Learning from Self-Reported Twitter Data by FastText with Sample Selection Bias Correction

Madison Cooley[1], Hui Hu[1], Dr. Chao Lan[1]
[1] University of Wyoming, Machine Learning Lab

ABSTRACT

Learning to predict Twitter user gender is an important task, but it is often challenged by the sparsity of manually labeled training data. Recently, Emmery et al proposed to construct automatically labeled training set, which only includes users who self-report their genders. They train and test fastText, a popular gender prediction technique, on this data set and show it achieves promising performance.

We argue Emmery's experiment design is unfair, as it ignores the potential bias of self-reporting users in representing the entire user population. In this project, we show fastText trained on the self-reported set generalizes poorly on a test set of a broader user population. To alleviate this problem, we propose to correct the sample selection bias for Fasttext by employing the kernel means matching (KMM) method. In our experiments, we show the proposed technique improves the generalization performance of fastText when it is trained on the self-reported data set.

INTRODUCTION

Gender information is important for many research applications because it plays a role in decision making, communication, and preferences. Gender influences the way an implementation strategy works on certain people [6]. For example, Twitter has been used to predict the onset of depression and mental illness [5], in this case, the gender of the user would be important in order to formulate a proper intervention strategy if the user was found to be depressed.

Using Twitter as a data source is popular due to its large amount of free and easily accessible data. Gender information is important for research, however, Twitter does not readily store this information about its users. This means demographics such as gender, race, etc, must be inferred from a user's timeline using supervised methods. These methods require large amount of labeled data, which is time consuming and costly to generate. A solution to labeling data is automatically generating labels by querying Twitter for users who self-report their own gender. This is done by querying tweets that contain the phrases "I'm a man," "I'm a woman," "I'm male," etc [2] (see figure 9 for an example). We will refer to this dataset as the self-reported dataset.

We employ the kernel means matching (KMM) [4] technique to correct the potential sample selection bias problem in the standard fastText model when it is trained on the self-labeled dataset.

Figure 9 . An example of a tweet which would have been queried to create self-labeled dataset.

METHODOLOGY

fastText Model Architecture:

The fastText [1] model, shown in figure 1, is a popular text classification model. It is a simple convolutional neural network with one hidden embedding layer which learns sentence representations. The output layer produces a probability distribution over the given classes using the softmax function.

$$-\frac{1}{N}\sum_{n=1}^{N} y_n log(f(BAx_n))$$

Figure 1. Shows fastText model architecture.

Kernel Means Matching (KMM):

KMM is a method of correcting sample selection bias. This method re-weights training points such that the means of the training and testing points are close in a reproducing kernel Hilbert space (RKHS). Figure 2 shows how the reweighting coefficient (beta) is optimized.

$$||\frac{1}{N^{tr}}\sum_{i=1}^{N^{tr}}\beta_i\Phi(x_i^{tr}) - \frac{1}{N^{te}}\sum_{i=1}^{N^{te}}\beta_i\Phi(x_i^{te})||^2$$
$$= minimize_\beta \frac{1}{2}\beta^T K\beta - k^T\beta$$
$$\text{s.t. } \beta_i \in [0, B] \text{ and } |\textstyle\sum_{i=1}^{N_{tr}}\beta_i - n_{tr}| \le N_{tr}\epsilon$$

Figure 2. Shows the optimization procedure for KMM re-weighting coefficient beta.

*fastKMMText Model Architecture:

We propose applying KMM to the fastText model to reduce bias, which we will refer to as fastKMMText. See figure 8 for possible reasons for the bias in the data. Figure 3 shows how the reweighting coefficient (beta) is applied.

$$-\frac{1}{N}\sum_{n=1}^{N}\beta_n y_n log(f(BAx_n))$$

Figure 3. Shows fastKMMText model architecture.

EXPERIMENT

Data Collection:

- *Self-reported Dataset:* This dataset was created by using the Twitter Streaming API. Tweets which contained the phrases "I'm a man," "I'm a woman," "I'm female,", "I'm male," etc. were collected. 200 of the most recent tweets from the timelines of the users who tweeted those phrases were then collected. The label of each user then became that user's self-report of their own gender.
- *Manually Labeled Dataset [3]:* This dataset was created by collecting tweets with the Twitter Streaming API. 10,000 tweets containing the words "and" were collected and 10,000 tweets containing the words "the" were collected. Querying tweets this way ensures that no particular type of person is represented more within the dataset. Human annotators then visited each profile and assigned a gender label. After collecting the tweets we were left with 624 manually labeled instances.

Experimental Design:

- Trained the original model, and the new model ten separate times to get an average classification error and standard deviation shown in figures 4, 5, 6 and 7. Each experiment was trained on a subset of 10,000 instances of the self-reported dataset, and tested on 9,000 instances of the self-reported training set, and the 626 instances of the manually labeled dataset.

	Total Users	Female	Male	Other	Active Users
Self-reported	6246	3986	2112	148	2956
Manually Labeled	9967	2837	3063	4067	3189

Figure 8. Shows the gender distribution of the self-reported and manually labeled datasets. The distribution in the self-reported has many more females than males, whereas the manually labeled set has a much more even distribution. This could be a possible reason for the bias.

RESULTS

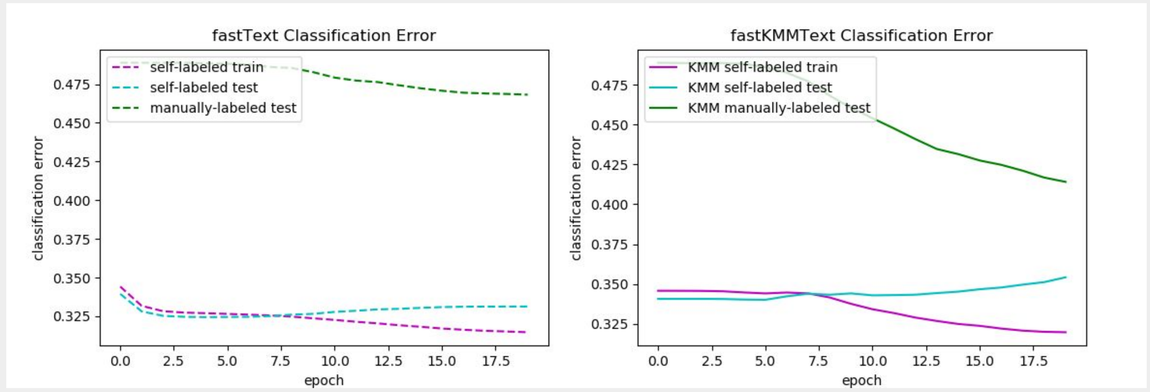

Figure 4. Average classification error for ten trials with the fastText model.

Figure 5. Average classification error for ten trials with the fastKMMText model.

Epoch	2	4	6	8	10	12	14	16	18	20	Avg. Std
Self-labeled Training Set	0.332	0.327	0.327	0.325	0.324	0.321	0.319	0.317	0.316	0.314	0.00364
Self-labeled Testing Set	0.328	0.325	0.325	0.325	0.327	0.328	0.329	0.33	0.331	0.331	0.00292
Manually Labeled Testing Set	0.489	0.489	0.489	0.486	0.483	0.477	0.474	0.470	0.469	0.469	0.00532

Figure 6. Average classification error and standard deviation for ten trials on fastText model.

Epoch	2	4	6	8	10	12	14	16	18	20	Avg. Std.
Self-labeled Training Set	0.346	0.345	0.344	0.344	0.337	0.332	0.327	0.324	0.321	0.320	0.01004
Self-labeled Testing Set	0.341	0.351	0.340	0.344	0.344	0.343	0.344	0.347	0.350	0.354	0.00672
Manually Labeled Testing Set	0.489	0.489	0.486	0.477	0.460	0.448	0.435	0.427	0.421	**0.414**	0.01748

Figure 7. Average classification error and standard deviation for ten trials on fastKMMText model.

REFERENCES

[1] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.
[2] Chris Emmery, Grzegorz Chrupała, and Walter Daelemans. 2017. Simple queries as distant labels for predicting gender on twitter. In Proceedings of the 3rd Workshop on Noisy User-generated Text, pages 50–55.
[3] CrowdFlower AI. (2015 November). Twitter User Gender Classification, 1. Retrieved May 2018 from https://www.kaggle.com/crowdflower/twitter-user-gender-classification/home
[4] Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Scholkopf, B. Covariate shift by kernel mean matching. In Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N.D. (eds.), Dataset Shift in Machine Learning. MIT Press, 2009.
[5] Reece AG, Reagan AJ, Lix KLM, Dodds PS, Danforth CM, Langer EJ (2016) Forecasting the onset and course of mental illness with Twitter data. arXiv:1608.07740
[6] C. Tannenbaum, L. Greaves, I.D. Graham, Why sex and gender matter in implementation research, BMC Med Res Methodol, 16 (2016), p. 145