Analysis in Twitter Gender Classification Final Report

Chuong Trinh
ctrinh@tamu.edu

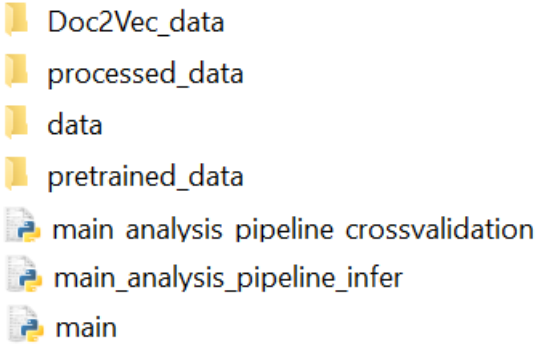

Fig. 1. The folders must be constructed exactly like the figure above

I. How to run

The following step and folder structure is needed to be able to run the experiment:

1) Have the file and folder's structures as the 1
2) Make sure the glove file is in "pretrained data" folder. Because the size is large, we do not put it in submission folder.
3) "data" folder contains .csv file dataset collected from [1]
4) "pretrained data" is where we store the glove 2 billions tweets model
5) Run "main.py" to retrieve and process data from original csv file.
6) Run either "main analysis pipeline crossvalidation" or "main analysis pipeline infer" to conduct experiments

II. Introduction

People are not comfortable to disclose their real genders in social public media such as Tweeter. In addition, a lot of tweets are not even come from human-beings such as spammers, advertises, automatic robot texts, etc. Moreover, human writing styles are changing through time and events; therefore, creating challenges for previous training models to adapt with current state. This also makes it even harder in reality where the trends of mixing genders increased.The goal of my project is to investigate machine learning techniques in natural language processing to predict the gender of authors of their tweets.

III. Motivation

Gender prediction has been long-time interest in research community due to its difficulties and benefits. In many recommendation tasks, user's gender is one of the most critical features to effectively learn a good model because gender of a person contributes greatly to his/her opinions, political stances,interests, preferences, etc, so that we can be able to predict gender identity in real-time to provide quality services or recommend products. However, such information is not easy to be found due to user's privacy concerns.

IV. Dataset

In this project, for Tweets evaluation, a dataset, provided by CrowdFlower, containing more than 20000 tweets with human labels can be found on Kaggle [1]. Dataset contains 20,000 tweets labeled by Amazon Turker and CrowdFlower's labeling system. About 75-percent of data can be used due because of non-English tweets. There are three types of labels such as male, female, and brand. "brand" tweets are considered to be spamming, advertising, business, or particular organizing accounts.

In addition, a recent query dataset including tweets from these users in the Kaggle dataset for comparison analysis along with their profile information for true labels.

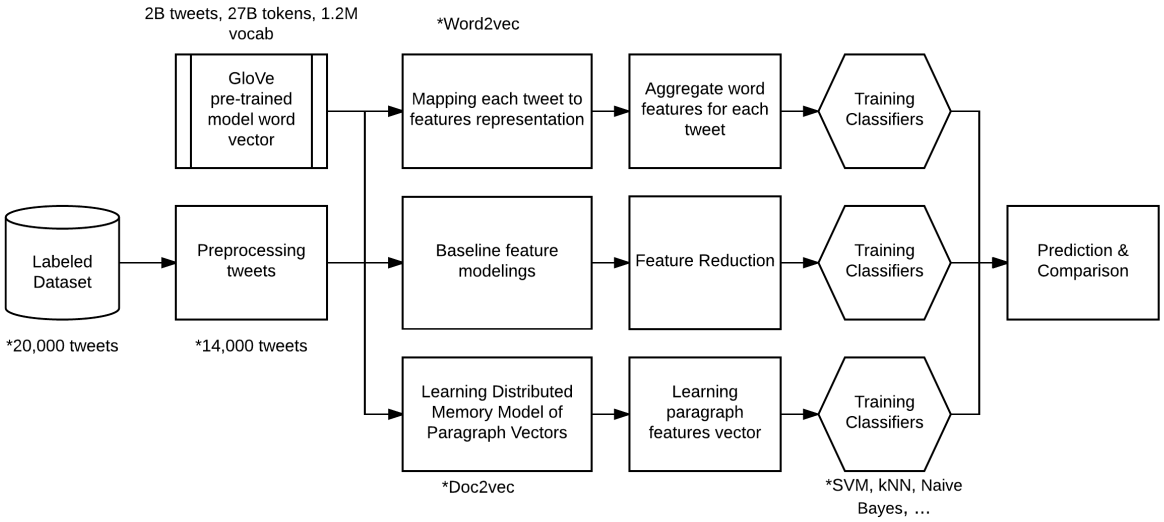

Fig. 2. Example of User interaction with current password manger applications

V. Methodology

A. Preprocessing

After first collecting data from CrowdFlower, the tweets are then pre-processed using typical tokenizer such as [**?**] and [**?**]. Basically, this converts all url or mention to an easy representation such as ¡url¿ or @¡user¿. In addition, about 14,000 out of 20,000 tweets are considered valuable due to its confidence value, which indicates how confident whether this tweet belongs to a specific gender, provided by human.

B. Model learning

In this work, we explore three different learning model: typical baseline vector representation through counting words in the dataset, using word embedding GloVe model [**?**] to learn our data, and building model through Doc2vec inspired by Distributed Memory Model of Paragraph Vectors [**?**]

1) GloVe: Global Vectors for Word Representation: This is an unsupervised learning algorithm for obtaining vector representations for words. The idea of this algorithm is to explore the word-word co-occurrence probabilities to explore some forms of meaning related to pair of words. For example, two words often appear together in the whole corpus can be seen as highly similar words. In this project, we will use pre-trained matrix model generated from 2 billions tweets with 27 billions tokens, and each word is represented with a 25 200 dimensional features vector.

2) Doc2Vec - Distributed Memory Model of Paragraph Vectors (PV-DM): An Unsupervised algorithm that converts variable-length text to fixed-length features representation. The intuition is similar to Word2vec where each word is mapped to high dimensional features vector based on word-word co-occurrence in training corpus. The task to predict the next work based on words in current window. However, Doc2vec also learn a paragraph vector that maps each paragraph into a high dimensional vector. The word vectors are shared among all documents; however, the paragraph vector is only specific used in its paragraph. By concatenating all word vectors in the current window size with paragraph vector, it then fits into classifiers to predict the next word. The process is updated through gradient descent with back propagation.

VI. Experiment Results

A. Male vs Female

GlOve performs the best among all of the methods based on its accuracy. The reason is that GloVe is already trained using large corpus; hence, it can model it well while Doc2Vec seems promising but

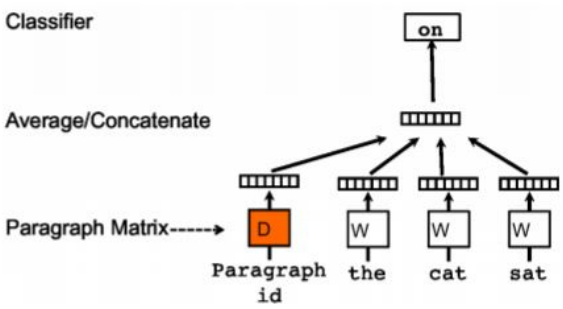

Fig. 3. Framework learning for Distributed Memory Model of Paragraph Vectors

		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

Fig. 4. Accuracy tables for both experiments

lack of pre-trained model so it mainly depends on the training data which may contains noises and errors.

B. Male vs Female vs Brand

Classifying brand or advertising accounts is much easy due to its use of words. However, male and female are still difficult to distinguish them.

VII. 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)

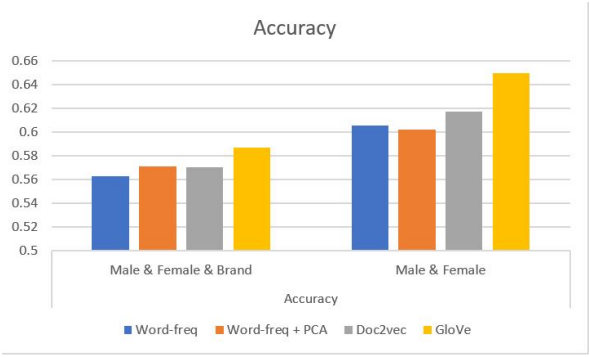

Fig. 5. Accuracy visualization

		Word-freq	Word-freq + PCA	Doc2vec	GloVe
Precision	Male	0.5811	0.5844	0.5880	0.6426
	Female	0.6238	0.6140	0.6425	0.6550
Recall	Male	0.5374	0.4974	0.5879	0.5552
	Female	0.6643	0.6932	0.6425	0.7322
F1 score	Male	0.5582	0.5374	0.5879	0.5957
	Female	0.6433	0.6512	0.6425	0.6914

Fig. 6. Precision, Recall,and F1 results

		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

Fig. 7. Precision, Recall,and F1 results

References

[1] CrowdFlower. http://aiweb.techfak.uni-bielefeld.de/content/bworld-robot-control-software/, 2015.

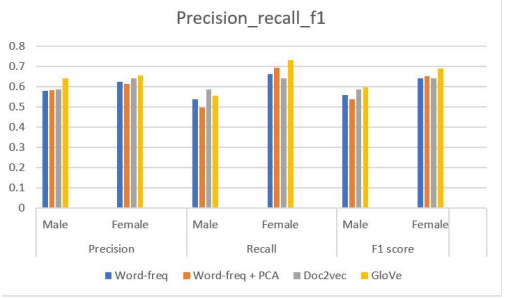

Fig. 8. Precision, Recall,and F1 visualization