Classification of Online Medical Discourse by Modified Co-training

University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Rana Alnashwan

Supervisors Humphrey Sorensen
Adrian O'Riordan

Princess Nourah Bint Abdulrahman University

Background

Most of the current sentiment and text classification approaches were developed using supervised learning approach, which requires labelled training data. We believe that the limited availability of labelled data has been challenging and limiting for most of the researchers who wish to build and evaluate sentiment analysis and classification systems. Furthermore, whereas the process of manually labelling data is expensive, tedious, time-consuming and, in some cases, unfeasible, unlabelled user-generated data are often readily available and cost-free. Therefore, semi-supervised learning as a strategy for machine learning has drawn considerable attention, as it aims to leverage the benefits of using unlabelled data in order to enhance the performance of learning approaches with limited labelled data. One question arises, however: can we achieve a high degree of accuracy in the classification of medical discourse using unlabelled data (i.e., posts related to Lyme and Lupus disease)?

Approach

Co-training learning method, which is one of the well-known approaches in semi-supervised learning, originally proposed by Blum and Mitchell*. we apply a modified version of the co-training model to assess whether it would be adequate for domain dependent multi-classifications model. This modified version is differentiated from the original co-training model by less restrictive assumptions. This affects two aspects of co-training: 1) building the classifiers, and 2) Determining auto-labelled data.

1) Building $C1$ and $C2$

$C1$ and $C2$ were built using two views: a Domain independent (DI) feature set and a Domain dependent (DD) feature set as the two different views of each post from online health communities. This is based on our previous work in [3].

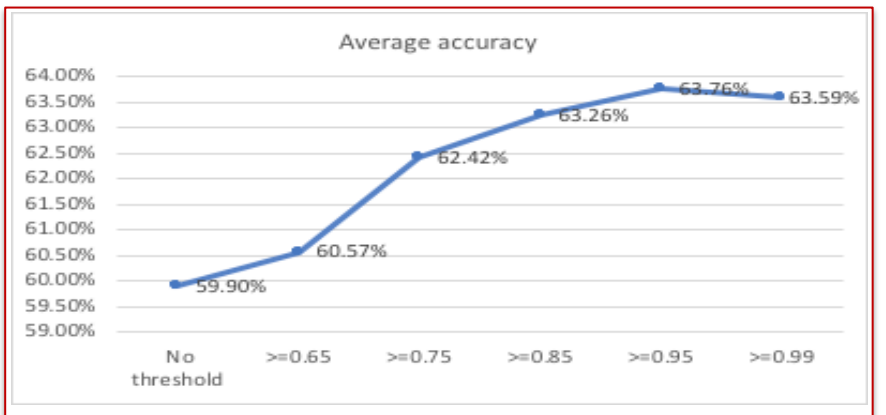
Algorithm The enhanced Co-training algorithm

Given:
- Set of L labelled training instance.
- Set of U unlabeled instance.
- Classifiers C1 and classifier C2

Parameters:
- Set an initial pool of U' that is created by randomly sampling u instances from U.
- K the number of iteration.
- **Loop for K iteration:**
  - Use L to train classifier C1 with considering only DI views
  - Use L to train classifier C2 with considering only DD views
  - Apply C1 to U' to label and assign confidence scores to each example
  - Apply C2 to U' to label and assign confidence scores to each example
  - Reduce disagreement between F(C1) and F(C2)
  - Select examples with highest confident that is not less that a threshold to add to L
  - Randomly choose u examples from U to replenish U'
- **End loop**
- **Output:** Classifiers C1, C2 and L.

2) Determining Auto-labelled Data

We applied co-training with a weak assumption by minimizing the auto-labelled data and selecting only examples on which both classifiers agreed. Then, we selected the top-N highest confidence examples from the comparable high confidence view classifiers according to the class distribution, subject to a threshold.

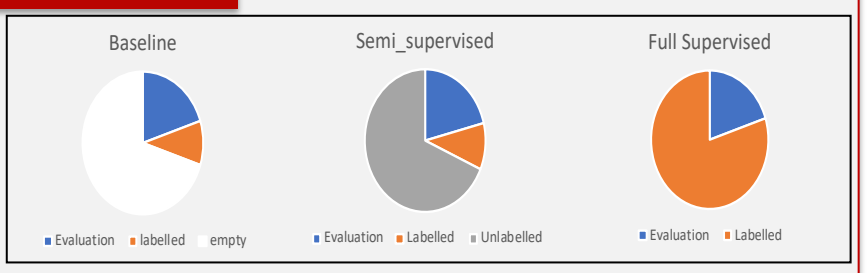


Average accuracy of different threshold values. When the confidence score is higher than the threshold □, the labelled dataset (U') will be selected as training data (L)

Result

Run type	Lyme Disease dataset		Lupus disease dataset	
	Labelled data		Labelled data	
	20%	25%	20%	25%
Baseline	51.01%	58.05%	54.38%	59.45%
Initial co-training	59.40%	61.75%	54.38%	58.53%
Modified co-training	62.08%	65.44%	62.21%	64.06%
Fully supervised	66.78%	69.80%	65.44%	66.82%

Baseline



20% of the data were reserved as an evaluation dataset and 80% were set as a co-training training dataset, which consisted of L and U.

*Blum, A., & Mitchell, T. (1998, July). Combining labeled and unlabeled data with co-training. In Proceedings of the eleventh annual conference on Computational learning theory (pp. 92-100). ACM.

Publications

1.Alnashwan, R., O'Riordan, A. P., Sorensen, H., & Hoare, C. (2016). Improving sentiment analysis through ensemble learning of meta-level features. In *KDWEB 2016: 2nd International Workshop on Knowledge Discovery on the Web*. Sun SITE Central Europe (CEUR)/RWTH Aachen University.
2.Alnashwan, R., Sorensen, H., O'Riordan, A., & Hoare, C. (2017, December). Multiclass Sentiment Classification of Online Health Forums using Both Domain-independent and Domain-specific Features. In *Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies* (pp. 75-83). ACM.
3.Alnashwan, R., Sorensen, H., O'Riordan, A., & Hoare, C. (2018). Accurate classification of socially generated medical discourse. *International Journal of Data Science and Analytics*, 1-13.
4.Alnashwan, R., Sorensen, H., O'Riordan, A. (2019). Classification of Online Medical Discourse by Modified Co-training. *In 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService). IEEE*.