Natural Language Processing Applications for Prediction of Violence in Gang-Related Social Media

Terra Blevins

Submitted in partial fulfillment of the
requirements for the degree of
Bachelor of Arts
in Columbia College

Columbia University

2017

Abstract

Violence among gang members is a well-known and growing concern. It has been exacerbated by the increasing popularity of social media, as members of rival gangs use the platforms to interact with and taunt each other. This thesis explores a number of natural language processing solutions towards this issue, focusing on users involved with gangs in the Chicago area. We work towards building a classifier that identifies tweets that are at a high likelihood of precipitating later real-world violence. This is done by integrating qualitative theories about how these users interact both online and offline with data-driven machine learning methods. We experiment with both supervised and semi-supervised learning approaches to solve this task. We also generate justifications of these predictions through explanation of the features most significant to them.

Acknowledgments

This thesis was completed with the help of the many fantastic researchers that I have had the pleasure of working with on the Gang Intervention project. I am deeply grateful to Kathleen McKeown; this work would not have been possible without the wonderful guidance and support she has given me. I would also like to thank the other researchers on the Gang Intervention project, especially Owen Rambow and Desmond Patton for their terrific help and advice on many of the projects that make up this thesis, as well as Robert Kwiatkowski, whose work was integral to the success of this project.

I would also like to thank my family and friends, who have provided me with so much support. I am especially grateful to my parents, who've always believed in me and pushed me to be my best self, and to Chris Luccarelli for his unwavering support in everything I do.

Table of Contents

CHAPTER 1

Introduction

Gun violence is a major issue in the United States, particularly in urban areas. Chicago saw a 40% increase in firearm violence in 2015, and these numbers continued to increase during 2016. In December of 2016 (while this study was underway), sixty people were shot and eleven killed over Christmas weekend in Chicago (Nickeas et al. 2016). Other cities in the US, such as Atlanta, Baltimore, and Detroit, have similar stories and statistics.

Much of this violence is gang-related, and a number of recent studies show that specifically gang violence has been worsened by rising usage of social media. Décary-Hétu and Morselli (2011) performed a qualitative analysis that found that gangs are increasing their presence on social media. Furthermore, this rise in social media usage has led to activity called "internet banging," in which gang members discuss gang affiliations and activities as well as disseminate violence (Patton, Eschmann, et al. 2013). Due to the visibility of these online interactions (many gang-involved youths do not have any privacy settings on their social media accounts), discussions of violence among gang members are becoming much more public.

The overall aim of this project is therefore to identify discussions of violence in social media posts and interactions before any actual violence occurs off-line. We currently focus on data from gang members in Chicago. This decision allows us to focus on the language and culture specific to Chicago gangs, while still developing tools that can also be applied to other cities (provided appropriate data from those areas). To our knowledge, no tools have yet been built to identify threatening language on social media, which makes our work a new task.

Despite the lack of previous work on this topic, it is an important task. If we are able to automatically identify the posts that are most likely to lead to violence, we will be able to provide a new set of tools for intervening *before* this violence occurs. There are community outreach programs that already perform these interventions, both on- and off-line, when they see a problematic post on social media. The process currently used to do this is completely manual, which means it is not very scalable. However, the amount of information people post on social media continues to grow rapidly, making automatic tools to augment manual monitoring of social media more relevant than ever.

The work in this thesis was conducted in close collaboration with other computer scientists as well as with collaborators in the Columbia School of Social Work's SAFE Lab. The Social Work researchers annotated and interpreted the social media data, while the computer science researchers gathered the (larger) datasets and developed systems that automatically process and classify the data. The research group also meet for frequent meetings to discuss the intersection and interactions of our work.[1] This collaboration with Social Work researchers is particularly helpful to the development of NLP tools for this specific task, because our collaborators provide (or obtain) expert domain knowledge on a topic about which the data scientists know very little.

The specific contributions of this thesis are as follows:

- We develop a number of fully supervised classifiers that identify expressions of both aggression and loss in a dataset from one Chicago gang member.
- We expand our fully supervised classifier to a semi-supervised (via self-training) model that learns from both our set of unlabeled tweets from gang members in Chicago and from our labeled data.
- We develop of system to generate explanations of individual predictions made by our classifiers, in order to make these models more accessible to a non-computer scientist end user.

Each of these contributions works towards automating the process for identifying social media posts at a high risk of leading to offline violence, either directly or indirectly. Making the assumption that these high-risk tweets express either loss or aggression (for reasons outlined in Section 3.2), our two classifiers automatically identify tweets that contain one of these two topics. Our fully supervised approach was a success; our best models significantly outperform a baseline when predicting both aggression and loss in tweets from Chicago gang members. However, our semi-supervised approach was less successful, because it performed about the same as our fully supervised classifiers (and any improvements over the supervised model were not significant).

The explanations of these predictions provide a more indirect contribution to this task. However, they are also important to the overall goal of this project, because they work to clarify our predictions to the users who do not know how ML models work. Without these

[1] This group follows a similar model to that presented in Ford (2014).

clarifications, our desired end user (who would use our system to more effectively monitor the social media of gang-involved people for high-risk tweets) would have no reason to trust our model's predictions.

An additional contribution derived from all three projects undertaken by this thesis is the application of NLP tools to the language that occurs in social media posts of gang-related users. This dialect of English (which contains elements from both mainstream social media language and African-American Vernacular English (AAVE), as well as gang-specific features) is relatively unstudied in NLP. It is also different enough from Standard English that standard NLP tools generally do not perform well on our data, requiring us to experiment with various knowledge bases in order to apply these tools to our data.

Each of the three areas examined by this thesis thus work towards facilitating preventative intervention in gang-related gun violence, which is the overall goal of this project. We are able to apply NLP tools to a rather unstudied dialect of English successfully. Models for predicting tweets that express loss or aggression, which we have hypothesized are the ones that are at a high-risk of precipitating real-world violence, were also developed. This work serves to lay the groundwork for a relatively new application of computational social science and NLP.

CHAPTER 2

Related Work

2.1 Computational Social Science

Data science tools have often been applied towards investigating questions in social science; this is also true for the specific question of gang violence in Chicago. Wijeratne et al. (2015) built a system which analyzed the social media posts of known Chicago gang members. They used standard emotion analysis tools (that were not tailored to the language of the dataset) as well as social network analysis in order to examine these social media interactions with the goal of gaining a better understanding of the structure of these gangs.

Prior research has also studied gang activity outside of Chicago by using similar methods; Radil et al. (2010) conducted a social network analysis to study the geographic relationships of gangs in Los Angeles using spatialized network data. Other studies have investigated automatically identifying gang members on Twitter, both inside and outside of Chicago (Balasuriya et al. (2016); Wijeratne et al. (2015)). They use text features (from the user's tweets and profile descriptions) as well as descriptive tags about users' profile pictures that are obtained from a standard image classification tool in order to predict if a user is affiliated with a gang or not.

Data science tools have also been used to study gun violence more broadly. Green et al. (2017) used social network analysis in order to model gun violence in Chicago as a social "contagion". They find that these social contagion factors play a significant role in the spread of firearm violence in their dataset. Other researchers have used information extraction techniques to identify shootings from news articles. These extracted events have been used to build a database of gun violence incidents across the United States (Pavlick et al. 2016).

2.2 Extracting Features from Twitter Data

A fair amount of work has been done on sentiment analysis of Twitter data, and Twitter has become a popular source of data on which to develop sentiment analysis techniques (Rosenthal, Nakov, et al. 2015). Rosenthal and McKeown (2013) used the Dictionary of Affect in Language (or DAL) as we do for emotion recognition on our dataset. They map the scores obtained from the DAL to a sentiment score instead of using the full set of three dimensions to model emotion (Whissell 2009). Other approaches to sentiment analysis of Twitter data experiment with feature based models as well as tree kernel ones (Agarwal et al. 2011).

Beyond sentiment analysis, other work has been conducted on emotion recognition in Twitter data. Mohammad and Kiritchenko (2015) used hashtags to automatically label a Twitter dataset with fine-grained emotion tags. These emotion labels differ from our approach since there is no similarity metric between the labels. Emotions are instead considered different, discrete states. With the DAL, we use a set of dimensions that attempt to quantify the emotion expressed in a tweet, rather than qualitatively describe it. Because of this, our emotion scoring methods are more similar to the sentiment approaches than the emotion recognition ones.

There has also been previous work developing POS taggers for Twitter data. Owoputi et al. (2013) built one such tagger for English tweets that deals with many of the non-Standard elements of tweets, such as emoticons and acronyms. Others have developed POS taggers for AAVE (Jørgensen et al. 2016). They used various methods of domain adaptation in order to learn a AAVE POS tagger from a partially labeled dataset. This is similar to our approach, as we use domain adaptation so that we can build NLP tools for our data using Standard English resources.

Other work has performed domain adaptation through feature set augmentation (and then learning the adaptation with a supervised ML model); this technique informed the development of the POS tagger trained for our specific dataset (Daumé III 2007). Another method of domain adaptation is seen in Agarwal et al. (2011). In this work, Internet slang is mapped to Standard English using online resources map between the two; this is similar to the methods of "translation" we use for emotion scoring, though existing resources for the language of our data are harder to come by than for mainstream Internet slang.

2.3 Semi-Supervised Learning Methods

Much of the work done on distant labeling exploits existing aspects of the data that are correlated with the labels they want to learn. This is seen most often with sentiment and emotion recognition; both Go et al. (2009) and Agarwal et al. (2011) use emoticons to automatically assign sentiment labels to their data.[1] Similarly, Purver and Battersby (2012) used both emoticons and emotion word hashtags (such as #angry or #happy) to automatically assign emotion labels to tweets. The noisy labels in all three studies were then used to directly train a supervised ML model.

In all of these cases, the features that are used to distantly label the data are manually chosen using *a priori* knowledge about the task. Ouyang and McKeown (2015) use a more complex distant labeling algorithm, in which the method for assigning the labels is learned by an SVM over a larger set of heuristics. Other researchers have exploited existing knowledge bases to perform distant labeling on tasks such as relation extraction and semantic parsing (Mintz et al. (2009); Berant et al. (2013)).

Other approaches to semi-supervised learning involve directly incorporating unlabeled data into the learning process. One example is co-training, where two different "views" of the same labeled data L (often meaning different feature sets derived from the same dataset) are used to train different models. These models are used to verify the other's predictions on an unlabeled dataset U and incorporate unlabeled points into L if verified (Blum and Mitchell 1998). This approach requires two different sets of features that can relatively accurately model the data for classification.

Self-training is another approach taken towards semi-supervised learning, in which a model is trained on a small seed dataset L and then used to label the larger unlabeled dataset U before retraining on both L and the automatically labeled U. Zhou et al. (2012) use a variant of self-training where only the unlabeled data points that improve the performance of the classifier are automatically labeled and added to the training data. Other researchers have performed self-training with unlabeled data in conjunction with distant labeling (such that they verify each other), similar to our approach (Ouyang and McKeown 2015).

2.4 Explanations of ML Models

Prior studies have experimented with a number of different methods of generating explanations or justifications of a classifier's predictions. Some have focused on explaining predictions

[1] Emoticons are facial expressions depicted by punctuation. In these studies, a tweet containing emoticon :) would be assigned a positive sentiment label

through visualization of features and how important they were to the prediction (Kononenko et al. (2010); Baehrens et al. (2010)). However, these visualizations rely on knowledge of how the underlying models work and so are not helpful when trying to clarify how a prediction was made to a user who is not a computer scientist.

Other researchers have worked on making ML models more understandable to human users by mapping from the probabilistic representation of these models to qualitative phrases (such as "likely" or "very unlikely" to represent different p-values) about how confident the model was about the prediction (Druzdzel 1996). This work generated basic text explanations that used these qualitative phrases to describe the prediction process. Biran and McKeown (2014) developed a system to justify a model's predictions to a non-computer scientist user; these text justifications were automatically generated using NLG tools. Here, the goal was to convince the user to trust the model, rather than just explain the prediction process.

Ribeiro et al. (2016) take a different approach to explaining their classifier's predictions for image classification: their system highlights the sections of the image that contribute most to the prediction. This is similar to our approach of highlighting the features that have the largest effect on a prediction. They also propose using their explanations to help users determine if a model is trustworthy or not, rather than to just build trust in their system. Since their explanations consist of fragments of the input image, it is easy for users who are not familiar with ML models to still judge if the model selected the correct areas of the image on which to base its prediction.

CHAPTER 3

Premises

This chapter details background information about our work on predicting violence in gang-related social media. We provide the specifics of the different datasets used in each of our tasks. Then we focus on the Social Work theories and techniques that motivate this project. All of the work in this thesis was conducted with close interaction and collaborative meetings with researchers in Social Work. Our data and a number of the choices we made while designing our experiments are therefore informed by these theories.

3.1 Data

This work involved a number of different but interrelated datasets. Details of the various datasets are given in Table 3.1. All of our data was obtained and annotated by our collaborators at Fairfield University and in the Columbia School of Social Work. The annotated data was labeled with codes that described the content or intent of each tweet. For the computational experiments, these codes were collapsed into more general themes of "loss", "aggression", and "other", or tweets that do not fit into either of the first two themes. These themes are used as labels for our learning tasks.

Our labeled dataset consists of 820 tweets; these are mostly from Gakirah Barnes, a gang member who was shot and killed in April 2014 at the age of 17. It also contains a smaller number of posts from users with whom she communicated.[1] This is the dataset used for the fully-supervised classification task described in Chapter 4; it was also used in Chapter 5. Our collaborators initially choose to focus on Gakirah due to her active presence on Twitter as well as her known affiliation with a Chicago gang. Due to her frequent posts, all of the tweets come from three months in 2014: January, March, and April. This data was labeled by two Social Work graduate students; the inter-annotator agreement on the portion of the data used for evaluation is κ=0.62, indicating moderate agreement.

[1] This data can be found online at http://dx.doi.org/10.7916/D84F1R07.

Dataset	Number of Tweets	Date Range	Annotated?
Gakirah	820	Jan. 2014, Mar.-Apr. 2014	yes
Top-Ten Comm.	47	Feb.-Apr. 2014	yes
Gang Network	1.6 million	Mar. 2010-Jul. 2016	no

Table 3.1: Summary of the datasets

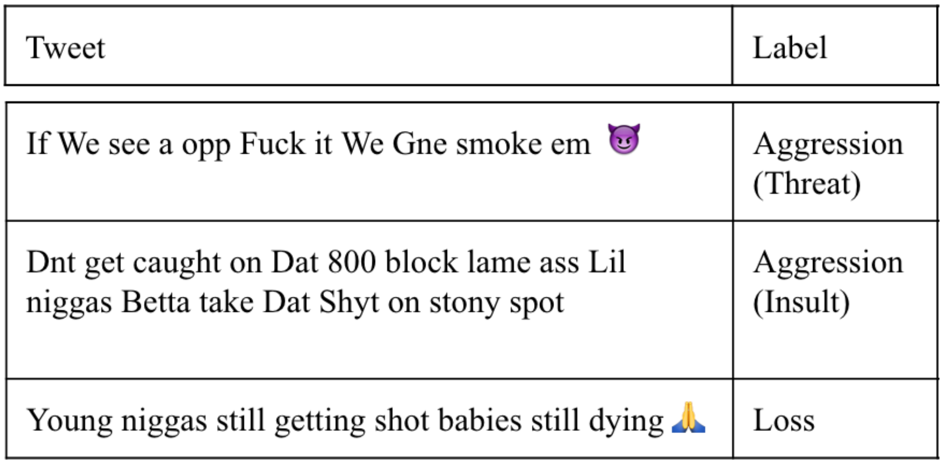
We also worked with tweets from Gakirah's "top-ten communicators", who are the ten users on Twitter with whom she communicated the most. These are determined by quantity of at-mention conversations, which are conversations where person A replies to a tweet from person B by including "@person A's username" in the response tweet. This data comes from the same time period as the Gakirah dataset. A portion of these tweets were labeled (with the same codes as the Gakriah dataset); only the labeled data from the top-ten communicators is detailed in Table 3.1. These labeled tweets from "top-ten communicators" were used for evaluation of the semisupervised learning task (5).

Finally, we continue to use Gakirah's network to gather a large, currently unlabeled dataset of tweets. Specifically, users whose tweets are included in this dataset are chosen by expanding on the "top-ten communicator" users and other people in Gakirah's network. Once these people are found, their 200 most recent tweets are added to the dataset. This process gave us a dataset of 1.6 million tweets. It covers a much longer time period than our other datasets, with tweets that date from March 2010 to July 2016. However, since the most recent tweets for each user are collected, the distribution of dates is skewed, so that there are many more recent tweets than earlier ones.

It is important to note that the tweets in our dataset contain language that differs both from Standard American English and from the mainstream "social media" language commonly used on Twitter. Rather, the language of our data contains aspects of African-American Vernacular English (AAVE) as well as subculture specific abbreviations, spelling, and vocabulary. Tweets from our data are provided in Figure 3.1 as examples of this dialect. This specific language is not yet studied in NLP, and deviates greatly from the Standard English used to develop most NLP tools. This complicates the tasks of text processing and understanding utilized in this work.

3.2 Social Work Theories

This work was conducted as part of a collaboration with researchers at the Columbia School of Social Work. All of the manually labeled data used to train and evaluate our systems were annotated by these collaborators, using the methods described below. Additionally, much



Tweet	Label
If We see a opp Fuck it We Gne smoke em 😈	Aggression (Threat)
Dnt get caught on Dat 800 block lame ass Lil niggas Betta take Dat Shyt on stony spot	Aggression (Insult)
Young niggas still getting shot babies still dying 🙏	Loss

Figure 3.1: Example tweets demonstrating the language of the data (and their associated label).

of our computational research was motivated and informed by the following theories in the field of Social Work.

The first step of the qualitative analysis of the data was to hire domain experts to explain the context needed to understand the data. For this study, our collaborators hired people from the Chicago neighborhoods on which this project focuses. These youths "translated", or explained, the tweets from Gakirah. These explanations included their reactions, the perceived message of the tweet, and an explanation of the language and context needed to understand the tweet (Patton, McKeown, et al. 2016).This information was not directly used to label the data for computational work. However, it provided specific domain knowledge to our collaborators with respect to the culture and language of the tweets, and therefore informed their annotation process.

The social work researchers then used a process called the *Digital Urban Violence Analysis Approach* (DUVAA) to annotate the data (Patton, McKeown, et al. 2016). The DUVAA is a systematic qualitative approach that proceeds by: discovering any offline "precipitating events" that could cause a threat or aggressive conversation on social media; identifying the user handle of the author of the tweet in question; analyzing the body of the tweet for the message, tone, and gang-related references. The content of the tweet and the surrounding context (such as retweets or replies) are also evaluated for evidence as to why this tweet might precipitate violence and for the tone of the tweet. Finally, the researchers identified if a given tweet is a "trigger event", which is a point where the overall tone of a conversation or a user's posts goes from neutral or positive to aggressive.

This technique was used to gain a rich, qualitative understanding of the data. The dataset was then annotated using all of the research of the tweets from the DUVAA and the domain knowledge from the hired experts. The specific, fine-grained codes came from a codebook developed during this process; they were later collapsed into general themes for the computational work (Patton, McKeown, et al. 2016).

Outside of data annotation, our work is also influenced by previous work done by our collaborators. Specifically, an existing theory on displays of aggression argues that threatening posts often follow posts expressing grief in a cycle of reactionary violence (Patton, Lane, et al. 2016). This theory was further substantiated by the DUVAA process while annotating Gakirah's tweets (Blevins et al. 2016). We designed our system with this theory in mind, by focusing on identifying expressions of grief and loss as well as on predicting aggression.

CHAPTER 4

Supervised Learning: A Case Study

Much of this work is focused on identifying the tweets at a higher risk of leading to violence. Our first experiment was a supervised approach to this task. Specifically, we trained fully supervised classification models to identify which tweets from this dataset express loss or express aggression. We focus on these two categories, loss and aggression, due to the theorized cycle of grief leading to violence discussed in the previous chapter (3.2). We hypothesized that, if we can identify the tweets that express loss and aggressive intent from the user, we can use this cycle to better understand when online aggression turns into offline violence.

With this in mind, we developed a number of different classifiers in order to predict the labels of “loss”, “aggression”, and “other”. These were trained and evaluated on the Gakirah dataset that was introduced in 3.1. Because of the limitations of this dataset, we focus on building a case study for a single individual’s language and content, and work towards providing a framework future work can build off of.

We noticed that our labels of “aggression” and “loss” are related to both the content and the tone of the tweets; this realization informed the features chosen to train our system. We used a number of language features to model the content. Also included as features for the models are the predicted emotion scores of this data.

The rest of this chapter proceeds as follows. First, the process of obtaining emotion scores for the tweets in our dataset is explained. We then cover the features used to train these models and how we find these features for each tweet, before walking through the design of each of the models for this experiment. The evaluation of how our systems perform on a held-out test set are then presented. Finally, we discuss these results and how it relates to future work towards predicting and preventing incidents of gang violence. This work was originally discussed in Blevins et al. (2016) and is expanded upon here.

4.1 Emotion Scoring

We use an emotion recognition process for the tweets in our dataset as part of our system for predicting aggression and loss. Our process for obtaining the emotion of a given tweet is based on the Dictionary of Affect in Language (or DAL) (Whissell 2009). The DAL gives scores for individual words, which we then combine to generate a tweet level score. Due to the language differences of the tweets from Standard English (which the DAL is created from), we use a number of adaptation techniques in order for this resource to work with our specific data set.

The DAL is a lexicon that maps English words to a three dimensional score. The three dimensions of this score are as follows: pleasantness, which is how positive the word (or connotation) is; activation, which is a measure of a word's intensity; and imagery, or how easy a word is to visualize. The DAL contains approximately 8,700 words, which (while sizable) does not cover all of Standard English. Our system augments the DAL with WordNet in order to get emotion scores for Standard English words that are not handled by the DAL, following the work done in Rosenthal and McKeown (2013). For each word that is not in the DAL but is in WordNet, the synonyms from the first (most common) synset of that word are searched against the DAL.[1] It is assumed that the emotion of a synonym will be similar to that of the original word. Therefore, if there is a match between the synonyms and the DAL, the emotion score of the synonym is used for the original word.

While this extension to the DAL helps with coverage of Standard English, the language found in our dataset is a different dialect of English. Another part of building an emotion recognition system for this data is therefore to find an adaptation that applies the DAL to the language of the tweets. First, we make the assumption that any word not found in the DAL or in WordNet is not a word in Standard English; we then try to map these non-standard words to a similar word or phrase in Standard English.We experimented with both Wiktionary, which is a large open-source dictionary from Wikipedia, and a phrasebook learned from our dataset to "translate" these tokens to standard English.[2] With Wiktionary, the definition of a word was considered to be its translation. The coverage and accuracy of these two resources are compared in Table 4.1. The accuracies for these lexicons were calculated based on a manual evaluation. It assessed translations for terms from tweets in the development set. Since the phrasebook has a coverage comparable to Wikipedia over

[1] We do not use any form of word-sense disambiguation (WSD) to choose the best synset for the specific context; adding WSD to this process could tested in future work on emotion recognition through this process.

[2] This phrasebook was automatically generated with a machine translation (MT) system by another member of our team, Robert Kwiatkowski. Details on how this system worked can be found in Blevins et al. (2016).

Resource	Coverage	Accuracy
Wiktionary	47.7%	45.1%
Phrasebook	43.6%	83.2%

Table 4.1: Comparision of two lexicons on the task of mapping from nonstandard English of our dataset to Standard English.

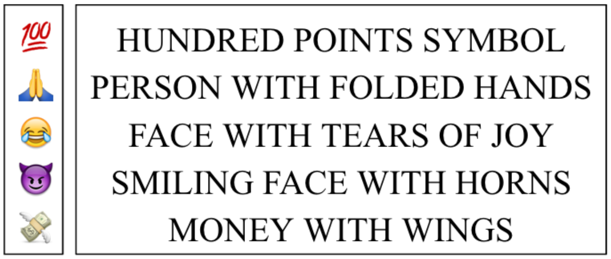


Figure 4.1: The most common emojis in our dataset and their unabbreviated descriptions.

the non-standard tokens, and a much higher accuracy, we use the phrasebook for mapping to Standard English in the final system.

Many of the tweets in the dataset contain emojis in addition to the other language features described. Emojis are Unicode symbols that are popular for online communication, and they function similarly to the older “emoticon”, which is a facial expression depicted by punctuation. Emojis add to the emotion of the tweets they appear in and are frequent in our dataset; 12.6% of non-stopword tokens in our data are emojis. Since emojis clearly contribute the overall emotional content of a tweet, we looked for ways to automatically ascertain a DAL-style emotion score of individual emojis, as this would allow them to act like any other token in the tweet when getting the final emotion score of the tweet.

In order to achieve this, we use a technique similar to the one for translation of non-standard tokens. We use a lexicon that maps each emoji to a representative English word or phrase. Our Emoji Lexicon uses abbreviated versions of the “names”, or descriptions given by the Unicode Consortium, as the definition of its respective emoji. Examples of these descriptions for the five most common emojis in our dataset are shown in Figure 4.1 (Blevins et al. 2016). We use this lexicon to obtain a “translation” of each emoji analogously to how non-standard words are handled.

The techniques described above were combined together into a emotion scoring system that scored each tweet in the following manner. First, the tweet was preprocessed in order to remove any stopwords or other tokens that do not contribute to the emotional meaning of the tweet; these included tokens such URLs and Twitter handles. For each nonstandard token found in the tweet, we search a translation lexicon (made up of the MT-generated phrasebook and the Emoji Lexicon) to obtain a Standard English translation. When a translation is found for the token, it is given to the DAL system described above to obtain an emotion score; any words in the DAL or Wordnet skip the translation step and are given directly to the DAL system.

Once the emotion scores for all tokens in the tweet are obtained, the scores are combined to represent the overall emotion of the tweet. We tested a number of different methods for combining the individual scores into a tweet-level representation. One method was to average the dimensions over all words in the tweet to obtain an "average pleasantness score", "average activation score", and "average imagery score" for each tweet. A second method we looked at was to get a score across the three dimensions for each tweet based on the deviation from the average value of that dimension: for a given dimension d and tweet T of length n, $\text{score}_d(T) = \frac{1}{n}\sum_{t \in T}(\text{score}_d(t) - \mu_d)$. Finally, we tried using the minimum and maximum scores across all tokens in the tweet, which gave six scores (a min and max for each dimension) per tweet. We found that the best results were obtained during the development process with the min/max scores.

4.2 Feature Selection

We trained our models using a number of features computed from the tweets in the dataset. Standard features that were used include unigrams and bigrams; for these n-grams, emojis are treated as regular tokens. This means that a single emoji is a unigram, and behaves analogously in bigrams extracted from the data. (However, the emojis are treated differently when extracting emotion scores, as described below.) Additionally, for unigram features, Twitter handles are mapped to a common token, and URLs are handled in the same manner.

Another feature we used to train this model is predicted POS tags. For part-of-speech tags, we experiments with both POS unigrams and bigrams. Like other NLP tasks discussed, the language of the dataset complicated the task of POS tagging. Therefore, the predicted tags were obtained from a POS tagger trained specifically for the domain of our dataset; this tagger was developed by another member of our team, Robert Kwiatkowski.

This was accomplished by performing domain adaptation on the CMU Oct27 dataset. The POS tagger was then trained on the adapted CMU data and our data, which was manually annotated in order to train the tagger (Owoputi et al. 2013). For these tags, emojis are again treated as regular tokens, that correspond to 'E' (the emoticon tag for the CMU Twitter tagger that we expanded to also cover emojis). This POS tagger had an accuracy of of 89.8% on our development set and 81.5% on the test set. This meant that it significantly outperformed both the CMU Twokenizer and the Stanford POS tagger for this specific data ($p < 0.0001$) (Owoputi et al. (2013); Toutanova et al. (2003)). More details on this process are given in Blevins et al. (2016).

We also provide the emotion scores computed for each tweets as features to the classifier. We found the score for each tweet by using, for each dimension (pleasantness, activation, and imagery), the minimum and maximum scores found across all tokens in that tweet. The process by which we obtain these scores and chose how to combine them into a tweet-level representation of emotion is detailed in the previous section.

4.3 System Design and Development

We experiment with a number of supervised classification systems to predict which tweets are aggressive or demonstrate loss. All of our systems are built from Support Vector Machines (SVM), as we found that they performed best on this dataset (Cortese and Vapnik 1995). A number are unmodified SVMs that perform binary classification; these are ternary classification on the full dataset (TCF) and binary classification on the aggression-loss subset (BCS). With the TCF classifiers we predict aggression versus the rest of the dataset, loss versus the rest of the dataset, and a combined label that looks for loss or aggression tweets out of the full dataset. For the BCS classifiers, we create a subset of the data by selecting all the tweets in our Gakirah dataset that contain tweets manually labeled as “aggression” or “loss”. We then train our BCS classifiers on this subset.

We also implemented an additional model, which we call a “cascading classifier” (CC). It uses two SVM models. One model is trained to find all aggression and loss tweets, and is the same model as the aggression+loss task for the TCF. This automatically generates a aggression/loss subset; then, the relevant label (loss or aggression, depending on the task) is chosen from the subset by a second SVM model. These second models are identical to the BCS on loss and aggression.

There were a number of motivations for designing the cascading classifier. In our preliminary work, we found that we achieved markedly better results when we worked with the subset of data that contained only loss or aggression tweets. However, this subset is not a realistic dataset; any data gathered from the real world will undoubtedly contain miscellaneous tweets that are not related to grief or aggression. We therefore built the CC in an attempt to recreate this subset with a realistic dataset.

During the development process, we also choose which subset of the potential features to include in the final model presented for evaluation. All of the feature types (ngrams, POS tags, and emotion scores) are helpful for the final models. Table 4.2 shows the results of an ablation study on the development set for our final models; the last row for each experiment/label pair lists the full feature set used in evaluation (Blevins et al. 2016). Here we only show results for the features that improve performance (by themselves) over the unigram baseline.

This study found a number of interesting associations between features and tasks. Bigrams were found to be helpful for predicting aggression (including the task of predicting if

Experiment	Label	Features	F-measure
TCF	**Aggression**	unigrams (baseline)	0.609
		unigrams, bigrams	0.674
		unigrams, POS-unigrams	0.674
		unigrams, emotion score	0.659
		unigrams, bigrams, POS-unigrams, emotion score	0.741
TCF	**Loss**	unigrams (baseline)	0.756
		unigrams, POS-bigrams	0.818
TCF	**Aggression + Loss**	unigrams (baseline)	0.727
		unigrams, bigrams	0.738
		unigrams, POS-bigrams	0.812
		unigrams, bigrams, POS-bigrams	0.821
BCS	**Aggression**	unigrams (baseline)	0.866
		unigrams, bigrams	0.884
		unigrams, emotion score	0.914
		unigrams, bigrams, emotion score	0.926
BCS	**Loss**	unigrams (baseline)	0.708
		unigrams, POS-unigrams	0.766
		unigrams, emotion score	0.723
		unigrams, POS-unigrams, emotion score	0.800

Table 4.2: A breakdown of the impact of the feature sets for each experiment. The first line given for each experiment and label is the unigram baseline, and the last line is the full feature set.

a tweet demonstrated aggression or loss in the full dataset). POS tags (either using unigram or bigram representations) were useful for most experiments, but they did not help with classifying aggression on the aggression/loss subset data. Emotion scores were useful for most experiments as well; the only exceptions were cases of predicting loss (including the task of predicting both loss and aggression) on the full dataset.

4.4 Evaluation

The results of these experiments on an held-out test set are shown in Table 4.3 (Blevins et al. 2016). All experiment abbreviations are defined in the previous section, and the averages listed for the TCF and CC classifiers are macro-averages over the listed labels. Overall, our models showed better recall performance than precision. The BCS classifiers performed much better than the other models; however, they are trained on an artificially constructed dataset. We therefore tried to recreate this performance on a real-world dataset using the cascading classifiers. This attempt was rather successful; the improvement the CCs gained over the

Experiment	Label	Precision	Recall	F-measure
TCF	**Aggression**	0.525	0.600	0.560
	Baseline (unigrams)	0.462	0.514	0.486
TCF	**Loss**	0.500	0.625	0.556
	Baseline (unigrams)	0.500	0.688	0.578
TCF	Average of **Aggression** and **Loss**	0.513	0.613	0.558
TCF	**Aggression** or **Loss**	0.588	0.800	0.678
CC	**Aggression**	0.471	0.923	0.623
CC	**Loss**	0.483	0.933	0.636
CC	Average of **Aggression** and **Loss**	0.477	0.928	0.630
BCS	**Aggression**	0.868	0.943	0.904
	Baseline (unigrams)	0.906	0.829	0.866
BCS	**Loss**	0.750	0.938	0.833
	Baseline (unigrams)	0.813	0.813	0.813

Table 4.3: Experimental Results on the test set.

TCFs was statistically significant (using randomization), with p = 0.023 for aggression and p = 0.039 for loss.

4.5 Discussion

Our system works on the novel task of predicting which tweets (and social media interactions in general) are likely to escalate to violence. It is a case study, focusing on the language of one person, that is a stepping stone towards future work in this area. Our focus on one user's data is thus beneficial, because it allows us to make the most of our limited dataset by focusing on the specific language of one person and build a coherent system that acts as a prototype.

One of the contributions this study makes is addressing the issue to using common NLP techniques on a non-standard, low-resource dialect of English. We find that the language of our data has enough similarities to Standard English (and the recently studied "social media language") that we can use a number of existing NLP datasets, namely the Dictionary of Affect in Language (DAL) and CMU Twitter Corpus for POS tagging, as long as we incorporate some adaptation techniques. These account for the ways in which the language of our data varies from Standard English and give us much better performance over use of standard NLP tools.

The high recall of our models is also important. The end goal of our work on this topic is to design a system that can help with the work already being done by non-profits such as CureViolence (Chapter 1). Specifically, we would like to enhance their work by narrowing down the amount of data a person needs to look at by highlighting the most high-risk posts on social media. However, this is a high-stakes task and we don't want to miss any threatening tweets; the user can judge if a post is truly threatening or not. Thus, our system has a high recall that catches almost all of the relevant tweets in order to better enhance the existing work done by nonprofits.

Future work on the task of threat classification on social media will be to generalize this model with a larger, more varied dataset. The goals with a more robust dataset would be to scale this system to work on a real world dataset while also improving its accuracy. There has also been recent success in other studies with the use of character ngrams for document classification, especially with smaller dataset.[3] Due to the amount of data considered here, one possible improvement on the feature set could be a character ngram language model, which may be less sparse than the token-level one considered for our classifiers.

[3] For example, in Weissenbacher et al. (2016), the character ngrams were found to be the most informative feature for their classifier by an ablation study.

CHAPTER 5

Semi-supervised Learning

This section discusses a semi-supervised learning approach to identifying tweets that express aggression. One of the major issues with a fully supervised approach to our task is that it does not take advantage of the large amount of unlabeled data we have available to us. One way we could integrate this data into our supervised model (chapter 4) would be to label this data. However, it is generally very expensive and time-consuming to label all or a significant portion of such a large dataset; the process our collaborators use for annotation (described in 3.2) is no exception.

Since it is not feasible to label all or most of the 1.6 million tweets in the Gang network dataset (described in Section 3.1) by hand, we altered our classification approach instead. We do this by building a semi-supervised system that can use both the labeled and unlabeled datasets to train a classification model. Our final semi-supervised system first uses a distant labeling algorithm to automatically label the unannotated dataset. The system then iteratively retrains a classifier using both the manually labeled and distantly labeled data, until a halting condition is met. In the next section, we discuss this process and alternate designs we considered while building the system.

By modifying our supervised classifier to a semi-supervised one, we hope to improve our performance on this task by learning additional information from our unlabeled dataset. We tested this hypothesis on a small test set and present the results in Section 5.2. The evaluation shows that we are only moderately successful on improving on the fully supervised baseline. However, some of the results indicate that this approach has potential with some minor changes with respect to the datasets used to train our models. We therefore conclude with a discussion of both the results and the system itself, considering how this experiment can be improved upon in future work.

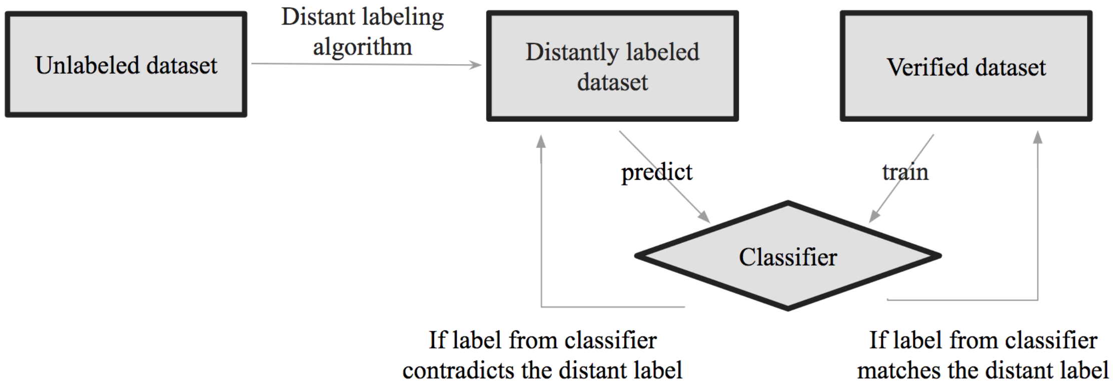

Unlabeled dataset
Distant labeling algorithm
Distantly labeled dataset
Verified dataset
predict
train
Classifier
If label from classifier contradicts the distant label
If label from classifier matches the distant label

Figure 5.1: Flowchart of the semi-supervised system

5.1 System Design

The overall goal of our semi-supervised model is this: to train a supervised learning model using both our small set of labeled data and the much larger set of unlabeled data. We achieve this by using a self-training paradigm similar to that found in Ouyang and McKeown (2015). The overall flow of how our system does this is shown in Figure 5.1. This system considers two datasets: the manually labeled Gakriah dataset (also referred to as the "seed dataset") and the unlabeled Gank Network dataset. The unlabeled data is first distantly labeled by an algorithm described below, and then repeatedly passed to a (retrained) classifier to verify these distant labels. If the distant label is verified by the classifier, it is added to the verified dataset (initially composed of only manually labeled tweets), and the classifier is retrained on the verified dataset. This process continues until the halting condition is met.

Thus, the first step in designing this system was to develop a robust distant labeling algorithm. We initialize this process by discovering a set of tokens, or "indicators", that are strongly associated with each class. This is a generalization of the distant labeling concept seen in Go et al. (2009) and Agarwal et al. (2011): in those papers, specific tokens (i.e., emoticons) were manually selected and associated with labels, whereas we choose our indicators automatically. These indicators are determined calculating the tf-idf score of the tokens in our seed dataset with respect to each class. We consider the same classes from the supervised learning task. This means that for a given token t, we have three scores: tf-idf$_{aggression}(t)$, tf-idf$_{loss}(t)$, and tf-idf$_{other}(t)$. For each class c, we select the top k tokens based on their tf-idf$_c$ scores and consider them to be "indicators" of class c.[1] If an indicator is in the top k for two or more classes, it is removed from all the classes it occurs in; this

[1] The number of indicators k for a given class is experimented with, and the results for a range of k values are presented in Section (5.2).

ensures that each indicator only indicates one class.

We assign the distant labels of "aggression", "loss", and "other" to each tweets in the unlabeled dataset based the occurrence of these "'indicators" in the tweet. A tweet is distantly assigned to a class c if it contains at least one indicator of c, and if it has more c indicators than indicators of any other class. Once a class is determined for a tweet, all indicators of that class are removed before being evaluated by or used to train a classifier. For example, if the distant labeling algorithm sees the tweet "Free my gang rip my gang I love my gang 🔓 🙏 💯" and considers the token "🔓" to be an indicator of loss, then it will assign the distant label of "loss" to that tweet and remove "🔓" from the original text. Removing these indicators is important, since we don't want the model to learn any biases we could introduce by heuristically labeling the data. If a tweet doesn't contain an indicator for any of the classes considered, it is removed from the dataset.

Once we have a tentative labeling for the Gang Network dataset from the distant labeling algorithm, we can use them to update our classifier. However, we do not use the unverified tf-idf labels to train the model. (If this approach was successful, we could simply label the entire dataset with the distant labeling algorithm and skip the ML models.) Instead, we iteratively train a classifier based on the distantly labeled data. On each iteration, an SVM model is trained by our "verified" dataset, which initially contains only the manually labeled tweets. We then predict labels for the "unverified", or distantly labeled, tweets. If the prediction from the SVM matches the assigned distant label for a tweet, that tweet is added to the "verified" dataset. However, if the predicted label and the distantly assigned label don't match, the tweet is put back into the "unverified" set.

This process iterates until a halting condition is met. We experimented with a number of potential conditions while designing this system. These included: stopping after a specific number of iterations, stopping when there was no update to the verified dataset, and stopping if less than a specific number of tweets were added to the verified set (a generalization of the previously stated condition). The final system uses a combination of these as its halting condition: it stops iterating if the verified dataset is not updated or if a maximum number of iterations is reached.

5.1.1 Alternative Designs

We considered a number of different techniques for incorporating our unlabeled data into a semi-supervised learning model. Specifically, we tested a number of different distant labeling algorithms in order to automatically label the unlabeled data. The method described in the previous section was selected after experimenting with these options, which are discussed here.

Both of the other options considered were nearest centroid clustering algorithms; they assigned a label to the tweets of the unlabeled dataset based on the clusters they are assigned to. The first clustering approach we tried was to cluster the labeled and unlabeled datasets together. Then the distant label of the unlabeled tweets in each cluster were then assigned by taking a majority vote over the labeled tweets that were located that cluster.

The next approach we tried to assign distant labels was to cluster only the labeled data together. Our goal here was to get (one or more) centroid representations of each label by clustering the data semantically. Then, for each unlabeled tweet, we found the nearest centroid and assigned the label of that centroid (based on the majority of tweets occurring in that cluster) to the unlabeled tweet. If the tweet was further than a threshold distance from any of the clusters, it was discarded from the considered data.

5.2 Evaluation

Evaluation is performed by training a SVM model on the verified dataset obtained from the self-training system, varying over a range of k values during the self-training process. We use the fully supervised model from Chapter 4 as a baseline for comparison. We evaluated on a different test set from the supervised experiment; we instead used a collection of tweets written by Gakirah's top-ten communicators.

We use a different test set from the supervised model experiment because the original Gakirah test set was notably more similar to the manually labeled data than the unlabeled data. Since we want to see if our model can learn to predict aggression and loss in general (instead of just on Gakirah's data), we choose a test set that was not drawn from Gakirah's tweets.[2] This means that the test set is not biased towards any of the models we evaluated.

Since we have such a small test set (47 tweets), we do not reserve any tweets for tuning, and kept the settings that were found to be best for supervised classification task for this evaluation. This favors the baselines and strengthens the case for instances where the semi-supervised model outperforms the baseline. We instead perform this primary evaluation and discuss methods of improving the datasets and process in the final section of the chapter.

We observe mixed results of this model on the test set when compared against the supervised model; the full results can be seen in Table 5.1. For the classification of the "loss" label and on classifying tweets as "loss" or "aggression", the fully supervised model outperforms any of the semi-supervised models. However, when predicting aggression, the semi-supervised

[2] The test set is also not drawn from the unlabeled dataset but from a totally different source: Gakirah's top-ten communicators.

Label	**Number of Indicators** k							
	Fully Supervised	1	2	3	4	5	6	7
Aggression	0.647	0.6	0.69	**0.692**	**0.692**	0.615	0.66	0.66
Loss	**0.839**	0.813	0.824	0.824	0.788	0.83	0.813	0.8
Loss+Aggression	**0.918**	0.862	0.862	0.862	0.857	0.833	0.852	0.833

Table 5.1: Semi-supervised evaluation results over a range of number of indicators k, compared to the supervised model baseline.

model outperformed the supervised baseline with five of the considered k settings; the models with k=4 and k=5 beat the baseline by 0.045 points. Overall, the f-scores for aggression are much lower than for the other two classification tasks; this could indicate that semi-supervised models improve performance on tasks where the fully supervised model performs poorly.

5.3 Discussion

There are a number of potential explanations for why the model only outperformed the supervised model on one class. One possibility is that the manually labeled data came from a different source than the unlabeled data. The manually labeled data differed from the unlabeled data both with respect to the users who wrote the tweets and the time frame represented in the data. In a similar vein, another data concern for this task is that the manually labeled data comes from a single individual. While this was good for developing the supervised classifier on a small dataset, it is hard to generalize from such a specific dataset.

One approach to try in future work in order to address both of these concerns is to manually annotate a random subset of the unlabeled data and use it as the initial verified dataset, instead of using an unrelated set. Using a subset of the larger unlabeled set of tweets would both diversify the "seed" dataset for the semi-supervised algorithm and make the "seed" dataset more similar to the unlabeled tweets we are trying to learn from. This would hopefully make it easier for the semi-supervised model to learn from the unlabeled data and make our model more accurate.

Another potential problem with our system is that we remove the indicators from the unlabeled tweets when assigning a distant label to them. This is necessary in order not to introduce bias into our model, but it also means the distantly labeled data are no longer natually occuring tweets. Rather, they are missing certain elements that may have been important to understanding if a given tweet was expressing loss, or aggression, or something else entirely. This is especially true since these indicators are closely correlated with their respective labels. This issue is therefore another to consider when attempting to improve this model in the future.

Semi-supervised learning systems such as this one are important to develop, because they are a potential solution for solving growing range of tasks (including the one addressed in this work) inexpensively. Getting large datasets is becoming easier with the quickly growing quantities of data on the Internet. However, it remains expensive and time-consuming to manually label data, even though large quantities of labeled data are often necessary to build a robust ML classifier. This is especially true for new tasks, which do not have data from past experiments to build upon. A system that can successfully learn from only partially labeled data would benefit our immediate goal of predicting tweets at a high risk of leading to violence as well as a broad range ML applications.

CHAPTER 6

Prediction Explanations

We implemented an explanation system that automatically generates a text that details why our classifier made a specific prediction. These explanations were designed with a user who is not computer scientist (and so, does not have an in depth understanding of how our classification system works) in mind. Because of this target user, the explanations focus less on the mechanics underlying the classification process, and more on the key features of the tweets that were important to the classifier for the prediction. Therefore, our explanations are designed to highlight the specific features from each tweet that were important to the prediction, in order to show the user what about a tweet indicated to the classifier to classify it a certain way.

In the next section, we discuss the design of our generation system. This system selects features important to the classifier for a specific prediction and then uses standard natural language generation techniques to create a text explanation of them. We then present the preliminary results of this explanation system. We also consider the importance of these explanations to the overall goal of our project, and consider ways in which these explanations could be helpful to users of our classification system.

6.1 Generation System Design

We present the design of our system to generate explanations for predictions over the "aggression" and "loss" classes on our dataset. This system was designed specifically for justifying predictions on this task to non-computer scientist users. Because of this, we abstract away specifics about the classification process (such as the weights for features in the trained SVM model) and instead focus on intuitively presenting the reasons the classifier made its prediction. The overall design of this system is based on standard NLG techniques and specifically work on justifying the predictions of ML classifiers by using these techniques (Reiter and Dale (1997); Biran and McKeown (2014)).

When presented with a tweet and corresponding prediction, the justifier first selects the content to be included in the explanation. Specifically, it determines which features from the tweet had the largest effect on the classifier's decision as well as the features that acted as strong evidence *against* the prediction. Since our classifier is a linear SVM model, the features we want to include in the explanation can be determined with the feature weight vector from the SVM and the value of the feature in the tweet.[1] The current settings for the system choose the three most important features for the prediction as well as the strongest counterevidence feature to be included in the justification (though the number of features included as evidence for a prediction could be experimented with to develop the most useful explanations). We provide features the both support and contradict the prediction, so that the user has a better understanding of the classification process for that tweet, and can see to most likely reasons the classifier would misclassify a tweet.

We also provide context for the features that are derived from the tweet during classification (as opposed to the the n-gram features that are occur directly in the tweet) in our explanations. This means we provide context for the POS unigrams, POS bigrams, and emotion scores that are highlighted by an explanation. This context maps a derived feature back to the token or set of tokens that feature is derived from, so that the user understands how these derived features were arrived at by the classifier from the original tweet. For example, the second example in Figure 6.1 contains a list of tokens from which we get the POS tag "verb". This aspect of our explanation thus demonstrates how the classifier gets the derived features from the original tweet.

Furthermore, we highlight the selected features that are particularly strong evidence for a certain class. While we do not present the user with the specific weights or feature values, we do emphasize if a feature deemed as strong evidence or counterevidence for a specific class across tweets. We consider this to be an important fact to convey to the user when justifying a prediction, since a feature that is highly weighted for a class is better evidence of a prediction than one that is not nearly as highly weighted, even if it is still the best evidence in the case of a particular tweet. Thus, our system includes a note that a feature is a particularly strong piece of evidence if it is in the top weighted group of features for that class. (This is currently with a weight cutoff of 0.5, but this should be determined empirically based on the model it is generating explanations for.) The information chosen in the content selection process (the selected features, associated context if relevant, and whether a feature is particularly strong evidence for a class) are then associated with sentence templates to

[1] The specific value that a feature has in the tweet depends on its type. Unigrams or bigrams from the tweet have a value equivalent to their occurrence frequency in the tweet; the emotion features' values are the numerical score for that dimension of the emotion score.

create “messages”. Each message contains a reference to a specific sentence template and a set of features to populate that template.

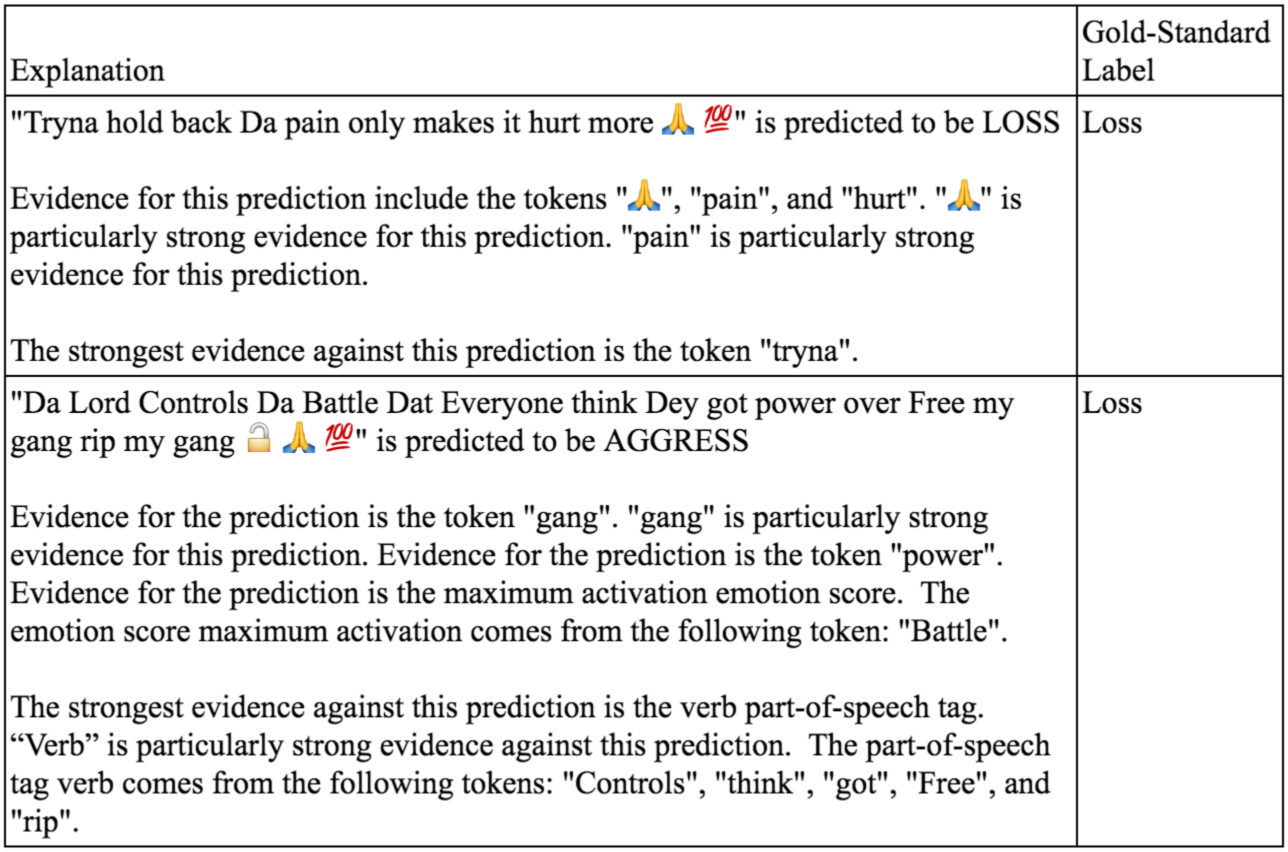
Once the content to include in the explanation is chosen, the system then arranges the messages into the desired ordering for the explanation. The messages are sorted such that the evidence supporting a particular prediction is first, followed (in a separate paragraph) by any counterevidence against the prediction. Within these two paragraphs, we group messages that reference the same entities. For example, if we have message #1 that says a specific feature f supports the predictions and a message #2 with supporting context about f, the discourse ordering step orders these messages to make sure message #1 appears right before message #2. The messages that indicate a feature is strong evidence for a specific class (based on its feature weight in the SVM) are dealt with analogously.

Finally, we render the ordered messages into a natural language explanation. Each of the entities (pieces of information chosen in the content selection step) are inserted into their respective template. The system then performs preprocessing on each of these new sentences in order to ensure the grammar and sentence structure is correct. Additional information (such as the prediction decision) are put into text and added to the explanation, which is then shown to the user.

6.2 Discussion

Figure 6.1 shows example output from the generation system. We present two examples: one for a correctly classified tweet (the top example) and another for an incorrectly classified tweet. Each message contains the tweet that is being labeled by the classifier, the classifier’s prediction, and the generated explanation of this prediction. Since this was built with a user who is not a computer scientist in mind, we abstract out the details of classification, with the goal of focusing on specific aspects of the tweet that the user can verify as good evidence (or not) of the predicted class.

Most modern natural language generation (NLG) tasks are done using ML techniques. However, there are currently no applicable datasets of model prediction explanations that we could use to develop a ML model; this is especially true for the specific task, predicting expressions of violence in social media data, with which our classifier deals. We instead use a manual approach to designing the system that generates these explanations. This means that we determine the parameters of the generation algorithm, such as the number of features included, by hand rather than with a data-driven approach. One potential avenue of research on explaining predictions of violence is thus to gather a dataset of manually



Explanation	Gold-Standard Label
"Tryna hold back Da pain only makes it hurt more 🙏 💯" is predicted to be LOSS Evidence for this prediction include the tokens "🙏", "pain", and "hurt". "🙏" is particularly strong evidence for this prediction. "pain" is particularly strong evidence for this prediction. The strongest evidence against this prediction is the token "tryna".	Loss
"Da Lord Controls Da Battle Dat Everyone think Dey got power over Free my gang rip my gang 🔓 🙏 💯" is predicted to be AGGRESS Evidence for the prediction is the token "gang". "gang" is particularly strong evidence for this prediction. Evidence for the prediction is the token "power". Evidence for the prediction is the maximum activation emotion score. The emotion score maximum activation comes from the following token: "Battle". The strongest evidence against this prediction is the verb part-of-speech tag. “Verb” is particularly strong evidence against this prediction. The part-of-speech tag verb comes from the following tokens: "Controls", "think", "got", "Free", and "rip".	Loss

Figure 6.1: Two example outputs from the justification system. The top one shows an explanation of tweet correctly predicted to express “loss” by the classifier; the bottom shows one incorrectly predicted to express “aggression”.

generated explanations of the classifier.[2] With that data, we could redesign the generation system to use an automatic approach of determining the best explanation design.

We have not performed an quantitative evaluation of the explanations. Due to the lack of applicable data, any automatic evaluation is out of the question. Additionally, manual annotation has not been attempted. Instead, these explanations have been discussed with our Social Work collaborators, who find that they would be confusing for an end user. One reason for this is lingering issues with the clarity and fluidity of the explanations. Beyond these issues, though, the explanations also suffer with respect to usefulness for our target end user because the method the classifier uses is significantly different from the one the human annotators use.

[2] If done manually, collecting this dataset would be expensive. One approach is to use descriptions written by the social work researchers during their annotation process; however, their explanations take into account more information that our classifier has access to. A method of gathering explanations that are relevant only to the features the classifier uses is an open question.

Therefore, one of the outcomes of generating these explanations is that they highlight the differences between the social work methods of labeling the tweets and the NLP models. Our process focuses on specific features or aspects of the tweet, albeit including some derived ones (specifically the emotion scores and POS tags). The social work approach is much more cohesive, because it takes into account much more context than our classification systems are able to and does not (usually) rely on single words or phrases. One approach that is currently being taken on this project, while not necessarily related to generation, is to incorporate more of the context that is used in the social work approach to classifying these tweets into our models, so that the NLP approach more closely resembles the social work one.

We have also found other unexpected applications of the explanations. The original goal of this system was to provide justifications for the predictions of our classifiers for predicting "loss" and "aggression" to the user. However, during the development of the generation system, we also discovered other, potentially more useful applications for it. One such application is for debugging of the system (especially if the system is debugged by a user unfamiliar with computer science). A user who is familiar with the domain of our data would be able to use this system to identify cases where certain features incorrectly act as evidence for a class. These explanations would be especially useful for that user in detecting bias in the classifier.

The second example in Figure 6.1 is a good example of how the explanations can be used to detect problems in the classifier. The strongest evidence from that tweet for the "aggression" label was the token "gang." However, since this data comes from a group that discusses their gangs often on social media, "gang" is often used in tweets that have nothing to do with violence (such as our example). Therefore, having this token act as particularly strong evidence for a prediction that a tweet is expressing aggression most likely worsens the reliability of the classifier overall. Perhaps more importantly, the word "gang" by itself as evidence of "aggression" for the classifier makes it less trust-worthy to a user who is familiar with the domain of our data. Making the classification process more transparent is thus the first step towards identifying these issues and improving on them in the classifier.

This application (of finding the issues with the classifier) can be combined with our initial goal of explaining how our models work to our social work collaborators, as well as the end users for the classification system. Since the explanations clarify what the models are doing when the predict a class for a tweet, they will make our collaboration easier, while simultaneously making it easier to correct any issues with our models. When used on real world data, they will also make it easier for the user to identify when the classifier is incorrect.

CHAPTER 7

Conclusion

In this thesis, we have discussed three contributions towards the task of predicting gang violence from social media posts. In this chapter we review the implications of this work and discuss next steps that can be taken to further the work presented here.

First, we developed a set of supervised classifiers to identify expressions of loss and aggression in gang data, as discussed in Chapter 4. These models were developed as a case study for future work on this topic, using a small labeled dataset from one Chicago gang member, Gakirah Barnes. We hypothesized that tweets in these two categories (but especially the aggressive tweets) are the social media posts that lead to incidents of gun violence. Our models were successful, and the best-performing supervised classifiers outperformed a baseline by 13.7 f-score points when predicting expressions of aggression in tweets and by 5.8 points when predicting expressions of loss.

We then built a second classification system, which is discussed in Chapter 5. This system uses a self-training approach, which allows us to train the model with a small labeled dataset and a larger, unlabeled one. This was done in order to utilize large number of unlabeled tweets we currently have without incurring the cost to fully annotate the dataset. While the results are not a significant improvement over those obtained from our best fully supervised models, we present a number of potential methods through with our semi-supervised learning system can be improved.

Finally, in Chapter 6 we present a system that, given a tweet and associated prediction from one of our classifiers, generates an explanation of that prediction. We designed these explanations in order to make our classifiers more accessible and trustworthy to a non-computer scientist user. However, we found that an unexpected (yet extremely useful) application for these explanations is that they make discovering problems and bias in our classifier much easier, especially when soliciting feedback from researchers or users outside of computer science.

There are also some limitations to this work. For example, we hypothesized that aggression and loss are the tweets that lead to violence; as of yet, though, we have not conducted an empirical study to back this up. Another general limitation of this task is the difficulty of applying NLP tools to the specific language seen in our datasets. Though we had success adapting NLP tools to this language, they still perform worse than state-of-the-art tools on Standard English data; this is mostly due to the much larger amount of data available to work with in Standard English. These limitations and those presented in earlier chapters can hopefully be addressed by future work on this project (discussed in the next paragraph).

We presented potential avenues of future work that applied specifically to each contribution in Sections 4.5, 5.3, and 6.2. Another more general task for future work on this project is to conduct an empirical study which examines if the "aggressive" and "loss" tweets we identify with our models correlate to incidents of real world violence. A study along these lines, or an alternative one that would identify which tweets are the ones leading to violence, would improve our focus and allow us train more helpful classifiers in the future.

In general, this thesis works towards developing a system that can facilitate community interventions before gun violence occurs. We strive to achieve this by focusing attention on the tweets that are most high-risk with our classifiers. Furthermore, by explaining why our models choose these tweets as high-risk, we aim to help the end user in deciding if they agree with the classifier's assessment.

References

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. “Sentiment analysis of Twitter data”. In: *LSM ’11 Proceedings of the Workshop on Languages in Social Media*. Portland, Oregon, 2011, pp. 30–38.

[2] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert MÃžller. “How to explain individual classification decisions”. In: *Journal of Machine Learning Research* 11.Jun (2010), pp. 1803–1831.

[3] Lakshika Balasuriya, Sanjaya Wijeratne, Derek Doran, and Amit Sheth. “Finding street gang members on twitter”. In: *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE. 2016, pp. 685–692.

[4] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. “semantic parsing on freebase from question-answer pairs”. In: *Proceedings of the Conference on Emperical Language Processing, EMNLP*. Seattle, Washington, 2013, pp. 1533–1544.

[5] Or Biran and Kathleen McKeown. “Justification narratives for individual classifications”. In: *Proceedings of the AutoML workshop at ICML*. 2014.

[6] Terra Blevins, Robert Kwiatkowski, Jamie Macbeth, Kathleen McKeown, Desmond Patton, and Owen Rambow. “Automatically Processing Tweets from Gang-Involved Youth: Towards Detecting Loss and Aggression”. In: *Proceedings of Coling 2016*. COLING, 2016.

[7] Avrim Blum and Tom Mitchell. “Combining labeled and unlabeled data with co-training”. In: *Proceedings of the eleventh annual conference on Computational learning theory*. ACM. 1998, pp. 92–100.

[8] Corinna Cortese and Vladimir Vapnik. “Support Vector Networks”. In: *Machine Learning* 20 (3 1995), pp. 273–297.

[9] Hal Daumé III. “Frustratingly Easy Domain Adaptation”. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 256–263. URL: http://www.aclweb.org/anthology/P07-1033.

[10] David Décary-Hétu and Carlo Morselli. “Gang Presence in Social Network Sites”. In: *International Journal of Cyber Criminology* 5.2 (2011), p. 876.

[11] Marek J Druzdzel. “Qualitiative verbal explanations in Bayesian belief networks”. In: *AISB QUARTERLY* (1996), pp. 43–54.

[12] Heather Ford. “Big Data and Small: Collaborations between ethnographers and data scientists”. In: *Big Data & Society* 1.2 (2014), p. 2053951714544337.

[13] Alec Go, Richa Bhayani, and Lei Huang. “Twitter sentiment classification using distant supervision”. In: (2009).

[14] Ben Green, Thibaut Horel, and Andrew V Papachristos. “Modeling contagion through social networks to explain and predict gunshot violence in Chicago, 2006 to 2014”. In: *JAMA internal medicine* (2017).

[15] Anna Jørgensen, Dirk Hovy, and Anders Søgaard. “Learning a POS tagger for AAVE-like language”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 1115–1120. URL: http://www.aclweb.org/anthology/N16-1130.

[16] Igor Kononenko et al. “An efficient explanation of individual classifications using game theory”. In: *Journal of Machine Learning Research* 11.Jan (2010), pp. 1–18.

[17] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. “Distant Supervision for Relation Extraction Without Labeled Data”. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*. ACL ’09. Suntec, Singapore: Association for Computational Linguistics, 2009, pp. 1003–1011. ISBN: 978-1-932432-46-6. URL: http://dl.acm.org/citation.cfm?id=1690219.1690287.

[18] Saif M Mohammad and Svetlana Kiritchenko. “Using hashtags to capture fine emotion categories from tweets”. In: *Computational Intelligence* 31.2 (2015), pp. 301–326.

[19] Peter Nickeas, Megan Crepeau, and Katherine Rosenberg-Douglas. “A violent Christmas in a violent year for Chicago: 11 killed, 50 wounded”. In: *The Chicago Tribune* (Dec. 2016).

[20] Jessica Ouyang and Kathleen McKeown. “Modeling Reportable Events as Turning Points in Narrative.” In: *EMNLP*. 2015, pp. 2149–2158.

[21] Olutobi Owoputi, Brendan O’Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. “Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters”. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 380–390. URL: http://www.aclweb.org/anthology/N13-1039.

[22] Desmond Patton, Robert Eschmann, and Dirk Butler. “Internet banging: New trends in social media, gang violence, masculinity and hip hop”. In: *Computers in Human Behavior* 29.5 (2013), A54–A59.

[23] Desmond Patton, Jeffrey Lane, Patrick Leonard, Jamie Macbeth, and Jocelyn Smith-Lee. “Gang violence on the digital street: Case study of a South Side Chicago gang member’s Twitter communication”. In: *New Media & Society* (2016). DOI: 10.1177/1461444815625949. eprint: http://nms.sagepub.com/content/early/2016/02/10/1461444815625949.full.pdf+html. URL: http://nms.sagepub.com/content/early/2016/02/10/1461444815625949.abstract.

[24] Desmond Patton, Kathleen McKeown, Owen Rambow, and Jamie Macbeth. “Using Natural Language Processing and Qualitative Analysis in Gang Violence: A Collaboration Between Social Work Researchers and Data Scientists”. In: *Proceedings of Bloomberg Data for Good Exchange*. 2016.

[25] Ellie Pavlick, Heng Ji, Xiaoman Pan, and Chris Callison-Burch. “The Gun Violence Database: A new task and data set for NLP”. In: *Proceedings of The 2016 Conference on Empirical Methods on Natural Language Processing (EMNLP)*. Austin, TX, Nov. 2016. URL: http://www.cis.upenn.edu/~ccb/publications/gun-violence-database.pdf.

[26] Matthew Purver and Stuart Battersby. “Experimenting with distant supervision for emotion classification”. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 2012, pp. 482–491.

[27] Steven M Radil, Colin Flint, and George E Tita. “Spatializing social networks: Using social network analysis to investigate geographies of gang rivalry, territoriality, and violence in Los Angeles”. In: *Annals of the Association of American Geographers* 100.2 (2010), pp. 307–326.

[28] Ehud Reiter and Robert Dale. “Building applied natural language generation systems”. In: *Natural Language Engineering* 3.01 (1997), pp. 57–87.

[29] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “”Why Should I Trust You?”: Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: ACM, 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778. URL: http://doi.acm.org/10.1145/2939672.2939778.

[30] Sara Rosenthal and Kathleen McKeown. “Sentiment Detection of Subjective Phrases in Social Media”. In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 478–482.

[31] Sara Rosenthal, Preslav Nakov, Alan Ritter, Veselin Stoyanov, Svetlana Kiritchenko, and Saif Mohammad. “Semeval-2015 Task 10: Sentiment Analysis in Twitter”. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver,CO, 2015.

[32] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network”. In: *Proceedings of HLT-NAACL 2003*. 2003, pp. 252–259.

[33] Davy Weissenbacher, Johnson A. Travis, Laura Wojtulewicz, Dueck Amylou, Dona Locke, Richard Caselli, and Graciela Gonzalez. “Towards Automatic Detection of Abnormal Cognitive Decline and Dementia Through Linguistic Analysis of Writing Samples”. In: *Proceedings of NAACL-HLT 2016*. San Diego, California, June 2016, pp. 1198–1207.

[34] Cynthia Whissell. “Using the Revised Dictionary of Affect in Language to Quantify the Emotional Undertones of Samples of Natural Language”. In: *Psychological Reports* 105 (2009), pp. 509–521.

[35] Sanjaya Wijeratne, Derek Doran, Amit Sheth, and Jack L Dustin. “Analyzing the social media footprint of street gangs”. In: *Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on*. IEEE. 2015, pp. 91–96.

[36] Yan Zhou, Murat Kantarcioglu, and Bhavani Thuraisingham. “Self-training with selection-by-rejection”. In: *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE. 2012, pp. 795–803.