CIS530 Final Project Literature Review: Bias in Sentiment Analysis

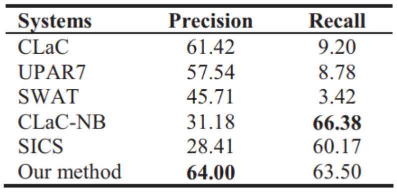
Luke Yeagley, Eduardo Ortuno Marroquin,
Arjun Lal, Sneha Advani, Tasya Tsygankova, Kanika Mohan

Improving Emotional Intensity Classification using Word Sense Disambiguation-

Carillo de Albornoz, Plaza, and Gervas.

These researchers from the Universidad Complutense de Madrid sought to automatically tag sentences with a emotional intensity value using the WordNet Affect Lexicon along with a word sense disambiguation algorthim. The purpose of the word sense disambiguation algorithm is to assign emotions to concepts instead of terms. This allowed for a unique approach to the Emotional Intensity Classfication problem that did not solely rely on the matching of words to a lexicon. The researchers' solution allows for the finding of emotional intensity and polarity, a task that proves to be very difficult. The WordNet Affect lexicon is used to identify those concepts which are "a-priori candidates" of certain emotions or feelings. The article states that most lexicons face a limitation given that they use words and stems as the primitive units, as opposed to the context in which the word is used. The WordNet Affect database however provides a list of 911 Wordnet synsets labeled with a higherarchical set of categories representing different emotions.

First the researchers preprocessed with Tokenizer, Part of Speech tagger and Sentence splitter modules in General Architecture for Text Engineering (GATE) along with a stop words list to eliminate generic and high freqency terms. Next the terms were mapped to the concepts in the WordNet lexical database using an implementation of the "lesk" algorithm in the WordNet Sense Relate Perl package. Then the WordNet concepts were mapped to emotional categories from the WordNet Affect Lexicon. Finally A Random Forest Model was used to classify the intensity of the emotions.



Systems	Precision	Recall
CLaC	61.42	9.20
UPAR7	57.54	8.78
SWAT	45.71	3.42
CLaC-NB	31.18	**66.38**
SICS	28.41	60.17
Our method	**64.00**	63.50

Evaluation was measured against a test set of 1000 manually labeled news headlines. The researcher's model had the best Precision score of 64.00 and the second highest Recall Score of 63.5.50 when compared to other systems from SemEval 2007.

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings-

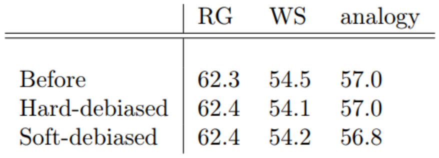
Bolukbasi, Chang, Zou, Saligrama, and Kalai

Researchers from Boston University and Microsoft sought to develop algorithms that de-bias word embeddings so that models trained using these word embeddings will no longer exponentiate the bias in the model when performing tasks. This paper focuses specifically on the analogy task. Word embeddings are used for many tasks given that they are a mathematical way to represent words and context. In this paper, the researchers realized that building the word embeddings based on the 3 million English words from the Google News texts produced visible bias when accomplishing tasks like solving an analogy. The namesake analogy clearly showing the bias in the current system was:

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

The researchers then go on to show that the bias stems from the word-embeddings which portray the gender biases in society in general.

The researchers then layed out their plan to debias the word embeddings. Their debiasing algorithms first sought to identify a direction or ”gender subspace” of embeddings that capture the bias. The second step was to either ”Neutralize” or ”Equalize” the bias present by setting gender neutral words to zero in the gender subspace or by making sure the gender neutral word is equidistant to all words in the equality set. Then the researchers ran a SVM classifier on the words to find Gender neutral words.



	RG	WS	analogy
Before	62.3	54.5	57.0
Hard-debiased	62.4	54.1	57.0
Soft-debiased	62.4	54.2	56.8

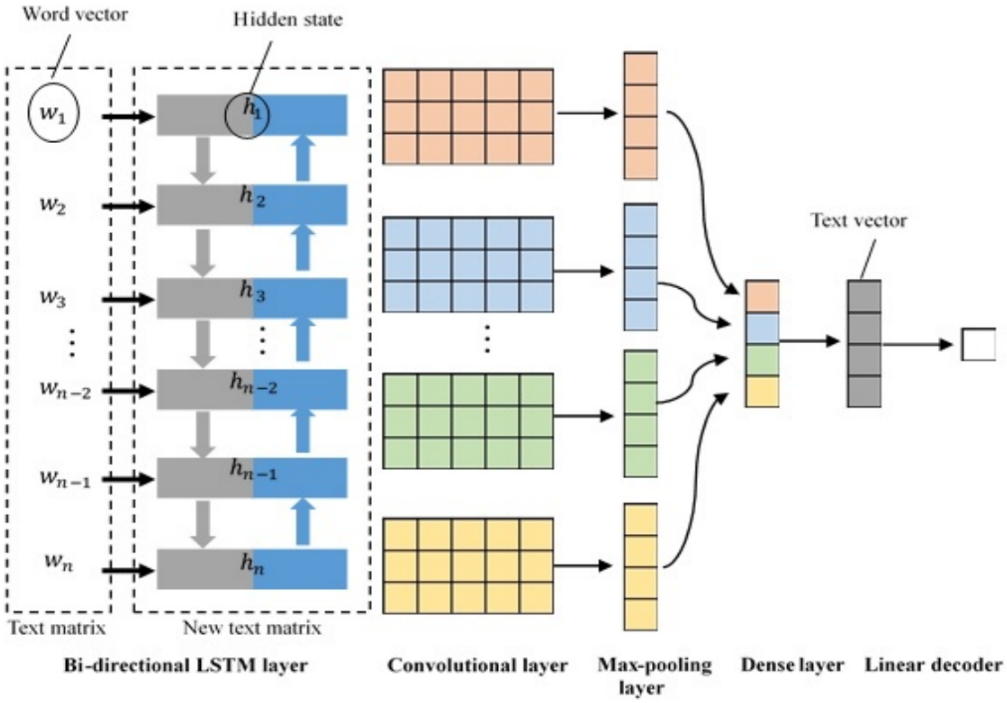
Evaluation consisted of comparing the accuracy of the analogy task using the normal embedding and the ”Hard-debiased” and ”Soft-Debiased” embeddings. The ”Hard-debiased” embeddings performed as well as the normal embeddings at 57.0 percent accuracy. The researchers also found that the debiased embeddings worked better at clearing up direct bias versus indirect bias.

YZU-NLP at EmoInt-2017: Determining Emotion Intensity Using a Bi-directional LSTM-CNN Model

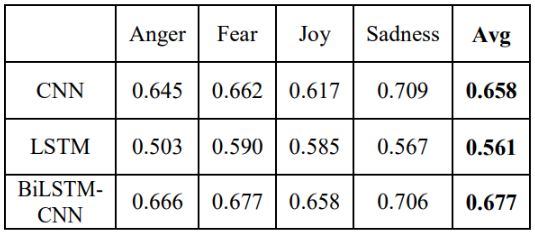
He, Yu, Lai, Liu.

Researchers from Yuan Ze University also tackled the emotional strength classification problem (EMOInt-2017) this time with a neural network. Furthermore, given a tweet, and an emotion X, the researchers sought to determine the strength from 0 to 1 of the emotion being expressed. The proposed system used word embeddings and a bi-directional LSTM-CNN model. The paper states the Convolutional Neural Network (CNN) is effective for finding features from input without including the global information of that text, i.e. context. However, the Long Short-term memory can leverage context between words in a text and feeding that input into a CNN.

First the researchers transformed tweets into text matrices and then applied the bi-directional LTSM to build new text matrices as input to the CNN to obtain text vectors for emotion intensity prediction. The following is a diagram of the setup.



The training and testing dataset was comprised of tweets hand labeled with one of the four emotions (Joy, Anger, Fear, and Sadness), along with a manually labeled intensity value from 0 to 1. The models' accuracy was evaluated using the Pearson and Spearman coorrelation coefficient. The following displays the results using the CNN, LSTM and the combined model.



	Anger	Fear	Joy	Sadness	**Avg**
CNN	0.645	0.662	0.617	0.709	**0.658**
LSTM	0.503	0.590	0.585	0.567	**0.561**
BiLSTM-CNN	0.666	0.677	0.658	0.706	**0.677**

The combined model had the best score.

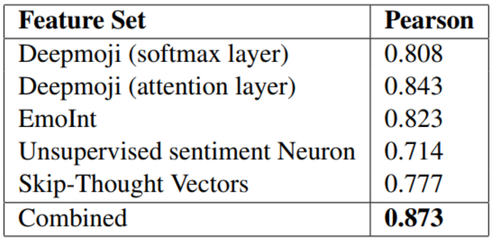
SeerNet at SemEval-2018 Task 1: Domain Adaptation for Affect in Tweets

Duppada, Jain, and Hiray.

The researchers at Seernet Technologies found the best performing system for the SemEval-2018 Affect in Tweets problem. The researchers focused on the classification and regression on four classes of emotion: anger, fear, joy, and sadness, as well as the valence of these emotions on a scale from -3 to 3. The created system conducts ”domain adaptation of 4 different models and creates an ensemble” for prediction.

First the researchers preprocessed the tweets using the **tweettokenize** tool and then used various deep learning techniques to extract features of these tweets. DeepMoji allowed for state-of-the-art results in

”downstream” tasks harnessing transfer learning. Skip-Thought vectors was then used to produce generic sentence representations which where then used as input to the Unsupervised Sentiment Neuron. Other features were added using the EmoInt package followed by various machine learning tools to predict. Like other researchers above, to evaluate performance for the 4 emotion type classification, the researchers used the Macro-averaged Pearson correlation.



Feature Set	Pearson
Deepmoji (softmax layer)	0.808
Deepmoji (attention layer)	0.843
EmoInt	0.823
Unsupervised sentiment Neuron	0.714
Skip-Thought Vectors	0.777
Combined	**0.873**

As you can see the ensemble of the various methods worked well for prediction for all tasks especially the V-reg task, which achieved an impressive score of .873. Some limitations of the model included contextual knowledge. For example, the tweet ”Your club is a laughing stock” followed by a laughing emoji does not convey joy in this case given that the laughing emoji and phrase seek to convey sarcasm.

Why We Chose This Published Baseline

Most importantly, we chose to reimplement this published baseline because it yielded the best results for our relevant shared task. Ultimately, it was the most informative guidance we could find for capturing emotional intensity and so we thought it would best inform our own approach if we were to follow it. Furthermore, we were drawn to the logical and multilayered approach that was taken by the authors; the focus on extracting different and varied features to build multiple models that are then combined is simple yet effective. It is also one that can be appropriately split up among multiple members and extrapolated off of to form extensions. Thus, we decided that this would be the best to implement.

Works Cited

Duppada, V., Jain, R., & Hiray, S. (2018). SeerNet at SemEval-2018 Task 1: Domain Adaptation for Affect in Tweets. *Proceedings of The 12th International Workshop on Semantic Evaluation,*18-23. doi:10.18653/v1/s18-1002

Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervs. 2010. Improving emotional intensity classification using word sense disambiguation. Special issue: Natural Language Processing and its Applications. Journal on Research in Computing Science, 46:131--142

He, Y., Yu, L., Lai, K. R., & Liu, W. (2017). YZU-NLP at EmoInt-2017: Determining Emotion Intensity Using a Bi-directional LSTM-CNN Model. *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis,*238-242. doi:10.18653/v1/w17-5233

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, Adam Tauman Kalai: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NIPS 2016: 4349-4357