CircumplexSentiNet - BiLSTM-CNN for Affective Circumplex Sentiment Analysis

Daniel Chen
danschen@stanford.edu

Abstract

Previous work on sentiment analysis has focused on different models of emotions, most notably Ekman's basic emotions (happy, sad, angry, etc.) [1]. However, one model of emotion that has not been used much in sentiment analysis is the affective circumplex, which studies emotions on the dimensions of valence (positive vs negative) and arousal (low energy vs high energy) [2]. The few studies that have used this model for sentiment analysis programs have been quite simple, using techniques such as weighted lexicons and bag-of-words linear regression [3]. The current work aims to expand the techniques used for this type of sentiment analysis using deep learning techniques as well. The current work also aims to improve results by taking word order into account, instead of using bag-of-words models. The current model uses combinations of LSTMs, CNNs, and regression to create a sentiment analysis program that aims to improve results on previous state-of-the-art. Current findings improve on previous state-of-the-art, increasing the correlation of predicting valence using multitask and self attention models.

1 Introduction

In psychology, there are two main models of measuring emotion - the basic emotion model and the affective circumplex. The basic emotion model is based on Ekman's work, which identifies several emotions that are considered as the discrete, fundamental building blocks of emotion across all human beings [1]. Examples of such emotions include happiness, sadness, and anger, and can be measured through methods such as facial expression and physiological measures. As one of the predominant models of emotion in the field of psychology, it provides a good way of measuring the emotions in sentiment analysis as well.

Another dominant model of emotion is the affective circumplex, a model of emotion that combines two dimensions: valence, and arousal [2]. Valence is a measure of the positivity of an emotion - whether it is something that is positive (e.g. happiness) or negative (e.g. sadness). Arousal is a measure of the amount of energy of emotion - whether it is something with high energy (e.g. excitement) or something with low energy (e.g. calm). Note that every emotion is a combination of these dimensions, and different emotions can have the same valence but different arousal. An example is excitement, which is a high-arousal positive emotion, whereas calm is a low-arousal positive emotion.

While these models exist within psychology, most sentiment analysis models do not use these dimensions to analyze emotion. As an example, Google's sentiment analysis program is only able to measure positive versus negative emotions [7]. Similarly, Stanford's Sentiment Treebank dataset also only measures along the dimension of valence, ranging from scores of 1-25 of highly negative to highly positive words [8]. These models for sentiment analysis have extended throughout the literature, mostly focusing on the single dimension of valence to determine emotion. However, there is also work being done on discrete emotions as a way of classifying emotion, mirroring Ekman's basic emotion model. Currently, it seems like most work is being done on either the dimension of valence, or on discrete emotions [9]. Thus, most sentiment analysis programs are highly lacking

in the ability to measure along both the dimensions of arousal and valence as part of the affective circumplex. Because of that, the current work seeks to use deep learning techniques to improve on the minimal amount of previous work with affective circumplex sentiment analysis.

Improvements in sentiment analysis to include the affective circumplex would be useful in a variety of domains. On one hand, this simple, 2-dimensional model of emotion would encapsulate a whole set of emotions at once, allowing for sentiment analysis programs to easily identify and analyze all of emotion. By being able to identify emotions by a combination of two dimensions (e.g. excitement as a combination of high positivity and high arousal), more emotions can be expressed instead of having to use a smaller set of discrete emotions. Thus, the affective circumplex offers a means of sentiment analysis that efficiently covers many types of emotion. On the other hand, psychological research would benefit from having a sentiment analysis program to quantify emotion in a way that matches with their research. Because there are very few sentiment analysis programs that quantify emotion along the affective circumplex, it is difficult to analyze lots of textual research for psychological insights. Currently, psychologists have had to adapt other sentiment analysis programs, or have had to measure emotion manually. Thus, creating a sentiment analysis tool for the affective circumplex would help psychologists analyze text automatically in the dimensions they are interested in, and would be a boon for psychological research.

2 Related Work

There are very few datasets that have training labels for valence and arousal, and as such, there are very few sentiment analysis programs for the affective circumplex. We could only find one dataset that is labelled along these dimensions, which consisted of two ratings of 1 to 9 for each dimension [3]. For more details about the dataset, see the Data section. The creators of the dataset also provided a simple model, using bag-of-words and linear regression in order to predict each dimension of valence and arousal. This model was compared to the target scores in the dataset, and had a correlation of 0.650 for valence, and 0.850 for arousal, and is the current state-of-the-art. For further details of the model, refer to Preotiuc-Pietro, 2016 [3].

Because of that, we looked to other sentiment analysis architectures in order to gain inspiration for our current model. One aspect of the architecture that we were looking to include was word order - the previous work only used a bag-of-words method, and it seems like methods using word order would benefit sentiment analysis programs for the affective circumplex. As such, we planned on using an LSTM in order to encode word order of the sentence. One previous model that the current work takes inspiration from is a LSTM that feeds into a CNN in order to classify sentiment analysis [4]. A variant of this model was chosen because we wanted to include the LSTM, but also wanted to use the CNN in order to change the variable-length input of the LSTM to a fixed size input. By using max-pooling, the CNN is able to turn the variable length into a fixed size based on the number of filters. Additionally, since this method of LSTM-CNN has been proposed before, and has reasonable results in other sentiment analysis tasks, this model was used as a baseline for our experimentation.

In the baseline model, one LSTM-CNN is used for each dimension of valence and arousal. However, instead of using one model each for the dimensions of valence and arousal, we also tested whether multi-task learning would help improve results of our model. While the affective circumplex represents all combinations of the two dimensions, research has shown that there appears to be a V-shaped correlation when actual emotion words are mapped onto the circumplex [10]. Emotion words neutral in arousal tend to be classified as neutral in valence as well. However, both highly positive and highly negative valence words tend to be high in arousal. Thus, there seems to be somewhat of a pattern between the various dimensions of the affective circumplex. This inspired the use of multi-task learning, because previous research has used multi-task learning to learn inter-related concepts, such as comorbidity in mental illnesses [11]. This previous research suggests that interrelations between valence and arousal would benefit from using multi-task learning. As such, we used a single model based on the LSTM-CNN, but ended added on multi-task learning to predict each dimension of emotion based on the LSTM-CNN core.

We also wanted to use self-attention mechanisms, since they have been highly effective in other sentiment analysis programs as well. We modeled our self attention layer after previous work, and applied the self attention to the hidden states of the LSTM [12]. This self attention layer serves various purposes in our model - first, we hope that by using this self-attention, we can replicate other

sentiment analysis program's success in prediction. Additionally, we think that the self-attention may help bypass some of the computational complexity of using a CNN, while also providing an effective means of turning the variable length LSTM output into a fixed size. Thus, one version of our model incorporates the previous literature by using self-attention mechanisms as well.

3 Approach

3.1 Baseline Model

This baseline model serves as the basis for the other models tested. As such, the following section will elucidate the architecture of the baseline model.

For this model, Facebook posts are first tokenized using NLTK's TweetTokenizer. This specific tokenizer ensures that each element of the message is properly parsed out, such as separating out punctuation from words, and extracting tokens such as emoticons (e.g. ':)').

Once the posts are tokenized, each token is turned into a 100-dimension pre-trained GloVe vector. Specifically, the GloVe vectors pre-trained on Twitter are used, since it is most similar to the current domain this model studies, and is likely to produce higher-quality embeddings. Each token is converted to lowercase, and if it is in the GloVe lexicon, it is turned into the 100-dimension word embedding. Otherwise, the token is turned into a 100-dimension padding vector of all zeros.

After creating embeddings from tokens, these embeddings are fed into the rest of the model. Two identical models were created in order to separately predict the dimensions of valence and arousal, although the inputs (embeddings) as well as the architecture of the models are identical. This section continues describing the design of this model.

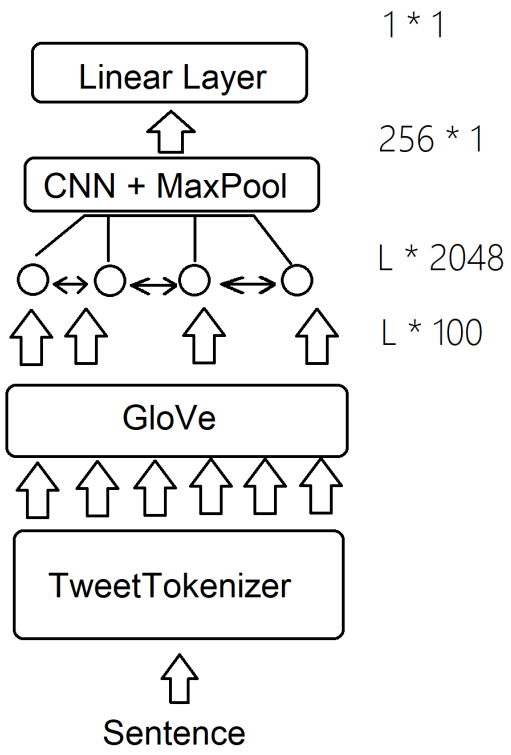


Figure 1: Overview of Baseline Model

At the start of the model, inputs are fed to a bidirectional LSTM. The LSTM is used in order to contain word-order information. Since this information was not included in previous models, this model aims to improve on previous work by incorporating this aspect of the messages. The BiLSTM we use for our specific model has a hidden layer dimension of 1024.

The outputs of the BiLSTM are then fed into a CNN, and then max-pooled across filters. Originally, the model solved a classification task, as it treated the sum of the two human labels as a class. For instance, if a message was labelled with a valence of 4 and 3, the class the model would try to predict was 7. Since the labels ranged from 1-9, the classes ranged from 2-18. Thus, the CNN had 17 filters, in order to classify the posts into one of the 17 target levels of emotion. The output of these 17 filters was then compared to the target level using cross-entropy loss. However, the results of this classification task was quite poor (as described in the Experiments section), so the model was switched to a regression task instead, which tried to predict the label as a continuous variable between 1 to 9.

In the regression task, the outputs of the BiLSTM are fed into a CNN with 256 filters and kernel size of 5. The outputs of the CNN are max-pooled across each filter. The CNN helps condense the dimensions of the LSTM output to a fixed size so it can be fed into a Linear layer for regression. The Linear layer takes 256 dimensions as input, and returns a one dimensional value as the predicted

emotion level for the input Facebook post. The loss of this predicted value is calculated using mean squared error.

All code was written from scratch, except for some starter code given by a PyTorch tutorial [5].

3.2 Multi-task model

Since previous research has shown that valence and arousal may be interrelated, we also tested a multi-task model. This model is a variant on the previous baseline model. Instead of training two separate models for valence and arousal, the same model is used for both. The LSTM and CNN are kept the same - however, in order to predict the two dimensions, two separate linear layer heads are used. Instead of feeding the output of the CNN into one linear layer, the output of the CNN is fed into one layer for valence, and one layer for arousal. These layers still take 256 dimensions as input (each), and return a one dimensional value representing the corresponding value for each emotion dimension. Again, the loss of this predicted value is calculated using the mean squared error. While this is a simple architecture change from the baseline model, we thought that adding two Linear layer heads to the same model would help the model learn the core features that underlie both valence and arousal, and allow for the Linear layer heads to differentiate between the two dimensions.

3.3 Self-attention model

We also incorporated a self-attention layer into one version of our model. The self-attention replaced the CNN layer, but kept everything before it. The self-attention layer mirrors the one used by Lin et al [12], and takes in the variable length LSTM to create a fixed length embedding matrix. For our self-attention layer, we used a hidden layer of 350, and an output layer of 100. Thus, while the LSTM would give us a matrix of size L by 2048, the output matrix from the self attention layer would be 100 by 2048. Afterwards, the self-attention model also incorporates multitask learning by splitting into two heads, predicting valence and arousal. Each head incorporated a linear layer that collapses the vector to dimension 2048 by 1, passes it through a ReLU layer, and then passes it to another linear layer that collapses it into a single value that represents the predicted value.

4 Experiments

4.1 Data

The current dataset used is one of the few existing datasets that labels on the dimensions of both valence and arousal [6]. It has around 3000 Facebook posts, and for each post there are two labels each for valence and arousal, coded by two different human subjects. For the regression task in this model, the labels for each dimension were averaged to make the target score for the model. For more information on the dataset, refer to refer to Preotiuc-Pietro, 2016 [3].

4.2 Evaluation method

In order to compare with previous work that measures sentiment analysis on the affective circumplex, the model was evaluated using the correlation coefficient between the target values of each dimension and the predicted value of each dimension. Thus, there were two correlation coefficients calculated, one for the valence dimension model, and one for the arousal dimension model.

4.3 Experimental details

We experimented on using the final layer as either a classifier or a regression. As described in the Approach section, a CNN was initially used as the output from the model, and had 17 filters for the 17 classes. When tested on a tiny dataset, after around 10 epochs, the model settled on a total loss of around 4.9. However, the loss given by chance would be 2 * -ln(1/17), giving a loss of around 5.6. Thus, the classifier model did not give good results, especially since the model should have overfit the dataset since it was so small.

Because of that, the model was switched to using a linear regression layer. When the model was run on the same tiny dataset, after 10 epochs, the model settled on a total loss (mean squared error) of

about 10. This was better than the previous result, since it suggested that the rating was only about 2 or 3 points off the target rating. However, it was still concerning that the model did not overfit the small dataset.

The learning rate was changed in order to optimize the model, and have it successfully overfit the small dataset. Initially, a learning rate of 0.001 was used. When it was increased to both 0.005 and 0.01, the model's loss would keep increasing, instead of decreasing and settling down. Thus, those learning rates were too high. Additionally, a learning rate of 0.0001 was used, and while it converged to the same loss as a learning rate of 0.001, it took many more epochs to converge. Thus, a learning rate of 0.001 was used, since it didn't cause the loss to explode, while also converging within a reasonable number of epochs.

Finally, the batch size was experimented with as well. Originally, a batch size of 32 was used. However, on closer inspection it seemed like the predicted values for each batch were all around the middle of the scale. This suggested that the averaging of the gradients to maximize the mean squared error might have removed differences between the more extreme target ratings. Thus, to remove this possible averaging of the gradients that would cause overly safe and middling predictions, the batch size was changed to 1. By doing so, the model was able to properly overfit the small dataset and have a loss near 0.

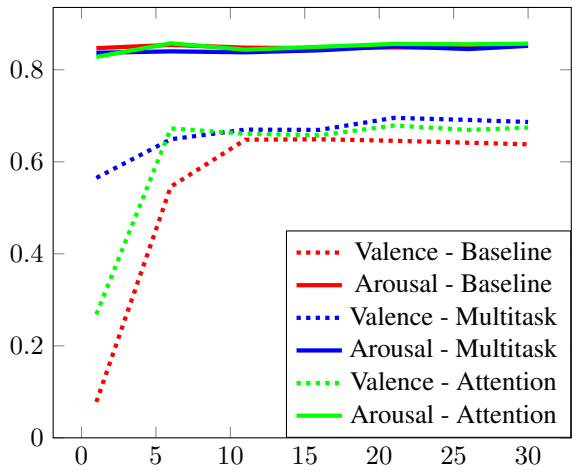
4.4 Results

For the baseline model, the maximum correlation was around 0.649 for Valence, and around 0.856 for Arousal. These results seemed to stabilize at different times depending on the dimension - for Arousal, one epoch was enough to be able to predict with fairly high accuracy. On the other hand, it took about 11 epochs to be able to predict Valence, and even then, the accuracy was much lower than that of Arousal.

All models were run for 30 epochs. For the multi-task model, the maximum correlation was around 0.696 for Valence, and around 0.852 for Arousal. These results converged much more quickly than that of the baseline model, especially for the Valence dimension. While the baseline model had a 0.08 correlation at epoch 1, and a 0.55 correlation at epoch 6, the multitask model had a 0.57 correlation at epoch 1, and a 0.65 correlation at epoch 6. The multitask model had around a 5 epoch advantage in training over the baseline model, and was able to have good results even at the first epoch.

For the self attention model, the maximum correlation was around 0.679 for Valence, and 0.857 for Arousal. While these results were overall better than the baseline model, the model also took a while to converge - starting with a correlation of 0.27 for Valence at epoch 1, before jumping up to 0.67 at epoch 6. Additionally, while it initially slightly improves over the multitask model at epoch 6, the multitask model eventually has the highest correlation in later epochs.

The following figure graphs these differences in model results:



Surprisingly, the baseline results are only as good as previous work, which used a simple linear regression bag-of-words model. Remarkably, their results for valence and arousal were 0.650 and

0.850, respectively, as was the result of the current baseline model. This suggests that the LSTM and CNN may not have as much of an effect on the data as the linear regression itself. Since both models end with a linear regression, it may just be that the linear model is able to predict the results, whereas the underlying LSTM and CNN architecture may just be mixing up the data without extracting any meaning. Thus, it indicates that a different underlying structure may be needed to improve results, instead of just replicating the results of previous, simpler models.

However, both the multitask and self-attention models show slight improvements, specifically to the Valence dimension. Since both methods use multiple heads, it suggests that using the same model to predict both dimensions may help incorporate information that underlies both valence and arousal. Nevertheless, the multitask model that uses a CNN instead of a self-attention layer still performs better. Thus, while the baseline model suggested that the underlying LSTM-CNN architecture did not improve results, the differences between the multitask and self-attention models suggests that the architecture behind the linear layers also matters. Thus, it remains to be determined what combination of methods leads to an architecture that can most effectively model the sentiment before it gets passed onto the linear layers.

5 Analysis

In order to perform qualitative analysis, the epoch that each model performed best on was chosen. Specifically, since the correlations for the Arousal dimension were virtually the same across all epochs and all models, the epoch with the maximal Valence correlation was chosen. For the baseline model, the 16th epoch was chosen, and for the multitask and self-attention models, the 21st epoch was chosen.

First, since we were previously worried about the model predicting middling values, we tested how the model handled extreme values. We picked the messages that had been rated by humans as less than or equal to 3, or greater than or equal to 7 for each dimension of valence and arousal. Contrary to our predictions, all three models actually had higher correlations on the partial dataset than the entire dataset. For valence and arousal, the baseline model had a correlation of 0.782 and 0.879, respectively, the multitask model had 0.812 and 0.882, and the self-attention model had 0.8039 and 0.880. Thus, the accuracy on the extreme values were actually more accurate than in general. This suggests that the model is good at predicting extreme valence and arousal, and actually may have more difficluty determining when a message is neutral. Additionally, it is interesting to note that while the arousal correlations were around the same for the partial dataset, the valence correlations followed the expected pattern.

We also looked at the messages that had particularly high losses, and qualitatively examined what similarities were in messages with high error. One thing we noticed was that many of the messages with high loss were present across all three models. Thus, our analysis continues without distinguishing between the different models.

One consistent error was that the models would predict a positive valence instead of a negative valence when many exclamation marks were used. For instance, when reading the message "My friends are going back home!!! I'll miss you guys so much!!! :((", all models gave a slightly positive prediction, when the human rating was quite low. Another example message is "Why can't exs just DIE after you are done with them!?!?!?LOL.....seriously!!!!!!!" This message clearly is highly negative - however, the model predicted it as slightly positive, most likely because exclamations are normally used in very positive messages. Thus, due to the large amount of exclamation marks in the sentence, the model may have expected a more positive emotion even though the exclamation was a negative one.

Another consistent error was that the models appeared to have difficulty understanding what the entailment of a situation would mean. For example, the message "will somebody out there please shoot me already" had a slightly negative predicted valence. However, to a human reader, this message would entail an extremely negative valence, since it suggests that a person may be suicidal and be at risk of ending their life. A sentiment analysis program could interpret this at face value - that someone wants to get shot, without understanding the entailment of suicide and the accompanying suffering that would cause suicide. An example can be seen for arousal as well - for the message "getting ready for the big move people", the models predicted a moderately positive arousal. However, the human ratings were much higher in arousal. This is likely due to the entailment of a 'big move' involves not

only moving the people themselves, but also moving possessions, departing from loved ones, and settling down in a new, unfamiliar place. To a human, moving is not just a big thing - it's often a large upheaval, and can even be a significant life event. Thus, the amount of energy that moving entails is more than can be taken from face value, and requires a human understanding of the entailment of moving.

6 Conclusion

The current work extends previous work on creating a sentiment analysis program for the affective circumplex. Previous work with sentiment analysis on the affective circumplex has been limited, even though it is a prevalent model of emotion in psychology. The current state-of-the-art is a simple bag-of-words linear regression. The current work extends this model by using deep learning methods to improve results. The current work uses three models - a baseline LSTM-CNN model, a multitask variant, and a self-attention variant. While originally formulated as a classification task, a regression task has shown to perform better at prediction. Furthermore, while the baseline model did not improve previous state-of-the-art, using both the multitask and self-attention variants allowed for increases in the correlation of Valence from 0.65 to around 0.69. Thus, the current work improves on previous state-of-the-art, and improves the prediction of the valence of messages.

Future models will aim at improving the results of the predicted values even more, most notably for the dimension of valence. Possible avenues for improvement may include using character representations of embeddings. Since many words in social media posts may be misspelled purposefully in order to indicate emotion (e.g. 'Happyyyyy'), some words may not show up using GloVe embeddings. This could be alleviated using character-based embeddings. Another benefit of using character-based embeddings is that it could also incorporate capitalization information. Currently, the model does not discriminate between lowercase and capitalized words, even though capitalization (especially in social media posts) may indicate various emotions.

Additionally, future models may improve on valence by taking semantic meaning and understanding of entailment into account. Although the BiLSTM sought to improve this using word-order, the results indicate that there was not much improvement in results than using a bag-of-words method. However, future models may incorporate other methods to improve semantic meaning, and improve the results of prediction as well.

7 Additional Information

We would like to thank Anand Dhoot for mentoring this project.

References

[1] Ekman (1999) *Basic Emotions*. Retrieved from `https://www.paulekman.com/wp-content/uploads/2013/07/Basic-Emotions.pdf`

[2] Posner (2005) *The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology*. Retrieved from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2367156/`

[3] Preotiuc-Pietro (2016) *Modelling Valence and Arousal in Facebook posts*. Retrieved from `https://wwbp.org/papers/va16wassa.pdf`

[4] Sosa (2017) *Twitter Sentiment Analysis using combined LSTM-CNN Models*. Retrieved from `https://www.academia.edu/35947062/Twitter_Sentiment_Analysis_using_combined_LSTM-CNN_Models`

[5] PyTorch *Training a Classifier*. Retrieved from `https://pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html`

[6] Preotiuc-Pietro (2016) *Valence and Arousal Facebook Posts*. Retrieved from `http://wwbp.org/downloads/public_data/dataset-fb-valence-arousal-anon.csv`

[7] Google (2018) *Sentiment Analysis Tutorial*. Retrieved from `https://cloud.google.com/natural-language/docs/sentiment-tutorial`

[8] Stanford (2013) *Sentiment Analysis*. Retrieved from `https://nlp.stanford.edu/sentiment/treebank.html`

[9] CodaLab *SemEval-2018 Task 1: Affect in Tweets (AIT-2018)*. Retrieved from https://competitions.codalab.org/competitions/17751#learn_the_details

[10] Kuppens (2012) *The Relation Between Valence and Arousal in Subjective Experience*. Retrieved from https://www.researchgate.net/publication/233900065_The_Relation_Between_Valence_and_Arousal_in_Subjective_Experience

[11] Benton (2017) *Multi-Task Learning for Mental Healthusing Social Media Text*. Retrieved from http://m-mitchell.com/publications/multitask-clinical.pdf

[12] Lin (2017) *A STRUCTURED SELF-ATTENTIVE SENTENCE EMBEDDING*. Retrieved from https://arxiv.org/pdf/1703.03130.pdf