Automatic Identification of Drugs and Adverse Drug Reaction Related Tweets

Segun Taofeek Aroyehun
CIC, Instituto Politécnico Nacional
Mexico City, Mexico
aroyehun.segun@gmail.com

Alexander Gelbukh
CIC, Instituto Politécnico Nacional
Mexico City, Mexico
www.gelbukh.com

Abstract

We describe our submissions to the Third Social Media Mining for Health Applications Shared Task. We participated in two tasks (tasks 1 and 3). For both tasks, we experimented with a traditional machine learning model (Naive Bayes Support Vector Machine (NBSVM)), deep learning models (Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Bidirectional LSTM (BiLSTM)), and the combination of deep learning model with SVM. We observed that the NBSVM reaches superior performance on both tasks on our development split of the training data sets. Official result for task 1 based on the blind evaluation data shows that the predictions of the NBSVM achieved our team's best F-score of 0.910 which is above the average score received by all submissions to the task. On task 3, the combination of of BiLSTM and SVM gives our best F-score for the positive class of 0.394.

1 Introduction

The emergence of social media platforms such as Twitter has led to the availability of huge amount of data for research purposes. Public health monitoring using this non-traditional mode of communication has received attention in recent times. The third edition of Social Media Mining for Health Applications (SMM4H) (Davy et al., 2018) shared task aims to facilitate pharmacovigilance research using social media data.

We participated in tasks 1 and 3. The purpose of task 1 is to identify tweets that contain drug name(s) while task 3 focuses on recognizing Twitter posts mentioning adverse drug reaction (ADR). Both tasks are binary classification tasks. The evaluation metrics for both tasks are the precision, recall, and F1 scores of the positive class.

In the following sections, we describe the data, our approach, results, and conclusion.

Task	Train set		Test set
	neg class	pos class	
1	4356	4700	5382
3	15326	1351	5000

Table 1: Number of Examples in the Train and Test Sets for Tasks 1 and 3

2 Data

The shared task organizers provided datasets consisting of tweet IDs and their corresponding label as well as a script to download the text of the tweets. Using the IDs, textual data was gathered from Twitter. For task 1, the tweets were annotated for the presence of at least one mention of drug name. The presence of ADR mention was equally annotated for task 3. We downloaded a total of 9056 tweets out of 9625 expected tweets as training data for task 1. Also, 16677 tweets were retrieved out of 25630 expected tweets for task3. Table 1 shows the number of examples per label in the training data for task 1 and task 3. For task 1, the number of examples per class is almost equal. For task 3, the number of examples per label are highly imbalanced with almost 92% of the examples belonging to the negative class (non-ADR) and approximately 8% of the training data are of the positive class (ADR). The blind test set consists of 5382 tweets and 5000 tweets for task 1 and task 3 respectively. We cleaned the datasets by removing special and repeated characters, numbers, URL, and hashtags. To handle mispellings, we ran a spell checker.

3 Method

Our approach to both tasks 1 and 3 is very similar. We experimented with NBSVM, deep learning models, and the combination of a deep learning model as feature extractor and SVM as classifier.

Proceedings of the 3rd Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task, pages 54–55

Task	Classifiers			
	NBSVM	CNN	LSTM	BiLSTM
1	**0.909**	0.877 (0.888)	0.848 (0.781)	0.876 (0.798)
3	**0.624**	0.619 (0.549)	0.591 (0.391)	0.622 (0.321)

Table 2: F1 Score of the Positive Class on our Development Split of the Training set using NBSVM and Deep Learning Models (For the deep learning models, the scores are the average of three runs and the values in parenthesis are for the corresponding character level model)

NBSVM is a strong baseline (Wang and Manning, 2012). The choice of the deep learning model to use as feature extractor was informed by the average performance across three runs on our development split. The train-development split used for task 1 is 90% for training and 10% for development. For task 3, the development split was generated after random undersampling of the majority class. We maintained class imbalance in the ratio 1:3 of the minority class to the majority class. As shown in Table 2, the best performing deep learning model for task 1 was CNN and BiLSTM for task 3.

In our experiments, the NBSVM model uses the log-count ratios over character n-grams ranging from 1 to 5 characters as features. In the deep learning models, we employed the pre-trained fastText word embedding [1]. The SVM model was trained using the RBF kernel.

For the deep learning models, we used the binary cross entropy loss function as our objective function. To optimize the loss function through backpropagation, we used ADAM optimizer with learning rate of 0.001. We ran the models for 100 epochs with earlystopping and dropout layers with probability of 0.2 in order to avoid overfitting.

4 Results

Table 3 shows the performance of our systems on the task 1 evaluation data. The NBSVM model achieved our best recall (0.899) and F1 (0.910) scores. These scores are above average. The average precsion, recall, and F1 scores are 0.8904, 0.872, and 0.880 respectively. The CNN model was marginally higher than the NBSVM by 0.002 on the precision score. For task 3, Table 4 shows that our BiLSTM+SVM model is our best submission reaching our best score on precision (0.314) and F1 (0.394) scores for the ADR class. The NBSVM model achieves a better recall on the ADR class, 0.795. The difference in recall scores suggests that an ensemble of classifiers might lead to a better F1 score.

[1] https://s3-us-west-1.amazonaws.com/fasttext-vectors/wiki.en.zip

System	P	R	F
NBSVM	0.920	**0.899**	**0.910**
CNN	**0.922**	0.786	0.848
CNN+SVM	0.909	0.803	0.853

Table 3: Scores on the Evaluation Data for Task 1 (P-Precision; R-Recall; F-F1 measure)

System	P	R	F
NBSVM	0.258	**0.795**	0.390
BiLSTM	0.293	0.586	0.390
BiLSTM+SVM	**0.314**	0.529	**0.394**

Table 4: Scores on the Evaluation Data for Task 3 (P-Precision for the ADR class; R-Recall for the ADR class; F-F1 measure for the ADR class)

5 Conclusion

In this paper, we describe our participation in tasks 1 and 3 of the SMM4H shared tasks. We developed three classifiers for both tasks using NBSVM, deep learning models (CNN, LSTM, and BiLSTM), and the combination of a deep learning model and SVM. For task 1, we achieved our best submission using the NBSVM. The BiLSTM+SVM model achieved our best F1 score for the ADR class on task 3 while the NBSVM model scores better in terms of recall.

As future direction, we would like to investigate the use of informed sampling techniques in handling class imbalance. Also, we will explore the enrichment of the training data with semantic and conceptual domain knowledge that could provide relevant priors for the classifiers.

References

Weissenbacher Davy, Sarker Abeed, Paul Michael, and Gonzalez-Hernandez Graciela. 2018. Overview of the third social media mining for health (smm4h) shared tasks at emnlp 2018. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics.