SAM Team Preliminary Results on Distinguishing Firsthand Self-reports of Breast Cancer from Other Tweets.

Seibi Kobara[1], Masoud Nateghi[1], and Alireza Rafiei[1]

[1]Department of Computer Science and Informatics, Emory University, Atlanta, GA, USA

Our initial task involves exploring the application of natural language processing (NLP) to differentiate firsthand breast cancer self-report tweets from other textual tweets. Of note, we investigate the capability of NLP techniques to extract meaningful features from the input text data. For this aim, we utilize an annotated dataset related to breast cancer, comprising 4,717 tweets (which was split into 3513 train, and 1204 test tweets) [1]. The developer team labeled the tweets as self-reports of breast cancer, reports about a family member or friend's breast cancer (F), or not relevant (NR) and calculated pair-wise inter-annotator agreements using Cohen's method. Subsequently, they combined the classes F and NR in the final published dataset. In tackling the supervised classification task, we adopt three distinct approaches. Firstly, we extract various feature sets from the text and construct machine learning classifiers. Secondly, we train a bidirectional long short-term memory (BLSTM) model from scratch. Thirdly, we fine-tune pre-trained models such as bidirectional encoder representations from transformers (BERT) and BERT_large for this classification task.

In the first approach, we explore the combination of a broad spectrum of features, including n-grams (ranging from 1 to 3), word clusters, word-to-vector representations, text length, term frequency-inverse document frequency (TF-IDF), latent Dirichlet allocation (LDA), sentiment score, and extracted features from a BERT model, as the input features for developing different machine learning models. These models are then optimized through a grid search method, involving an extensive range of parameters and 5-fold cross-validation on the training dataset. The second approach consists of training a two-layer BLSTM model (parameters: unit = 100, dropout = 0.2, recurrent dropout = 0.2), followed by a dance layer (parameters: unit = 100, dropout = 0.2). Regarding the third approach, we use the pre-trained architecture and weights of the aforementioned models and fine-tune them on the available dataset over 10 epochs. For the neural network-based models, we incorporate GloVe word embedding representations, designed to capture word meanings, with a maximum length of 100 words [2]. We develop the classifiers of these three approaches using the available training set and subsequently assess their performance metrics using the unseen test set. Table 1 summarizes the accuracy, F1-score for the positive class, and F1-score for the negative class of the developed models.

Table 1. Performance metrics of the developed models on the test set.

Model	**Accuracy**	**F1-score (S)**	**F1-score (NR)**
Naive Bayes	0.66	0.58	0.71
Support Vector Machine	0.84	0.64	0.90
Logistic Regression	0.83	0.64	0.89
Random Forest	0.84	0.61	0.90

XGBoost	0.86	0.69	0.91
Light Gradient-Boosting Machine	**0.87**	**0.72**	**0.92**
CatBoost	0.86	0.68	0.91
BLSTM	0.86	0.72	0.91
BERT_base	0.90	0.78	0.94
BERT_large	**0.92**	**0.84**	**0.95**

References

[1] Al-Garadi, M.A., et al. Automatic Breast Cancer Cohort Detection from Social Media for Studying Factors Affecting Patient-Centered Outcomes. In Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18 (pp. 100-110). Springer International Publishing.

[2] Pennington, J., et al. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.