Classification of Healthcare Forum Messages

Janu Verma, Bum Chul Kwon, Yu Cheng, Soumya Ghosh, Kenney Ng
IBM T.J. Watson Research Center,
Yorktown Heights, NY, USA
Email: {jverma, bumchul.kwon, chengyu, ghoshso, kenney.ng}@us.ibm.com

***Abstract*—We studied various approaches for categorizing healthcare forum messages into one of seven different categories, based on the message text. We investigated the performance of several standard text classification algorithms on this task. We also explored recently proposed convolution neural network architectures that have been demonstrated to be effective at classifying short texts. We found that an ensemble of the explored models performed better than any of the individual models. Our ensemble was ranked first among the 56 runs submitted by 12 participating teams in the ICHI Data Challenge 2016.**

I. Introduction

Classification of textual documents into different categories is an interesting albeit hard problem. In this paper, we study the problem of classifying questions posted on healthcare forums. Information seekers post their questions on forums, and other members of communities provide their suggestions or answers. Tagging the posts with relevant tasks helps draw attention from the members with relevant expertise. Usually the tagging is done by the original poster or by other members of the forum. Automatic tagging on the question based on the contents of the post is a very important step to ensure relevant and timely responses.

We performed a thorough exploration of several standard methods, naive Bayes, multinomial regression, support vector machines and random forests, that are known to be effective at document classification. We also explored convolutional neural networks, motivated by their recent success in image and text classification [1] problems. In [2], the authors demonstrate promising sentence classification results with a particular convolutional neural network architecture. We adopted and modified their convolutional neural network-based sentence classifier to the classification of healthcare forum messages.

II. Dataset and Problem Description

The data used for this work is provided as a part of Healthcare Data Analytics Challenge at ICHI 2016, which contain real messages posted on a health discussion forum. Two different data files were provided in tab separated format for training and testing, respectively. The training data has 8000 messages each with the title text, contents text, and a category. The challenge provided seven different types of categories (tags):

1) Demographic (DEMO): Forums targeted towards specific demographic sub-groups characterized by age, gender, profession, ethnicity, etc.
2) Disease (DISE): Forums related to a specific disease
3) Treatment (TRMT): Forums related to a specific treatment or procedure
4) Goal-oriented (GOAL): Forums related to achieving a health goal, such as weight management, exercise regimen, etc.
5) Pregnancy (PREG): Forums related to pregnancy, including forums on difficulties with conception and concerns about mother and unborn childs heath during pregnancy
6) Family support (FMLY): Forums related to issues of a caregiver (rather than a patient), such with support of an ill child or spouse.
7) Socializing (SOCL): Forums related to socializing, including hobbies and recreational activities, rather than a specific health-related issue.

The test data contains 3000 messages without any category label, and the task is to build a classifier to automatically tag the messages with their most probable category.

We used the training data for both training the models and tuning their hyperparameters. We assigned 80% of the training data to a training set and the rest to a held-out validation set, uniformly at random. We create 10 such train/validation splits and tune the hyper-parameters based on their average performance on the validation sets.

III. Methods

A. Preprocessing

We performed a high-level preprocessing and cleaning of the data before building a classification model. As a first step in this direction, we removed all the stop-words, hyperlinks, and special characters from the text of the titles and the questions. We, then, built features for each question entry by extracting unigram and bigram tokens. We transformed the raw tokens through term frequency-inverse document frequency (tf-idf) transformation [3]. The tf-idf vectors provide a representation of the message tokens in a high dimensional vector space, accounting for their frequency in and across documents in a corpus. These tf-idf vectors were used by most of our approaches.

For our convolution neural network model, we adopted a different approach, and instead of tf-idf, we used word2vec [4] features. The word2vec framework provides a vectorial representation of words and then the representation of the text can be obtained by aggregating (e.g concatenating) the word2vec vectors. We used word2vec vectors pre-trained on 100 billion words from Google News. The vectors lie in a vector space of dimension 300, and were computed by the continous bag-of-words approach [4], [5], [6]. The word2vec data can be downloaded freely in binary form (https://code.google.com/p/word2vec/).

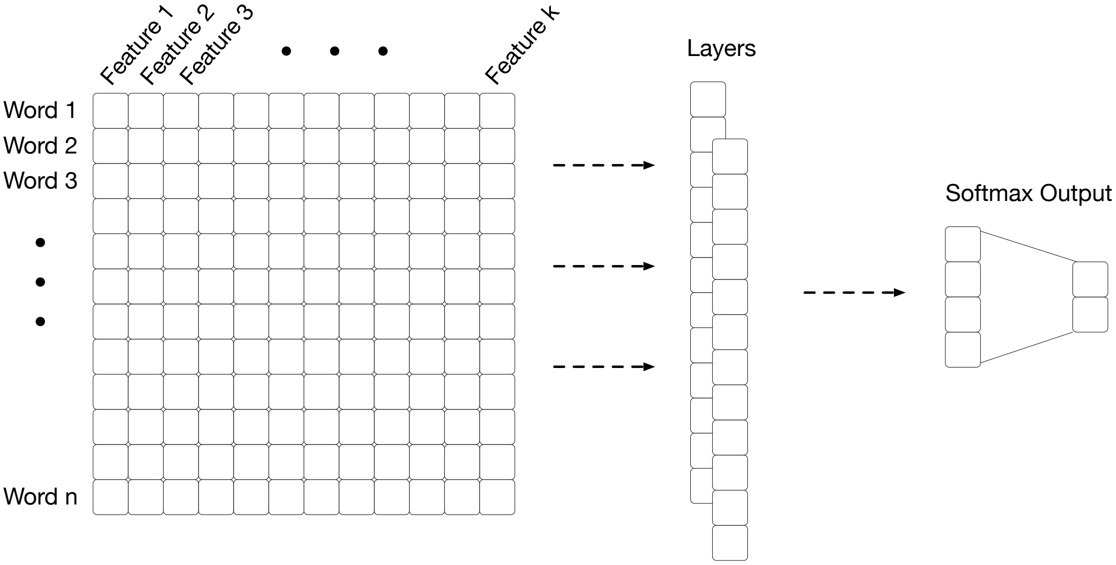

Fig. 1. The Architecture of the Convlutional Neural Network.

B. Models

We experimented with a number of approaches that have previously been shown to be effective for text classification problems. We adapted the models to the task at hand, with emphasis on hyperparameters tuning. We also used a neural network based model. For each of these models, we experimented with two versions. In the first version, we pooled the body text and title text together and trained a single classification model. Then, we tried another version where the title and body texts are separately feature transformed (either tf-idf or word2vec) and fed into separate classifiers. Predictions from the two independent classifiers were then averaged to generate the final classification.

1) **Baseline model - Naive Bayes:** Multinomial naive Bayes models are generative models that fit Multinomials to the class conditional distributions under the simplifying assumption that all features are independent conditioned on class membership. Combined with a prior over classes, a posterior over class membership is inferred and used for classifying data points. Despite their simplicity, they have been shown to be surprisingly effective at text classification tasks, and we find them to be a sensible baseline for the task at hand. This model was used to establish a baseline for our models, and the results of this model are not a part of our submission.

2) **Logistic/Multinomial regression:** Logistic regression models are discriminative models, that are fit by maximizing the class membership probabilities with respect to model parameters (feature weights). We train two versions of one-vs-all logistic regression models, one with l2 regularization and *lbfgs* solver, and another with l1 regularization with *liblinear* solver. The *liblinear [7]* is a standard method for solving a linear optimization problem, while the *lbfgs* [8] a limited memory quasi Newton method for general unconstrained optimization problems. Our submission **Run 1** (Table I) contains the results of logistic regression model with l2 regularization. We used a validation set to tune the hyperparameters, and finally used regularization strength value of 1.0.

3) **Support Vector Machines:** We also explored l2 regularized linear support vector machines. These models are similar to the logistic regression models, but optimize the hinge loss function. This sometimes leads to improved performance. SVMs are very effective in high-dimensional data, especially when the number of dimensions is greater than that of samples. For the current data, the tf-idf vectors are points in a very high-dimensional vector space with dimensions equal to the number of tokens (unigrams/bigrams) in all the training examples. The penalty parameter of the error term was chosen to be 1, and a linear kernel was used. The results are in **Run 2** (Table I) of the submission.

4) **Random Forests:** Random forests operate by ensembling high variance but low bias decision trees fit to bootstrapped subsamples of the data. The averaging operation reduces the variance exhibited by the individual estimators while retaining their low bias. They can effectively capture complex nonlinear decision boundaries between classes and serve as a strong baseline. We fed tf-idf vectors to a random forest containing 100 trees with the splitting criterion based on gini coefficient. The submission **Run 3** (Table I) is obtained from this model.

5) **Convolutional Neural Networks:** Finally, given the recent success of deep learning techniques, we also explored a convolutional neural network [2] trained in a one-vs-all fashion. These models can better incorporate long range context in the text to be classified, so they can outperform the baseline competitors. Here, we used a network with convolutional filter window sizes of 3, 4, and 5 for layer 1, 2, and 3 respectively. The number of hidden units was chosen to be 100. The Rectified linear unit (RELU) non-linear activation function was used. The outputs of these were max pooled over time and fed into a logistic output layer. The parameters of the network were learned in mini-batches of size 50 using *adadelta* [9], a method for gradient descent which adapts dynamically and has minimal computational expense, with the value of decay parameter 0.95. To guard against overfitting, we also employed dropout ($p = 0.5$) and l2 regularization with strength 3. The architecture of the convolution neural network is shown in the figure 1:

The results are contained in the submission **Run 4** (Table I).

6) **Ensemble Model:** We aggregated the results from the previous classifiers into an ensemble model by taking the arithmetic average of the probabilities provided by the different classifiers. We computed the probabilistic predictions from the previous models and used them to compute the aggregate probability. The submission **Run 5** (Table I) contains outcome of this model.

TABLE I. RESULTS OF THE MODELS ON THE TEST SET

Runs	Model	Accuracy (in %)
Run 1	Logistic Regression	67.47
Run 2	Support Vector Machines	67.47
Run 3	Random Forests	58.93
Run 4	ConvNet	64.4
Run 5	Ensemble model	68

IV. RESULTS

Our predictions for the categories of the messages in the test set obtained from these models were submitted to the ICHI Challenge committee for evaluation. Table 1 shows the results of the five runs. The ensemble model (**Run 5**) was ranked first among the 56 runs submitted by 12 participating teams.

V. CONCLUSION

We studied the problem of classifying text messages posted on a healthcare forum. We tuned some of the known methods for the current problem. In addition, we provided a model based on convolution neural network for sentence classification. These models were tested against the data provided as a part of ICHI Data Challenge. All of these models obtained good accuracy, and the ensemble of these methods was ranked first among the 56 runs submitted by 12 participating teams. The results are very promising, and we believe, that with further tuning of the hyperparameters, and using more training data, much better results can be obtained.

ACKNOWLEDGMENT

The authors would like to thank Shuangfei Zhai for useful discussions.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[2] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *arXiv:1408.5882 [cs]*, Aug. 2014, arXiv: 1408.5882.

[3] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to information retrieval*. New York: Cambridge University Press, 2008.

[4] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 2013. [Online]. Available: https://research.google.com/pubs/pub41224.html

[5] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *CoRR*, vol. abs/1310.4546, 2013. [Online]. Available: http://arxiv.org/abs/1310.4546

[6] G. Z. Tomas Mikolov, Scott Wen-tau Yih, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics, May 2013. [Online]. Available: https://www.microsoft.com/en-us/research/publication/linguistic-regularities-in-continuous-space-word-representations/

[7] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*

[8] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Math. Program.*, vol. 45, no. 3, pp. 503–528, Dec. 1989. [Online]. Available: http://dx.doi.org/10.1007/BF01589116

[9] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012. [Online]. Available: http://arxiv.org/abs/1212.5701