Automating Classification of Survey Data Using Few Labeled Documents and Human Feedback

Bhavana Ganesh
Amazon
ganesh@amazon.com

Arushi Prakash
Amazon
arusp@amazon.com

Abstract

Companies rely on large-scale surveys, interviews, and focus groups to gauge customer sentiment about their products or programs, which contain free form text data rich in information. Researchers currently use a manual, time consuming processing which delays the time to get actionable insights. This paper presents a scalable solution where researchers can interact with a custom UI to annotate text data faster by learning from sparsely annotated data by the researchers, using natural language processing.

1 Motivation

Qualitative researchers analyzing text data traditionally use an inductive approach which is a cyclic process requiring a constant interplay between the researcher and the data(Williams and Moser, 2019). To address this bottleneck, we propose a weakly supervised learning approach to produce coded text with high accuracy while requiring only 5-20% of coded data.

2 Proposed Solution

Our proposed solution has two stages - (1) **Weakly Supervised Learning** to code text documents for the first time using few labeled documents, and (2) **Active Learning** to refine the trained model with signal from newly labeled documents.

1. Weakly Supervised Learning
   (a) Pseudo Document Generation - Representative keywords for each class are extracted from labeled documents using tf-idf weighting. These keywords are embedded in a p-dimensional space using GloVE vectors that were trained on Twitter data(Pennington et al., 2014). The vectors are normalized to project them into a common unit sphere. The semantics of each class are modeled using a von Mises Fisher vMF distribution (Banerjee et al., 2005). The vMF distribution is fit using maximum likelihood estimates. The pseudo documents are created per class by sampling the document vector from the vMF distribution for the class (Meng et al., 2018). The same process is used to generate pseudo documents for multi label scenario. We generate pseudo documents for single label and combine all possible combinations of labels to obtain pseudo documents containing more than one label.
   (b) Pre-training - Pseudo documents and the labeled documents are combined to train a CNN based neural network. The input to the CNN is a concatenation of all word vectors in a documents. The sigmoid activation function is used with cross-entropy loss to allow for multi label classification.
   (c) Self Learning - Use high confidence predictions on the unlabeled documents to fine-tune the model until there is no change in the labels between training epochs.
2. Active Learning - Users provides additional labeled samples from the original corpus with the same target classes. The trained model is fine-tuned by freezing the feature extracting CNN layers and tuning only the classification layers. A decay rate is introduced in this stage to avoid catastrophic forgetting and over-fitting on new labeled samples. In each iteration, predictions with the lowest average cross-entropy are pre-chosen for the user to label.

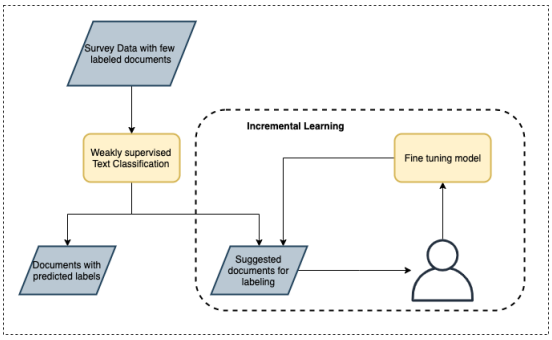

Figure 1: Flowchart of the proposed algorithm to incorporate labels from qualitative researchers and model predictions to create a more accurate and scalable coding process

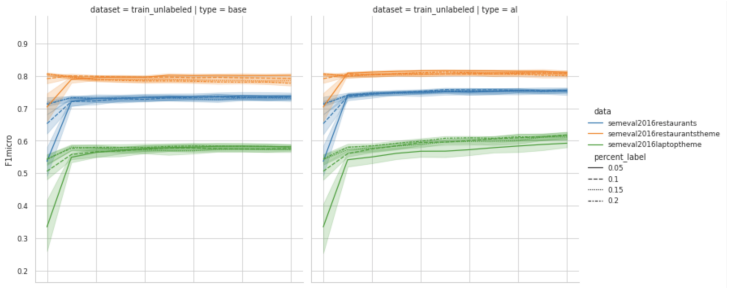

Figure 2: Performance of the algorithm during the iterative learning process with different sampling strategies. (Left) Randomly chosen, (Right) chosen based on low confidence predictions from the model. It is observed that active learning is able to learn more and predict better on the unlabeled and test datasets, as evidenced by the increasing F_1-score with increasing number of iterations

3 Experiments

This algorithm will be used to analyze qualitative survey data within Amazon but due to privacy restrictions we cannot test on this data. Instead, we use the Semeval-2016 Task 5 (Pontiki et al., 2016) that contains reviews for laptops (530 reviews) and restaurants (439 reviews). This data set is reflective of the qualitative analysis data, both are multi-label, multi-class, and class-imbalanced.

To demonstrate the value of the proposed algorithm, we use a bag-of-words method with a perceptron classifier as the baseline model. We use term frequency inverse-document frequency features to generate features (maximum features $= 50$) to feed into a single layer perceptron for this multi-labeled classification task.

To show the value of using active learning to add new samples, we create a baseline where labeled samples are created from randomly sampled data from the original unlabeled data set. The proposed algorithm is used on the new data set without modifications.

4 Results

Figure 2 and Figure 3 show the results from the experiments for the two stages.

5 Conclusion and Future Work

This paper presented a multi-class, multi-label weakly supervised text classification model that can iteratively learn from additional labeled samples. We want to extend the model by using multilingual embeddings or embeddings trained on in-domain data might provide a better user experience. We might consider using the user-provided data to refine the embeddings before pre-training. Since these datasets suffer from class imbalance, we want to test alternative sampling strategies to improve performance for rare classes. Finally, we want to allow users to provide new classes in iterations, rather than constraining to fixed classes.

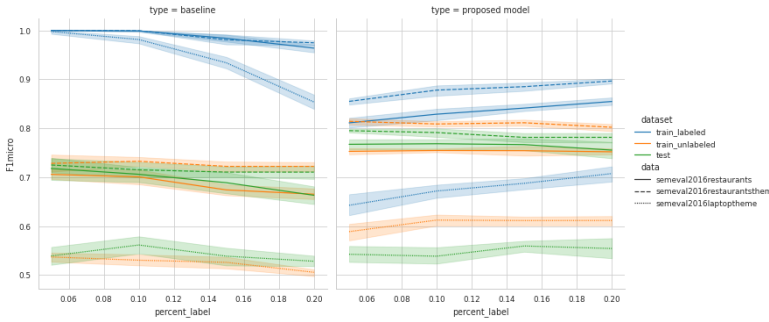

Figure 3: Performance of the proposed algorithm over the TF-iDF baseline (Left) TF-iDF baseline model, (Right) Proposed algorithm. It is observed that the proposed algorithm outperforms the TDF-iDF baseline for all the datasets considered in this study. This shows that the proposed algorithm is able to learn from small labeled datasets and classify the output correctly.

References

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. 2005. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(46):1345–1382.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2018. Weakly-supervised neural text classification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 983–992.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.

Michael Williams and Tami Moser. 2019. The art of coding and thematic exploration in qualitative research. *International Management Review*, 15(1):45–55.