Detecting Tweets Reporting Birth Defect Pregnancy Outcome using Two-View CNN RNN based Architecture

Saichethan Miriyala Reddy
Indian Institute of Information Technology
Bhagalpur, Bihar, India
miriyala.cse.1725@iiitbh.ac.in

Abstract

This research work addresses a new multi-class classification task (fifth task) provided at the fifth Social Media Mining for Health Applications (SMM4H) workshop. This task involves distinguishing three classes of tweets that mention birth defects. We propose a novel two view based CNN-BiGRU based architectures for automatic tweet classification task. Experimental evaluation of our proposed architecture over the validation set gives encouraging result as it improves by approximately 7% over our single view model for the fifth task. Code of our proposed framework is made available on Github[1]

1 Introduction

Twitter is a micro-blogging system, which allows its users to publish tweets of up to 280 characters in length to tell others what they are doing, what they are thinking, or what is happening around them. Over the past few years twitter has become very popular, for every second, on average, around 6,000 tweets are tweeted on Twitter. Twitter is also considered as "It's what's happening". Social media can be considered as an enormous corpus since last decade many researcher started exploring different tasks such as sentiment analysis(Kouloumpis et al., 2011) (Wang et al., 2012), Named Entity Recognition(Baldwin et al., 2015) (Suman et al., 2020) and Disambiguation(Dredze et al., 2016) using this data.

More recently twitter data is used to identify and study a small cohort of Twitter users whose pregnancies with birth defect outcomes (Klein et al., 2019). Due to unique characteristics of social media there are few challenges associated with using social media data for health care research(Weissenbacher et al., 2019), including informal texts, colloquial expressions and misspellings of clinical concepts, noisy text, data sparsity, ambiguity, and multilingual posts.

2 Proposed Framework

Hashtags are one the most important aspects of Twitter. Hashtag lets users to apply dynamic, user-generated tagging that helps other users easily find posts on twitter with a specific theme or content. Users create and use hashtags by placing a hash symbol in front of a word or unspaced phrase in a tweet. The hashtag may contain symbols, digits and letters(capitalized & uncapitalized). Many tweets contain at least one hashtag, and often they provide context which motivated us to incorporate them in our framework.

2.1 Two View

We propose a novel two view CNN-RNN framework for this multi label classification task. The illustration of the Two view CNN-RNN framework is shown in Fig. 2. It contains two parts: The tweet part extracts semantic representations from clean text; the hashtag part which extracts semantic representation from hashtags. Concatenation of these semantic representations are fed to a dense layer and its output

[1] https://github.com/Saichethan/SMM4H-2020

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 125–127
Barcelona, Spain (Online), December 12, 2020.

is used as an input to softmax classifier. We also used single view for comparison which has only tweet part as shown in Fig. 1, same hyperparameters are used for both models.

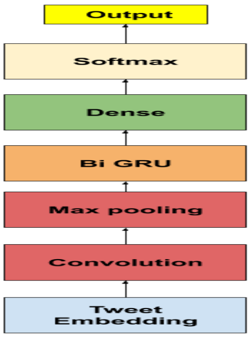

Figure 1: Single View

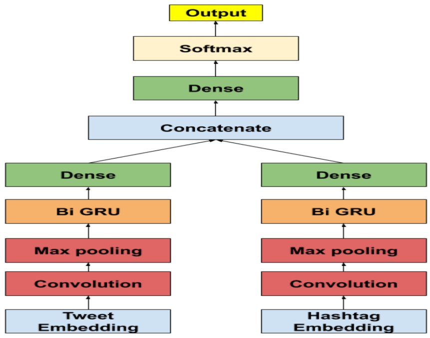

Figure 2: Two View

3 Experimental Evaluation

3.1 Experimental Set-up

We used Glove 300 dimension embeddings (Pennington et al., 2014) for both tweet and hashtags, maximum length of clean text and hashtags are set to 75 and 3 respectively. For convolution layer we used 256 filters with kernel size of 3 and pool size of 2. GRU's use less training parameters and therefore use less memory, execute faster and train faster than LSTM's, due to this we used GRU's for extracting semantic representation, we used 256 units and recurrent dropout of 0.1. We used regex to remove all unwanted symbols and to identify hashtags. Some of the lines of test data are corrupted due to which length is more than 280 for those we assigned length as 0.

	TRAIN	**DEV**	**TEST**
max len of tweet	62	61	77
avg no. of hashtags in a tweet	2.10	2.18	2.20
no. of tweets with atleast one hashtag	4277	1086	1303
total number of tweets	14717	3680	4372

Table 1: Dataset Statistics

3.2 Results

Two show the efficacy of our proposed approach, results on validation data are shown in the Table 2. From the Table 2 it is evident that our Two View system performed better than Single view system. Performance of our Two view system on the test set achieved a micro-averaged F-score of **0.58**, a micro-averaged precision of **0.54**, and a micro-averaged recall of **0.64**.

	P	**R**	**F1**
Single View	**0.52**	0.63	0.57
Two View	0.50	**0.78**	**0.61**

Table 2: Results on Validation set

4 Conclusion

Our main goal is to show the potentiality of our Two View framework which employs semantic views of both tweet and hashtag. Our approach work better when the percentage of tweets having atleast one hashtag are high. We also expect to improve the performance of our system by employing intra-attention between two views and through parameter tuning.

References

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135.

Mark Dredze, Nicholas Andrews, and Jay DeYoung. 2016. Twitter at the grammys: A social media corpus for entity linking and disambiguation. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 20–25.

Ari Z Klein, Abeed Sarker, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2019. Towards scaling twitter for digital epidemiology of birth defects. *NPJ digital medicine*, 2(1):1–9.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Fifth International AAAI conference on weblogs and social media*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Chanchal Suman, Saichethan Miriyala Reddy, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Why pay more? a simple and efficient named entity recognition system for tweets. *Expert Systems with Applications*, page 114101.

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 system demonstrations*, pages 115–120. Association for Computational Linguistics.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 21–30, Florence, Italy, August. Association for Computational Linguistics.