Multi-modal sentiment Classification on Memes

Project in Neural Network and Fuzzy Logic (BITS F312)

SuhasPrasanna
2017*A*7*PS*0002*G*

NahushHariharKumta
2017*A*7*PS*0930*G*

SujaySharma
2017*A*7*PS*0012*G*

1 INTRODUCTION

Sentiment analysis can be broadly characterized as the field which deals with computationally identifying and characterising opinions/emotions expressed in a medium. It's one of the fastest growing fields in machine learning and nearly 7,000 papers on this topic have been published and, more interestingly, 99% of the papers have appeared after 2004[1]. Moreover, there is a shift in focus towards researching on Social Media such as Facebook,Twitter,etc.. since 2016[1]. It's safe to say at this point that a large part of the shared communication that occurs on Social Media occurs in the form of memes. Given, the shift in sentiment analysis to studying social media, it becomes important to consider one of the most popular forms of discourse on Social Media. Currently, machine learning techniques are the most popular techniques used in supervised sentiment analysis. However, hybrid approach with lexicon based techniques provide the best results in the area of supervised sentiment analysis[2]. In this case we will be using deep learning combined with natural language processing techniques to tackle the problem. The tasks at hand are to gauge the offensiveness of a meme, the overall sentiment of a meme as well as find out if it is motivational. All these tasks can be broadly placed under the category of sentiment classification. The data given is memes which have both a visual aspect as well as a text based aspect to it. Hence this falls under the category of multi-modal sentiment classification, the two modes being the image of the meme as well as the text of it.

2 RELATED WORK

Text-based sentiment analysis has long been the standard bearer in the area of sentiment analysis and only recently has sentiment analysis from other modalities, such as speech and vision, began to be considered. More recently in the field deep recurrent neural networks have found great success[9] and word vector based approaches have been very successful too keeping up with the general trends in Natural Language Processing[7]. Application of visual sentiment analysis in Social Media has been a relatively recent development in the field with papers such as [6] being a pioneering one. They used data consisting of images from social media and assigned adjective-noun pairings to each one. They used this to train a computer vision based machine learning model to associate images with these adjective-noun pairs and then used these pairs to generate sentiment. This opened the door for a lot of computer vision based models to generate adjective-noun pairs for visual sentiment analysis[7]. With the success of Convolutional Neural Networks in the recent past, they have been used as a basis for training on these adjective-noun pairings to achieve good results[8].

The field of multi-modal sentiment analysis is a relatively nascent one. The most well researched part of multi-modal sentiment analysis is in the field of spoken reviews and blogs due to the rich data that they generate[3][4][5]. Given the problem at hand, we will focus on image-text based multi-modal sentiment analysis. When it comes to multi-modal sentiment analysis since we need to incorporate different modalities, there are three broad perspectives.[10]

Early Fusion :- This involves combining the data from both the modalities at the input level itself and this is called early fusion. However, the early fusion methods cannot fully exploit the complementary nature of the modalities involved and may produce very large input vectors that may contain redundancies. And one would need to ensure that the domain and range of the modalities are compatible to combine without one rendering the other pointless.

Intermediate Fusion:- This is mostly used with neural network based models. It involves combining the modalities at an intermediate level in the network. This achieves the best performance of the three, but is the hardest to conceptualise and build a working model and their performance may be affected when a portion of multi-modal contents are incomplete

Late Fusion:- This involves using two independent models to train on each modality separately and combining them at the end. These are the simplest to build and provide reasonably accurate results. But it operates under the assumption that different modalities are independent in the feature space, which is not necessarily true, since they tend to be correlated.

Given the constraints of time, we will specifically focus on late fusion techniques as they tend to do better on multi-modal tasks than early fusion[11]. Hu,Flaxman 2018[12] use late fusion to get good results to infer the latent emotional state of the user. They focus on predicting the emotion word tags attached by users to their Tumblr posts, treating them as "self-reported emotions.". They used a simple but effective architecture to achieve decent success. We will be using their architecture on the task at hand.

3 METHODOLOGY

The model which we have decided to use involves the concept of late fusion. What we did here was that we took two separate models, one being a Convolutional Neural Network (CNN) that takes the images as the input and the other being an LSTM which operates on word embeddings which takes the text as the inputs. These two models are then coupled at their outputs using a dense fully connected layer which gives the final output. **Figure 1** depicts our model although it does not exactly depict the number of layers that were used in either the CNN or the LSTM layers.

The inspiration for such a model came from the ideas presented in the paper by Anthony Hu and Seth Flaxman. 2018 [12] which depicted how a model involving late fusion worked better as opposed to only a CNN or only a RNN or even early fusion [11] for

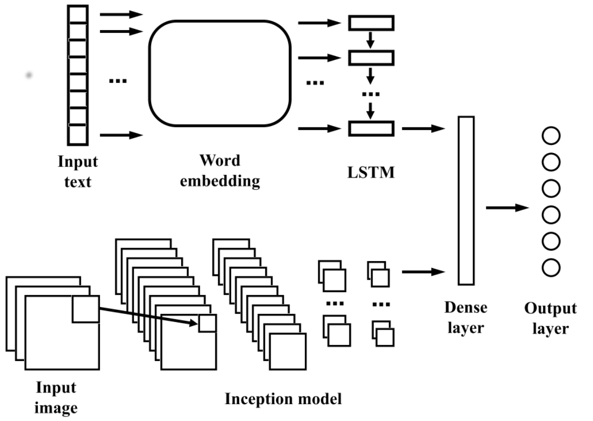

Figure 1: Architecture that we used, based on (Hu,Flaxman 2018)

multimodal analysis. Moreover we also tried various different models (as will be explained) before fully accepting this approach. Our architecture also makes use of pre-trained word embeddings for the RNN layers and a ResNet50 for the CNN layers. The choice of using pre-trained models for the respective layers is based on the success of transfer learning as presented in the following paper by Chuanqi Tan et al 2018 [13], which explains how ideas from a previously trained model can be used and fine tuned on the present dataset to obtain credible predictions. An overview of what approach we adopted to reach the above mentioned model involved a large chunk of trial and error techniques along with the ideas obtained off various resources and papers as mentioned. We started by running a basic Long-Short term Memory (LSTM) model which did not use any pre-trained embeddings which relayed a result that showed it was overfitting on the given dataset. A graph of validation accuracy and validation loss versus number of epochs has been shown in **Figure 2** to depict the overfitting that was observed. **Figure 2** is only plotted for the output of degree of offensiveness but this trend was observed for the other two tasks as well.

In light of the observed overfitting, we decided to run a LSTM with pre-trained word embeddings. We first went with glove-6B-200d.txt embedding and then decided to go with the glove-twitter-27B-200d.txt word embedding which was based off the twitter dataset. It has been shown that domain adapted word embeddings tend to do better than generic word embeddings if the domains have some correlation[15]. It makes sense to domain transfer from twitter since a lot of memes are propagated on twitter. Thus it should work best for the given meme dataset. **Figure 3** depicts its performance has been shown below. Again these depict validation accuracies and validation losses for the degree of offensiveness (the other tasks being more or less similar).

Although a very erratic output, this model definitely performs better than the non-pretrained model. Moreover, we also tried a normal CNN on the given images of the dataset which gave us the following output in **Figure 4**(again for the degree of offensiveness with the other tasks being more or less the same).

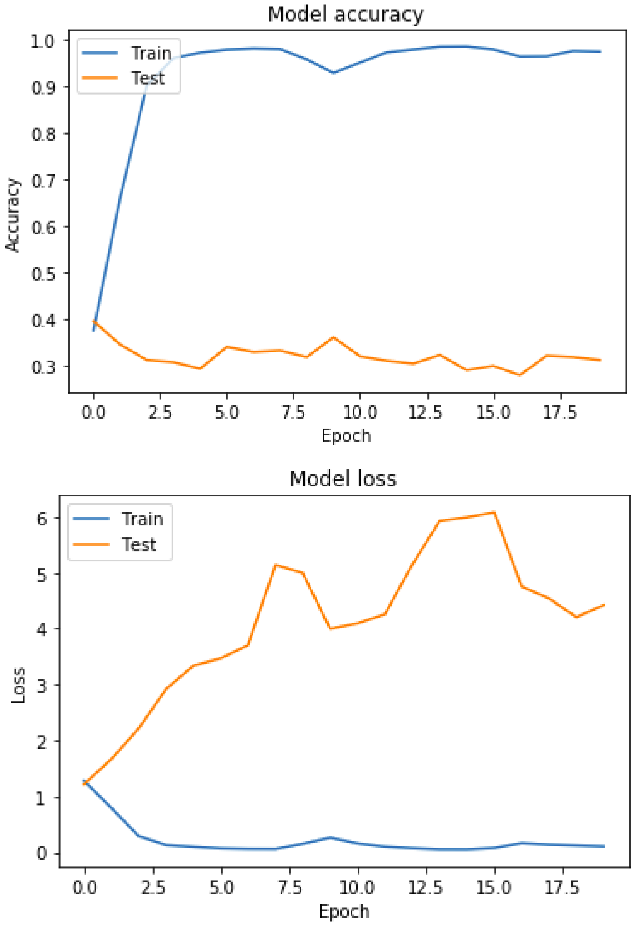

Figure 2: These figures depict the overfitting that was observed when the non-pretrained LSTM model was run on the degree of offensiveness output.

We decided to then use a pre-trained(on imagenet) Resnet to fit the data. Resnet50V2 has very good accuracy[16] and is not too bulky and hence makes it a good choice to train on. **Figure 5** shows the results we achieved with the Resnet.

Although it isn't that much better than the simple CNN model, the marginal improvement around epoch 5 is the best we can use at this point. We can see that this output is also very erratic as it is not learning uniformly. Taking into account the erratic nature of the output that was obtained (but definitely a better result), we decided to use the best of both worlds and adopted the late fusion technique. We trained the ResNet based pre-trained CNN and the twitter word embeddings based LSTM networks independently on the data first. Then we combined the models using late fusion by feeding their intermediate output into a 64 unit dense layer to generate the output. Only the final layer is trainable since both the other networks have been independently trained. The output for the late fusion technique has been shown in **Figure 6**. As we can

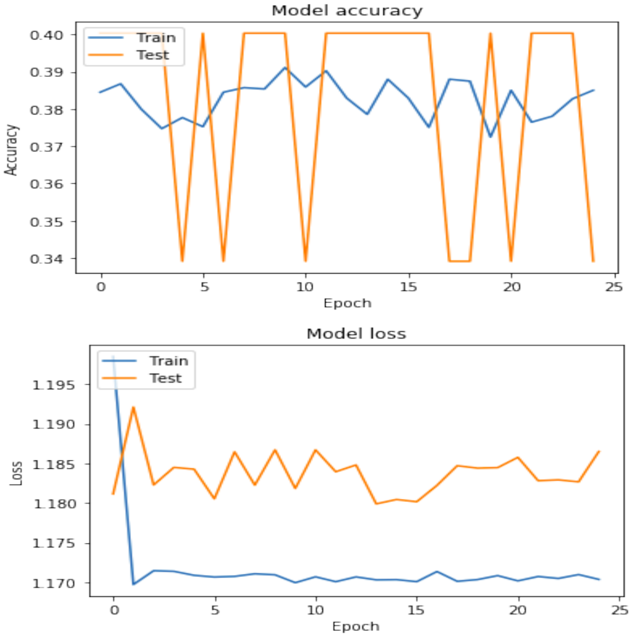

Figure 3: These figures depict the LSTM model running with twitter word embeddings.

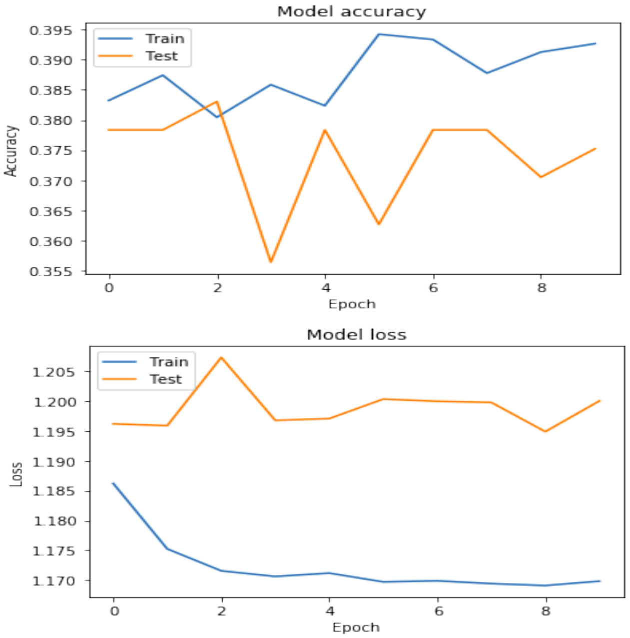

Figure 4: These figures depict a very simple CNN model running on the task of offensiveness

see for the task of motivation, it clearly was unable to learn and started to gradually overfit the data.

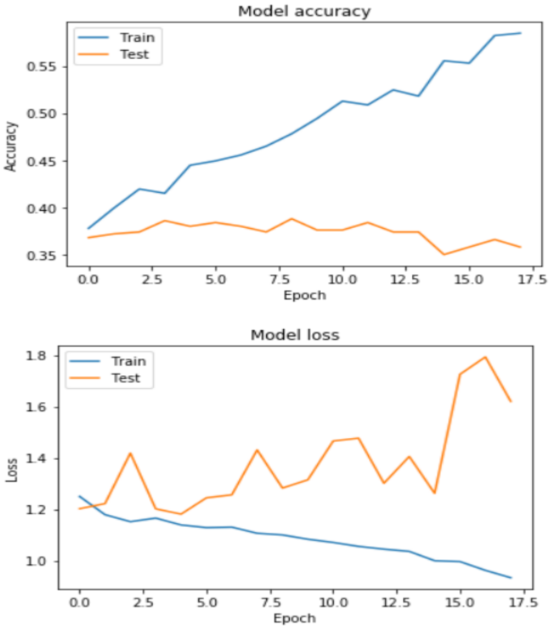

Figure 5: Pre-trained ResNet on the task of offensiveness, note the clear overfitting as opposed to the simple CNN model

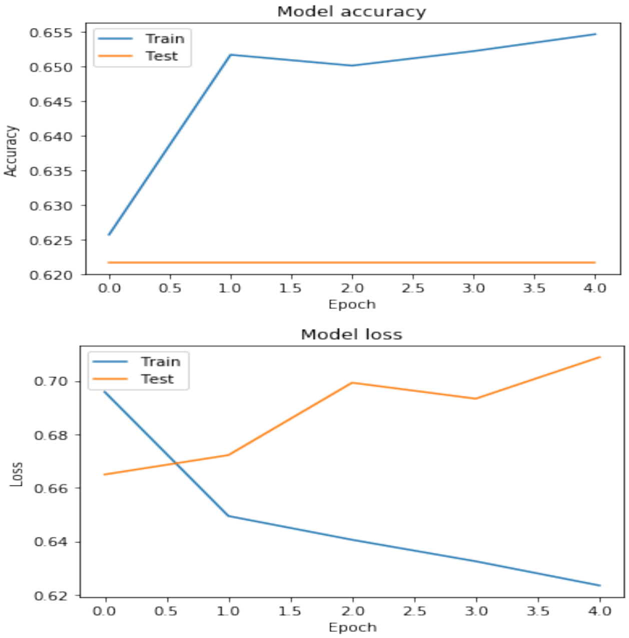

Figure 6: These figures depict the final results of our Late Fusion network on the task of motivation

The results so far as we can see, are not exactly very satisfying. This could be attributed to the lack of data as well as the inconsistencies and lack of underlying patterns in the dataset that are hampering the learning process. These issues have been presented in detail in the following section.

3.1 Other attempts

We also tried to apply other State of the Art models to perform the individual sentiment analysis, however restrictions such as compute resources and time as well as unfamiliar frameworks made it them an inevitable failure.

BERT: BERT(Bidirectional Encoder Representations from Transformers) is a state of the art model in Natural Language Processing and generally achieves the best results in most NLP tasks[17]. It was developed by Google and contains a massive number of parameters. It seemed to require too many computational resources for us to use it

Imavis: A 2017 paper[18] called "From Pixels to Sentiment: Fine-tuning CNNs for Visual Sentiment Prediction" trained a CNN on a twitter dataset and achieved decent results while predicting the overall sentiment of the image as 'positive' or 'negative'. They also have the code as well as the pre-trained weights up on Github, unfortunately for us it was written in Caffe. We attempted to convert it to Keras, but it was to no avail.

4 FINAL MODEL

Figure 7 shows the final model that we are using for the task of offensiveness. A very similar model will also be used for the other two tasks as well.

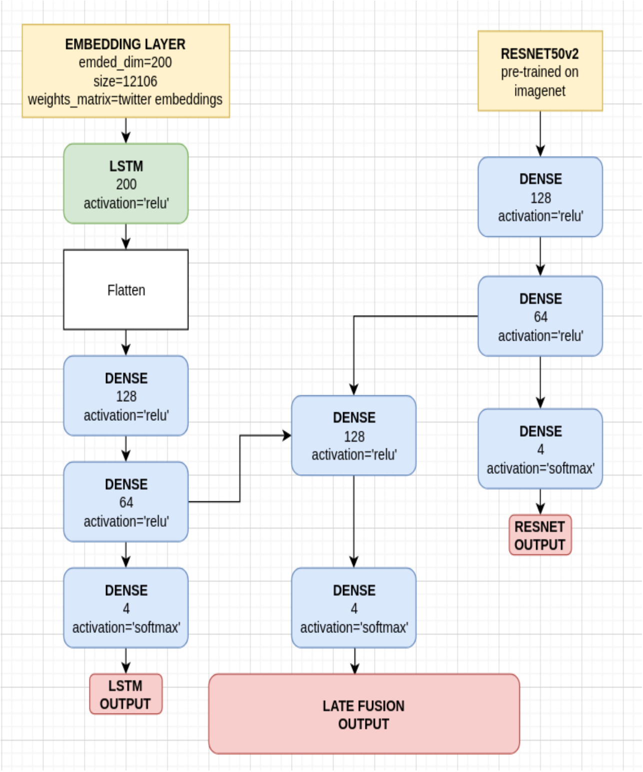

Figure 7: Final model that we are using

5 PROBLEMS IN THE DATA

On a manual inspection of the data, we found several problematic samples. Given the small number of examples that are there in the data, it will be difficult to learn if there are so many obvious mistakes. On doing a simple analysis of memes that explicitly contain the words 'n****r' or 'f*****t', this is the distribution we found. In **Figure 8**, we can clearly see that more than half of them are classified as either 'not offensive' or 'slight'. These memes should be classified as offensive, regardless of content as they use pejorative terms that are extremely offensive to some communities. Moreover, even a visual inspection of the actual content of the memes show that they are clearly offensive. Some of them are even labelled positive and motivational which makes it difficult to understand what those categories are really about. With the most obvious examples being clearly mislabelled it becomes difficult to see how much can really be learned from this dataset. We have compiled to show the clear problems with the dataset. We have emailed them to the Teaching Assistants of this course. Especially given the ongoing problems that already exists when it comes to classifying offensive language or other subjective issues such as the apparent racist predictions of many machine learning algorithms[14]. We have many questions on the methodology of collecting the data given the already existing issues with datasets like these. The mislabelling of the most obvious examples and the low sample size probably shows why it's difficult to learn from this dataset.

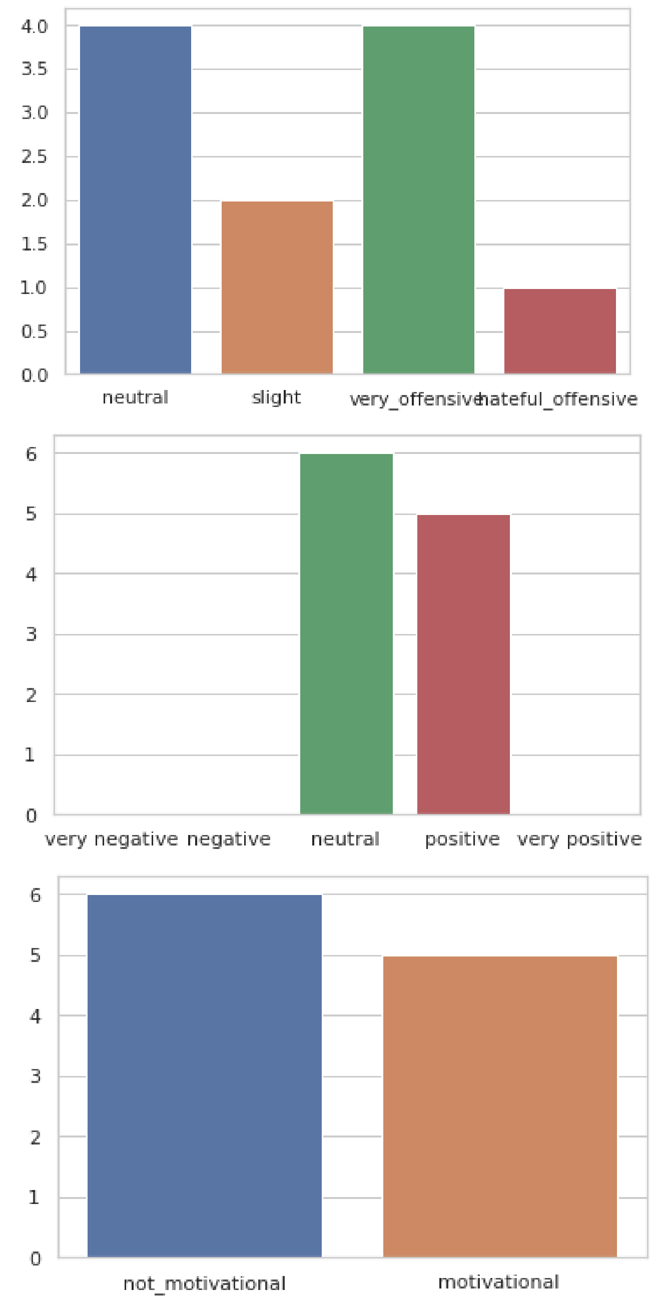

Figure 8: These figures depict the distribution of memes which contain obviously offensive words such as 'n**r' or 'f*****t'**

REFERENCES

[1] Mika V. Mäntylä Daniel Graziotin, Miikka Kuutila. *The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers.* February 2008 https://doi.org/10.1016/j.cosrev.2017.10.002

[2] MR. S. M. VOHRA, PROF. J. B. TERAIYA *A COMPARATIVE STUDY OF SENTIMENT ANALYSIS TECHNIQUES* https://www.semanticscholar.org/paper/A-COMPARATIVE-STUDY-OF-SENTIMENT-ANALYSIS-1-Vohra-Teraiya/3f10b006bab60c7f363bc03e72ad405d264b8d42

[3] Morency, Louis-Philippe and Mihalcea, Rada and Doshi, Payal Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web https://dl.acm.org/citation.cfm?doid=2070481.2070509

[4] M. Wöllmer et al. *YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context* https://ieeexplore.ieee.org/document/6487473

[5] V. Pérez Rosas, R. Mihalcea and L. Morency *Multimodal Sentiment Analysis of Spanish Online Videos* https://ieeexplore.ieee.org/document/6419687

[6] Borth, Damian and Ji, Rongrong and Chen, Tao and Breuel, Thomas and Chang, Shih-Fu *Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs* https://dl.acm.org/citation.cfm?doid=2502081.2502282

[7] Mohammad Soleymani and David Garcia and Brendan Jou and Björn Schuller and Shih-Fu Chang and Maja Pantic *A survey of multimodal sentiment analysis* https://doi.org/10.1016/j.imavis.2017.08.003

[8] Chen, Tao and Borth, Damian and Darrell, Trevor and Chang, S. *DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks* https://arxiv.org/abs/1410.8586

[9] *The Unreasonable Effectiveness of Recurrent Neural Networks* http://karpathy.github.io/2015/05/21/rnn-effectiveness/

[10] Soujanya Poriaa,Erik Cambriac, Rajiv Bajpaib, Amir Hussaina *A review of affective computing: From unimodal analysis to multimodal fusion* https://doi.org/10.1016/j.inffus.2017.02.003

[11] Cees G.M. Snoek, Marcel Worring, Arnold W.M. Smeulders *Early versus Late Fusion in Semantic Video Analysis* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.78.5928rep=rep1type=pdf

[12] Anthony Hu,Seth Flaxman *Multimodal Sentiment Analysis To Explore the Structure of Emotions* https://arxiv.org/pdf/1805.10205.pdf

[13] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, Chunfang Liu *A Survey on Deep Transfer Learning* https://arxiv.org/abs/1808.01974.pdf

[14] https://www.technologyreview.com/f/614144/googles-algorithm-for-detecting-hate-speech-looks-racially-biased/

[15] Prathusha Kameswara Sarma, Yingyu Liang, Bill Sethares *Domain Adapted Word Embeddings for Improved Sentiment Classification* https://www.aclweb.org/anthology/W18-3407/

[16] *Keras pre-trained models page* https://keras.io/applications/

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* https://arxiv.org/pdf/1810.04805.pdf

[18] Campos, Victor and Jou, Brendan and Giro-i-Nieto, Xavier *From Pixels to Sentiment: Fine-tuning CNNs for Visual Sentiment Prediction* https://doi.org/10.1016/j.imavis.2017.01.01