**Modelling Valence and Arousal in Facebook posts**

Shane Talbot, Nathan Kelderman

# **Introduction**

Sentiment analysis is a large area of research, in which researchers are studying and implementing methods with which to determine the implications of text. This topic has multiple use cases ranging from product feedback to political campaigning and is a great task for machine learning. There are several types of models that can be utilized to process the given context of a piece of text more effectively and efficiently. Sentiment analysis specifically seeks to analyze the underlying attitude and emotion expressed in a given text. In general, there are two main theories regarding emotion in psychology: the first assumes that emotions are discrete and finite, in contrast, the second suggests that emotions lay on a range of different scales (Daniel et al.).

After learning about the basics of this topic in class, our group decided to look at a more complicated example of sentiment analysis. Specifically, we chose to replicate the works of H. Andrew Schwartz et al. in the paper *Modelling Valence and Arousal in Facebook posts*. Since this paper primarily utilizes a simple bag of words model, our group also decided to investigate alternative methods to help improve our understanding of other methodologies mentioned over the course of the class. Specifically, we chose to utilize an RNN. An RNN was chose over other models for several reasons. Firstly, it was a model that was mentioned in class. Secondly, our initial research indicated that an RNN might in fact be a very effective way to perform multi-value sentiment analysis. According to Nhan Cach et al, RNN’s show a a relatively high overall accuracy when used for sentiment analysis. In fact, it is even observed that RNN models could possibly overcome some issues with short text seen in other deep learning models (Nhan Cach et al.).

# **Dataset**

The dataset used consists of 2895 Facebook posts. Two independent annotators, each with a background in psychology, were tasked with rating the valence and arousal of each post on a scale of one to nine. The annotators reached an agreement correlation of 0.768 for valence and 0.827 for arousal (Daniel et al.).

Valence and arousal represent the sentiment, positive or negative, and intensity of the post respectively. The values are annotated separately; however, according to Schwartz et al., there is some evidence to suggest that the values may be weakly related (Daniel et al.). A valence score of one represents a negative sentiment, while a valence score of nine represents a positive sentiment. Neutral sentiments are expressed by a value near five. Arousal is weighted similarly, with a value of one indicating the lowest level of intensity and nine representing the highest.

To utilize the data provided, a large amount of pre-processing was done. Specifically, as can be seen in figure one, which is an image provided in the original paper by Daniel et al, the data set is very Chart, histogram

Description automatically generatedimbalanced for both classification categories. Our group originally attempted to train models on this data alone; however, for both of our models this led to a rather large issue.

Figure 1: Valence and Arousal Distribution

Figure 2 is a heatmap that indicates which classifiers in our valence model were most frequently guessed. Of course, seeing as the majority of our training data had a value of five, this value was guessed more than others. The same was true of our arousal models; however, those predictions were centered closer to one.

This is especially problematic for the minor classifications. In this case, one and nine are almost never properly identified. To adjust for this, our team did research and found that a common way to deal with imbalanced data classes is to perform under-sampling of high-density classes, over-sampling of low-density classes or some combination there-of (Peter et. al.). In this case, we chose to do both. This was accomplished by duplicating examples for low-density classes and randomly omitting examples from high-density classes.

In addition to addressing our imbalanced data, we also chose to do several steps of pre-processing to help ensure our training process was as effective as possible.Chart, histogram

Description automatically generated First each post was converted to lowercase. This created an even playing field for all words to have a face value as well as reduced the number of unique words. Secondly, regex was utilized to remove anything that wasn’t a letter or a number. The third and final technique was to remove all common English stop words from the data.

Figure 2 Original Valence Prediction Distribution

For the Bag Of Words (BOW) model, the next step was to create a bag of words that the model could then use. This was done using the sklearn CountVectorizer class.

For the RNN, we tokenized and only kept the top 2000 most popular words. The reason for picking out the top 2000 words is that at a certain point, words can be meaningless to a model in training. If, for example, a word is only used once in the whole dataset, the importance of that word cannot be determined with a significant amount of certainty. Whereas if a word is used many times in a dataset, then that word can be linked to a specific output with a higher degree of certainty.

# **Models**

We decided to attempt two models in our study. Specifically, an RNN and a BOW model were constructed. This resulted in four models total, as valence and arousal were both represented by separate models.

The first model is an RNN model. The network was created using Keras and is composed of a sequential network with the following layers: embedding layer, a dropout layer with a dropout of 0.3, 2 LSTM layers and a dense softmax activation layer. The model was compiled using the Adam optimizer and used the categorical cross entropy loss function. The BOW used a linear regression model with l2 regularization. In addition, both models utilized a 10 fold-cross validation setup. The choice to include this cross-validation setup was based on the inclusion in the paper written by Schwartz et. al. This cross validation was performed on the training data generated to help our models improve performance when predicting values for the testing data.

Table 1: Model Prediction results based on Pearson r correlation.

# **Results**

Prior to balancing the sample data, the accuracy was typically around forty percent for valence and twenty percent for arousal. In each case, the Pearson r correlation value was around 0.1. Numerous different tactics to increase this, ranging from preprocessing the data differently to adding/removing layers and hyper-parameter tuning; however, these only resulted in marginal changes to the output.

After some research, it was concluded that balancing the data as described above would be extremely helpful.

After balancing the data, our results improved significantly. Table 1, shows how our models compared to the BOW model created by Schwartz et. al, as well as several other benchmarks utilized by Schwartz et. al. These benchmarks are all measured using an r-value to determine the level of correlation between the predicted results and the actual results.

|  |  |  |
| --- | --- | --- |
| **Method** | **Valence** | **Arousal** |
| **ANEW** | .307 | .085 |
| **Aff Norms** | .113 | .188 |
| **MPQA** | .385 | --- |
| **NRC** | .405 | --- |
| **BOW (Schwartz et. al.)** | .650 | .850 |
| **RNN** | .930 | .810 |
| **BOW(Ours)** | .940 | .760 |

# **Conclusion**

In conclusion, we attempted to use the data from Schwartz’s study to recreate the results and try to push them even further. We spent a lot of time trying to get our models to work better. We realized though that it wasn’t specifically our models that weren’t working, but that the data was skewed and disproportionate. After adjusting the data, we were able to achieve much better results and even in some cases better than the results we were trying to match. Specifically, our models performed extremely well on the valence data sets; for arousal however, in relation to the results achieved by Schwartz et. al., they were noticeably lower. This could be due to several factors, including different pre-processing steps, or even different data balancing steps. Were we to continue this project, our future work would be focused on improving our results in relation to arousal score.

# **References**

Dang, Nhan Cach, et al. “Sentiment Analysis Based on Deep Learning: A Comparative Study.” *Electronics*, vol. 9, no. 3, 2020, p. 483., https://doi.org/10.3390/electronics9030483.

Gnip, Peter, et al. “Selective Oversampling Approach for Strongly Imbalanced Data.” *PeerJ Computer Science*, vol. 7, 2021, https://doi.org/10.7717/peerj-cs.604.

Preoţiuc-Pietro, Daniel, et al. “Modelling Valence and Arousal in Facebook Posts.” *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2016, https://doi.org/10.18653/v1/w16-0404.