Semi-Supervised Language Models for Identification of Personal Health Experiential from Twitter Data: A Case for Medication Effects

Minghao Zhu[1], Keyuan Jiang[2]
[1]School of Computer Science and Technology, Donghua University, Shanghai 201620, China
[2]Department of Computer Information Technology and Graphics,
Purdue University Northwest, Hammond, Indiana 46323, U.S.A.
minghao.zhu0@gmail.com, kjiang@pnw.edu

Abstract

First-hand experience related to any changes of one's health condition and understanding such experience can play an important role in advancing medical science and healthcare. Monitoring the safe use of medication drugs is an important task of pharmacovigilance, and first-hand experience of effects about consumers' medication intake can be valuable to gain insight into how our human body reacts to medications. Social media have been considered as a possible alternative data source for gathering personal experience with medications posted by users. Identifying personal experience tweets is a challenging classification task, and efforts have been made to tackle the challenges using supervised approaches requiring annotated data. There exists an abundance of unlabeled Twitter data, and being able to use such data for training without suffering in classification performance is of great value, which can reduce the cost of laborious annotation process. We investigated two semi-supervised learning methods, with different mixes of labeled and unlabeled data in the training set, to understand the impact on classification performance. Our results from both pseudo-label and consistency regularization methods show that both methods generated a noticeable improvement in F1 score when the labeled set was small, and consistency regularization could still provide a small gain even a larger labeled set was used.

1 Introduction

First-hand experience related to any changes of one's health condition and understanding such experience can play an important role in advancing medical science and healthcare. What has happened since the COVID-19 pandemic started demonstrates potential values and applications of such experiential knowledge, ranging from understanding the symptoms of the viral infection, to learning the effects after vaccination – personal experience shared on social media pertaining to symptoms of infection and side effects of vaccine may help us gain insight into the virus and vaccine, and ultimately advance medical science and clinical practice. Post-market surveillance is an important activity of pharmacovigilance, and experiential information from the users of the therapeutic products can help supplement the knowledge of medication effects gathered with other data sources. Many nations recognized importance of patient reporting of drug effects and its scientific value (van Hunsel et al., 2012), and potential benefits of patient reported drug events were studied (de Langen et al., 2008; Blenkinsopp et al., 2007; Avery et al., 2011; Anderson et al., 2011). Patient reporting could help identify new adverse drug effects sooner than that by healthcare professionals alone (Egberts et al., 1996). A study by McLernon and colleagues found that patient reports contained a higher median number of suspected adverse drug reactions (ADRs) per report, and described reactions in more detail, and they were richer in descriptions of reactions than those from healthcare providers (McLernon et al., 2010). One study showed that consumers reported seven categories of ADRs unreported by the other sources, and the investigators recommended that consumers should be included in systematic drug surveillance systems (Aagaard et al., 2009).

It is a primary concern of monitoring the health conditions as well as the safe use of pharmaceutical products to find a rich and accessible data source and build an efficient system to process and analyze the data.

Proceedings of the BioNLP 2021 workshop, pages 228–237
June 11, 2021. ©2021 Association for Computational Linguistics

People share their personal health experience on social media thanks to their prevalence. As such, social media have been considered an alternative and active data source for studying health surveillance. Platforms like Twitter allow users to express their health condition freely online. There exist many studies of using social media for health surveillance, such as influenza outbreak detection (Culotta et al., 2010), public health analyzing (Paul et al., 2011), dental pain surveillance (Heaivilin et al., 2011).

Personal experience tweets (PETs) related to medication use are defined as Twitter posts expressing one's first-hand personal encounters or observations about their health conditions after the administration of pharmaceutical drugs. Medication effects can be undesirable feelings caused by medication's side-effects which exacerbate one's health condition, or beneficial effects which help alleviate one's health condition after medication intake. Below are examples of personal experience tweets (PETs) pertaining to medication use (the medication names are in boldface and experiences are underscored):

> "**codeine** got me feeling sloooow **xanax** got me sleeping"
>
> "this **vicodin** is putting me to sleep"
>
> "**morphine** actual makes ur face so itchy think ive scratched ma whole face off"

As a general purpose social media platform, Twitter contains posts on almost all thinkable topics and many of them are unrelated to health, let alone misspellings, incorrect grammas, and creative short texts found in the posts. Therefore, differentiating personal experience tweets (PETs) from other irrelevant or noisy tweets is challenging. Efforts have been made in previous endeavors. Personal pronouns were chosen as the feature to distinguish PETs from irrelevant tweets such advertisements, news, even spams (Jiang and Zheng, 2013). An effort was made to engineer features including Twitter specific features, n-grams, punctuation elements, and topics, but the topic feature was discarded because of its significant efforts required to achieve minimum merit of classification performance improvement (Alvaro et al., 2015). Later, a set of 22 Twitter features including textual data and metadata was engineered by Jiang and his colleagues (2016) and conventional machine learning methods such as decision tree were applied to predict PETs. The concept of deep grammulator was proposed to include a textual feature with expressions in one class but not in the opposite class, to enhance the discriminatory power of the classification (Calix et al., 2017). In recent studies, application of neural embedding and recurrent neural network (LSTM) was investigated to improve the classification performance (Jiang et al., 2018). In the latest development, pre-trained attention-based language model approaches based on BERT and RoBERTa language models were explored to have achieved even better classification performance (Jiang et al., 2019, Zhu et al., 2020).

However, all previous attempts are based on fully supervised learning mechanisms, requiring the laborious effort of annotation which can be cost-prohibitive if a large amount of accurately labeled data is needed with a limited budget.

Unlike labeling text data in formal writing, annotating Twitter posts can be especially challenging, because of various complexities associated with the data such as misspellings, use of nonstandard language, and lack of sufficient context within the limited space. In addition, supervised methods are widely used on social media data in health-related tasks due to their higher accurate than unsupervised approaches, requiring manual annotation of large corpora of data. Furthermore, the subjectivity of labeling social media data is of concern. Inter-annotator agreements tend to be relatively low for social media–based annotation tasks even with domain experts as annotators (O'Connor, 2020).

On social media, there exists an abundance of unannotated data, and being able to use such large amount of unlabeled data for training may improve the classification performance, without spending a significant amount of resources in annotation. In this study, we investigate how the classification performance in predicting personal experience tweets related to medication use will be affected using a relatively small amount of annotated data instances in training an attention-based language model.

2 Background

Data and features determine the upper limit of machine learning, while models and algorithms can only approach this upper limit. For most machine learning tasks, the amount of labeled data directly affects the final learning performance.

In order to obtain a large set of labeled data, researchers usually need to spend a significant amount of time to annotate data, and the cost of annotation process can be drastic and sometimes unaffordable. On the other hand, there is abundance of unlabeled data which are easily accessible. Semi-supervised learning is a promising approach in machine learning which uses the combination of both of labeled and unlabeled data. Studies have shown that it can achieve considerable improvement for various tasks with a small labeled dataset in conjunction with a large set of unlabeled data.

Based on the cluster assumption, semi-supervised learning methods are mainly classified into two different categories: proxy-label and consistency regularization. The proxy-label method uses the supervised model or its variants to generate proxy labels for the unlabeled data, and the proxy labels are mixed with true labels to provide additional features to benefit training process. A typical implementation of such approach is the pseudo-label method as described below (Lee 2013).

Consistency regularization is a relatively new method. In consistency training, models are regularized to be invariant to a small amount of noise applied to inputs or hidden neurons. The invariance can be all or parts of hidden states in the network, or the outputs of the model. Common methods include Temporal Ensembling (Laine et al., 2016), Mean Teachers (Tarvainen et al., 2017), and Unsupervised Data Augmentation (Xie et al., 2019).

3 Method

The pipeline of data processing and analysis of our methods is depicted in Figure 1. Our approach of identifying personal experience tweets was based upon the two semi-supervised learning methods mentioned above: (1) Pseudo-Label which generates pseudo labels for unlabeled tweets and trains the model with labeled tweets together in a supervised behavior, and (2) Consistency Regularization which does not generate any labels for unlabeled data but tries to keep the consistency of the model outputs with the same inputs injected with some stochastic noise.

Our language model is based upon the Google's attention-based Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). Although the Google team highly recommends using its pre-trained language

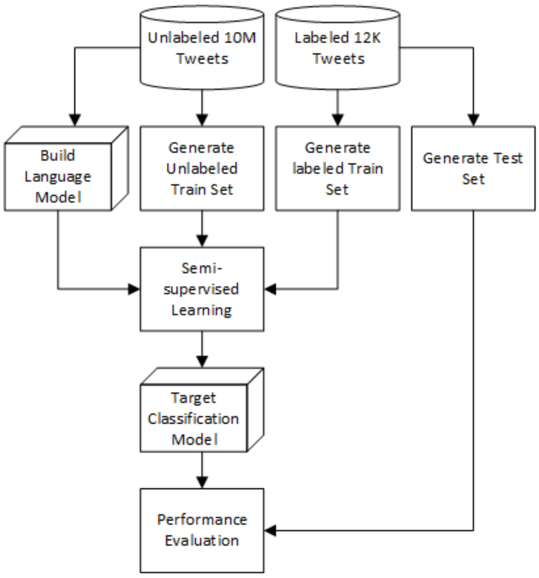

Figure 1. Pipeline of Data Processing and Analysis.

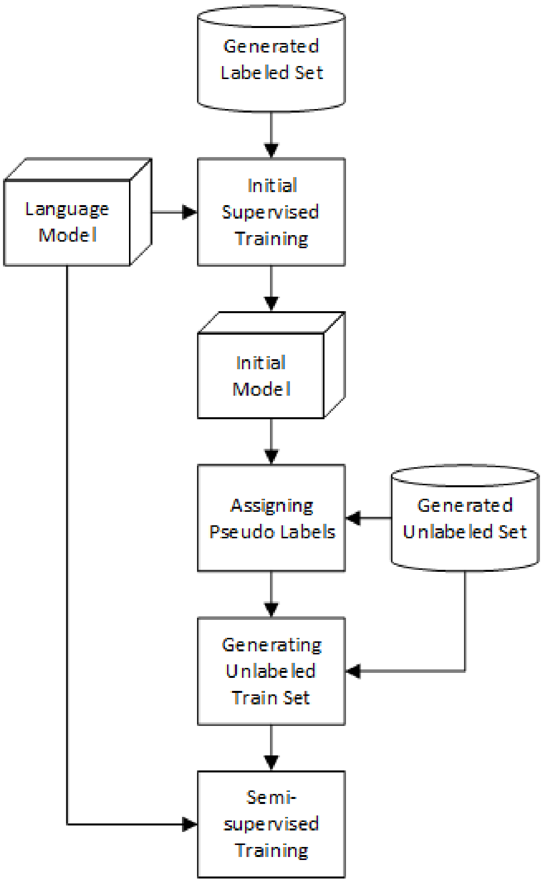

Figure 2. Setup of Pseudo-label Based Semi-supervised Learning.

model, a Microsoft team demonstrated that the domain-specific BERT language model can perform better (Gu et al., 2020), and we adopted Microsoft's approach in this work.

3.1 Pseudo-Label

Figure 2 shows our pseudo-label based method. A naive but efficient semi-supervised learning structure was implemented in this method by combining both labeled and unlabeled (pseudo-labeled) sets of data. First, the model was trained in a fully supervised manner with the labeled set, and the trained model was used to assign pseudo labels to unlabeled tweets. Later, a subset of unlabeled tweets was chosen for prediction by the initial model. From the prediction results, each unlabeled tweet was assigned a pseudo-label whose class has the maximum predicted probability. Due to the class imbalance of the corpus of labeled tweets (Table 1 below), the composition of the unlabeled train set was made up of the classes which were inversely proportional to the annotated corpus (PET: non-PET = 3:1). Finally, both sets of labeled and pseudo-labeled tweets were combined to train the model.

3.2 Consistency Regularization

For consistency training, the framework of Π-model (Laine et al., 2016) was used for reference. Figure 3 and Algorithm 1 shows our method. In this approach, two parts of the loss function were considered: (1) the classification loss, which is usually the cross entropy, and (2) the consistency loss. The classification loss was only applied to the labeled tweets while the consistency loss was for all. During training, each input was evaluated twice with hidden noise injection, and the difference between the two evaluation results was calculated by the squared error. In combining these two parts of loss, a weight variable was applied to scale the consistency loss. The weight variable was initialed to zero, allowing the training loss to be dominated by classification so that the model could learn from labeled data first, and later the weight variable was recalculated in the training epochs to reflect the consistency loss consistent with the data. The value of this weight would be adjusted to a fixed level according to the number of labeled and unlabeled data used in training. It was an important and tricky step. This is because if the value is too small, the training is more likely to be supervised and prone to overfitting, and on the contrary, the model trained with an overly large weight will noticeably deteriorate, making the predictions less meaningful. Refer to Appendix A for the details of training parameters.

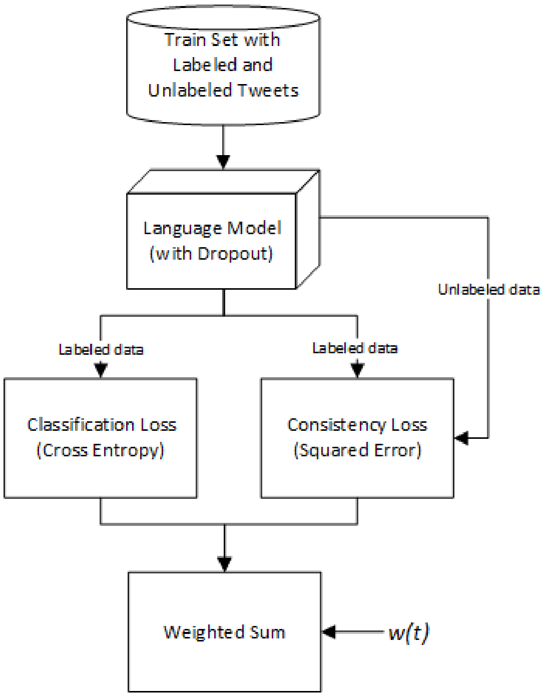
Train Set with Labeled and Unlabeled Tweets
Language Model (with Dropout)
Unlabeled data
Labeled data
Labeled data
Classification Loss (Cross Entropy)
Consistency Loss (Squared Error)
Weighted Sum
w(t)

Figure 3. Setup of Consistency Regularization

```
Require: $x_i$ = training stimuli
Require: L = labeled set
Require: $y_i$ = labels of labeled data
Require: w(t) = weight ramp-up function for consistency loss
Require: $f_\theta(x)$= neural network with dropout and parameter θ

for t in [1, num_epochs] do
  for each minibatch B do
    $z_{i\in B} \leftarrow f_\theta(x_{i\in B})$
    $z'_{i\in B} \leftarrow f_\theta(x_{i\in B})$
    $loss \leftarrow -\frac{1}{|B|}\sum_{i\in(B\cap L)}(y_i \log z_i + (1 - y_i)\log(1 - z_i)) + w(t)\frac{1}{|B|}\sum_{i\in B}|z_i - z'_i|^2$
    Update θ by using Adam
  end for
end for
```

Algorithm 1. Pseudo code of consistency regularization

Dropout regularization was chosen as the noise injection method in the hidden layer of the model. Because the dropout performed stochastically, the model outputs were different from training even with the same inputs. Therefore, there were two different evaluation results for each input, and the expected goal was to minimize them.

3.3 Network Structure and Baseline

Pre-trained language models such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019) and Generative Pre-trained Transformer (GPT) (Radford et al., 2019) have achieved the state-of-the-art performances in many NLP benchmarks and downstream tasks. Efforts have been made to apply the pre-trained BERT and RoBERTa (and its continuous pre-training) language models in identifying PETs (Jiang et al., 2019; Zhu et al., 2020) which demonstrated significant improvement in all classification measures. However, these language models are all pre-trained using general-domain texts such as news, Wikipedia and/or BookCorpus which may not have significant relevance to Twitter posts, especially pertaining to personal health experience. Gu and colleagues (Gu et al., 2020) found that the language model learned with domain specific data can provide substantial gains over the general domain language model, and domain-specific learning from the scratch is better than the one that starts with the general domain model and is updated with the specific domain data. Therefore, we decided to pre-train a domain-specific BERT language model from scratch to investigate its performance. The pre-training process used 10M unlabeled medication-related tweets we collected, and a new set of in-domain sub-word vocabulary was generated. See Appendix A for detail settings. This newly pre-trained language model was utilized as a network backend for both semi-supervised learning approaches as well as the baseline method.

For the baseline, supervised learning was considered. Following the official transfer learning guideline, the domain-specific BERT was transferred for binary classification and trained in a fully supervised way with the same labeled data as semi-supervised methods used.

3.4 Data

A set of 22 million raw tweets was collect with Twitter Streaming APIs from 25 August 2015 to 7 December 2016, and another set of 52 million tweets posted between 2006 as 2017 was collected using a home-made crawler which followed the crawling policy documented in the Twitter.com's robots.txt file. To clean the collected raw data, both sets were filtered by a set of brand and generic medication names, and duplicate as well as non-English tweets were all eliminated. The above pre-processing yielded a total 10 million tweets, among which a collection of 12,331 (12K) tweets was selected and annotated according to the annotation guideline which defines what is a PET and a non-PET. Table 1 lists the composition of labeled tweets.

First, a corpus of 10 million unlabeled tweets was used to build a sub-word vocabulary and pre-training domain-specific BERT language model. To avoid any possible data leakage, the 12K annotated tweets were excluded from the 10 million set. The set of 12K labeled tweets was used for both supervised baseline method and the labeled part of semi-supervised methods. As for the unlabeled part of semi-supervised learning, a stochastic subset of was randomly generated from the 10 million unlabeled corpus.

	PETs	Non-PETs	Total
Count	2,962	9,369	12,331

Table 1. Composition of annotated tweets.

3.5 Implementation

Our two semi-supervised learning methods were evaluated in the task of identifying personal experience tweets related to medication effects. To simulate the situation of the lack of labeled data and investigate how semi-supervised learning would perform for our task, both of our methods were tested with different percentages of the labeled data in our training set, and we evaluated the performance of the semi-supervised methods along with the fully supervised settings as the baseline.

Ten-fold cross-validation was applied to both supervised and semi-supervised approaches, and the mean of each classification measure was collected. For each fold, 10% of labeled tweets was partitioned as the test set which was only used for testing the classification performance. The labeled training set was randomly selected from the remaining 90% tweets by a proportion – we set six different proportions: 10%, 30%, 50%, 70%, 90% and 100% to investigate how the size of labeled data would affect the performance. Note that the initial training of pseudo-labeling method used the

%[1]	Method[2]	Acc. (PET)	Recall (PET)	Prec. (PET)	F1 (PET)	AUC/ROC
10	C.	**0.8597**	0.6239	**0.7500**	0.6786	**0.9079**
	P.	0.8390	**0.7512**	0.6500	**0.6916**	0.8944
	S.	0.8466	0.6607	0.7038	0.6706	0.9048
30	C.	**0.8723**	0.6894	**0.7559**	**0.7206**	**0.9235**
	P.	0.8528	**0.7802**	0.6661	0.7180	0.9155
	S.	0.8638	0.6904	0.7325	0.7081	0.9199
50	C.	**0.8765**	0.6945	**0.7689**	0.7294	**0.9287**
	P.	0.8591	**0.7927**	0.6798	**0.7303**	0.9211
	S.	0.8657	0.7474	0.7198	0.7280	0.9266
70	C.	**0.8797**	0.7134	**0.7678**	**0.7392**	**0.9313**
	P.	0.8646	**0.7765**	0.6963	0.7338	0.9239
	S.	0.8755	0.7171	0.7556	0.7338	0.9289
90	C.	**0.8829**	0.7316	**0.7680**	**0.7488**	**0.9338**
	P.	0.8685	**0.7846**	0.7037	0.7415	0.9266
	S.	0.8758	0.7552	0.7383	0.7452	0.9315
100	C.	**0.8855**	0.7245	**0.7802**	0.7508	**0.9344**
	P.	0.8678	**0.7937**	0.6990	0.7427	0.9278
	S.	0.8792	0.7620	0.7447	**0.7519**	0.9335

[1] the percentage of labeled data used.

[2] 'C' for consistency regularization, 'P' for pseudo-label, 'S' for supervised baseline.

Table 2. Classification performance

%	Method	F1	AUC/ROC
10	C.	2.749x10^{-1}	5.534x10^{-2}
	P.	**3.676x10^{-2}**	5.188x10^{-5}
30	C.	**3.653x10^{-2}**	**6.923x10^{-3}**
	P.	**3.828x10^{-2}**	3.372x10^{-4}
50	C.	4.115x10^{-1}	**2.538x10^{-2}**
	P.	3.002x10^{-1}	4.669x10^{-5}
70	C.	7.326x10^{-2}	**1.331x10^{-2}**
	P.	4.965x10^{-1}	5.089x10^{-4}
90	C.	1.979x10^{-1}	**4.581x10^{-3}**
	P.	6.620x10^{-2}	9.336x10^{-4}
100	C.	3.724x10^{-1}	1.018x10^{-1}
	P.	3.222x10^{-2}	3.619x10^{-4}

Table 3. T-test results (*p*-values) between baseline and semi-supervised learning methods (C. and P.) Boldface figures: < 0.05

same proportion of labeled data as the supervised baseline as well as the consistency regularization method to initial the model for assigning pseudo labels of unlabeled tweets. A fixed random seed was used to ensure that all methods have the same partition of data.

For both semi-supervised learning methods, a collection of 10K of unlabeled tweets was used as the unlabeled training set and mixed up with labeled ones. More specifically, in consistency training, the unlabeled train set was simply generated from the 10M unlabeled corpus by a stochastic sampler. But for pseudo-label, it was time consuming to predict for 10M tweets, and it was observed that the prediction of unlabeled data was imbalanced – the number of non-PETs was about ten times more than that of PETs. In other words, the training set would be more imbalanced if the 10K unlabeled set were to be used before assigning pseudo labels. To address the issue, and keep the training time tolerable, a set of 100K unlabeled tweets was chosen and assigned with pseudo labels, and afterwards, a training set of 10K unlabeled tweets was composed from the 100K tweets with pseudo labels. To balance the labeled set with a 1:3 ratio for PET: non-PET (Table 1), our pseudo-labeled training set was made up of 7,500 tweets with the PET pseudo-label and 2500 the non-PET pseudo label.

4 Results and Discussions

Listed in Table 2 are the measures of classification performance of our semi-supervised methods along with the supervised baseline in different proportions of labeled data (the highest values are in boldface).

As can be seen in Table 2, no single method achieved the best performance in all classification measures. The pseudo-label-based method achieved the best recall but showed the poorest accuracy, precision and AUC/ROC, in all the proportions of the labeled tweets used. On the contrary, the consistency regularization approach demonstrated the completely opposite

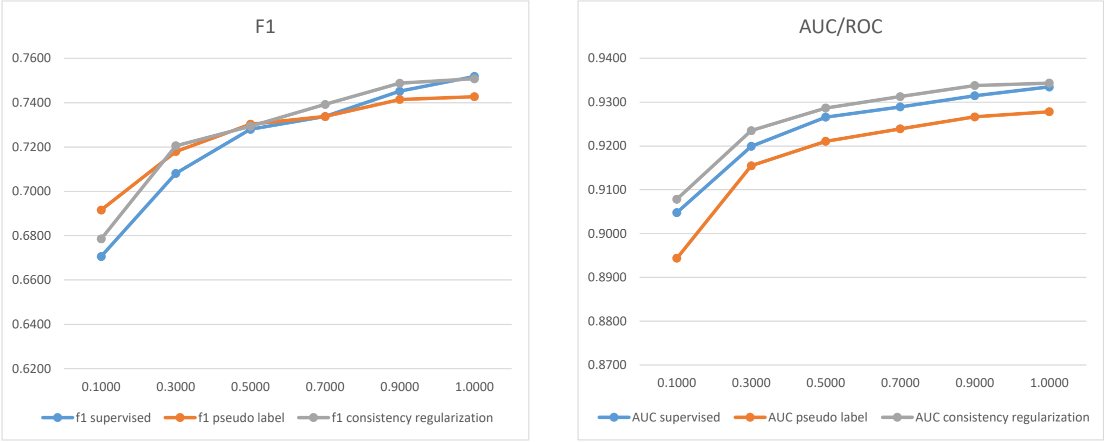

Figure 4. F1 and AUC/ROC Measures for Each Method.

performances, and notably its AUC/ROC, a more comprehensive measure for discriminability, is consistently higher than other two methods (Fawcett, 2006) – this may indicate that the consistency regularization method performed better in correctly predicting each class of data.

It seems to be inconclusive to state which one is the winner on a single classification measure. A higher accuracy does not necessarily indicate that the method is better because the accuracy is calculated based upon the prediction results of both positive and negative classes. The class imbalance in our test set may contribute to a higher accuracy if the majority class dominates. Recall and precision are of equal importance but neither of them alone can measure a network independently because recall focuses on the sensitivity of positive class while precision just measures the percentage of true positive samples in the predicted class. The F1 measure represents the harmonic mean calculated from both recall and precision, along with AUC/ROC, they have been considered as the most comprehensive measure among these five measures. Figure 4 shows the changes of F1 value for each method along with the proportions of the labeled data used. To confirm if the improvement difference does exist, we conducted statistical analysis (paired t-test) on the results between semi-supervised learning methods and baseline. We set the null hypothesis to that the difference between a pair of method does not exist while the labeled data remain the same. Table 3 shows the results of statistical analysis on F1 and AUC/ROC between the baseline and two semi-supervised learning methods. We set the *p*-value threshold to 0.05, meaning that any p-value less than 0.05, and if its corresponding value of performance measure larger than that of baseline, it indicates that the improvement difference does exist with statistical significance and it is *not* due to chance (these values are shown in boldface).

As shown in Figure 4, it is clear that both semi-supervised learning methods performed better than supervised baseline when a small amount of labeled data was used. In other words, semi-supervised learning may help us build a more robust PET prediction network in the situation where only a limited or small amount of the labeled data is available.

More specifically, according to Figure 4 and its corresponding *p*-values, the pseudo-label based method showed an outstanding F1 performance in tiny labeled sets (about 10% and 30% of the labeled data), and its improvement is of statistical significance. However, its performance appeared to deteriorate when the training set contained a large amount of annotated tweets (about more than 70% of labeled data). A possible explanation of this phenomenon might be the mislabeling of unlabeled tweets, which may mislead the training process if it has more labeled data than unlabeled one.

The consistency regularization method appears to be more stable than the pseudo-label method. It demonstrates consistently good performance even with the larger labeled sets (about 90% of labeled data), except that it shows a slight but not significantly lag behind the supervised baseline in F1 when the training set contains all of the labeled data. Although the statistical analysis seems does not show that the improvement in F1 is significant, the constant outperformance in AUC/ROC could be confirmed by the t-test – the larger AUC/ROC values in 30%, 50%, 70% and 90% of labeled data demonstrated their statistical significance. In the case of 10% of labeled data, the t-test results do not confirm the existence of performance differences. This may be attributed to the fact that too many unlabeled instances dominate the training set, which confused models in training. It may be concluded that Consistency Regularization is consistently better if the training data have more than 10% but less than 100% labeled instances and performs equally well with 100% labeled data in this task. However, it seems not performing well as the pseudo-label one in F1 when very few tweets are labeled (about 10%). This may indicate that labels are important in contributing to the performance, and the pseudo-label approach may be more suitable for the training with an extremely small labeled set, whereas consistency regularization seems to perform well in other situations.

5 Conclusion

In this study, we investigated classification performance using semi-supervised learning in identifying personal experience tweets. Two methods of semi-supervision were studied: pseudo-label and consistency regularization. Our results show that either of the methods performs outstandingly well in individual classification measures, in comparison with the supervised baseline method. However, the F1 and AUC/ROC scores show that both could enhance the network performance when a small size of the labeled set was used, and consistency regularization performed consistently well even with the datasets containing high number of labeled instances. In summary, either semi-supervised method performed well in predicting PETs with a small amount of labeled instances in the training set, which could significantly reduce the annotation effort. Although this study focused on the personal experience pertaining to medication effects, it is conceivable that our semi-supervised approach can help other health-related studies where personal experience is needed.

6 Acknowledgement

Authors wish to thank anonymous reviewers for their critiques and constructive comments which helped improved this manuscript, and College of Technology at Purdue University Northwest for providing funding to support this work.

References

Aagaard L, Nielsen LH, Hansen EH. Consumer reporting of adverse drug reactions: a retrospective analysis of the Danish adverse drug reaction database from 2004 to 2006. Drug Saf. 2009;32(11):1067-74.

Alvaro, N., Conway, M., Doan, S., Lofi, C., Overington, J. and Collier, N., 2015. Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use. Journal of biomedical informatics, 58, pp.280-287.

Anderson C, Krska J, Murphy E, Avery A, Collaboration YCS. The importance of direct patient reporting of suspected adverse drug reactions: a patient perspective. Br J Clin Pharmacol. 2011;72(5):806-22.

Avery AJ, Anderson C, Bond CM, Fortnum H, Gifford A, Hannaford PC, Hazell L, Krska J, Lee AJ, McLernon DJ, Murphy E, Shakir S and Watson MC. Evaluation of patient reporting of adverse drug reactions to the UK 'Yellow Card Scheme': literature review, descriptive and qualitative analyses, and questionnaire surveys. Health Technol Assess. 2011;15(20):1-234, iii-iv.

Blenkinsopp A, Wilkie P, Wang M, Routledge PA. Patient reporting of suspected adverse drug reactions: a review of published literature and international experience. Br J Clin Pharmacol. 2007;63(2):148-56.

Calix, R.A., Gupta, R., Gupta, M. and Jiang, K., 2017, Novem-ber. Deep gramulator: Improving precision in the classification of personal health-experience tweets with deep learning. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1154-1159). IEEE.

Culotta, A. (2010). Detecting influenza outbreaks by analyzing Twitter messages. arXiv preprint arXiv:1007.4748.

de Langen J, van Hunsel F, Passier A, de Jong-van den Berg L, van Grootheest K. Adverse drug reaction reporting by patients in the Netherlands: three years of experience. Drug Saf. 2008;31(6):515-24.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Egberts TC, Smulders M, de Koning FH, Meyboom RH, Leufkens HG. Can adverse drug reactions be detected earlier? A comparison of reports by patients and professionals. BMJ. 1996;313(7056):530-1.

Fawcett, T. An introduction to ROC analysis." Pattern recognition letters 27.8 (2006): 861-874.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J. & Poon, H. (2020). Domain-specific language model pretraining for biomedical natural language processing. arXiv preprint arXiv:2007.15779.

Heaivilin, N., Gerbert, B., Page, J. E., & Gibbs, J. L. (2011). Public health surveillance of dental pain via Twitter. Journal of dental research, 90(9), 1047-1051.

Jiang, K., Calix, R., & Gupta, M. (2016, August). Construction of a personal experience tweet corpus for health surveillance. In Proceedings of the 15th workshop on biomedical natural language processing (pp. 128-135).

Jiang, K., Chen, T., Calix, R. A., & Bernard, G. R. (2019, July). Prediction of personal experience tweets of medication use via contextual word representations. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 6093-6096). IEEE.

Jiang, K., Feng, S., Song, Q., Calix, R. A., Gupta, M., & Bernard, G. R. (2018). Identifying tweets of personal health experience through word embedding and LSTM neural network. BMC bioinformatics, 19(8), 67-74.

Jiang, K., & Zheng, Y. (2013, December). Mining twitter data for potential drug effects. In International conference on advanced data mining and applications (pp. 434-443). Springer, Berlin, Heidelberg.

Laine, S., & Aila, T. (2016). Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242.

Lee, D. H. (2013, June). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Workshop on challenges in representation learning, ICML (Vol. 3, No. 2).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

McLernon DJ, Bond CM, Hannaford PC, Watson MC, Lee AJ, Hazell L and Avery A. Adverse drug reaction reporting in the UK: a retrospective observational comparison of yellow card reports submitted by patients and healthcare professionals. Drug Saf. 2010;33(9):775-88.

O'Connor, Karen, Abeed Sarker, Jeanmarie Perrone, and G. Gonzalez Hernandez. "Promoting Reproducible Research for Characterizing Nonmedical Use of Medications Through Data Annotation: Description of a Twitter Corpus and Guidelines." J Med Internet Res 22, no. 2 (2020): e15861.

Paul, M., & Dredze, M. (2011, July). You are what you tweet: Analyzing twitter for public health. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 5, No. 1).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Schuster, M., & Nakajima, K. (2012, March). Japanese and korean voice search. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5149-5152). IEEE.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780.

van Hunsel F, Härmark L, Pal S, Olsson S, van Grootheest K. Experiences with adverse drug reaction reporting by patients: an 11-country survey. Drug Saf. 2012;35(1):45-60.

Xie, Q., Dai, Z., Hovy, E., Luong, M. T., & Le, Q. V. (2019). Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848.

Zhu, M., Song, Y., Jin, G., & Jiang, K. (2020, November). Identifying Personal Experience Tweets of Medication Effects Using Pre-trained RoBERTa Language Model and Its Updating. In Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis (pp. 127-137).

A Setup and Training Parameters

All networks were trained by Adam optimizer. For supervised baseline and the supervised initial training of pseudo-label method, we used the parameters suggested by the BERT's official fine-tuning guideline with learning rate being 1e-5, and training with a batch size of 32 for two epochs. For the semi-supervised learning, both methods were trained with a batch size of 128, started with a learning rate of 1e-5, then progressed with linear decay in four (for pseudo-label) and five (for consistency regularization) epochs.

The unsupervised weight variable used in consistency training was set by following the setup of Π-model (Laine et al., 2016). A Gaussian curve $\exp[-5(1 - T)^2]$ was used to ramp-up the weight, where T was increased linearly from zero to one during the ramp-up period. We set the first three epochs as the period to ramp-up the weight, and the maximum value of this weight variable was set to w_{max} * M / N where M is the number of labeled tweets and N is the total number of train set. w_{max} was manually set to 1 in this task.

The structure of domain-specific language model was based on BERT which has 12 layers, 768 hidden neurons and 12 self-attention heads. The pre-training process used masked language model task only and the next sentence prediction was discarded. Following the official guideline of pre-training settings, fifteen percent (15%) of words in each tweet were masked by special [MASK] tokens and the model is trained to predict the masked token correctly. We trained the language model with 10M unlabeled tweets for 400K steps with a batch size of 512. The learning rate started with zero, and warmed up to 1e-4 in first one thousand steps and then linearly decayed to zero in the rest of training steps.

A sub-word vocabulary with about 50K tokens was generated by applying WordPiece algorithm (Schuster et al., 2012) in our 10M unlabeled tweets.

Our implementation was based on TensorFlow (`www.tensorflow.org`) and Transformers (`huggingface.co/transformers`).