# NLP@VCU: Identifying adverse effects in English tweets for unbalanced data

**Darshini Mahendran, Cora Lewis and Bridget T. McInnes**
Department of Computer Science
Virginia Commonwealth University
mahendrand@vcu.edu corammlewis@gmail btmcinnes@vcu.edu

## Abstract

This paper describes our participation in the Social Media Mining for Health Application (SMM4H 2020) Challenge Track 2 for identifying tweets containing Adverse Effects (AEs). Our system uses Convolutional Neural Networks. We explore downsampling, oversampling, and adjusting the class weights to account for the imbalanced nature of the dataset. Our results showed downsampling outperformed oversampling and adjusting the class weights on the test set however all three obtained similar results on the development set.

## 1 Introduction

This paper describes our participation in Task 2 of the Social Media Mining for Health Application (SMM4H) 2020 challenge to automatically identify Adverse Effects (AE) in English tweets. To address this challenge, we explored a supervised binary classification system to automatically identify the AEs using Convolutional Neural Networks (CNNs). In order to deal with the unbalanced nature of the data set, we explored downsampling, oversampling, and utilization of the class weights.

## 2 Methods

In this section, we discuss our AE identification system. Our system can be found here[1].

*Feature Representation*. We pre-process the data as follows: 1) the HTML symbol *&* are removed as per (Úbeda et al., 2019); 2) double-quotes are removed; 3) hashtags, links, and usernames are replaced with the string "hashtag", "link" and "username" respectively as per (Cortes-Tejada et al., 2019); 4) emojis are substituted with a phrase that represents that emoji as per (Vydiswaran et al., 2019); 5) tweets are lowercased. Each word in the tweet is represented as an embedding. We evaluated three embedding types, GloVe (Pennington et al., 2014), word2vec (Mikolov et al., 2013) and FastText (Godin, 2019), trained over different corpora types. Our resulting system described here uses GloVe trained on Twitter.

*Algorithm*. We evaluated using Convolutional Neural Network (CNNs) as per (Úbeda et al., 2019). One of the beneficial properties of CNN is that it preserves the spatial orientation, in this case, the sequence of the words in the tweet. CNNs consist of four layers: 1) embedding layer - to encode words in sentences by real-valued vectors; 2) convolution layer - to get local features from each part of the input; 3) pooling layer - to extract the most relevant features, and 4) feed-forward layer - a fully connected layer to perform classification. For this, we first feed each tweet into CNN to learn the AE representation of the tweet. Second, we apply the convolution layer to learn the local features from the embedding vectors obtained from each word of a tweet. Next, we apply the max-pooling layer to extract the most important feature. Next, we unstack the volume into a flat vector and feed it into the fully connected feedforward layer. Finally, the fixed-length vector is fed into a softmax layer to perform the classification. For training, the classification error is back-propagated and the model is re-trained until the error is minimized. The weights of the matrix and bias are the parameters that get tuned until the optimized model is obtained.



[1] https://github.com/NLPatVCU/SMM4H



*Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 158–160
Barcelona, Spain (Online), December 12, 2020.

*Imbalance data*. Due to the imbalanced nature of the dataset, we evaluated our algorithm using three methods to reduce the imbalance: downsampling, oversampling, and adjusting Keras class weights:

**Run 1:** To downsample, non-AE tweets are removed in order to reach an input ratio of AE tweets to non-AE tweets. Through experimentation, we discovered that the best downsampling ratio is 1 AE tweet for every 4 non-AE tweets.

**Run 2:** To oversample, the AE tweets are repeated a specific number of times. For example, if the data set is oversampled 3 times, there would be 3 copies of each AE tweet. Through experimentation, we learned that oversampling 6 times is the most effective number of oversamples.

**Run 3:** We used the class weights option in Keras which incorporates the ratio of how much an AE tweet should be valued compared to a non-AE tweet.(Chollet and others, 2015) For example, if class weight is 1 for the non-AE tweets and 20 for the AE tweets, Keras would treat each AE tweet like its worth twenty non-AE tweets. Through experimentation, we found that a class weight of 1 for non-AE tweets and a class weight of 10 for AE tweets is most effective.

## 3 Data

The training set contains 20,544 tweets: 1,903 tweets contain AEs and 18,641 tweets do not contain AEs. The development set contains 5,134 tweets: 4,660 tweets are negative and 474 are positive. The training and development data sets are both highly imbalanced containing a 1:10 ratio of AE tweets to non-AE tweets. The test set contains 4,759 tweets.

## 4 Results

In this section, we report the Precision (P), Recall (R), and F-measure (F) of our system on the SSM4H Task 2 data set for the three runs described above. Table 1 shows the results obtained evaluated over the development data (Development Results), and the reported results evaluated over the test data (Reported Results). The results were achieved using Train-Test. The precision and recall were returned only for the best performing run. The results on the development set showed that using downsampling (Run1) or oversampling (Run 2) obtained a higher precision and lower recall whereas using class weights (Run 3) obtained a higher recall and lower precision. The F-1 scores achieved with the development data were all very similar, with downsampling and class weights achieving identical F-1 scores. This was unlike in the test data results where downsampling (Run 1) showed a higher recall than precision and obtained the highest F-1 score of all the runs.

| Run | | **Development Results** | | | **Reported Results** | | |
|---|---|---|---|---|---|---|---|
| | **Dimension** | **Precision** | **Recall** | **F1 Score** | **Precision** | **Recall** | **F1 Score** |
| Run 1 (downsample 1:4) | 100 | 0.59 | 0.43 | 0.50 | 0.28 | 0.36 | 0.35 |
| Run 2 (oversample 6) | 50 | 0.56 | 0.44 | 0.49 | - | - | 0.34 |
| Run 3 (class weights 10:1) | 50 | 0.43 | 0.58 | 0.50 | - | - | 0.31 |

Table 1: Development and Evaluation Results

## 5 Conclusions and Future Work

Our CNN model achieved a reported F1 score of 0.35 on the test dataset and 0.50 on the development data. Our experimentation with pre-trained word embedding showed that GloVe trained on Twitter is optimal for this task. Our results also show that downsampling, oversampling, and Keras class weights achieve similar F1 scores with the development data, though downsampling outperformed oversampling and class weights on the test data. In the future, we would like to experiment further in the pre-processing stage. We also plan to explore the possibility of character embeddings or utilizing Recurrent Neural Networks (RNNs) in depth. We would like to investigate the usage of additional word embeddings such as Bidirectional Encoder Representations from Transformers (BERT) for this task as well.

## References


François Chollet et al. 2015. Keras. https://keras.io.

Javier Cortes-Tejada, Juan Martinez-Romo, and Lourdes Araujo. 2019. Nlp@ uned at smm4h 2019: Neural networks applied to automatic classifications of adverse effects mentions in tweets. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 93–95.

Fréderic Godin. 2019. *Improving and Interpreting Neural Networks for Word-Level Prediction Tasks in Natural Language Processing*. Ph.D. thesis, Ghent University, Belgium.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jeffrey Pennington, Richard Socher, and Christopher 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.

Pilar López Úbeda, Manuel Carlos Díaz Galiano, M Teresa Martín-Valdivia, and L Alfonso Urena Lopez. 2019. Using machine learning and deep learning methods to find mentions of adverse drug reactions in social media. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 102–106.

VG Vinod Vydiswaran, Grace Ganzel, Bryan Romas, Deahan Yu, Amy Austin, Neha Bhomia, Socheatha Chan, Stephanie Hall, Van Le, Aaron Miller, et al. 2019. Towards text processing pipelines to identify adverse drug events-related tweets: University of michigan@ smm4h 2019 task 1. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 107–109.