, low confidence, or unreadable is ignored)
    - Labels: male + female + brand

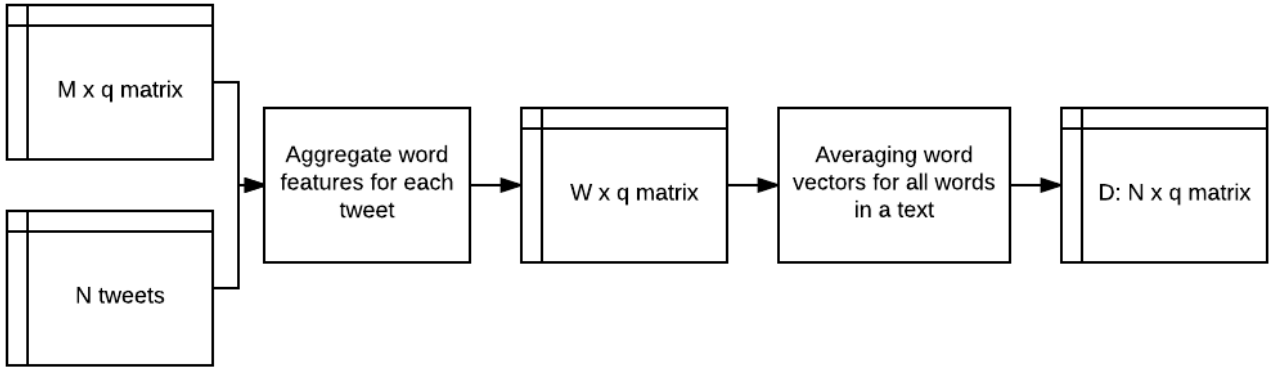




- Men are more likely to talk at another account
- Women are more likely to use emoji
- Current accuracy: ~60%

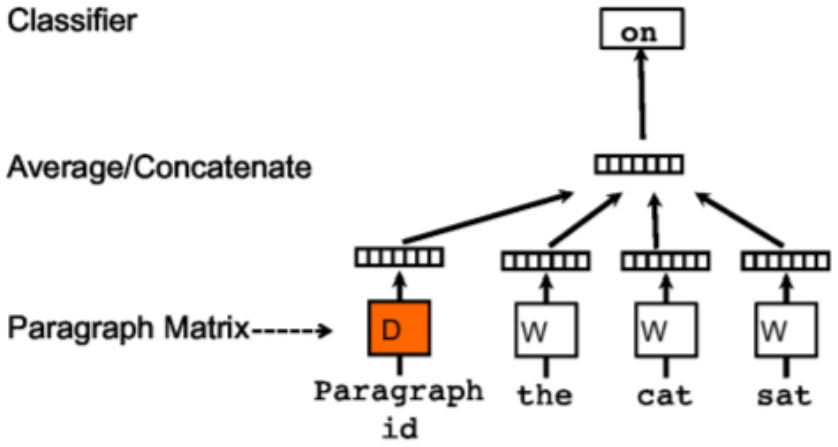
GloVe: Global Vectors for Word Representation

- Unsupervised learning algorithm for obtaining vector representations for words
- Ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning
- Pre-trained matrix model: Twitter – 2 billions tweets, 27 billions tokens , 25 to 200 dimensional features



Doc2Vec - Distributed Memory Model of Paragraph Vectors (PV-DM)

- Word2vec : Converts a word into a vector → losing ordering of the words
- Doc2vec: Learn word features + aggregate all the words in a sentence into a vector
  - Unsupervised algorithm that converts variable-length text to fixed-length feature representation.



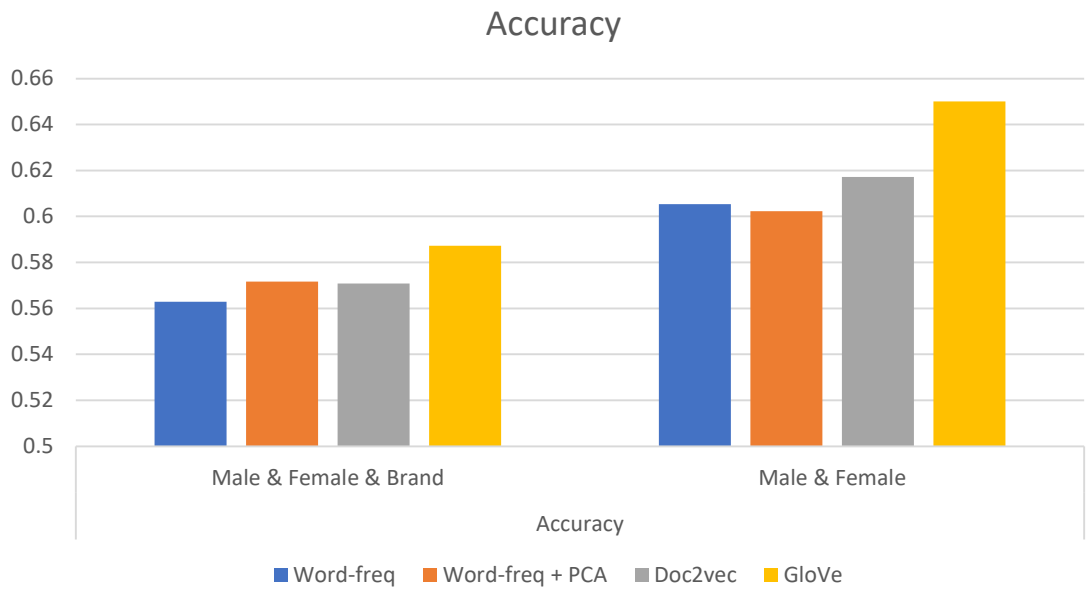
D: N x p matrix paragraph vector (each paragraph is mapped to p-dimensional features vector)

W: M x q matrix word vector (each word is mapped to q-dimensional features vector)

Q. Le, T. Mikolov. 2014. Distributed Representations of Sentences and Documents. In Proceedings of ICML 2014

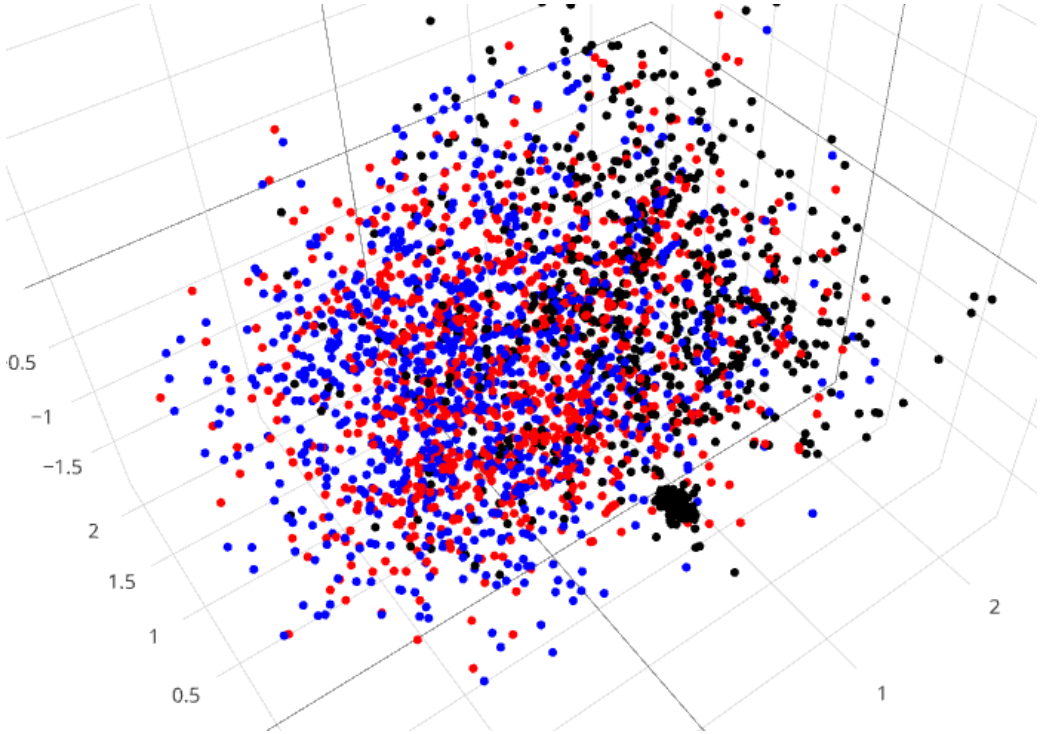
Analysis & Evaluation

		Word-freq	Word-freq + PCA	Doc2vec	GloVe
Accuracy	Male & Female & Brand	0.5629	0.5716	0.5708	0.5872
	Male & Female	0.6054	0.6023	0.6172	0.6500



Analysis & Evaluation

		Word-freq	Word-freq + PCA	Doc2vec	GloVe
Precision	Male	0.4888	0.5131	0.4898	0.5342
	Female	0.5678	0.5838	0.6043	0.5930
	Brand	0.6341	0.5961	0.6027	0.6294
Recall	Male	0.4359	0.3564	0.4183	0.4312
	Female	0.6060	0.6132	0.6050	0.6798
	Brand	0.6580	0.7770	0.7096	0.6477
F1 score	Male	0.4608	0.4203	0.4512	0.4771
	Female	0.5862	0.5981	0.6046	0.6334
	Brand	0.6457	0.6745	0.6516	0.6383



First 3 principal components

Black: brand; Red: female; Blue: Male

Conclusion

- After all, we’re not all that much different. We use a lot of the same words
- GloVe performs best because its underlying concept that distinguishes man from woman, i.e. sex or gender, or king and queen.
- Doc2vec performs weaker than GloVe because it could be the lack of its pre-trained model from very large corpus (only unsupervised learning on training data)

Thank you