Adverse drug effect and personalized health mentions CLaC at SMM4H 2019, Tasks 1 and 4

Parsa Bagherzadeh, Nadia Sheikh, Sabine Bergler
CLaC Labs, Concordia University
Montreal, Canada
{ p_bagher, n_she, bergler } @encs.concordia.ca

Abstract

CLaC labs participated in Task 1 and 4 of SMM4H 2019. We pursed two main objectives in our submission. First we tried to use some textual features in a deep net framework, and second, the potential use of more than one word embedding was tested. The results seem positively affected by the proposed architectures.

1 Introduction

The ongoing SMM4H challenge tasks define evolving challenges defined on Twitter data (Weissenbacher et al., 2019). The intention of epidemiologists is to detect mentions of health issues early on Twitter. One of the challenges is to detect real reports of personally experienced health issues and to distinguish them from generalizations, hypotheticals, news, and institutional advice.

Task 1 of SMM4H 2019, "Automatic classification of adverse effects mentions in tweets", asks to distinguish tweets that report an adverse drug effect (AE) from those that do not. Training data consists of 25,672 tweets with imbalanced distribution: 2,374 positive and 23,298 negative labels. An example of an adverse effect mention in a tweet is:

> *saphris gives me a mad appetite omg i hate this*

Task 4 is on "Generalizable identification of personal health experience mentions". Two specialized training sets were released , "flu vaccination" and "flu infection", comprising approximately 6,200 and 1,100 tweets. Task 4 training data was balanced. A sample positive tweet from this task is:

> *I must say that flu shot packed a punch. #WorstInoculationEver*

The CLaC submission to SMM4H 2019 had three general goals: first, to experiment with architectures that can address both tasks, second, to compare different word embeddings for their individual, but also their combined effectiveness, and third, to test whether we can augment the basic word vectors input with additional local and global knowledge from word lists and text preprocessing. The experiments remain inconclusive, due to an error in our submission pipeline.

2 Word embeddings

We experimentd with three types of word embeddings: BERT (a Transformer-based Bidirectional representation) (Devlin et al., 2018) (BERT-Base, Uncased)[1]; Word2Vec (Mikolov et al., 2013) trained on Sentiment140 [2] as well as training data from SMM4H 2018 and 2019 (all tasks) using Gensim package (Řehůřek and Sojka, 2010); and Glove word embeddings, pretrained on tweets (Pennington et al., 2014).

3 Textual features

Use of textual features as external source of knowledge has recently been the topic of interest (Sennrich and Haddow, 2016), (Ebert et al., 2015). We preprocess the tweets using the ANNIE Twitter Tokenizer (Cunningham et al., 2002), the Hashtag Tokenizer (Maynard and Greenwood, 2014), and the Stanford Part-Of-Speech Tagger with a model trained on tweets (Toutanova et al., 2003). We determine negation and modality spans using (Rosenberg et al., 2012). We use the Diego Lab ADR wordlist (Nikfarjam et al., 2015) to annotate terms appropriate for negative effects and health concerns.

[1] https://github.com/google-research/bert

[2] http://help.sentiment140.com/for-students

Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 123–126
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

User mentions (@) were removed from the tweets. URLs are annotated, as are the first person personal pronouns *I, my, mine*.

Negation and modality The span of negation and modality is determined using (Rosenberg et al., 2012) and projected onto the token representation: tokens present in the span of a negation or modality are indicated by a binary flag appended to the respective word vector (see Figure 1). The presence of negation and/or modality might reflect uncertainty in a given tweet and it may not convey facts.

URL Tweets about a personal experience do not usually include a URL. Specifically for Task 4, 80% of the tweets including a URL are negative. A binary URL feature encodes presence or absence of a URL in the tweet.

POS embedding We experimented with the notion of part of speech embeddings to address sparsity. Here, a representation for each POS tag is obtained using Word2vec by training on a POS tagged corpus (instead of words themselves). We use the Penn tree bank tag set (36 tags) with a window size of 5.

ADR lexicon Terms from the Diego Lab adverse drug reaction lexicon (Nikfarjam et al., 2015) are indicated as a binary, tweet level feature, in order to increase recall.

First person personal pronoun First person pronouns *I, my,* and *mine* are indicated at token level by a separate binary feature. In both tasks, a personal experience is more likely to be a positive sample, therefore, enhancing recall.

	I	should	n't	have	gotten	that	flu	shot
W2V	...	...	...		...	...	...	...
Neg.	0	0	0	1	1	1	1	1
Mod.	0	0	1	1	1	1	1	1
1st	1	0	0	0	0	0	0	0

Figure 1: Feature vector encoding

3.1 System architecture

Our system has two parallel branches and is trained in two stages. One branch works only with BERT word embeddings, the other branch works on our concatenated token level features plus word embeddings (Word2Vec/Glove) shown in Figure 1. The input vectors of each branch are fed into Bi-LSTMs and are followed by attention and finally two softmax decision neurons.

After optimizing each branch with binary cross-entropy loss, the parameters of the networks are frozen for the second stage of training. We train an SVM on the input vector that concatenates class probabilities provided by the softmax neurons with the tweet level features, ADR and URL.

The network is optimized using the Adam optimizer (Kingma and Ba, 2014) with learning rate $lr = 0.001$ for 5 epochs (for both tasks). For Task 1, the class weights of $cw_{pos} = 1$ and $cw_{neg} = 0.4$ are used as thresholds for positive and negative samples respectively. For the SVM, the RBF kernel is used with $\gamma = 0.001$. The hyper-parameters have been chosen by cross validations. The first stage deep net learning is implemented using Keras [3] and the second stage SVM classification is implemented using Scikit-learn (Pedregosa et al., 2011).

4 Development phase

During the development phase we considered a number of different features and performed an ablation study with more than 130 different configurations. For this phase, 22,000 and 3,672 samples were considered for training and test sets respectively.

An interesting observation was the different behavior of word embeddings in the presence of language features. For Task 1, Glove embeddings usually performed higher, whereas in Task 4, Word2Vec embeddings were generally superior. In Task 1, adding textual features to Word2Vec embeddings resulted in a decrease in performance, however, adding the same features to Glove re-

Table 1: Development results for Task 1. Submitted configurations are indicated by *

	Prec.	**Rec.**	**F1**
Glove	0.41	**0.73**	0.52
BERT	0.56	0.50	0.53
Glove+ADR	0.46	0.67	0.55
Glove+BERT	0.49	0.64	0.55
Glove+Mod+BERT	0.53	0.57	0.55
Glove+Neg+BERT	0.48	0.61	0.54
Glove+Neg+Mod+BERT	0.58	0.55	0.56
Glove+BERT+ADR	0.53	0.64	0.58
Glove+Neg+Mod+ADR	0.49	0.65	0.56
* Glove+Neg+Mod+ADR+BERT	0.54	0.64	**0.59**
W2V	0.42	0.65	0.51
W2V+ADR	0.39	0.67	0.49
* W2V+BERT	**0.59**	0.53	0.56
W2V+1st	0.48	0.62	0.54
* W2V+1st+BERT	0.52	0.63	0.57

[3] https://keras.io

sulted in increased performance. This effect was small, but persistent across ablation of the other features, and we concluded that the different behaviors of the embedding vectors could be leveraged in an ensemble situation.

For Task 1, the ADR word list generally increased recall in our ablation studies, demonstrating that domain specific gazetteer lists can effectively supplement training data. In combination with Glove, textual features such as negation and modality increased precision, but diminished recall. Adding ADR to this combination (*Glove+Neg+Mod+BERT*) compensates for the drop in recall without significantly decreasing precision. The results also corroborates the hypothesis that the *1st* feature enhances the recall (*W2V+1st* and *W2V+1st+BERT* compared to *W2V* and *W2V+BERT*).

Looking at the confusion matrix reveals that the model (specifically *Glove+BERT*) associates drug mentions in the subject position with positive labels, incurring a considerable amount of false positives, see for instance:

> *this lozenge has my sore throat fading*

> *paxil makes you susceptible to sunburns?*

The ADR feature (*Glove+ADR+BERT*) reduces these false positives while it causes other instances of false positives. As mentioned before, ADR generally increases recall, but in some configurations with Glove it has increased precision which is interesting and we will study it in more detail.

Modality reduces false positives and is the most effective token level textual feature. Two instances of false positives (in *Glove+BERT*) which are correctly classified in the presence of modality are:

> *i'm sucha psycho when i study already if i ever took adderall i* ***would*** *probably explode*

> *seroquel* ***can*** *have potential fatal effects when taken & being in direct sunlight for extended periods. can i get you a bottle a tanning bed?*

When combined with Glove, we observed that the negation feature degrades the F1 score, however, it inter-plays well with the modality feature.

For Task 4, combining textual features with Word2Vec increases precision. The URL feature by itself increases precision even more, but incurs a larger drop in recall.

Table 2: Development results for Task 4. Submitted configurations are indicated by *

		Prec.	Rec.
F1			
W2V	0.70	**0.88**	0.78
BERT	0.78	0.82	0.80
W2V+BERT	0.76	0.85	0.80
W2V+Mod	0.72	0.87	0.79
W2V+POS	0.76	0.81	0.79
W2V+URL	0.76	0.84	0.80
* W2V+URL+BERT	**0.83**	0.79	0.81
W2V+1st+URL	0.77	0.83	0.80
* W2V+1st+URL+BERT	0.81	0.81	0.81
W2V+Mod+POS+URL	0.78	0.85	0.81
* W2V+Mod+POS+URL+BERT	0.81	0.84	**0.83**

5 Evaluation phase

Task 1 We submitted three configurations to Task 1: Glove with our textual features, W2V alone, and W2V with the first person pronoun feature (all used in an ensemble with BERT). These were not our top performing configurations during development, rather we included W2V to bridge to Task 4 and we included two runs with different textual features and one without. The performance of our system in the competition is provided in Tables 3, the competition performance of all three models is commensurate with our development results with ±2% in F1 measure. Moreover, the three configurations performed near identically and all three were above the competition mean.

It is interesting to note that the Word2Vec embeddings trained on Sentiment140 data proved as effective on this data set as Glove with the textual features, in contrast to our development experiments. We interpret the fact that W2V in an ensemble with BERT lies above the competition's mean to confirm the importance of our genre selection for Word2vec training.

Table 3: CLaC competition results for Task 1

	Prec.	Rec.	F1
W2V+ BERT	0.54	0.60	0.57
Glove+Neg+Mod+ ADR+BERT	0.52	0.60	0.56
W2V+ 1st+BERT	0.51	0.59	0.55
Competition mean	0.53	0.50	0.50

Task 4 Our three submissions for Task 4 were all based on Word2vec and the URL feature. Results, however, diverge drastically from our development runs, where runs scored between 75-85%

Table 4: CLaC competition results for Task 4

	Prec.	Rec.	F1
Condition 1			
W2V+Mod+POS+URL+BERT	0.84	0.32	0.47
W2V+1st+URL+BERT	0.83	0.42	0.56
W2V+BERT+URL	0.75	0.29	0.42
Condition 2			
W2V+Mod+POS+URL+BERT	0.42	0.19	0.26
W2V+1st+URL+BERT	0.44	0.12	0.20
W2V+BERT+URL	0.44	0.12	0.20
Condition 3			
W2V+Mod+POS+URL+BERT	0.71	0.26	0.38
W2V+1st+URL+BERT	0.62	0.26	0.37
W2V+BERT+URL	0.62	0.26	0.37
Overall			
W2V+Mod+POS+URL+BERT	0.70	0.28	0.40
W2V+1st+URL+BERT	0.75	0.29	0.42
W2V+BERT+URL	0.74	0.33	0.46
Competition mean	0.90	0.58	0.70

F1 measure. The official results in Table 4 demonstrate.

6 Conclusions

We participated in the SMM4H 2019 shared task with two major ideas. First, we tried to use textual annotations in a deep net architecture and specifically proposed encodings for negation, modality, and use of a gazetteer list. Our observations during the development phase showed that textual features are effective for enhancing the performance of the system but that standard embedding vectors without additional textual features give comparable performance on these datasets.

Our second idea was to have more than one type of embedding in our system to have an ensemble and try to aggregate the predictions using a support vector machine rather than using a simple majority voting. This worked well, but again, on the datasets of this challenge, the computational overhead seems questionable for the degree of improvement achieved.

References

H Cunningham, D Maynard, K Bontcheva, and V Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

J Devlin, M-W Chang, K Lee, and K Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv:1810.04805v1 [cs.CL].

Sebastian Ebert, Ngoc Thang Vu, and Hinrich Schütze. 2015. A linguistically informed convolutional neural network. In *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 109–114.

DP Kingma and J Ba. 2014. Adam: A method for stochastic optimization. ArXiv:1412.6980 [cs.LG].

DG Maynard and MA Greenwood. 2014. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In *LREC 2014 Proceedings*.

T Mikolov, I Sutskever, K Chen, GS Corrado, and J Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*.

A Nikfarjam, A Sarker, K Oconnor, R Ginn Rachel, and G Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, (12).

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

R Řehůřek and P Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

S Rosenberg, H Kilicoglu, and S Bergler. 2012. Clac labs: Processing modality and negation. *Working Notes for QA4MRE Pilot Task at CLEF*.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. *arXiv preprint arXiv:1606.02892*.

K Toutanova, D Klein, ChD Manning, and Y Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL-HLT*.

D Weissenbacher, A Sarker, A Magge, A Daughton, MJ Paul, and G Gonzalez-Hernandez. 2019. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In *Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop and Shared Task*.