CSE/ISYE 6748

Piper Gradient Practicum Final Report

**Sentiment Analysis and Emotion Inference of Twitter Data**

**Using Convolutional Neural Networks**

Dennis A. Morozov

22 July 2021

Contents

[Introduction 4](#_heading=h.30j0zll)

[Challenge Fundamentals 5](#_heading=h.1fob9te)

[Related Work 5](#_heading=h.3znysh7)

[Data Exploration 6](#_heading=h.tyjcwt)

[Analysis Report 9](#_heading=h.2s8eyo1)

[Models 9](#_heading=h.17dp8vu)

[Benchmarks 11](#_heading=h.26in1rg)

[SpaCy 11](#_heading=h.lnxbz9)

[Vector Derived Labels 11](#_heading=h.35nkun2)

[Results 12](#_heading=h.1ksv4uv)

[Model Overfitting 12](#_heading=h.44sinio)

[Bias 12](#_heading=h.2jxsxqh)

[Result Comparisons 12](#_heading=h.z337ya)

[Benchmark Comparison 14](#_heading=h.4i7ojhp)

[Ethics Report 14](#_heading=h.2xcytpi)

[Ethics Discussion 14](#_heading=h.1ci93xb)

[Motivation 15](#_heading=h.3whwml4)

[Data 16](#_heading=h.2bn6wsx)

[Stakeholders 16](#_heading=h.qsh70q)

[Fairness 17](#_heading=h.3as4poj)

[Transparency 18](#_heading=h.49x2ik5)

[Flexibility 19](#_heading=h.2p2csry)

[Accountability: Model Life Cycle 19](#_heading=h.147n2zr)

[Conclusion 19](#_heading=h.3o7alnk)

[Appendix 22](#_heading=h.23ckvvd)

[CNN -1 22](#_heading=h.ihv636)

[CNN-2 22](#_heading=h.32hioqz)

[CNN-4 23](#_heading=h.1hmsyys)

[Sample Activation Functions Used in Deep Learning 23](#_heading=h.41mghml)

[Hyperparameter Optimization Mesh Plot of Training Iteration with Different Combinations of Hyperparameters. 24](#_heading=h.2grqrue)

[CNN-4 Model Training and Validation Plots 25](#_heading=h.vx1227)

[CNN-1 Best Performing Model Label Distribution with Most Misclassified Normalized Labels. 26](#_heading=h.3fwokq0)

[Benchmark Review 27](#_heading=h.1v1yuxt)

[Twitter samples with their bench mark predictions and model predictions. 28](#_heading=h.4f1mdlm)

# Introduction

Over millennia, humans have developed reliable tools to communicate with our voice, smell, touch, and body language. We infer a great deal of emotional information using our social toolkit to help us reinforce trustworthiness, build alliances, or perhaps strengthen our socio-economic standing. Social media and social networks opened new dimensions for social interaction and presented some challenges. Understanding the emotional content of text as it was meant by the writer is difficult even when the writer curates every word in the sentence. It is especially challenging to infer the appropriate emotions from text that was written in the same breath of air as it was shared with the online community. By reading any Twitter chain, it quickly becomes clear that emotional content abounds. The challenge is how does the emotional information spread and is it properly understood by the receiving party?

Sentiment analysis is the exercise of mining opinions, attitudes, or emotional states from human language. Sentiment analysis has both supervised and unsupervised learning approaches and can be applied to images, sounds, text or other recorded means of expression. This report will focus on sentiment analysis of twitter text data for supervised deep learning applications. Additionally, the report will briefly survey how emotions change and flow within a single message thread [1].

The dense connectivity on social media and the lack of emotional insight may antagonize people to engage in more volatile conversations. Additionally, misinformation can spread fast on social media and have tangible impacts on the real world, as was witnessed by the events within the recent years of 2020-2021. Political division and inflammatory discord along with the algorithms used by social media companies can further silo individuals in their affiliated groups and polarize conversations. Adding a layer of emotional context to social media conversations may provide crucial social clues that humans use in their physical interactions. The emotional context can guide users about the emotional atmosphere in the conversations they choose to participate in.

This Piper Gradient capstone project attempts to add emotional context to tweets, and to their underlying relational graphs (retweet, mentions, hashtag networks). Several Convolutional Neural Net (CNN) architectures are trained with pre-labeled text dataset. CNN models are then used to predict a sentiment value for each tweet passed to the model. Emotional insights in each tweet can be used by analysts to better understand how emotions are passed around with each reply, how emotions transform over the lifetime of a thread, and what emotions are dominant in different clusters (users or topical clusters).The gradient tool can use sentiment analysis on political affiliations, twitter topical and categorical clusters, and review public reactions to political or emergency events.

Artificial neural networks were chosen for sentiment analysis because of their promising performance in natural language processing. CNNs are an especially exciting configuration of artificial neural networks of high interest natural language processing (NLP) and image processing. CNN can be used for numerous applications such as voice recognition, time series analysis,and image recognition among others. CNNs are particularly good at learning in context because during each convolution, collected features look at a single data point as well as the datapoint neighbors to adjust weights for the features. Three CNNs are reviewed in this paper: (i) single layer CNN, (ii) two layer CNN, (iii) and a four layer CNN. Each model was trained with a pre-labeled set of text with the following sentiment values: fear, anger, sadness, love, joy, and surprise. Most reviewed papers performed sentiment analysis with only three labels: positive, neural, negative, which makes the analysis in this paper novel and exploratory.

# Challenge Fundamentals

## Related Work

Twitter sentiment analysis spans from traditional methods in clustering like support vector machine (SVM) to novel approaches with deep neural networks. Twitter is an invaluable data source for data scientists as evidenced by the sheer quantity of research papers that use twitter data. A majority of the literature reviewed on sentiment analysis covered simple categories such as positive, negative and neutral. It proved to be a challenge to find research on nuanced sentiment labels like fear, anger, happiness, joy, surprise for supervised learning. Research on nuanced sentiment analysis was readily available in the unsupervised learning domain where traditional ML methods as well as more interesting models such as latent semantic analysis (LSA) were used [1]. Naïve Bayes, SVM and Maximum Entropy (maxent) require rigorous data cleaning and transformation with models showing high sensitivity to feature selection and data pre-processing. These traditional models perform fair with simple categories like negative, neutral and positive with an average accuracy of 85% [2].

Artificial Neural Networks (ANN) and underlying theory has been around since the mid 1940’s. Critical research breakthroughs in ANN during the 80’s and 90’s laid the foundation for a renaissance in machine learning and AI within the last two decades [3]. Yann LeCun developed LeNet-5, a seminal deep learning model, that sought to solve a problem for USPS in the 1980’s. The model used image processing to recognize handwritten numbers and automatically organize letters by their zip code [4] [3]. LeNet-5 and MNIST, the famous handwritten digits dataset, are considered staples in the deep learning communities and often used as a “hello world” introduction for students.

Although Convolutional Neural Networks (CNN) have been around since the 1980s, modern computing power and vast amounts of data have brought new life to CNN based architectures. CNNs are flexible models in that they perform well in many domains of interest to modern research, such as natural language processing, image recognition, voice recognition and other context-based learning [3]. Aside from pure excitement around CNNs, their performance in natural language processing made them a prime candidate for this project. Since the foundational 80’s and 90’s, many different models have been developed for different purposes. Alfredo Canziani, Eugenio Culurciello and Adam Paszke compared various deep learning architectures with the ImageNet classification challenge. The goal of each model was to maximize accuracy in a multi label classification problem of a collection of images. Their results are shown in Figure 1 [5]. The size of the x axis represents how many operations the model must go through for a single forward pass through the network. This is indicative of network depth but may not always correspond. The size of the bubble represents the number of parameters generated by the model. Deeper models (with more hidden layers) generally result in more parameters, and a longer training time.

|  |
| --- |
| *Figure 1. Top Accuracy Results Comparison for Various Deep Learning Model Architecture. The x axis plots the number of operations required for a single forward pass through the network. The size of the bubble represents the number of network parameters [5].* |

## Data Exploration

Two main sources of data were used: Kaggle and Twitter.

**Twitter**: The data was collected using Twitter API by the Piper team and it was filtered around the topic of American Jobs Act and the Infrastructure Bill proposed by the Biden administration. The data was filtered around specific topics; thus, the user pools, sentiments, and general topics of the data are biased toward the query parameters and do not represent the general Twitter user-pool. Three datasets were provided for the project:

Small – 47,000 rows, used for developing and testing

Medium – 923,000 k rows, used to test and validate algorithms

Large – 93 million rows, used to perform final analysis

Each tweet was cleaned to match the training and testing dataset and improve classification. Some of the text cleaning may have removed critical emotional content (lowercases, punctuations) in order to have a comparable set of text among all datasets used in the training, testing, and benchmarking. The cleaning steps followed the outlinedprocedures:

    1. Replace emojis (Unicode) with words. ex: “\U0001F60A” -> [smiling](https://codepoints.net/search?na=SMILING%20FACE%20WITH%20OPEN%20MOUTH%20AND%20SMILING%20EYES) face with smiling eyes

    2. Split joined words. ex: SpamEmail -> spam email

    3. Make all words lowercase. ex: HAPPY -> happy

    3. Remove hash sign from hashtags ex: #happy -> happy

    4. Remove space \n and \r characters.

    5. Remove mentions. ex: @michael published a book -> published a book

    6. Remove websites.

    7. Remove numbers and punctuations.

    8. Remove non-English words.

    9. Remove single letters and empty spaces.

    10. Lemmatize words.

**Kaggle Datasets with Emotional Labels:** Kaggle datasets were chosen based on their Kaggle usability scores, content, data acquisition methods and traceability.

*Emotions dataset for NLP*.

This dataset was organized for a study of emotional content in text and effect of context on emotional recognition. The dataset and the study can be traced to several authors and a conference where the study was presented [6]. Kaggle usability score for this dataset was a 10 out of 10. The dataset consists of sentences and a corresponding sentiment. For example, one row in the dataset looks like this:

I have the feeling she was amused and delighted; joy

The data was used to train a collection of deep learning models to recognize sentiment from a short passage of text. Sentiment distribution in the data is skewed and biased toward emotions joy and sadness as shown in Figure 2. The emotion of surprise is so under-represented, that most trained models managed to optimize on the accuracy score by simply guessing that a sentence is either joy or sadness instead of surprise. Sampling the dataset with replacement didn’t improve the results. To properly train a deep learning model, a bigger set of data with better sentiment representation can make a significant difference. Text data with multiple emotion labels was difficult to find; a dedicated effort to compose such a dataset is warranted for an actionable text classification model.

|  |
| --- |
| *Figure 2. Frequency Distribution of Sentiment Labels in the Training and Testing Datasets.* |

*Emotions Sensor Data Set.*

The second dataset from Kaggle was used mostly for benchmarking. The dataset has a Kaggle usability score of 8.2 and consists of words and a probability mapping to a set of emotions. A sample of rows from the dataset are shown in Figure 3.

|  |
| --- |
| *Figure 3. Sample Rows from Kaggle Dataset Used for Benchmarking* |

The benchmark dataset has some similar sentiments as the main dataset (surprise, anger, fear), a few different emotions (neutral, love) and several that can be mapped between the two datasets (joy: happy, sadness: sad). The distribution is not even, as shown in Figure 4, however the distribution is consistent with the training dataset and is deemed sufficient for benchmarking purposes.

|  |
| --- |
| *Figure 4. Frequency Plot of the Sentiment Labels from the Kaggle Benchmark Dataset.* |

# Analysis Report

## Models

Several models were trained with the Kaggle dataset, model architectures are included in the appendix:

1. CNN-1: architecture with a single convolution – max pooling iteration.
2. CNN-2: architecture with two iterations of convolution – max pooling
3. CNN-4: architecture with four convolution layers in the following architecture.

Each model was performed with a combination of the following layers.

**Input Layer**: initial layer and entry point into the model. The layer accepts a numpy array of text and transforms the array into a Keras tensor object.

**Vectorizer Layer:** vectorizer layer takes the text tensor object,and transforms text data into integers based on a vocabulary map. It’s the feature extraction step of the model. For example, if the vocabulary is {“day”:0, “rocks'':1, “beautiful”:2, “sun”:3, “shine”:4, “river”:5, “kayaking”:6}, then the sentence “The sun is shining, it’s a beautiful day for river kayaking” will be vectorized to [3,4,2,0,5,6]. The vectorizer step can also hold 1 or 0 if a word at index i is present in a certain sentence, in which case the vector will be [1,0,1,1,1,1,1]. Another two common examples for vectorizations are TFIDF and one hot encoding.

**Embedding Layer:** embedding layers optimize deep learning by reducing the sparse vectors produced by the vectorizer layer into low dimension vectors. If n is the number of words in a vocabulary, the input vector is then 1xn in size. The embedding layer transforms the sparse vector to be represented by a predefined k dimension (where k is much smaller than n). An example is included in the appendix with a more detailed explanation.

**Convolution Layers:** convolution layers allow the model not to focus on single words or full text, but the layer allows for different maps to search for different word combinations and treat them as features. Figure 5 shows how a traveling window of size dxm captures combinations of words into feature maps. Feature maps are smaller matrices with varying weights that slide along the word embedding matrix and transform the embedding into feature maps. These weights are part of the mechanism by which the model learns [7].

|  |
| --- |

*Figure 5. CNN Model Convolution and Embedding Visual Representation [7].*

**Pooling**: pooling occurs when a map from the convolution step is “pooled” or all values in the map are either averaged, or the maximum value is extracted, to form a new vector of 1xn, where n is the number of words in the sentence [3].

**Dense Layers**: are fully connected artificial neuron layers that simply perform a dot product of the input vector with a weight vector plus bias. Dense layers can also transform outputs with an activation function (tanh, reLu, sigmoid). Softmax output layer, is a dense layer with a softmax activation function. Examples of activation functions are included in the appendix[7].

**Regularization Layers**: deep learning models have tendency to overfit training data. Regularization layers help reduce overfitting. Two types of regularizations are used: dropouts and batch normalization. Dropout is a method to randomly “drop” certain nodes by setting their softmax values to zero. The dropouts force the model to develop better feature paths through the network rather than rely on individual nodes. Batch normalization transforms the original vector to a vector that has a mean close to zero and a max close to one.

**Softmax Output Layer**: softmax output layer is a dense layer with a softmax activation function. The output layer is a fully connected layer that uses a softmax activation function to transform a vector of 1xn to 1xp where p is the number of desired outputs or number of labels. The sum of the 1xp vector equals one, thus the output can be understood as the probability that a certain sentence belongs to any one of the p labels.

The hyperparameters for each layer were optimized using Keras Hyperband class from which the best model was extracted and used to retrain on full data with best hyperparameters. Reference the appendix for a sample of hyperparameter values tried during the model optimization step.

## Benchmarks

### SpaCy

SpaCy library was used to predict positive, negative and neutral polarity scores from the twitter dataset. To compare model prediction to spacy prediction, each sentiment value was mapped to one of the spacy sentiments:

"pos":["happy","surprise","joy"]

"neg":["fear","disgust","anger","sadness"]

 "neu":["neutral"]

### Vector Derived Labels

The labels derived from Kaggle Emotions Sensor Data Set followed vector multiplication steps. Please reference the appendix for an example of how twitter emotion labels were extracted with the word sentiment vectors.

## Results

### Model Overfitting

The main theme of the observed performance is that all CNN models overfit the data. During training, most of the models had 95% accuracy on the validation dataset, however during testing the majority model performance was anywhere between 37%-60% precision score specifically for sentiment joy - the most common sentiment – but much lower for other labels. The best model performance with the least amount of overfitting was achieved by the simplest CNN-1 models. Several different runs returned consistent results with a range of precision score between 60% for sentiment surprise to 94% for sentiment joy. Review Figure 6 and Table 1 for full results.

### Bias

The training dataset label distribution, reference Figure 2, is highly skewed toward sentiments of sadness and joy. During optimization steps, most models that were tested, zeroed in on local accuracy optimums and tended to overpredict labels of joy and sadness. Surprise, the least represented sentiment, was hardly predicted by most model iterations. Trial runs with evenly distributed data were performed by sampling the dataset with replacements, however the results in performance were similar.

### Result Comparisons

The best model performance was achieved with a simple CNN-1 architecture, the results are presented in Figure 6. CNN-1 was tested in two different configurations, one with a single dropout layer and another model with two dropout layers. Just by adding a dropout layer, the model performance improved considerably. Although CNN-1 had significantly better results than other models, the same pattern of misclassifying the least represented labels – surprise and love - was observed. The Normalized distribution of misclassified sentiments for CNN-1s best performing model is included in the appendix.

|  |  |
| --- | --- |
| *Figure 6. CNN-1 Best Model Performance Results.* | |

Final results show just how much neural networks tend to overfit data, with additional hidden layers model performance degrades. The takeaway is that additional regularization steps must be included with each additional layer. The precision scores in *Table 1* show how often the model predicted the label as correct. For example, from the total times the model predicted the sentiment anger, 91% of the time the predictions were correct. Recall score can be understood as follows, from all the times the model predicted the sentiment fear, 86% of the time it was correct. Recall score gives additional insight into how likely it is for a label to be assigned to another label. For example, how likely is that anger will be assigned as either fear, joy, love, sadness or surprise. Anger was mislabeled 14% of the time. Surprise was mislabeled as other categories the most. Love was least accurately predicted by the model.

*Table 1. Final Results and Model Comparison Scores for Precision and Recall*

| Model Test Performance (%) | Anger | | | Fear | | Joy | | Love | | | Sadness | | | Surprise | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pre | Rec | Pre | | Rec | Pre | Rec | Pre | Rec | Pre | | Rec | Pre | | Rec |
| CNN-1 one dropout layer | 87 | 87 | 83 | | 83 | 88 | 88 | 79 | 79 | 93 | | 93 | 65 | | 65 |
| CNN-1 two dropout layers | 91 | 86 | 84 | | 92 | 94 | 90 | 73 | 92 | 92 | | 94 | 88 | | 55 |
| CNN-2 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 | 29 | | 1 | 0 | | 0 |
| CNN-4 | 29 | 24 | 18 | | 20 | 54 | 62 | 19 | 11 | 42 | | 47 | 0 | | 0 |

Multiple regularization layers were not explored in depth for CNN-2 and CNN-4 models in this report. Additional optimization steps in reducing model overfitting may yield interesting results with perhaps better performance.

### Benchmark Comparison

Benchmark comparison plots are included in the appendix for the best performing model. The benchmark models have neural sentiments, while the trained model only includes positive and negative sentiments. Trained model tends to classify tweets predicted to be neutral by SpaCy as positive. The model diverges from the benchmarks in the negative ratings.

# Ethics Report

## Ethics Discussion

Before exploring machine learning (ML) and artificial intelligence (AI) ethics, we will clarify why this topic is important. What are ethical dilemmas with AI becoming ubiquitous? Let’s review a few AI and ML related incidents reported by news agencies.

Hill Kashmir published an article in several major newspapers, such as New York Times, Seattle Times, and Chicago Tribute, that tells the story of an African American man who was arrested by police after facial recognition software falsely identified him as a wanted thief. The article points out that the facial recognition models used by the police misidentify people of color at higher rates than white people [8].

Samuel Gibbs published an article in the Guardian in 2015 citing a study that google search results and ad targeting are biased. The study revealed that given male and female job seeker profiles with similar credentials, 1,852 men receive advertising for high-paying executive jobs, while only 318 women receive the same lucrative advertising [9].

An article published in ProPublica reported on US courts' use of algorithms to help sentence cases. Yes, algorithms are used in the judicial system. In the article titled *Machine Bias*, ProPublica exposed a case where a judge charged a black girl, who had a prior misdemeanor as a juvenile, with theft and assigned a fee of $80 and jail time for a minor incident. The decision was guided by an algorithm that assigned the girl a high “future risk” score. While a 42-year-old white male, with prior criminal record and prison time had a similar case, but the algorithm assigned the man a low “future risk” score. Two years later, the white man, with a low risk score, was back in jail on an 8-year sentence, while the black girl, with a high risk score, was staying out of trouble. [10]

The articles above should bring awareness to how integrated AI is in everyday life and the real impact it can have on people. There is no debate that AI will continue to play a major role in society; the conversation mostly focuses on how we can use AI and ML to improve the human condition.

The quick pace of AI integration may instill fear and trepidation into some people, especially after reading the incidents presented above. It’s important to remember also that algorithms are designed by humans, the bias is in the data and in humans. Contrary to our sense of control, federating decision making to an algorithm may improve fairness and reduce bias in our society. When you think about it, human bias is difficult to change, it takes years or maybe even generations. Algorithm bias can be scrutinized by different parties, without the algorithm getting defensive. The next day, a data scientist can get a team together, update and normalize the data, update the model, and redeploy the algorithm. The flexibility to iterate algorithms has a promising outlook for ML and AI’s role in our society. The impetus is on us to define what is important to us and what roles and functions we want to delegate to AI.

It’s not whether AI systems alone produce questionable ethical outcomes, it’s whether the AI systems are less biased than the current human systems and their known biases [11].

AI development must undergo rigorous ethical scrutiny: what value can we gain from integrating AI into our society; what are the risks associated with ML and AI; and a reflection on our ethics, what is important to us, so we can define what ethics we want to reflect in our AI and ML systems. [12] AI and ML ethics are in a sense “a wild west” of ethics, where each organization and government body is catching up with the ethical implications of ML technologies. To fill the ethics vacuum, most organizations are left to define their own AI ethics and principles. Despite the maze of ethics guidelines, most resources reviewed for this report converged on several key points defined below: [13] [14] [15] [16] [17]

1. *Motivation*: define a clear purpose and need for the model.
2. *Data*: the old trope “garbage in, garbage out”.
3. *Stakeholders*: consider the human element and all stakeholders, protect their rights and privacy.
4. *Fairness*: minimize bias, strive for fairness.
5. *Transparency*: strive to reduce complexity, build systems that are transparent, and explainable.
6. *Flexibility*: build models that are flexible for improvements and updates, allow room for debate and conversations.
7. *Accountability*: maintain a clear chain of ownership and accountability.

## Motivation

The model was developed for the capstone project, the final class for a Masters of Analytics at Georgia Tech. Piper Gradient, is the sponsor for this project. The main direction was guided by the Pipe Gradient Objectives. The Pipe Gradient team is developing a tool called Gradient, which will ingest social media posts and process the text in several ways: the initial step is to perform clustered analysis on the data dividing the posts into clusters representative of the real world groups or categories; each cluster will undergo further analysis on the main topics, text summaries, semantics, and sentiment analysis.

This specific project attempts at adding a layer of emotional context to tweets that can be applied while reviewing topics or text summaries and their general emotional mapping, clusters emotional mapping, or dedicated sentiment review of tweets.

For example, given an initial tweet by a public figure, one can follow each retweet and reply to the original message and monitor the emotional evolution of the thread as different people add to the conversation. Imagine a tree with each branch representing a sub thread and each having its own color scheme representing different emotions. A similar type of sentiment analysis can be performed on topics, and what topics are preferred by different groups… etc.

Sentiment analysis provides deep insight and productive utility to data scientists. Bad actors can also take advantage of the insight and publish content that rates to be divisive and polarizing. Strictly profit driven corporations can also abuse the tool by crafting their marketing campaign to appeal to more users. Such a use of a sentiment analysis will be unethical if the company’s products and services diverge from the marketing appeal and promises. Furthermore, foreign and domestic bad actors, bots or otherwise, can also abuse the insight from sentiment analysis to create content that can spread faster. All the mentioned risks are more general to sentiment analysis, which is very prevalent in less sophisticated packages. For example, numerous pretrained libraries already exist in most NLP packages (NLTK, SpaCy, PyTorch, TensorFlow) and will label text as positive, neutral or negative with minimal additional programing. Machine learning has a democratic direction where all parties have access to powerful ML tools. The sentiment analysis models for this project are not in deployable packages and are experimentative in nature. The models do not sway the ubiquity of natural language processing, however ethical implementation and improvements of the sentiment analysis tool must consider cases of misuse and proper remediation. The Piper Gradient team may or may not decide to reuse the models or the code in Gradient development. Ideally, the Piper team will establish clear channels of communication with social media platforms and their users in order to inform involved parties of any malpractice or misuse of their tool.

## Data

Two main sources of data were used: Kaggle and Twitter. *Emotions dataset for NLP* dataset from Kaggle was used to train the CNN models, while the second *Emotions Sensor Data Set* was used as a benchmark. Both datasets have an uneven distribution of sentiment labels as shown in Figure 2 and Figure 3. Twitter data was used to classify and review results of the trained model. Additionally, Twitter data was used to build conversation trees, tweets and tweet replies, and map emotional labels on those tweets to review how emotions change with each reply. An in-depth review of the data is covered in the Data Exploration section.

## Stakeholders

Twitter

Twitter the company, and Twitter users are participants of this study and must be considered as stakeholders. Twitter has a specific policy that governs how developers should use Twitter data. Since the dataset is used for a school project, *Commercial Use Restrictions* are not applicable [18]. If the Piper team decides to pursue commercial interests with any of the models trained on the Twitter data, this section of the agreement must be revisited. Section III part A and B request that developers update twitter datasets in order to comply with twitter policy. If a user decides to remove their public tweets, twitter requires that the datasets used by third parties are also updated to reflect the changes [18]. The length of the semester, and the training nature of the models make daily updates impractical with limited educational resources. If the Piper team deploys a model to production, a pipeline for regular data and model updates will ensure compliance with Twitters policy and improve model integrity.

*Data Science Practitioners*

The data science practitioner, in this case the author, intends to clearly document the process and methods for developing the deep learning models in this report. The goal of the author is to leave a traceable path for reproducibility and transparency. The author is not a subject matter expert, but a student practitioner, and developed the deep learning models as part of a capstone project.

Piper Gradient Team

Piper Gradient team holds multiple roles as stakeholders. The Piper team are part developer owners because they are the sponsors and acted as instructors throughout the semester. The Piper team influenced the direction of the project, area of focus, provided datasets for model development, and had a commentary period earlier in the semester. The Piper team are also decision makers and ultimately the owners of any version of the project they choose to implement into the Gradient tool. The guidance provided by the Piper team is responsible for such an in-depth ethics review. Resources provided by the Piper team throughout the semester fostered the development of a nuanced perspective on AI ethics.

Regulators

The US government has many regulations in place for privacy, property and free speech protections. The application of regulations to social media is in a constant flux as the regulatory agencies decide the best way to protect user privacy online. Sentiment analysis model developed for this project may need to be retrained with a different set of data in the future as regulations change.

The Twitter Legal Team may also update policies regarding the use of Twitter data. The updates may be a response to legal policy updates or may be independent. If Twitter's legal team reviews the sentiment analysis and decides that the model infers too much psychological insight from user tweets, Twitter may decide that the model is in violation of Twitter developer policy [18]. The proactive approach will be to request commentary fromwitter before the model is deployed to production.

## Fairness

The dataset was filtered around hashtags related to the Biden Administration’s American Jobs Plan. Twitter users that participate in the specific topics can be assumed to make up the demographic for the collected data, which at the least meet the following requirements: must have a twitter account and are likely to be politically engaged. The initial criteria alone segments a large set of groups out of the analysis. For example, a busy working parent may not have the time to engage in politics on Twitter. Pew Research Center reported that lower income families are less likely to use Twitter [19]. Furthermore, the topic reduces the diversity pool to the most politically engaged twitter users, as the topics around American Jobs Plan are current and hotly debated in politics. As Figure 7 suggests, there are subtle demographic differences in people who use Twitter. Similar differences can be expected for different topics and hashtags on Twitter [19].

The deep learning models were trained with an independent dataset, composed from a collection of tweets filtered around certain hashtags (339 in total) which targeted references of emotions in the hashtag themselves. The research team that organized the dataset did not present further insight into the general groups of users the final dataset included, which can be difficult to infer. The hashtags cover a wide range of expressions, for example #worried or #eww, and the sheer number of hashtags used to collect the data reduces bias of the dataset overall.

Using Twitter data for emotional mapping of text does have its limitation, it excludes non-twitter users, which fall along demographic or socio-economic lines. For example, Pew Research reported that the top 10% of most engaged tweeters produce 80% of tweets and are mostly younger, educated and democratic leaning people engaged in political conversations. Women make up a slight majority of the top 10%, while are in a slight minority when considering all Twitter users. [19] Additional demographic mapping is presented in Figure 7.

|  |
| --- |
| *Figure 7. Twitter User Demographics as Reported by Pew Research Center* |

## Transparency

The main motivation for transparency is to provide a reproducible set of results, with thorough documentation of data acquisition and methods. Another critical aspect of transparency is model interpretability and expandability. Deep learning models are notoriously hard to explain. The author attempts at documenting model architecture and provides several benchmark comparisons to improve explainability.

## Flexibility

The model was trained in a google colab notebook organized by sections and with ample documentation. Models are organized so that they can easily be loaded and retrained with new data. The neural network architecture can also be easily updated within the colab notebooks. All the necessary code, from data import and cleaning, to model training, testing and benchmarking are included in a single notebook.

## Accountability: Model Life Cycle

The initial model training, testing, and benchmark comparison are prepared by the author and documented in this report. The Piper Gradient team will own the deployment and maintainability of any reused code from this project. The author doesn’t have any immediate plans to use the models trained for this project for any production implementations. The author may continue to use the code from this project and the models developed for educational purposes.

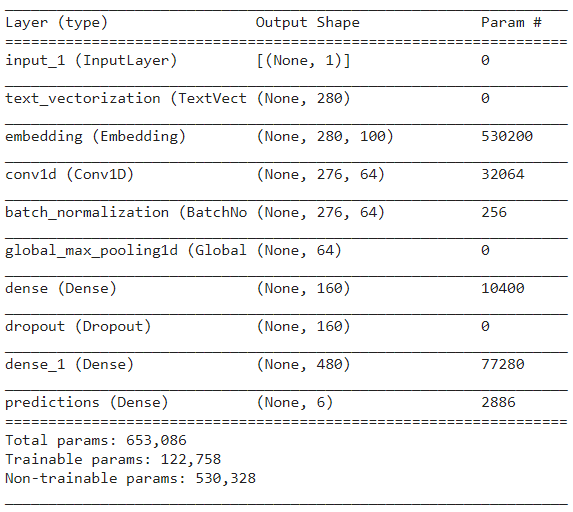
# Conclusion

Works Cited

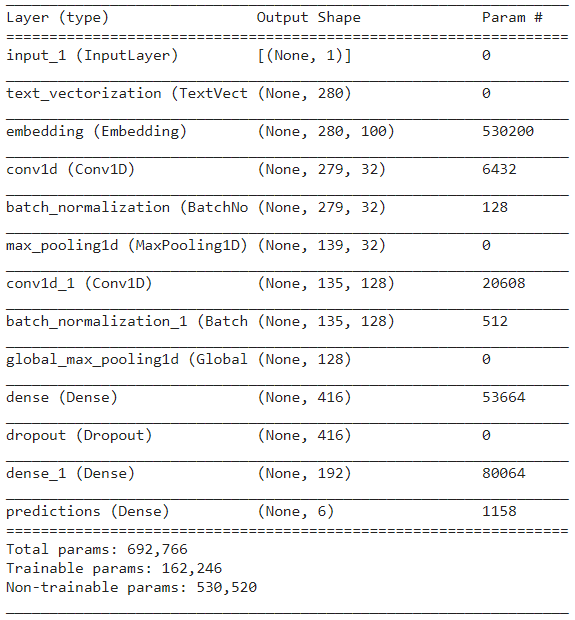
| [1] | J. B. E.Y. BANN, "The Conceptualisation of Emotion Qualia: Semantic Clustering of Emotional Tweets," in *Proceedings of the Thirteenth Neural Computation and Psychology Workshop*, University of Bath, 2013. |
| --- | --- |
| [2] | K. G. Jeffery, "Technical Infrastructure and Policy Framework for Maximising the," in *Conference on Electronic Publishing*, Vienna, 2007. |
| [3] | A. Géron, "Introduction to Artificial Neural Networks, Convolutional Neural Networks," in *Neural networks and deep learning*, Boston, O'REILLY, 2021. |
| [4] | S. G. a. M. K. Anirudh Koul, "Exploring the Landscape of Artificial Intelligence," in *Practical Deep Learning for Cloud, Mobile, and Edge*, Sebastopol, O’Reilly Media, Inc, 202. |
| [5] | A. P. E. C. Alfredo Canziani, "An Analysis of Deep Neural Network Models for Practical Applications," *arXiv,* 14 April 2017. |
| [6] | H.-C. T. L. Y.-H. H. J. W. Y.-S. C. Elvis Saravia, "CARER: Contextualized Affect Representations for Emotion Recognition," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* p. 3687–3697, 4 November 2018. |
| [7] | Z. Z. Yuling Chen, "Research on text sentiment analysis," in *2018 13th IEEE Conference on Industrial Electronics and Applications*, 2018. |
| [8] | K. Hill, "Wrongfully Accused by an Algorithm," 24 June 2020. [Online]. Available: https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html. [Accessed 7 24 2021]. |
| [9] | S. Gibbs, "Women less likely to be shown ads for high-paid jobs on Google, study shows," 08 July 2015. [Online]. Available: https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study. [Accessed 07 24 2021]. |
| [10] | J. L. S. M. L. K. P. Julia Angwin, "Machine Bias," 23 May 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. [Accessed 24 July 2021]. |
| [11] | J. A. E. A. V. LEE RAINIE, "Experts Doubt Ethical AI Design Will Be Broadly Adopted as the Norm Within the Next Decade," 16 June 2021. [Online]. Available: https://www.pewresearch.org/internet/2021/06/16/experts-doubt-ethical-ai-design-will-be-broadly-adopted-as-the-norm-within-the-next-decade/. |
| [12] | L. C. J. B. M. C. R. C. P. D. V. L. C. M. R. P. U. R. F. S. B. V. P. &. V. E. Floridi, "AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and machines, 28(4), 689–707.," *Minds and Machines,* vol. 28, no. 4, p. 689–707, 2018. |
| [13] | C. M. U. S. E. INSTITUTE, "Designing Ethical AI Experiences: Checklist and Agreement," [Online]. Available: https://resources.sei.cmu.edu/asset\_files/FactSheet/2019\_010\_001\_636622.pdf. [Accessed 24 7 2021]. |
| [14] | A. D. o. Industry, "Australia’s AI Ethics Principles," [Online]. Available: https://www.industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles. [Accessed 24 July 2021]. |
| [15] | U. I. Community, "PRINCIPLES OF ARTIFICIAL INTELLIGENCE ETHICS FOR THE INTELLIGENCE COMMUNITY," 23 July 2020. [Online]. Available: https://www.intelligence.gov/principles-of-artificial-intelligence-ethics-for-the-intelligence-community. [Accessed 24 July 2021]. |
| [16] | R. N. M. S. S. E. E. J. B. S. M. Ziad Obermeyer, "Algorithmic Bias," Chicago Booth The Center of Applied Artificial Intelligence, Chicago, 2021. |
| [17] | L. Hogenhout, "A Framework for Ethical AI at the United Nations," UN Office for Information and Communications Technology, 2021-15-03. |
| [18] | Twitter, "Developer Agreement and Policy," [Online]. Available: https://developer.twitter.com/en/developer-terms/agreement-and-policy. [Accessed 24 July 2021]. |
| [19] | A. H. STEFAN WOJCIK, "Sizing Up Twitter Users," 24 April 2019. [Online]. Available: https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/. [Accessed 24 July 2021 ]. |

# Appendix

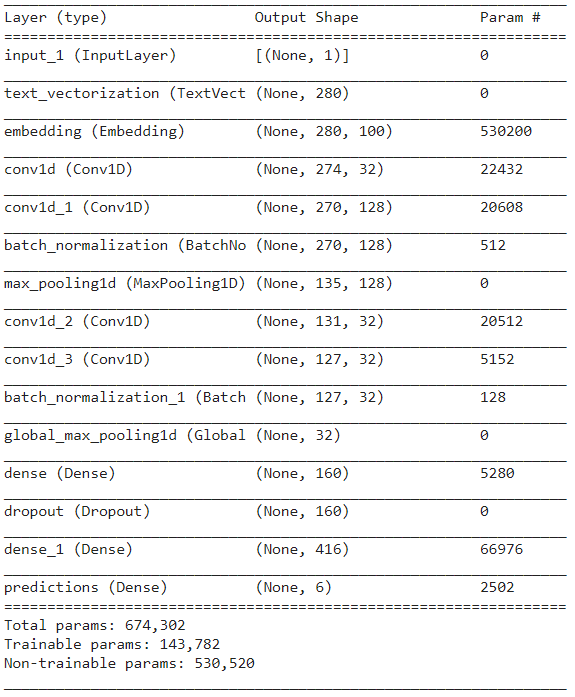
## CNN -1



## CNN-2



## CNN-4



## Embedding layer example

For example, while the word “kayak” will be represented by the following vector from the vectorization step [0,0,0,0,0,0,1], the embedded version may just be in two dimensions of bool values [2.345, 1.332]. The embeddings can either be learned during model training, or a pretrained embedding can be added to the model. Embedding vectors have shown to capture syntactic and contextual relations of words. For example, the vector for word “man” is closer to the vector for word “kind” than to the vector for word “woman”. Besides capturing word similarity in the vector space, embedding layers also improves deep learning convergence since neural network models perform poorly with sparse matrices. GloVe pretrained embedding libraries developed by Stanford University include an embedding trained on a twitter dataset.

## Vector Derived Labels

For example, the table below is a snippet of the dataset with each sentiment value mapped to a word. Values of 0 represent that the word is missing from the dataset.

| Word | Disgust | Surprise | Neutral | Anger | Sad | Happy | fear |
| --- | --- | --- | --- | --- | --- | --- | --- |
| The | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| bird | 0.011 | 0.0207 | 0.002 | 0.016 | 0.0207 | 0.0622 | 0.0207 |
| is | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Beautiful | 0 | 0.014 | 0 | 0.006 | 0.003 | 0.1129 | 0.005 |

The vocabulary is extracted from the dataset, an in our case it’s simply [The, bird, is, beautiful]. We will now take two sentences and calculate their “happy score” as an example. The final score is extracted by getting the sentiment with the highest score.

Sentence: The sentence: The bird is beautiful.

The result from the above vector matrix multiplication is the vector below. The sentiment with the highest weight is assigned as a single label to the sentence, in this case the label is happy.

sentiment vector =

sentiment value = max(sentiment vector) = 0.175 = Happy

The result from the above vector matrix multiplication is the vector below. The sentiment with the highest weight is assigned as a single label to the sentence, in this case the label is happy.

Let’s look at another example sentence.

Sentence: What a beautiful day.

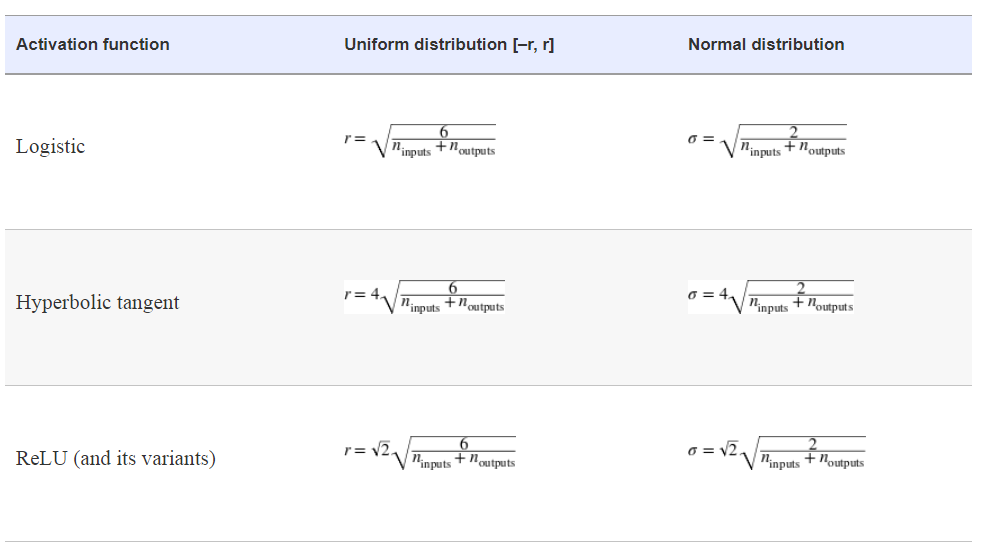
sentiment vector =

sentiment value = max(sentiment vector) = 0.1129 = Happy

The same steps were performed on the twitter dataset to extract a single label for each tweet.

## Sample Activation Functions Used in Deep Learning

[3]



| Hyperparameter Optimization Mesh Plot of Training Iteration with Different Combinations of Hyperparameters. |
| --- |
|  |

| CNN-4 Model Training and Validation Plots |
| --- |
|  |

| CNN-1 Best Performing Model Label Distribution with Most Misclassified Normalized Labels. |
| --- |
|  |
| Misclassified Labels Frequency Plot |
|  |

| Benchmark Review |
| --- |
| Vector sentiment derived labels as True Labels, and model predicted labels as predicted. |
| SpaCy benchmark as true labels, and model predictions as predicted labels. |

| Twitter samples with their bench mark predictions and model predictions. |
| --- |
|  |