Using Data Augmentation to Improve the Classification of Moral Outrage

Mark Torres
[1]Department of Statistics and Data Science
Yale University
New Haven, CT
mark.torres@yale.edu

William J. Brady
Department of Psychology, The Crockett Lab
Yale University
Postdoctoral Fellow, Supervisor
william.brady@yale.edu

Molly J. Crockett
Department of Psychology, The Crockett Lab
Yale University
Supervisor
molly.crockett@yale.edu

Abstract

Moral outrage is an emotion that results from perceived transgressions of moral norms. In the era of social media, the expression of moral outrage has been amplified with implications for spread of information across various social and ideological groups. Past work has attempted to design empirically-validated tools in order to study the spread of moral outrage on these platforms. However, these methods- like other supervised deep learning classification algorithms - are limited by the availability of labelled training samples. In this thesis, I explored several methods from the domain of data augmentation and semi-supervised learning that can artificially increase a training sample without the costly process of seeking more hand-labelled data. We found that utilizing a pre-trained model to label new data and then re-training a classifier on the larger set achieved up to a 92.55% accuracy on an out-of-sample test set. This accuracy represented a 26.71% improvement over the original classifier performance. I conclude with applications of our newly augmented classifier for testing psychological hypotheses on social media regarding how moral outrage is propagated over social media outlets.

[1] This paper, in collaboration with the Crockett Lab, serves as a senior thesis as part of the Yale Statistics and Data Science requirements.

1 Introduction

Moral outrage is an emotion elicited by stimuli that a person may perceive as transgressing against a moral norm. Our exposure to events that trigger moral outrage (e.g., the plight of immigrants, stories of atrocities in foreign countries, marches for women's rights) has increased with the advent of social media platforms that expand our access to current events anywhere on Earth.

Social scientists have become interested in further understanding how moral outrage is elicited, spread, and acted upon, in the context of these modern digital platforms where the costs of expressing outrage are greatly reduced (e.g., Crockett, 2017; Brady & Crockett, 2019). For example, the increased visibility of events that elicit moral outrage can promote a "common sense" morality, in which certain events are clearly considered unjust because a large group of people on social media deem it as unjust (Shteynberg, 2018; Spring et al., 2018).

Because these platforms provide a new venue to answer psychologically relevant questions, social scientists have developed tools for measuring emotional expressions such as moral outrage, through the lens of social media messages. (e.g., Weber et al., 2018; Brady et al., under review)

This thesis builds on previous work to develop a moral outrage classifier for social media and aims to (1) propose methods that can improve the accuracy and generalizability of such tools and (2) to apply the improved tools to test psychological hypotheses relevant to the domain of moral outrage.

2 Literature Review

In the domain of social science, the availability of computational tools has already greatly expanded the types of questions that can be answered (Lazer et al., 2009). These tools allow social scientists to expand beyond the one-on-one interactions and survey-style methods that previously built the bedrock of established research methodology (Lazer et al., 2009). Now, social scientists have access to a wider breadth of data points that provide them access to people's behaviors outside of a laboratory context and at a scale that is impossible to recreate through past participant recruitment methods. For example, one tool now available to social scientists is the social media platform Twitter. It is a useful medium for those interested in, among other topics, studying wide-scale social behaviors (e.g., Leitch & Sherif, 2017; Hassan et al., 2013) and examining the dissemination of information (e.g., Crooks et al., 2012; Kwon et al,. 2014; McCormick et al., 2017).

These new computational tools and resources can be invaluable for the further study of moral outrage on social media platforms. For example, Crockett (2017) proposed a theoretical framework for understanding how moral outrage is propagated on social media. This work has been complemented with research performed by Sawaoka and Monin (2018), who found that behaviors which might be praised in isolation are instead seen as bullying when propagated in a group setting on social media. These findings suggest that social media is an invaluable platform for understanding how moral outrage is manifested. In light of this, Brady et al. (under review) developed a deep learning classifier trained specifically for the task of studying moral outrage in social media posts. These developments

are the start of what will continue to be a promising area of research in the social sciences that studies the intersection of human behaviors with the advent of new technological platforms.

In these studies, deep learning is a method often used to take advantage of the availability of large quantities of data in order to further research pursuits. Deep learning is part of a family of machine learning methods that apply neural networks in a variety of research settings, ranging from speech recognition to text analysis to image classification (Deng et al., 2013). Among many other applications, deep learning has been useful in sentiment analysis, where models attempt to understand the “emotion” of a certain text, because of their ability to ascertain underlying features or combinations of features (e.g., Severyn & Moschitti, 2015; Tang et al., 2015; Zhang et al., 2017). Brady et al. (under review) has extended this research by demonstrating that deep learning can be particularly useful for classifying moral outrage expression in social media text.

A current limiting factor in the domain of deep learning is the availability of pre-labelled training data. Deep learning methods require large quantities of high-quality training data, and the process of obtaining this data is often expensive and complex (Tan et al., 2018). For instance, Brady et al. (under review) built their original classifier using 16,000 labeled observations, which suggests there is room for improvement by increasing the training data set size.

One possible solution to resolve the lack of pre-labelled data is to supplement it with unlabelled data. As a result, a great deal of current research is exploring the use of data augmentation in order to supplement a labelled training dataset with synthetically created samples (Wong et al., 2016). A variety of data augmentation methods have been studied, primarily in the domains of image classification and text analysis. In the context of text analysis, current data augmentation methods often operate by either determining linguistic properties of text data through the implementation of deep learning networks (e.g., Hayashi et al., 2018) or by using some other features intrinsic to the training data, such as synonyms of words in the training set (e.g., Wei & Zou, 2019) or their embeddings in high-dimensional space (e.g., Pennington et al., 2014). These methods tend to improve classifier performance due to some mixture of increasing sample size, performing simple transformations, or by creating local perturbations to the sample (Ratner et al., 2017), which both increases the span of the sample space available to the classifier and provides it with additional samples for training.

In the space of moral outrage research, data augmentation can serve to further current efforts. Like in other deep learning application areas, this domain suffers from a lack of labelled training data, while obtaining additional labelled samples at the scale required by deep learning models is expensive. Therefore, the purpose of this project is to leverage several methods in data augmentation and see how they can improve a classifier previously designed to detect moral outrage on social media (Brady et al., under review). We would like to learn whether recent advances in data augmentation research can be transferred to the domain of social science research in order to promote further research on the classification and study of moral outrage on social media.

3 Description of the data

3.1 Training set

Our labelled training set was composed of 16,000 hand-labelled tweets, gathered by the Crockett Lab, and covering the Brett Kavanaugh Supreme Court nomination congressional hearings in 2018. This set of data was chosen because it would provide a sample that was representative of the entire political spectrum, due to the sexual assault claims made against Brett Kavanaugh by Christine Blasey Ford and by Kavanaugh's conservative judicial record and relatively young age potentially changing the idealogical composition of the Supreme Court. The Crockett Lab recruited 640 Academic Prolific workers to rate a random block of 250 tweets to determine if the tweets contained outrage or not. When 7/10 raters judged that a tweet expressed outrage, it was deemed as having outrage. From this method, 52.41% (8,386 out of 16,000) tweets were marked as having outrage.

Classification	**Example Tweets**
Has moral outrage	@SenBlumenthal Tell me how she is being bullied. You are a real piece of work considering your own background
	@SenBlumenthal Don't recommend...DEMAND that they listen to women who are sexually assaulted! So NO to #PERVanaugh
	@KatiePavlich @SenBlumenthal Blumenidiot is a left-wing Socialist hack!
Does not have moral outrage	What would @SenBlumenthal say if numerous women were produced that claimed he paid them for sex ? He would say I did not. HaHa
	@SenBlumenthal thinks the Constitution is a suggestion
	@JeromehartlF @KatiePavlich @SenBlumenthal If he has PTSD, it's from the stress of Toys for Tots, not the Tet Offensive

Table 1: Examples of Kavanaugh tweets, both with outrage and without outrage. Labels for these 16,000 tweets were determined by human labellers.

3.2 Corpus of unlabelled tweets

We supplemented our training set with 1,095,360 unlabeled tweets. These tweets were gathered by the Crockett Lab and spanned a range of politically charged topics selected for their high potential for containing moral outrage expression: Nancy Pelosi when she was being pushed to impeach President Trump (249,774 tweets), the border wall between Mexico and the United States (446,394 tweets), the 2018 March for Our Lives protests regarding gun violence (73,084 tweets), the 2019 admissions bribery scandals revolving around Rick Singer (6,433 tweets), and the Central American migrant caravans approaching the United States in 2019 (319,675 tweets).

We used this corpus of unlabelled tweets during our experimentation with data augmentation methods (explained in greater detail below).

3.3 Generalization set

The performance of the classifier was tested on a set of 10,000 hand-labeled tweets gathered by the Crockett Lab and spanning 4 different topics: the Covington Catholic High School Lincoln Memorial confrontation in 2019, the United States airlines incident of dragging a passenger off a plane in 2017, the Jussie Smollett attack claim in 2019, and President Trump's announcement of the transgender military ban in 2019. As was the case with the training set, when 7/10 raters judged that a tweet expressed outrage, it was deemed as having outrage. From this procedure, 30.51% of tweets (3,051 out of 10,000) were deemed to have outrage. These tweets served as the out-of-sample test set because its diversity in topics allows us to evaluate the generalizability of our classifier.

4 Methodology

4.1 Preprocessing steps

We parsed the tweets in order to transform them into a format conducive for analysis. We followed a procedure that has been empirically and heuristically shown to remove features that would minimally (and in cases, negatively) contribute to sentiment analysis research (Anand et al., 2018).

We first filtered out features such as hashtags (#), @ symbols, URLs, RT (retweet) and CC (carbon copy), white spaces, and punctuation marks. We then removed stopwords, such as articles (e.g., "the", "a"), pronouns (e.g., "my", "mine"), and prepositions (e.g., "of", "under") which are common during English discourse but add little connotation to the message. This improves our classification task by substantially reducing the feature space (Hassan et al., 2013).

Our last preprocessing step aimed to standardize words such as "run", "ran", and "runner", which are different lexical variations with a shared root meaning (that is, "run"). To perform this task, we used a process known as "lemmatization", which uses the WordNet corpus and morphological analysis to reduce words to their "standardized" forms, or "lemmas" (Manning et al., 2008; Miller, 1995).

4.2 Overview of the classifier architecture used

The classifier that we used in this study was a modification of a recurrent neural network (RNN) known as a deep gated recurrent unit (GRU). A deep GRU extends the principle of a RNN, which captures sequential dependencies in a deep learning framework, by providing a method to circumvent the problem of vanishing gradients (Chung et al., 2014). It is similar to an LSTM (Long Short Term Memory) model in many respects, and both outperform typical RNNs in tasks with long-term dependencies (e.g., machine translation tasks), in which vanishing gradients become more

prominent, by implementing gating mechanisms in order to control which pieces of information are maintained and used in the network. Past work (Brady et al., under review) demonstrated that, in the context of classifying moral outrage within a corpus of Twitter data, a deep GRU model outperformed other classifiers such as Random Forest, LSTMs, and Google's bidirectional ELMo model.

4.3 Outline of approach

We outline our research methodology in its entirety in Figure 1 below.

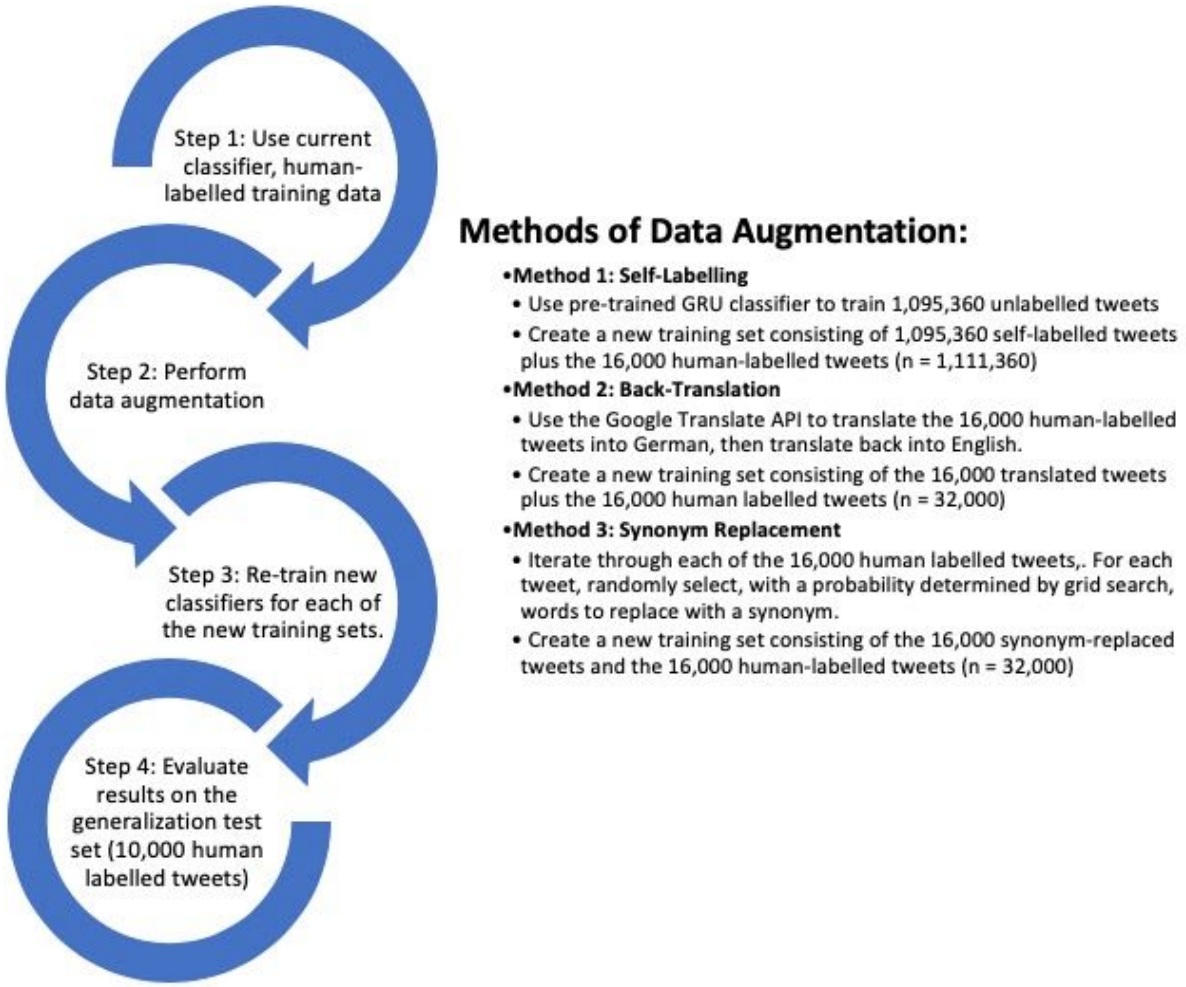

Figure 1: An illustration of our research procedure.

We first performed a preprocessing phase in which we eliminated features of the corpus that contained little semantic value, such as stopwords and punctuation.

After the preprocessing step, we then subjected the training data to several different data augmentation methods (outlined above).

After the data augmentation methods were completed (and any additional preprocessing steps performed), we then re-trained the deep GRU classifier for each of the different methods applied. We then tested the deep GRU classifier's performance on a generalization set of 10,000 labeled tweets regarding a wide array of topics. To evaluate our classifier's performance when trained with each set of augmented training data (and comparing the results to the control with no data augmentation), we recorded each classifier's accuracy, F1 score, precision, recall, and AUC.

Finally, we used the highest-performing classifier to answer a psychologically relevant question regarding how moral outrage propagates through social media networks.

5 Data augmentation methods

5.1 Choosing data augmentation methods

We searched for data augmentation methods that were dissimilar from each other, in order to capture multiple distinct approaches towards the problem of data augmentation. We avoided methods that seemed to rely on the same underlying features of the dataset (e.g., we avoided using multiple methods that relied on word embeddings). We experimented between methods that relied on combining labelled and unlabelled data (in the case of semi-supervised learning) and methods that relied on generating samples based solely on pre-labelled data (in the case of back-translation and synonym replacement).

5.2 Data Augmentation Method 1: Semi-Supervised Learning and Self-Labelling

Semi-supervised learning is an approach that trains a classifier with both a small set of labelled training data and a large set of unlabelled training data. It is currently a promising of deep learning research due to the dearth of high-quality pre-labelled training data and the wide swaths of unlabelled data that exist (Schwenker & Trentin, 2014). The various schools of semi-supervised learning approaches aim to use some information known a priori (e.g., pre-labelled training/test data) and combine it with previously unlabelled data in order to bridge supervised learning approaches, in which a classifier maps the space of input vectors into pre-defined categorical labels (Bishop, 2006), with unsupervised learning, which attempts to capture probabilistic or geometric interpretations of unlabelled data (Jain, 2010).

Self-training is a subset in the domain of semi-supervised learning in which a classifier trained with pre-labelled classifiers is used to label unlabelled data, and then retrained on the combination of pre-labelled and newly labelled data (Chawla & Karakoulas, 2005). Past research suggests that incorporating information from previously unlabelled data can improve the quality of classifiers (Zhang & Oles, 2000). These new self-labelled classifiers have been used in applications such as medicine, where they exceed the performance of traditional supervised learning classifiers (Bull et al., 2018; Bull et al., 2020; Nartey et al., 2020). The improvements in classification performance that come from self-training methods seem to come from both the expansion of the feature space of the training data and the increase in training data size (Nartey et al., 2020).

In the present study, we used a classifier trained on pre-labelled tweets (n = 16,000) and used it to label 1,095,360 previously unlabelled tweets. We then trained a new classifier on a new training data set of the previously labelled tweets and the newly labelled tweets (total n = 1,111,360) and evaluated its performance on the generalization set. Some past implementations of self-supervised classifiers involve iteratively labelling certain chunks of unlabelled data and then adding to the training data the observations for which the classifier had the highest probability prediction for its classification (e.g., Schwenker & Trentin, 2014). This process was repeated until a convergence criterion was met. However, in our implementation, we chose to incorporate the entire training corpus in our dataset because our classifier uses a sigmoid activation function that pushed class label probabilities to approach 0 or 1 (Han & Morag, 1995).

5.3 Data Augmentation Method 2: Back-Translation

Back-translation was another method we used in our data augmentation. In back-translation, we translate the English tweets into another language and then re-translate the tweets back into English.

Previous research indicates that combining monolingual (here, English) training data with back-translated data, in a 1:1 ratio, can outperform baseline metrics (Senrich et al., 2016), syntax-based models (Senrich & Haddow, 2015), and other neural machine translation models (Jean et al., 2015; Luong et al., 2015). These gains were attributed to the new dataset slightly expanding the sample space of the training data, rather than improving the classifier's accuracy from merely increasing the size of the training data (Senrich et al., 2016).

We chose to perform back-translation with the English - German language pair because this language pair has been used in past research to evaluate the efficacy of neural machine translation (NMT) models (Senrich et al., 2016) and has been shown to be robust in large-scale translation projects (Edunov et al., 2018).

For our project, we used the Google Translate API in order to translate our original 16,000 tweet training set into German, translate the newly synthesized tweets back into English, and add them to our original training set, giving us a total of 32,000 tweets.

5.4 Data Augmentation Method 3: Synonym Replacement

We also tested the effectiveness of augmenting a training data set with samples in which certain randomly chosen words have been replaced with a synonym. We follow the procedure outlined by Wei & Zou (2019), in which they demonstrate that synonym replacement improves classifier performance on text classification tasks spanning a wide variety of topics.

Moreover, Wei & Zou (2019) also demonstrate that synonym replacement is particularly amenable to classifier improvement when the size of the training data is small (in their case, they were able to show that the procedure was robust even with n = 500). In our study, with a sample size of 16,000

labelled tweets, this finding suggested that a low sample size would not inhibit the performance of the synonym-augmented classifier.

Additionally, we felt that a synonym-based approach would be appropriate for improving the accuracy of a moral outrage classifier because Wang & Yang (2015) found that a synonym-based data augmentation approach (rooted in examining word embeddings) was effective in classifying and examining annoyed behaviors on Twitter.

To implement our synonym replacement, we iterated through each preprocessed tweet and randomly selected, with probability of 0.4 for GloVe embeddings and p = 0.6 for NLTK WordNet replacements (these probabilities gave the best results in a grid search), whether a given word would be replaced with a synonym. We excluded words from replacement that we determined were proper nouns that would not have appropriate synonyms and were important to the essence of the message (e.g., “Trump” or “United States” or “Congress”).

5.4.a Data Augmentation with Word Embeddings

We considered using embeddings as a means of representing words in higher dimensional space. By doing so, we are able to use either distance or co-occurrence metrics to represent the “similarity” of two words or phrases, which gives us a more dynamic definition of “synonym” than hard-coded methods (Turian et al., 2010).

Most word vector methods rely on the distance or angle between two word vectors for determining their “similarities”. This is the approach adopted by Word2Vec, which is based on word analogies across multiple dimensions of difference (Mikolov et al., 2013). For example, in the Word2Vec implementation, the analogy “France is to Paris as Spain is to Madrid” would be encoded in the vector space by *France - Paris = Spain - Madrid* (Mikolov et al., 2013; Pennington et al., 2014). Word2Vec leverages a neural network to learn the context of a given word by examining the words around it.

Other methods use different ways to measure similarity between words. For example, researchers at Stanford, in 2014 (Pennington et al., 2014) proposed a method called “Global Vectors for Word Representation” (GloVe, for short), that determines similarity between words by factorizing a co-occurrence matrix and leveraging statistical information. In contrast to Word2Vec, which uses local information to determine which words are likely to be found in similar contexts, GloVe embeddings use a global metric, the co-occurrence matrix, in order to determine similarity.

Methods like these have been applied in past social science research. For example, Wang & Yang (2015), in their evaluation of data augmentation methods to analyze annoyed behaviors on Twitter, made use of Word2Vec to train their various lexical and embedding-based methods. Zhao et al. (2019) demonstrate that, in the context of examining gender bias in high-dimensional word vector representations, that GloVe embeddings were less likely to make systematic gender biases than ELMo, a bi-directional LSTM model implemented by Google, even though ELMo was designed to improve upon GloVe embeddings by accounting for word context.

For our purposes, research suggests that the performance of Word2Vec and GloVe is comparable (Levy et al., 2015). Other work suggests that, more generally, the performance of different word embedding methods is more dependent on the corpus domain and composition than the method itself (Lai et al., 2015). Therefore, we used only one of these methods in our current study. We chose to use pre-trained GloVe embeddings because we felt that the results from those embeddings would prove to be more generalizable (since they are derived from global, rather than local, metrics) than Word2Vec. However, we would not be surprised if Word2Vec embeddings were to provide results similar to GloVe, citing past literature (e.g., Levy et al., 2015).

In our current implementation, we used pre-trained GloVe embeddings from Python's "gensim" package. For each tweet, we kept words that we felt were proper nouns that did not have appropriate synonyms (e.g., "United States", "Trump", "Pelosi"). For the rest of the words, we iterated through each word and replaced, with probability = 0.4 (which gave the best results in a grid search), that word with another word that had the highest similarity to it based on pre-trained GloVe embeddings.

5.4.b Data augmentation with NLTK WordNet Lexical Database

As an alternative to a word-embedding based approach, we chose to also explore the use of a hard-coded lexical database of synonyms. Unlike word embedding methods, which are dependent on the nature of the data used to train the embeddings, using a lexical database of hard-coded synonyms better assures that a word is replaced with a term more easily seen as a "synonym" than in the case of GloVe embeddings, in which a word might be replaced with a term that isn't a synonym but rather is found to occur frequently with that word.

For the present study, we chose to use WordNet, a lexical database of semantic relations (i.e., synonyms, antonyms, etc.) that has implementations in 200 languages, but whose English database contains 155,327 words and 175,979 synsets (sets of words grouped together to be synonyms) - the number of synsets is larger than the number of words since a word can carry multiple meanings and thus be part of multiple synsets (Miller et al., 1990; WordNet Statistics, 2018). WordNet has been heavily used in text classification tasks as the lexical database of choice. In particular, WordNet is popular for tasks in word sense disambiguation (WSD), in which a word's context specific meaning, rather than its dictionary meaning, is determined (Navigli, 2009). WordNet is known to, in certain applications, encode distinctions in word senses (i.e., what a word means in two different contexts) that is more granular in detail than many human observers. For example, WordNet might examine two different sentences that use the word "happy" and determine that the use of "happy" in one sentence is different than the use of "happy" in the other sentence, even when human observers might see them as one and the same (Agirre & Lopez, 2003; Snow et al., 2007) - clustering methods are now used to work around this feature of WordNet, in certain contexts.

To implement our synonym replacement with the WordNet lexical database, we iterated through each preprocessed tweet and randomly selected, with probability of 0.6 (this probability gave the best results in a grid search), whether a given word would be replaced with a synonym, if it was encoded

in the database. We excluded words from replacement that we determined were proper nouns that would not have appropriate synonyms and were important to the essence of the message (e.g., “Trump” or “United States” or “Congress”). When we replaced words with a synonym, we aggregated the words in all the synsets within which the word in question, and we randomly picked a synonym. Though this approach may have added some noise since the synonym chosen may not have been chosen from a synset that represented the “context” of the original word, we also felt that this approach at the very least increased the probability that a word would be replaced with a word that could, to a human observer, be described as a “synonym”.

5.4.c A Comparison of the Two Synonym Methods

Original Preprocessed Tweet	**Tweet After GloVe Synonym Replacement**	**Tweet After NLTK Synonym Replacement**
@SenBlumenthal Go back to VietNam. Oh thats right, you never went. I think that makes you a credible liar.	senblumenthal go here oh shit though never went n't makes credible liar	senblumenthal start backwards oh thats right ne'er drop_dead reckon make credible prevaricator
@SenBlumenthal Don't recommend...DEMAND that they listen to women who are sexually assaulted! So NO to #PERVanaugh	senblumenthal include allow hear women mentally harassed pervanaugh	senblumenthal commend necessitate hear woman sexually assault pervanaugh
@SenBlumenthal Why do u believe Dr Ford & not Kavanaugh? She has no believable witnesses. She does not remember where or when? It sounds very political to us.	senblumenthal b remember comical murders remember kinda politics us	senblumenthal u believe believable witnesses recall sound political us

Table 2: A comparison of the two synonym replacement methods. The original preprocessed tweets (*left*) are compared to the GloVe implementation (*middle*) and the NLTK WordNet (*right*).

6 Results

We evaluated each of the classifiers on a human-labelled generalization test set (n = 10,000). For each of the classifiers, we evaluated their results based off the following metrics:

- Accuracy: *number of correct classifications / total number of classifications*
- Precision: *true positives / true positives + false positives*
- Recall: *true positives / true positives + false negatives*
- F1 Score: *(2 x precision x recall) / (precision + recall)*

- AUC: *Metric comparing true positive rate and false positive rate. AUC of 1 = perfect discrimination, no false positives. AUC of 0.5 = probability of true positive equals probability of false positives.*

Table 3 (see below) summarizes all results..

Data Augmentation Method (with sample size)	**Accuracy**	**F1 Score**	**Precision**	**Recall**	**AUC**
Benchmark Classifier (n = 16,000)	*65.84%*	*61.93*	*46.92*	***91.08***	*72.92*
Self Supervised Classifier *(n = 1,111,360)*	**92.55%**	**87.79**	**87.80**	87.77	**91.21**
Back Translation Classifier *(n = 32,000)*	66.84%	59.49	47.42	79.81	70.48
Word2Vec Classifier *(n = 32,000)*	65.06%	58.94	45.94	82.20	69.87
NLTK WordNet Classifier *(n = 32,000)*	67.56%	59.75	48.07	78.92	70.75

Table 3: **Classifier performance, when trained with different regimens of data augmentation and tested on a hold-out set of 10,000 tweets of various political topics.**. The classifier performances were evaluated on an external generalization test set of 10,000 tweets labeled for whether they express moral outrage. The generalization set spanned 4 different topics: the Covington Catholic High School Lincoln Memorial confrontation, the United States airlines incident of dragging a passenger off a plane, the Jussie Smollett attack claim, and President Trump's announcement of the transgender military ban. The performances of each of the augmented classifiers were compared to the benchmark classifier (*first row, italicized*) trained on the 16,000 human-labelled tweets. **Bold** indicates the highest values for a given evaluation metric. Notably, the self-supervised classifier outperformed the other classifiers across each of the metrics observed.

7 Discussion

Table 3 (*see above*) contains performance metrics for the classifiers, trained with the different data augmentation methods and evaluated on the generalization set. Our experiments demonstrated that the self-labelled data augmentation dramatically improved the performance of the moral outrage classifier, while the other data augmentation methods only marginally improved performance, if at all.

The classifiers differed in their tendency to classify tweets as having outrage, leading to differences in performance

It is interesting to observe that, though the accuracies for some of the augmented classifiers exceed that of the original GRU classifier, their recall scores are noticeably lower. This is true even for the self-supervised classifier, which greatly outperforms the original GRU classifier on the other metrics but doesn't quite achieve the same degree of recall. A high recall score indicates that, out of all the tweets that were hand-labelled by humans as having outrage, the classifier was able to correctly identify most of them as having outrage. It is possible for a model to have a high recall score if it classifies every single tweet as having outrage, since it would label as having outrage all the tweets that have outrage. Yet, this approach would also lead to a high number of false positives, where the classifier predicts that the tweet has outrage when it does not, and this would lead to a low precision score. Because we see the original classifier obtains a higher recall score than the other classifiers, we might suppose the original classifier merely classifies more tweets as having outrage. This may play a small role, since we do observe that the other classifiers have a slightly higher precision score than the original classifier. However, because we observe that the augmented classifiers have a similar precision score to the original classifier but a greatly reduced recall score, we can conclude that the slight decrease in accuracy observed in the augmented classifiers is due to the augmented classifiers' propensity for incorrectly identifying outrage tweets as having no outrage.

The self-labelled classifier outperforms the other classifiers, after accounting for sample size

A feature shared by the original human-labelled classifier and the classifiers augmented with back-translation and synonym replacement is that they maintain high recall scores, but at the expense of lower precision scores. In contrast, the self-labelled classifier, trained on both the human-labelled Kavanaugh tweets and the tweets labelled by the original classifier, exhibited both high recall and high precision scores. This feature of the self-labelled classifier holds true even when holding sample size constant and training on 16,000 tweets that were labelled by the original classifier (see Appendix, Figure C).

The success of the self-labelled classifier could be explained, in part, by the proportion of moral outrage tweets in the training set as compared to the generalization test set

This phenomenon could be attributed to imbalance in the representation of outrage tweets within the datasets. For the human-labelled Kavanaugh training set, 52% of tweets were classified as having outrage. Since the back-translation and synonym augmentation methods were permutations of the same Kavanaugh dataset, 52% of their tweets were also classified as having outrage. However, for the unlabelled data that was then used to augment the self-supervised classifier, only 38.32% of the tweets were classified as having outrage. The generalization test set used to evaluate the performance of the classifiers had 30.51% of its tweets classified as having outrage. Therefore, it is possible that the classifiers trained exclusively on the human-labelled Kavanaugh tweets, and permutations thereof (the original, back-translation, and synonym-augmented classifiers) may have been biased to classify more tweets as having outrage, while the classifier trained on the previously

unlabelled data was more prone to classify tweets as not having outrage, closer in line with the true frequency of outrage tweets in the generalization set. This hypothesis is supported by Foster (2011), who noted that if there is an imbalance in class priors within the training set, it will create biases in the posterior probabilities of the model. More support for this hypothesis comes from our comparison of the original classifier, trained on 16,000 human-labelled tweets, and a version of the self-labelled classifier that was trained on 16,000 unlabelled tweets labelled by the original classifier (see Appendix, Figure C). Holding sample size constant, we observed that the original classifier outperformed the self-labelled classifier in recall, but the self-labelled classifier had superior precision. Since a majority of the tweets in the generalization test set were classified by human labellers as having no outrage, we observe that in this instance, a model with high precision performs at a higher accuracy than a model with high recall (in this case, the model with high precision performed with 13.69% greater accuracy than a model with high recall).

The increase in sample size improved the accuracy of the self-labelled classifier

However, we would be remiss to exclude the impact of increased sample size in improving classifier performance. It is known that an increase in training data size tends to improve classifier performance, even in a self-supervised learning context (Schwenker & Trentin, 2014). We observed above that when sample size was held constant, the self-labelled classifier outperformed the original classifier and the other augmented classifiers, for reasons explained above. However, the full self-labelled classifier, with a training sample size of 1,111,360 tweets, greatly improves upon the self-labelled classifier trained on 16,000 tweets, which also does suggest that an increase in training sample size can improve classifier performance.

The possibility of propagating false labels in the self-labelled classifier was not a large cause for concern in our study

In our study, the self-labelled classifier achieved the best performance in classifying moral outrage tweets. However, a potential limitation in the use of a self-labelled classifier is that self-labelled data may be assigned false labels, and the accumulation of incorrectly labelled data may degrade classifier performance (Schwenker & Trentin, 2014). However, when we re-trained the classifier on a random subset of the self-labelled data, we found an improvement in performance, which suggests that in this scenario, the matter of class balance (that is, making sure that the proportion of outrage tweets in the training sample matches that in the test set) seems to carry more weight than issues regarding false labels (see Appendix, Figure C). Moreover, classifier performance improved as training size increased, which suggests that there was no significant accumulation of incorrectly labelled data that negatively affected classifier performance.

Classifier performance was improved by both matching the distribution of the training set to that of the test set and by increasing sample size

From the analyses above, we can see that, all else equal, two factors improve classifier performance on identifying moral outrage: (1) the proportion of moral outrage tweets in the training set and test

sets and (2) the size of the training set. We observed that, when size is constant across models, the model that performed the best (the self-labelled classifier) was the one whose proportion of moral outrage tweets was most aligned with the proportion in the test set. We observed that this was driven in part by the fact that as the proportion of moral outrage tweets in the training set increased, the resulting model tended to classify more tweets in the test set as having outrage. Additionally, when looking at the full self-labelled classifier, we saw that an increase in sample size dramatically improved classifier performance across all metrics. We also observed that data augmentation improves the model when the model is fed new data (in the case of using unlabelled tweets), rather than using permutations of existing data (in the case of back-translation and synonym replacement).

8 Applications to classifying moral outrage on Twitter

We then sought to use our improved classifier to answer a question relevant to the study of moral outrage in social media. In particular, we were interested in studying the phenomenon of “pluralistic ignorance”, in which a majority of group members privately reject some norm, but publicly support the norm because they incorrectly assume that most others are in favor (Prentice & Miller, 1993). A classic example of this is the observation that college students, when asked individually, tend to rate themselves as uncomfortable with college drinking culture but believe that their peers are comfortable with the drinking culture (and so, even when many individuals disapprove of the norm, they believe that everyone else is in support, so the individuals help reinforce the norms) (Prentice & Miller, 1993). In a societal context, pluralistic ignorance has been used to explain the longevity of racial segregation in the United States, under the rationale that individuals, in private, were more against the idea of racial segregation than the general public, but they perceived others as supportive of the existing social structures (O’Gorman, 1975), so they acted “in line” with their perception of social norms.

In the context of moral outrage, we were interested in observing whether people’s perceptions of outrage expressed by a person in a post or tweet were aligned with the actual degree of outrage felt by the person who originally posted the message. Because of factors such as lack of signals that are normally available in real-life conversations (e.g., facial expressions, tone) and the lack of barriers to making inflammatory or emotional comments (e.g., anybody can post a curse-laden, inflammatory tweet without fear of direct repercussions), it is possible that people, when they post a message that contains moral outrage, feel a degree of outrage that is different than the people who read their post (e.g., it is possible that viewers of the post might assume that the original writer of the post feels more incensed about a topic than they were in reality) (Crockett, 2017).

This application of our research is currently in development. We interacted with the Twitter API through the use of Python’s “tweepy” package. We created scripts to automatically scrape posts on Twitter and implement our classifier on the tweets. We then created additional scripts that send direct messages to users who posted tweets classified as having outrage so that we can ask them to rate their true levels of outrage on a Likert Scale from 1 to 7. We aimed to complement these ratings with ratings provided by subjects gathered by the Crockett Lab who were shown these same tweets.

We are still currently implementing a pilot version of the study. We hope that in the near future, we can aggregate these results in order to better understand how moral outrage is propagated on social media platforms.

9 Future directions

Improvements to the self-labelled classifier

To improve our classifier, we could try other data augmentation methods, such as using more complex deep learning models to model linguistic traits of the training set (e.g., Deng et al., 2013). Additionally, we can learn more about the self-labelling data augmentation method. We observed that accuracy was at 79.53% when we used 16,000 self-labelled tweets to train the classifier, while the accuracy increased to 92.55% when we used 1,111,360 self-labelled tweets. Future work could iteratively add self-labelled tweets to the training set and track how accuracy changes as the number of self-labelled tweets in the dataset changes (a methodology proposed by Wei & Zou, 2019).

Exploring the effect of class imbalance on classifier performance

Additionally, future work could also explore the notion of class balance and its effect on classifier bias. It is possible that the self-labelled classifier outperformed the original classifier, when sample size was constant, because the proportion of moral outrage tweets in the training set of the self-labelled classifier was closer to the proportion in the generalization test set than was the training set for the original classifier. Therefore, future work could modify the composition of the original Kavanaugh human-labelled dataset in order to better match the proportion of moral outrage in the generalization set, through means such as removing some tweets classified as moral outrage or artificially resampling or augmenting tweets classified as not having moral outrage. It would be interesting to explore whether classifier performance improves merely due to recomposition of class labels to remove any imbalances, even when the samples are a subset of the full dataset.

Generalization to other domains and sets of data

Finally, it would be interesting to learn how the classifier performs on other sets of data. Given that we are interested in how the classifier detects moral outrage in a wide variety of contexts on social media, it would be interesting to learn if there are characteristics of certain topics that influence classifier predictions in unexpected ways. This procedure would also give more insight into the embeddings within the neural network that influence its decisions. For example, it might be interesting to learn if there is any difference in classifier performance for events seen as racist as opposed to those that aren't seen as racist, since more recent work has shown that deep learning algorithms can underperform on topics that touch on race and ethnicity and magnify existing social divides (e.g., Garcia, 2016; Zou & Schiebinger, 2018; Robinson et al., 2020). Future work can explore how well this classifier generalizes to understand how moral outrage manifests in other contexts on social media.

10 Conclusion

In this thesis, we explored ways in which data augmentation could be used to improve a classifier trained to identify moral outrage in social media posts. We demonstrated that a self-labelled classifier, trained on a small set of hand-labelled tweets combined with a large set of unlabelled tweets classified by a pre-trained classifier, showed improved performance compared to a classifier trained on only hand-labelled training samples. This method also outperformed the other data augmentation methods tested, all of which performed similarly to the original classifier. We demonstrated that this result might have been due to two main factors: (1) the presence of class imbalance in the training and test sets and (2) the increase in the size of the training sample. From this, we show that it is possible to improve classification of moral outrage without incurring the cost of additional human-labelled training samples. Finally, we began to apply this improved classification to further explore how moral outrage is propagated on social media.

11 Acknowledgements

This thesis serves as the capstone to a long four-year academic journey at Yale, during which I switched between 8 (possibly more?) majors and struggled to find my place and sense of self before eventually discovering my interests at the intersection of statistics, data science, and social science.

I would like to thank my advisors Professor Molly Crockett, William Brady, and Killian McLoughlin, who provided consistent support, mentorship, and guidance throughout this process. Through them, I both learned how to question the trends in the world around me and developed the tools and experiences necessary to tackle those questions.

I'd also like to acknowledge my fellow first-generation, low-income college students. I was there with you, wondering if we deserved these opportunities or if instead, we were merely token diversity students. I was there with you, working 3 jobs during the school year because our parents couldn't financially support us. I was there with you, lying to other people about what our parents did because they weren't doctors or lawyers, but were instead waitresses and cashiers. I'm glad to have been one of you, and it's up to us to show the world that we belong at places like Yale.

Next, I dedicate this thesis to my family. Throughout college, I was motivated to constantly push further because I wanted to break the cycle of poverty and blaze a new path for you, one defined not by the circumstances of your birth, but instead by the size of your dreams and by your hard work and hustle. I wouldn't be the person that I am without your love and support. To Mom, Mikee, Molly, Max, and my countless aunts and uncles and cousins in the Philippines, this is for you.

Most of all, I thank God for being by my side, through all the peaks and troughs of this journey and the journey which is to come, shaping me to become a better man each and every single day. None of this would've been possible without His grace and guidance and strength.

12 Appendix

Tweet Topic	Proportion of Tweets Labelled as Having Outrage	Raw Counts (Tweets with Outrage / Total Tweets)
Kavanaugh	52.41%	8,386 / 16,000
Impeachment	41.91%	104,691 / 249,774
Emergency Wall	39.25%	175,228 / 446,394
Admissions	66.22%	4,260 / 6,433
March for Our Lives	52.66%	38,489 / 73,084
Migrant Caravan	30.36%	97,041 / 319,675

Figure A: The proportion of tweets marked by human labellers as having outrage (*top row, Kavanaugh)* and labelled by the pre-trained classifier as having outrage (*see rest of rows*).

Tweet Topic	Examples of Tweet Marked as Having Outrage	Examples of Tweet Marked as Having No Outrage
Impeachment	The Donald J. Trump lied to us, and the American people on fixing "America's Broken Healthcare System" riddled with Corruption.	@lookner What happened after the phone call is what will determine impeachment, and is likely in the whistleblower complaint.
Emergency Wall	@DonaldJTrumpJr A coup is an unconstitutional transfer of power, usually executed by the military. The 25th Amendment is a constitutional means to remove a president from office, the opposite of a coup..... Your mind numbing ignorance, people should be paid to read your tweets.	@WhiteHouse History will show you were the greatest president of all time 😎 so many books and movies will be made about this, go get em Mr president 😍
Admissions	Admissions scandal reveals 'aristocracy masquerading as a meritocracy https://t.co/2EQpFVBEWy via @CNNPolitics > #bospoli #mapoli #maedu #collegeadmissions #admissionscams #inequality #cheating #bribery #racketeering #wealth #privilege	https://t.co/4yixEbdyTN USC puts holds on accounts of students linked to college admissions cheating scandal so they can't register for classes or get transcripts USC said they may expel students caught up in the college-admissions cheating scandal after it completes a reviews
March for Our Lives	RT ComicsConform "RT voiceblm: The #NRA perhaps the most toxic force in U.S. politics !! The NRA is a terrorist organization!	if you're walking out today, please let it be to show Marjory Stoneman Douglas that we're by their side!! No more gun violence! #NeverAgain
Migrant Caravan	What specific concrete proof or evidence - not guesses or opinion - of ANY kind do you have to support this claim @realDonaldTrump? I'm an #Army #veteran & citizen and I demand to know the source of your fear-stoking information about the #Caravan. https://t.co/AWA72RtftR	It's all peace and love and the desire to mow lawns. Yup, that's all! #caravan https://t.co/dp1icZpp2i via @BreitbartTexas

Figure B: **Examples of tweets for each topic in the unlabelled dataset** (*left*), marked as having outrage (*middle*) or as not having outrage (*right*).

Classifier	Accuracy	F1 Score	Precision	Recall	AUC
Control Classifier	65.84%	61.93	46.92	91.08	72.92
Self-Labelled Classifier	79.53%	63.78	69.30	59.06	73.79

Figure C: Results of the control classifier and self-labelled classifier, after controlling for the size of the training set. A comparison of the performance of the control classifier (without augmentation) and a classifier trained on a random subset of 16,000 self-labelled tweets

Classification	Tweet
Outrage	Cannay believe some of mutuals have liked this ,, imperialist bullshit serve Ur country muck ,, it's a hard no to anyone who wants to join the military LGBT or otherwise western militaries just be out here commiting genocide for personal gain
	can you imagine wanting to defend your stupid country that is always involved in wars and other people's conflicts and still, they decide to deem you unfit to do so because you refuse to not be who you are? US really are something else.
No Outrage	@foxandfriends Wow. Could their stock get any worse? ❤ 2 watch them self implode. Always hated flying #united.
	@them Just to alert, the ACLU is asserting otherwise, so it might still be blocked.\nhttps://t.co/q1gXPel7AW

Figure D: **A comparison of tweets in the generalization set**, labelled by the final classifier as having outrage (*top*) or not having outrage (*bottom*)

13 References

Agirre, E., Lopez, O. (2003). Clustering WordNet Word Senses. In *Proc. of the Conference on Recent Advances on Natural Language (RANLP'03)*, *Borovetz, Bulgaria,* 121–130.

Anand, N., Goyal, D., & Kumar, T. (2018). Analyzing and Preprocessing the Twitter Data for Opinion Mining. *Lecture Notes in Networks and Systems*, *34*, 213–221. https://doi.org/10.1007/978-981-10-8198-9_22

Brady, W. J., & Crockett, M. J. (2019). How Effective Is Online Outrage? *Trends in Cognitive Sciences*, *23*(2), 79–80. https://doi.org/10.1016/j.tics.2018.11.004

Brady, McLoughlin, Doan & Crockett (under review). Social reinforcement of moral outrage in online social networks

Bull, L., Worden, K., Manson, G., & Dervilis, N. (2018). Active learning for semi-supervised structural health monitoring. *Journal of Sound and Vibration*, *437*, 373–388. https://doi.org/10.1016/j.jsv.2018.08.040

Bull, L. A., Worden, K., & Dervilis, N. (2020). Towards semi-supervised and probabilistic classification in structural health monitoring. *Mechanical Systems and Signal Processing*, *140*, 106653. https://doi.org/10.1016/J.YMSSP.2020.106653

Chawla, N. V., & Karakoulas, G. (2005). Learning from labeled and unlabeled data: An empirical study across techniques and domains. *Journal of Artificial Intelligence Research*, *23*, 331–366. https://doi.org/10.1613/jair.1509

Chung, J., Cho, K., & Bengio, Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, *3*, 1693–1703. https://doi.org/10.18653/v1/p16-1160

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. Retrieved from https://arxiv.org/abs/1412.3555

Crockett, M.J. (2017). Moral outrage in the digital age. *Nat Hum Behav* 1, 769–771. https://doi.org/10.1038/s41562-017-0213-3

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, *17*(1), 124–147. https://doi.org/10.1111/j.1467-9671.2012.01359.x

Deng, L., Hinton, G., & Kingsbury, B. (2013). New Types of Deep Neural Network Learning for Speech Recognition and Related Applications: an Overview. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, *Vancouver,* 8599–8603. Retrieved

from http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6639344&isnumber=6637585%0A

Garcia, M. (2016). Racist in the Machine: The Disturbing Implications of Algorithmic Bias. *World Policy Journal 33*(4), 111-117. https://www.muse.jhu.edu/article/645268.

Han, J., & Moraga, C. (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *930*, 195–201. https://doi.org/10.1007/3-540-59497-3_175

Hassan, A., Abbasi, A., & Zeng, D. (2013). Twitter sentiment analysis: A bootstrap ensemble framework. *Proceedings - SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013*, 357–364. https://doi.org/10.1109/SocialCom.2013.56

Hayashi, T., Watanabe, S., Zhang, Y., Toda, T., Hori, T., Astudillo, R., & Takeda, K. (2019). Back-Translation-Style Data Augmentation for end-to-end ASR. *2018 IEEE Spoken Language Technology Workshop, SLT 2018 - Proceedings*, 426–433. https://doi.org/10.1109/SLT.2018.8639619

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/J.PATREC.2009.09.011

Kwon, S. J., Park, E., & Kim, K. J. (2014). What drives successful social networking services? A comparative analysis of user acceptance of Facebook and Twitter. *Social Science Journal*, *51*(4), 534–544. https://doi.org/10.1016/j.soscij.2014.04.005

Lai, S., Liu, K., He, S., & Zhao, J. (2016). How to generate a good word embedding. *IEEE Intelligent Systems*, *31*(6), 5–14. https://doi.org/10.1109/MIS.2016.45

Lazer, D., Brewer, D., Christakis, N., Fowler, J., & King, G. (2009). Life in the network: the coming age of computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742.Life

Leitch, D., & Sherif, M. (2017). Twitter mood, CEO succession announcements and stock returns. *Journal of Computational Science*, *21*(0), 1–10. https://doi.org/10.1016/j.jocs.2017.04.002

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225. https://doi.org/10.1162/tacl_a_00134

Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 1412–1421. https://doi.org/10.18653/v1/d15-1166

McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & Spiro, E. S. (2017). Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. *Sociological Methods and Research*, *46*(3), 390–421. https://doi.org/10.1177/0049124115605339

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1–12.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, *3*(4), 235–244. https://doi.org/10.1093/ijl/3.4.235

Mogotsi, I. C., Manning, C. D., Raghavan, P., & Schütze, H. (2010). Introduction to Information Retrieval. *Inf Retrieval*, *13*(2), 192–195. https://doi.org/10.1007/s10791-009-9115-y

Nartey, O. T., Yang, G., Wu, J., & Asare, S. K. (2020). Semi-Supervised Learning for Fine-Grained Classification with Self-Training. *IEEE Access*, *8*, 2109–2121. https://doi.org/10.1109/ACCESS.2019.2962258

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, *41*(2). https://doi.org/10.1145/1459352.1459355

Prentice, D. A., & Miller, D. T. (1993). Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm. *Journal of Personality and Social Psychology*, *64*(2), 243–256. https://doi.org/10.1037/0022-3514.64.2.243

Robinson, W. R., Renson, A., & Naimi, A. I. (2020). Teaching yourself about structural racism will improve your machine learning. *Biostatistics (Oxford, England)*, *21*(2), 339–344. https://doi.org/10.1093/biostatistics/kxz040

Sawaoka, T., & Monin, B. (2018). The Paradox of Viral Outrage. *Psychological Science*, *29*(10), 1665–1678. https://doi.org/10.1177/0956797618780658

Schwenker, F., & Trentin, E. (2014). Pattern classification and clustering: A review of partially supervised learning approaches. *Pattern Recognition Letters*, *37*(1), 4–14. https://doi.org/10.1016/j.patrec.2013.10.017

Sennrich, R., & Haddow, B. (2015). A joint dependency model of morphological and syntactic structure for statistical machine translation. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, (September), 2081–2087. https://doi.org/10.18653/v1/d15-1248

Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, *1*, 86–96. https://doi.org/10.18653/v1/p16-1009

Severyn, A., & Moschitti, A. (2015). Twitter Sentiment Analysis with deep convolutional neural networks. *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 959–962. https://doi.org/10.1145/2766462.2767830

Shteynberg, G. (2018). A collective perspective: shared attention and the mind. *Current Opinion in Psychology*, *23*, 93–97. https://doi.org/10.1016/j.copsyc.2017.12.007

Snow, R., Prakash, S., Jurafsky, D., & Ng, A. Y. (2007). Learning to merge word senses. *EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (June), 1005–1014.

Spring, V. L., Cameron, C. D., & Cikara, M. (2018). The Upside of Outrage. *Trends in Cognitive Sciences*, *22*(12), 1067–1069. https://doi.org/10.1016/j.tics.2018.09.006

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11141 LNCS*, 270–279. https://doi.org/10.1007/978-3-030-01424-7_27

Tang, D., Qin, B., & Liu, T. (2015). Deep learning for sentiment analysis: Successful approaches and future challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *5*(6), 292–303. https://doi.org/10.1002/widm.1171

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word Representations: A Simple and General Method for Semi-Supervised Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 384–394). USA: Association for Computational Linguistics.

Wang, W. Y., & Yang, D. (2015). That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods*

in Natural Language Processing, (September), 2557–2563. https://doi.org/10.18653/v1/d15-1306

Weber, R., Mangus, J. M., Huskey, R., Hopp, F. R., Amir, O., Swanson, R., Gordon, A., Khooshabeh, P., Hahn, L., & Tamborini, R. (2018). Extracting Latent Moral Information from Text Narratives: Relevance, Challenges, and Solutions. *Communication Methods and Measures*, *12*(2–3), 119–139. https://doi.org/10.1080/19312458.2018.1447656

Wong, S. C., Gatt, A., Stamatescu, V., & McDonnell, M. D. (2016). Understanding Data Augmentation for Classification: When to Warp? *2016 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2016*. https://doi.org/10.1109/DICTA.2016.7797091

WordNet Statistics. Wordnet.princeton.edu.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), 1–25. https://doi.org/10.1002/widm.1253

Zhang, T., & Heights, Y. (2000). The Value of Unlabeled Data for Classification Problems. *Proceedings of the Seventeenth International Conference on Machine Learning*, 1191--1198. Retrieved from http://scholar.google.com/scholar?q=value+of+unlabeled+data&hl=en&btnG=Search#0

Zou, J., & Schiebinger, L. (2018). Design AI so that it's fair. *Nature*, *559*(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8