8-2015

# Analyzing Domestic Abuse using Natural Language Processing on Social Media Data

J Nicolas Schrading

# Analyzing Domestic Abuse using Natural Language Processing on Social Media Data

by

**J Nicolas Schrading**

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Science
in Computer Engineering

Supervised by

Dr. Raymond Ptucha
Department of Computer Engineering
Kate Gleason College of Engineering
Rochester Institute of Technology
Rochester, New York
August 2015

Approved by:

---

Dr. Raymond Ptucha, Assistant Professor
*Thesis Advisor, Department of Computer Engineering*

---

Dr. Cecilia Ovesdotter Alm, Assistant Professor
*Co-advisor, College of Liberal Arts*

---

Dr. Christopher Homan, Associate Professor
*Committee Member, Department of Computer Science*

---

Dr. Shanchieh Jay Yang, Professor
*Committee Member, Department of Computer Engineering*

# Acknowledgments

Dr. Ray Ptucha, I can confidently state that without your guidance I would not have accomplished my most proud recent achievements. Your tireless dedication to supporting me (all while helping numerous others, running a successful research lab, and supporting your family) has impacted my future for the better. Thank you for everything. Sidenote: thank you Kodak for going bankrupt and giving RIT such a valuable faculty member!

Dr. Cissi Alm, I am so thankful that you reached out to me and provided me with opportunities for research. You have taught me an enormous amount and your meticulous edits of and great advice on papers and this thesis were invaluable. You have made my future much brighter. Thank you.

Dr. Chris Homan, without your support I would not have been able to write and submit my research papers, and I could not have attended their conferences. Your expertise in research, your connections, your guidance and advice, have all been invaluable to me. Thank you so much.

To everyone involved with this work, thank you for your time and effort.

My friends and family, thank you for putting up with me, for supporting me, for spending time with me unrelated to work, and for generally keeping me sane.

# Abstract

## Analyzing Domestic Abuse using Natural Language Processing on Social Media Data

**J Nicolas Schrading**

Social media and social networking play a major role in billions of lives. Publicly available posts on websites such as Twitter, Reddit, Tumblr, and Facebook can contain deeply personal accounts of the lives of users – and the crises they face. Health woes, family concerns, accounts of bullying, and any number of other issues that people face every day are detailed on a massive scale online. Utilizing natural language processing and machine learning techniques, these data can be analyzed to understand societal and public health issues. Expensive surveys need not be conducted with automatic understanding of social media data, allowing faster, cost-effective data collection and analysis that can shed light on sociologically important problems.

In this thesis, discussions of domestic abuse in social media are analyzed. The efficacy of classifiers that detect text discussing abuse is examined and computationally extracted characteristics of these texts are analyzed for a comprehensive view into the dynamics of abusive relationships. Analysis

reveals micro-narratives in reasons for staying in versus leaving abusive relationships, as well as the stakeholders and actions in these relationships. Findings are consistent across various methods, correspond to observations in clinical literature, and affirm the relevance of natural language processing techniques for exploring issues of social importance in social media.

# Contents

**Acknowledgments** . . . . . **ii**

**Abstract** . . . . . **iii**

**1 Introduction** . . . . . **1**

1.1 Motivation . . . . . 1

1.2 Hypotheses . . . . . 3

1.3 Methods . . . . . 4

1.4 Contributions . . . . . 5

**2 Previous Work** . . . . . **6**

2.1 Studies in Computational Social Science . . . . . 6

2.2 Properties of Domestic Abuse . . . . . 9

2.3 Deriving Useful Features from Unstructured Text . . . . . 13

2.3.1 Tokenization . . . . . 14

2.3.2 Morphology and Lemmatization . . . . . 15

2.3.3 Part of Speech Tagging . . . . . 17

2.3.4 Dependency Parsing . . . . . 19

2.3.5 Semantic Role Labeling . . . . . 22

2.3.6 Context-Based Word Representations . . . . . 25
2.3.7 Feature Vectors . . . . . 29
2.4 Machine Learning Algorithms . . . . . 31
2.4.1 Perceptron . . . . . 31
2.4.2 Support Vector Machines . . . . . 33
2.4.3 Neural Networks . . . . . 36
2.4.4 Long Short-Term Memory . . . . . 37
2.5 Dimensionality Reduction . . . . . 42
2.5.1 Latent Semantic Analysis . . . . . 43
2.5.2 Supervised Locality Preserving Projections . . . . . 43
2.6 Model Evaluation and Error Metrics . . . . . 44
2.6.1 $K$-fold Cross-validation . . . . . 46
2.6.2 Accuracy, Precision, Recall, and F1 Metrics . . . . . 46

**3 Datasets and Data Analytics . . . . . 49**
3.1 Twitter . . . . . 49
3.1.1 Preprocessing . . . . . 49
3.1.2 Extracting Gold Standard Labels . . . . . 50
3.1.3 Annotation Study . . . . . 52
3.1.4 Lexical Usage . . . . . 53
3.1.5 Analysis of Subject-Verb-Object Structures . . . . . 56
3.2 Reddit . . . . . 59
3.2.1 Preprocessing . . . . . 61

3.2.2 Corpus Characteristics . . . . . 62
3.2.3 Lexical Usage . . . . . 63
3.2.4 Semantic Role Attributes . . . . . 67
3.2.5 Analysis of Subject-Verb-Object Structures . . . . . 69

**4 Twitter Data Experiments . . . . . 72**
4.1 Classification Experiments . . . . . 72
4.1.1 SVO Features Only . . . . . 72
4.1.2 Full Feature Set . . . . . 74
4.2 Long Short-term Memory Experiments . . . . . 77
4.2.1 LSTM Training Set 1 . . . . . 78
4.2.2 LSTM Training Set 2 . . . . . 79

**5 Reddit Data Experiments . . . . . 82**
5.1 Classification Experiments . . . . . 82
5.1.1 Combinations of Features . . . . . 83
5.1.2 Comment Data Only . . . . . 84
5.1.3 Comment and Submission Predictors Cascaded . . . 86
5.1.4 Comment and Submission Text Combined . . . . . 87
5.1.5 Uneven Set of Submissions . . . . . 88
5.1.6 Testing on Completely Held Out Subreddits . . . . . 89
5.1.7 Dense Features Experiment . . . . . 92
5.2 Long Short-term Memory Experiments . . . . . 97

**6 Conclusion and Future Work . . . . . 99**

6.1 Limitations . . . . . 101

6.2 Future Work . . . . . 103

**Bibliography . . . . . 106**

**A Lemmatization Rules . . . . . 117**

**B Dependency Relations . . . . . 118**

**C Dense Feature Sets . . . . . 120**

C.1 Actors . . . . . 120

C.2 Acts . . . . . 120

C.3 Sympathy . . . . . 120

C.4 Abuser onto Victim Verbs . . . . . 121

C.5 Victim as Subject Verbs . . . . . 121

C.6 Top Features . . . . . 121

**D Experiment Summary . . . . . 122**

**E LSTM Output . . . . . 124**

# List of Tables

2.1 Google Universal Tagset for Part of Speech . . . . . 18
2.2 Example Confusion Matrix . . . . . 47
3.1 Twitter Annotation Confusion Matrix . . . . . 53
3.2 Basic Lexical Statistics on the Tokens and Types in the Twitter Dataset . . . . . 53
3.3 Top 10 Most Frequent Unigrams with their Respective Frequencies in the Twitter Dataset . . . . . 54
3.4 Top 10 Most Frequent Bigrams with their Respective Frequencies in the Twitter Dataset . . . . . 55
3.5 Top 10 Most Frequent Trigrams with their Respective Frequencies in the Twitter Dataset . . . . . 55
3.6 Discriminative Verbs for *Abuser onto Victim* and *Victim as Subject* Structures . . . . . 58
3.7 Collected Domestic Abuse and Control Subreddits . . . . . 60
3.8 Basic Descriptive Statistics of the Reddit Dataset . . . . . 62
3.9 Basic Lexical Statistics on the Tokens and Types of the Reddit Dataset . . . . . 64
3.10 Top 10 Most Frequent Unigrams with their Respective Frequencies in the Reddit Dataset . . . . . 65

3.11 Top 10 Most Frequent Bigrams with their Respective Frequencies in the Reddit Dataset . . . . . 65

3.12 Top 10 Most Frequent Trigrams with their Respective Frequencies in the Reddit Dataset . . . . . 66

3.13 Top 10 Unique Role Labels of the Abuse and Non-Abuse Classes with their Respective Frequencies in the Reddit Dataset 68

3.14 Top 10 Unique Predicates of the Abuse and Non-Abuse Classes with their Respective Frequencies in the Reddit Dataset . . . 68

3.15 Discriminative *Victim as Subject* Verbs in the Abuse Class . 70

3.16 Discriminative *Abuser onto Victim* Verbs in the Abuse Class 71

4.1 Top 10 SVO Features for both #WhyIStayed and #WhyILeft in the Twitter Dataset . . . . . 73

4.2 Feature and Preprocessing Parameter Ablation Results on the Twitter Dataset . . . . . 75

4.3 Top 10 Features with their Respective Linear SVM Weights Using Ngrams, Retweet Count, and Informal Register Replacement in the Twitter Dataset . . . . . 76

5.1 Classification Accuracies of Attempted Classifiers on the Reddit Data Using Combinations of Features . . . . . 83

5.2 Top 10 Features Based on Linear SVM Weights Using only Ngrams from Reddit Submissions . . . . . 84

5.3 Confusion Matrix for the Abuse/Non-Abuse Comment Text Classifier Trained on an Even Set of Data . . . . . 86

5.4 Confusion Matrix for the Abuse/Non-Abuse Classifier Trained on an Uneven Set of Data . . . . . 89

5.5 Collected Held-out Subreddits . . . . . 89

5.6 Feature Ablation Results on the Dense Feature Reddit Data . 95

5.7 Confusion Matrix for the Abuse/Non-Abuse Classifier Trained on an Uneven Set of Dense Data . . . . . 97

A.1 Rules for Lemmatizing Tokens . . . . . 117

B.1 Dependency Relation Types and their Descriptions . . . . . 119

D.1 List of Experiments with their Properties and Results . . . . 123

E.1 LSTM-generated Text of the Twitter Dataset . . . . . 124

E.2 LSTM-generated Text of the Reddit Dataset . . . . . 125

# List of Figures

2.1 Full Dependency Parse of an Example Sentence . . . . . 20
2.2 Original Skip-gram Network Architecture . . . . . 28
2.3 Contexts Derived From Dependency Parsing when Learning Word Vectors . . . . . 29
2.4 Multi-class Perceptron . . . . . 31
2.5 Example of an SVM Decision in 2-dimensional Space . . . . 34
2.6 Example of a 3-class *One-Versus-Rest* Problem . . . . . 35
2.7 Example of 4-layer Neural Network . . . . . 36
2.8 Long Short-Term Memory Block . . . . . 38
2.9 Comparison of the First 3 Dimensions Computed Using PCA and SLPP on a High-dimensional Facial Expression Recognition Task . . . . . 45

3.1 Tweet Count per Hour of the First Week of the Twitter Dataset 51

# Chapter 1

# Introduction

Social media websites, such as Twitter, have frequently been used as a source of information for predicting and characterizing various societal and health issues [11, 12, 15, 32, 69, 72]. It is clear that social media is an effective tool for gathering high volumes of data quickly, and its use in previous research is indicative of its effectiveness. However, analyzing the dynamics of abusive relationships using social media data is largely unexplored. In this thesis, new datasets discussing abuse are collected and developed. Computational methods are applied on these data to integrate quantitative results with findings from clinical literature for a qualitative understanding of the characteristics of domestic abuse.

## 1.1 Motivation

Globally, 30% of women 15 and older have experienced physical and/or sexual intimate partner violence at some point in their life [20]. While domestic abuse tends to have greater prevalence in low-income and non-western

countries, it is still endemic in regions like North America and Western Europe. In the United States, by an intimate partner, 9.4% of women have been raped, 16.9% of women and 8% of men have experienced sexual violence other than rape, and 24.3% of women and 13.8% of men have experienced severe physical violence [3]. This translates to an estimated economic cost of $5.8 billion for direct medical and mental health care services, along with lost productivity and reduced lifetime earnings [17]. Economic costs are calculable and provide concrete metrics for policy makers, but the physical and psychological effects felt by victims of domestic abuse are the true costs. Domestic abuse is the $12^{th}$ leading cause of years of life lost [52], and it contributes to health issues including frequent headaches, chronic pain, difficulty sleeping, anxiety, and depression [3].

The data used to calculate such statistics are often derived from costly, time-consuming, and potentially dangerous to participate in population-based surveys that primarily seek to obtain insight into the prevalence, consequences, and risk factors of domestic abuse. Due to the safety concerns of having victims of abuse answer survey questions while potentially being in the relationship in question, these surveys follow strict guidelines set by the World Health Organization [25]. Great care must be taken by the researchers to ensure the safety of the participants, and therefore the number of participants is often quite small [7]. One way to avoid the cost of largescale surveys while still maintaining appropriate research conditions is to leverage the abundance of data publicly available on the web. Such data

provide researchers with an opportunity to better understand domestic abuse in order to provide resources for victims and efficiently implement preventative measures. While the age groups 0-17 and 55+ will be significantly underrepresented based on user demographics of these websites [21, 22], the prevalence of intimate partner violence acts is most prominent between the ages of 18 and 24 [3], aligning with the most active social-media using ages.

## 1.2 Hypotheses

1. Using unstructured[1] social media input from relevant sources of language data, natural language processing (NLP) methods and machine learning classifiers can detect language related to domestic abuse.

2. Analysis of these classifiers, along with data inspection, can reveal meaningful structural and semantic, linguistic, and textual characteristics, including actions, stakeholders, and situations involved in abusive relationships.

[1] Unstructured is used here to mean that the data is not organized by strict columns and datatypes like the structured data of relational databases, nor organized by lax key-value pairs as in semi-structured formats like XML and JSON. While some data used in this work may be organized by title and body, it is largely free text.

## 1.3 Methods

Two corpora are utilized to study these hypotheses. The first, a Twitter dataset detailed in Section 3.1, contains tweets with the hashtags #WhyIStayed and #WhyILeft. Tweets with these hashtags give reasons for staying in and leaving abusive relationships, respectively. Analysis of the linguistic structures embedded in these tweets provides insight into the critical reasons that victims of domestic abuse report for choosing to stay or leave these relationships. Trained classifiers agree with these linguistic structures, adding evidence that these social media texts provide valuable insights into domestic abuse.

The second, a Reddit dataset detailed in Section 3.2, contains Reddit submissions from various domestic abuse forums (called subreddits) and control subreddits. A classifier is developed to detect submissions discussing domestic abuse. Analysis of the features used in detecting abuse discourse provides insight into the dynamics of abusive relationships.

This thesis will be evaluated and considered successful via two methods: achieving high scores in standard machine learning metrics and by matching the findings of clinical literature with the findings of corpus-driven statistical methods.

## 1.4 Contributions

1. A large, new corpus derived from Twitter, containing #WhyIStayed and #WhyILeft labeled datapoints.

2. A classifier that predicts whether a tweet contains a reason for staying in an abusive relationship or a reason for leaving.

3. A large, new corpus derived from Reddit, with domestic abuse and control submissions.[2]

4. A classifier that detects Reddit text discussing domestic abuse.

5. Comprehensive analyses of discussions of domestic abuse in these social media texts and comparisons to clinical literature.

[2]Control submissions contain text discussing anything or discussing potential side-effects of abuse like anxiety and anger, not specifically the dynamics of abuse.

# Chapter 2

# Previous Work

This thesis falls under the domain of *computational social science*, which involves a multi-disciplinary application of computational methods to study issues from the social sciences – methods including NLP, social network analysis, machine learning, and big data. The following is a brief description of select studies and topics that have been essential to completing this work.

## 2.1 Studies in Computational Social Science

Social media sites are an emerging source of data for public health research. These sites provide less intimidating and more accessible channels for reporting, collectively processing, and making sense of traumatic and stigmatizing experiences [32, 70]. Several previous works have studied public health issues intersecting with domestic abuse, including depression and post-traumatic stress disorder [11, 12, 15, 32]. Many researchers have focused on Twitter data, due to its prominent presence, accessibility, and the characteristics of tweets (short texts, timestamped, trend-associated properties like retweets, hashtags, and user mentions, and potentially geotags).

For example, in De Choudhury et al. [12], the authors examined a set of tweets to predict the onset of depression. Using Amazon Mechanical Turk, gold-standard labels of *depression* and *non-depression* were applied to Twitter users. The depressed users' tweets were collected for a year before the onset of their self-reported depression. Using various statistical and machine learning models, the significant features used in predicting the onset of depression were determined, contributing a radial basis function (RBF) support vector machine (SVM) classifier, with principal component analysis (PCA) dimensionality reduction, that achieved 70% classification accuracy with a precision of 0.74. Features included the presence of known depression terms in tweets, social network features, prevalence of medication terms, tweet volume over time, the frequency of 1st, 2nd, and 3rd person pronouns, linguistic inquiry and word count (LIWC) scores, and the prevalence of swear words. Using the model for finding depression-indicative tweets on a corpus of millions of tweets within the United States, the authors then created a *Social Media Depression Index* (SMDI) for calculating levels of depression within regions of the United States. They found high correlation with depression statistics reported by the Centers for Disease Control and Prevention (CDC) [11].

Related to the above study is an analysis of high and low *distress* tweets in the New York City area [32]. Distress was examined as it has been shown to be a key risk factor for suicide, and is observable in the writing of microblog users. An SVM trained on uni-, bi-, and trigrams appearing in their corpus

achieved a precision of 0.59 and a recall of 0.71 using expert-annotated tweets in predicting distressed versus non-distressed tweets. While a precision of 0.59 in binary prediction is low, erring on the side of caution with a high recall score is beneficial due to the goal of discovering at-risk individuals. This task was challenging, considering the difficulty of recognizing conceptually subjective *distress* from a few informal tweets.

Other researchers have focused on different health issues including Post-Traumatic Stress Disorder [15], early detection of epidemics [40, 69], and bullying tweets [72, 73]. These studies use ngram bag-of-word models as features, and attempt to improve upon them with additional feature engineering or further lexical or semantic features. Adding part of speech tags to ngrams is often attempted, as well as creating word classes via data inspection, using morphosyntactic features, and exploiting the sentiment of text instances. In Xu et al. [73], linear models with ngrams are recommended for their simplicity and high accuracy, though in Lamb et al. [40] word classes, Twitter-specific stylometry (retweet counts, hashtags, user mentions, and emoticons), and an indicator for phrases beginning with a verb were found to be helpful over ngrams on two different tasks.

Reddit has been studied less in this area, with work mainly focusing on mental health. In Pavalanathan and De Choudhury [56], a large number

of subreddits on the topic of mental health were identified and used to determine the differences in discourse between throwaway[1] and regular accounts. They observed almost 6 times more throwaway submissions in mental health subreddits over control subreddits, and found that throwaway accounts exhibit considerable disinhibition in discussing sensitive aspects of the self. This motivates the present work in analyzing Reddit submissions on domestic abuse, which can be assumed to have similar levels of throwaway accounts and discussion. Additionally, in a study by Balani and De Choudhury, the authors used standard ngram features, along with submission and author attributes to classify a submission as high or low self-disclosure with a perceptron classifier [2]. They achieved 78% accuracy, 0.74 precision, and 0.86 recall.

## 2.2 Properties of Domestic Abuse

There are several terms used to describe relationships in the area of domestic abuse, and some terms are used interchangeably in standard conversation. In this work the following definitions from Black et al. and the World Health Organization [3, 55] are used:

[1] Reddit does not require personally identifiable information when registering for an account. Often, users wishing to remain anonymous will create one-time accounts, called throwaway accounts, with anonymous names to submit a single (often personal or sensitive) submission or comment. In doing so, their history cannot be examined, preventing their real-life identity from being discovered.

1. **Abuse**: Physical violence, sexual violence, stalking, psychological aggression, controlling behavior, and/or neglect.
   (a) Physical violence: Acts such as slapping, hitting, kicking, and beating.
   (b) Sexual violence: Rape and sexual coercion.
   (c) Stalking: Unwanted obsessive attention that directly or indirectly communicates threats and places the victim in fear [8].
   (d) Psychological aggression: Insults, belittling, humiliation, intimidation, and threats.
   (e) Controlling behavior: Isolation from friends or family, monitoring, and restricting finances, education, or medical care.
   (f) Neglect: Failing to provide adequate care for a dependent [42].
2. **Domestic Abuse**: Abuse of an intimate partner or family member (children and elders especially).
3. **Domestic Violence**: The same as domestic abuse but sometimes restricting consideration to only the violent aspects of abuse.
4. **Intimate Partner Violence (IPV)**: Abuse specifically of an intimate partner.

In this thesis, domestic abuse is focused on to cast a wide net over the dynamics of abuse. However, by far the most prevalent in the data, and most

studied in general, is IPV. Additionally, rape and sexual violence in IPV is much more prevalent for women than men, however IPV in general, when considering psychological aggression, occurs in equal proportions (48.4% of women and 48.8% of men) [3]. Data has shown that significant negative pressure on men exists in reporting their victimization, which may affect reporting in social media. It is taboo and considered to be emasculating to report abuse for males. 84.2% of women disclosed their abuse to someone, while only 60.9% of men did [3]. Additionally, when men do disclose their abuse, they report that doing so is *very helpful* to them significantly less frequently than to women who disclose. Of these reports, only 21.1% of women and 5.6% of men reported their victimization to a doctor or nurse [3].

In addition to prevalence statistics, research has characterized factors associated with IPV. An ecological model proposed by Heise et al. [30] and expanded on by the World Healh Organization [55] suggests four different levels that increase the likelihood that a man will abuse his partner.

1. **Individual**: Experiencing abuse as a child; witnessing abuse as a child; having an absent father; low levels of education; alcoholism or drug addiction; personality disorders; acceptance of violence as a means of punishment or solving issues.

2. **Relationship**: Control of finances and decision making; marital conflict; economic stress; infidelity; disparity in education levels.

3. **Community**: Women's isolation; lack of support by peers, friends, or family; a prevalence of social groups that condone abuse in the community; high rates of poverty; weak legal consequences for IPV; high rates of violence in the community.

4. **Societal**: Socially accepted defined gender roles, with a link between *masculinity* and *toughness* or *dominance*; socially acceptable violence as a means to settle disputes or punish; a concept of ownership of women when married or dating.

Additionally, Heise et al. [30] suggest that women are often *not* passive victims of abuse. The abused actively attempt to maximize the safety of themselves and their children, while struggling to navigate the often insufficient support structures in secret. Researchers outline several reasons women may choose to stay in an abusive relationship: fear of retaliation; lack of financial independence; concern for their children; emotional dependence; lack of support from friends and family; fear of divorce and the potential to lose custody of their children; and/or an optimistic hope through love that their abuser will change. Children play a huge role in abusive relationships: even if the victim has been in the abusive relationship for years, many will leave after their children have grown. Many of these reasons, along with others, are discussed in work by Buel [6].

Heise et al. [30] also suggest several reasons that victims of abuse leave their relationships: an increase in violence that triggers a realization that their

abuser will not change, that it is only going to get worse, that the violence is going to affect their children, or that they may be killed. Additionally, an increase in support from friends, family, or society often allows the abused to leave.

In any case, the victim must frequently go through a difficult process to leave. It usually involves a cycle of denial, self-blame, and doubt, and many women go back to their abuser several times before leaving permanently [30].

These studies on the prevalence, risk factors, and dynamics of abuse are usually done with population-based surveys with high costs and risk factors for the participants and researchers. In this thesis, an alternative means of gathering and analyzing relevant data is pursued by applying computational models to the abundance of online social media.

## 2.3 Deriving Useful Features from Unstructured Text

Natural language data collected from social media websites are characterized by a lack of structure. The texts' organization and length are constrained only by conventions of the particular online venue and the writing style of the author. Text may contain Unicode symbols, emoticons, hyperlinks, website-specific tokens and markers, non-standard language, and a nearly limitless variability in lexicon. The inherent lack of structure befits the freedom of discussion present on the internet, but creates challenges for

NLP.

Analysis tools of NLP, some described below, provide opportunities to make meaningful inference from information embedded in free text and allows for extraction of useful features for machine learning classification tasks.

### 2.3.1 Tokenization

Individuals raised under normal conditions, who learn to read, understand that text can usually be decomposed into individual words, and that each word has a meaning (or multiple context-dependent meanings, as in the case of homographs). However, natural language is not black-and-white. It is characterized by ambiguity and variation across the language system. In writing, it is even occasionally difficult to determine whether a group of characters should be considered as a single word or multiple words. Take for example the multi-word expression *black-and-white*. It could be considered as a single multi-word or as three words: *black*, *and*, and *white*. Contractions also pose problems, e.g., *couldn't*. Should *couldn't* be considered a single word, or as two words: *could* and *n't* (or *not*)? There is not necessarily a right answer for these ambiguities, and approaches to word tokenization in NLP use different philosophies. Finally, sentences may contain tokens with non-alphabetic characters, such as symbols and emoticons. These contain meanings just like words, but are not thought of as words in the sense of a lexicon. As such, it is important to keep these characters,

which is why the automated process of breaking up strings into word-like units is called *tokenization*.

In English, tokenization is a fairly straightforward process. The tokenizer used in this work, implemented in spaCy [35], splits whitespace-delimited chunks by attempting to match special cases like contractions, slang, emoticons, and abbreviations. If none are found, a prefix is removed (if one exists), and the matching of special cases is performed again. If there is still no match, a suffix is removed (if one exists) and the process repeats [35]. This tokenizer is often referred to as a Penn Treebank tokenizer, since this is what was used to develop the Penn Treebank [46], but it is an improved version that handles data from the internet such as URLs and emoticons. One downside is that this tokenizer does not consider multi-word expressions as single tokens. For example, idioms like *kick the bucket* are split into individual tokens, and therefore steps to correct these errors using hard-coded rules, named entity recognizers, or higher-order ngrams should be considered. In this work, bigrams and trigrams are used to attempt to correct for these errors. In the end, a list of tokens is provided, and the work of deriving meaning from them can begin.

### 2.3.2 Morphology and Lemmatization

In linguistics, the study of words or affixes as meaningful building blocks is called morphology. A morpheme is the smallest unit of meaning in a word,

and does not necessarily have to be a valid word. For example the word *cats* contains the morphemes *cat* and *-s*. The morpheme *cat* is the furry, domesticated feline and is called a free morpheme because it can stand alone as a word, while *-s* is called a bound morpheme because it needs to attach to a free morpheme, here indicates plurality. Combining these two morphemes, English users understand that there are multiple furry, domesticated felines. The morpheme *-s* is also called an inflectional affix because it adds grammatical information to an existing word. Derivational bound affixes, like *-able*, attach to root words to create entirely new words, and potentially change the part of speech of a word.

A lexeme is a base form linked to a word sense and to the entire set of its potential forms. For example the lexeme *go* is linked to all of its other forms: *goes*, *went*, and *going* [19].

A common step in using natural language data is to *lemmatize* all tokens. The process of lemmatization converts tokens to their base dictionary form. In doing so, dimensionality reduction is achieved, which may help to improve applications. Lemmatizing can also introduce ambiguities, as the inflectional morphemes are removed.

The lemmatizer in this work, implemented in spaCy [35], takes a list of tokens, along with their parts of speech, and applies rules based on the endings of the tokens to convert them to their lemma (see Appendix A). The lemmatizer applies the rules in-order from first entry in the table to last entry, and

only converts a token if, after the affix is changed, it is still the same part of speech and a valid token in its dictionary. For tokens that undergo noticeable form changes, e.g. *was* to *be*, rules are followed [35]. For example, given the word *facing*, tagged as a verb, the lemmatizer will look up the verb rules in Table A.1. The verb *facing* does not end in *s*, *ies*, *es*, or *ed*, so those rules are skipped. It reaches the rule for the ending *ing*, which matches, causing it to be stripped and replaced by *e*. The potential new word is *face* which happens to be a known verb. Therefore the lemmatizer adds *face* as a potential lemma and continues. If instead the word was *meeting*, also tagged as a verb, the first *ing* rule would fail, resulting in the non-existant word *meete*. The lemmatizer therefore would move to the last *ing* rule, resulting in *meet*.

### 2.3.3 Part of Speech Tagging

In many parsing tasks, including lemmatization, part of speech (POS) tags are useful. These tags can also be helpful in determining meaning for individual tokens, because tokens can have multiple definitions depending on their context. For example, the homograph *recall* has distinct word senses - a thought or recollection (a noun) or calling back or revoking (a verb). By examining the context of the word, and the parts of speech of preceding words, these word senses can often be disambiguated and the correct part of speech can be assigned.

POS taggers assign different granularities of parts of speech, depending on

tag set and language. The POS tagger in use in this work, from spaCy [35], uses the Google Universal Tag Set [58], which is a coarse-grained set of tags to provide broad parts of speech to tokens of a universal set of languages. The POS tags, along with their description, are shown in Table 2.1.

Table 2.1: The part of speech tags available in this version of the Google Universal Tagset

| **POS Tag** | **Description** |
|---|---|
| NOUN | A noun |
| VERB | A verb |
| ADJ | An adjective |
| ADV | An adverb |
| PRON | A pronoun |
| DET | A determiner or article |
| ADP | A preposition or postposition |
| NUM | A numeral |
| CONJ | A conjunction |
| PRT | A particle |
| PUNCT | A punctuation mark |
| EOL | An end of line marker |
| NO_TAG | A temporary marker |
| X | Anything else |

An overview of the implementation of this POS Tagger is next. In broad strokes, the algorithm can be described as a multi-class averaged perceptron using greedy decoding (see Section 2.4.1). It utilizes the following features for training: the Brown cluster ID of the token (see Section 2.3.6), the token lowercased, the orthographic shape of the token[2], the first character of the token, the last 3 characters of the token, the POS tag (if this particular token can only ever have one POS tag, or if this is a context word that has been

[2]An orthographic transform of the original token. All characters from a-z =>x, A-Z =>X, 0-9 =>d. 4 or more of the same consecutive mapping are truncated to length 4. E.g., 42 =>dd, Golgafrinchans =>Xxxxx

tagged already), the token's lemma (if it has already been lemmatized), and if the token is alphanumeric, known punctuation, a URL, or numeric. It also contains features of the previous two tokens and following two tokens as context, each given the above descriptors (as applicable) as well. This algorithm is greedy because it does not utilize any sort of search algorithm to correct errors made in the previous tags it has predicted. While this may sound like a poor method, it actually works quite well. Many tokens can only ever take on one part of speech, and by tagging them immediately when they crop up, you start off with a high baseline accuracy. Rarely will the POS tagger make a mistake. A search method will potentially increase accuracy slightly, but will slow down performance significantly [33, 35].

### 2.3.4 Dependency Parsing

Dependency parsing is the technique in which words in a given text are parsed to find their underlying asymmetrical relations, called dependency relations. A dependency relation exists when a subordinate word (a dependent) depends on another word (a syntactic head or root). A root is an artificial word introduced for convenience which allows every word to have a syntactic head [54]. Figure 2.1 shows an example of a full dependency parse of the sentence, *Economic news had little effect on financial markets.* Arrows travel from the head to the dependent, and are labeled by the type of their relation, called a dependency type. Appendix B contains a description of the dependency types in use in the dependency parser utilized here [35].

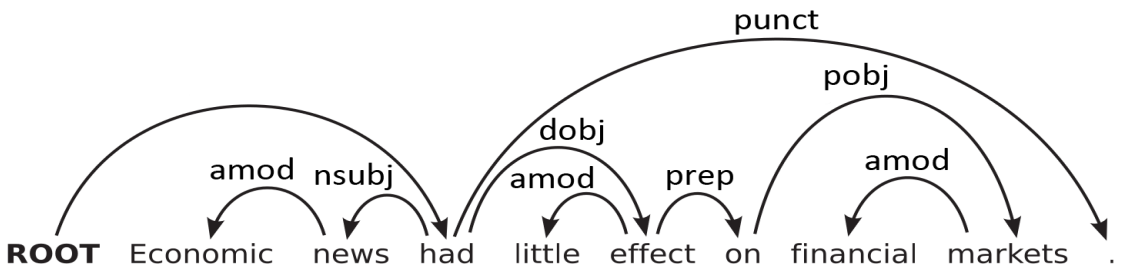



Figure 2.1: A full dependency parse of the sentence, *Economic news had little effect on financial markets.*. ROOT is added to allow *had* to have a syntactic head. Modified from Nivre [54].

This particular set of dependency types is from the ClearNLP project [9].

Dependency parses are typically represented as trees, where every word is a node, and each node can only have one parent node (except for root). A simplification to this tree when connecting the dependency relations is to define it as *projective*, meaning dependency relations cannot cross each other. Notice in Figure 2.1 that no lines cross. In some English sentences, long distance dependencies exist, which may cause dependency relations to cross if using a *non-projective* dependency parser. For example, in the sentence *A hearing is scheduled on the issue today.*, the prepositional phrase *on the issue today* is a dependent of the word *hearing*, so an arrow exists between the two. Additionally, a temporal dependency exists between *scheduled* and *today*. This dependency necessarily crosses the prepositional phrase dependency because *scheduled* comes before *on* and *today* is after *on*. More formally, if a dependency exists between words $a_1$ and $a_2$ and a different dependency exists between words $b_1$ and $b_2$, and $a_1$ comes before $b_1$ and $b_2$, then to be *projective*, $a_2$ must come after $b_2$. The dependency parser in use in

this work, implemented in spaCy [35], is a *transition-based* parser, meaning it is a state machine with a set of possible transitions to use to construct a dependency parse. Transition-based dependency parsers contain a machine learning classifier that take features from the current state to predict the best transition to use to go to the next state [34, 54]. The current state can be defined by the triple $c = (\sigma, \beta, A)$, where $\sigma$ is a stack of nodes (words), $\beta$ is a buffer of remaining nodes, and $A$ is a set of labeled dependency relations. The initial state has a stack with ROOT, all the words of the current instance in the buffer, and an empty $A$. The final state has an empty buffer and a full $A$. Only three possible transitions are defined:

1. **Shift**: Remove the first node from the buffer and place it on the stack.
2. **Right-Arc**: Create a dependency relation going from left to right by removing the top two nodes from the stack, modifying $A$, and placing the head node back on the stack.
3. **Left-Arc**: Create a dependency relation going from right to left by removing the top two nodes from the stack, modifying $A$, and placing the head node back on the stack.

As in its POS tagging algorithm (Section 2.3.3), spaCy's [35] dependency parser uses a perceptron learning algorithm and many features are considered. These include the first *n* words of the buffer, the top *m* words of the stack, the *p* leftmost and rightmost children of the node at the top of the

stack, and the $q$ leftmost children of the first word in the buffer. All of these nodes again have multiple attributes like orthographic shape and POS tag [54].

English dependency parsers tend to be trained on standard English datasets, leading to low reliability on social media corpora which contain informal and nonstandard forms. Some dependency parsers have been trained on specific domains, like Twitter, in order to overcome these problems. TweeboParser, an open-source tool developed at Carnegie Mellon [39], is capable of parsing tweets for dependencies accurately. The dependency parser in use in this thesis (from Honnibal's spaCy [35]) has been trained to handle noisy data from social media by adding data with corruptions (randomly swapped capitalization and replaced spaces with newline characters) and by training on data from different domains.

### 2.3.5 Semantic Role Labeling

Semantic role labeling (SRL) seeks to improve upon the syntactic knowledge derived from dependency parsing by providing semantic knowledge of the agents, actions, and patients within a given text instance. For each identified predicate within a text instance, the constituents involved with the predicate (e.g., an agent, patient, or instrument) are identified [63]. These constituents are labeled as A0, A1, and so on, where A0 represents arguments understood as agents, causers, or experiencers, A1 reflects patients,

and A2 is usually an indirect object like a semantic instrument or beneficiary [1]. The exact definitions are dependent on the specific predicate verb sense in any given instance [48]. Adjuncts within the text are also identified, e.g., AM-DIR for directions, AM-LOC for locations, AM-NEG for negations, etc. After performing SRL to identify the arguments, predicates, and their senses in a sentence, a lookup in Proposition Bank [47] with an argument number, predicate and sense can be performed. This will yield unique role labels for each argument. For example, in the sentence *Usually John agrees with Mary on everything*, an SRL program using PropBank as a source of role labels would provide this output: [AM-TMP— Usually] [A0— John] agrees [A2— with Mary] [A1— on everything], with A0 asigned the role *Agreer*, A1 *Proposition*, and A2 *Other entity agreeing* [48].

The SRL system in this work [63] uses four stages: *pruning*, *argument identification*, *argument classification*, and *inference*. The pruning and argument identification steps pick possible argument candidates for given verb predicates in a parsed sentence. The argument classification step independently labels the best identified arguments as A0, A1, etc. The final step, inference, then uses global linguistic and structural constraints to make sure that the independently labeled arguments are consistent with known language rules [63].

The pruning step's goal is to reduce training and run time by eliminating

constituents that cannot possibly be semantic arguments to a given predicate. It uses the heuristic rules defined in Xue and Palmer [74]. With a full parse of the sentence, and starting at a given predicate, it gathers the siblings of the predicate and considers them as candidates. If a sibling is a prepositional phrase, that phrase's siblings are gathered in a recursive manner. It then recursively collects the parent of the predicate until it hits the root.

The argument identification step takes these pruned candidates and applies a binary classifier to them, predicting whether they are good argument candidates or not. Features include the predicate's lemma, its POS tag, passive or active voice, the phrase type, the head word and its POS tag, the position of the constituent relative to the predicate, the full path from constituent to predicate, and the phrase structure around the predicate's parent. These features are described fully in Gildea and Jurafsky [26], per Punyakanok et al. [63].

The identified arguments are then passed to the argument classification step, which uses a multi-class classifier to apply type labels to the arguments. This classifier can apply *null* to an argument to indicate false positives from the previous steps. The features used are the same as in the argument identification step, with one additional feature: the sequential pattern of the noun phrases and the predicate.

The final step takes the confidence scores of the classifier's labels along with a list of known language constraints such as *arguments cannot overlap*

and *duplicate argument types for a single predicate verb cannot occur*. It applies a constrained integer linear optimization program to give an optimal solution to the possible labels, maximizing the linear sum of the confidence scores subject to the constraints [63].

### 2.3.6 Context-Based Word Representations

Tokens by themselves are not necessarily good features. Polysemy - the ability for a token to have multiple semantically linked senses, and synonymy - the ability for multiple tokens to hold the same or nearly the same meaning throw wrenches at the typical bag-of-words model. Such models make the incorrect assumption that a token has one meaning. Word representation or word clustering algorithms seek to correct this assumption by providing a mechanism for relating tokens syntactically and semantically.

One such an algorithm is the Brown clustering algorithm [5] (trained Brown clusters are built into spaCy [35]). This algorithm takes text as input, and provides a binary tree of output. The leaves of this binary tree are unique words ($w$) it has encountered in the text, the internal nodes are called clusters, and each word in the corpus can only be assigned to a single cluster. The algorithm works by starting with each word in an individual cluster, repeatedly merging clusters such that the merge (called a *clustering*, $C$) maximizes *quality*. This repeats while there are at least 2 clusters left.

$$Quality(C) = \sum_{c,c'} P(c, c') \log \frac{P(c, c')}{P(c)P(c')} + \sum_{w} P(w) \log P(w) \quad (2.1)$$

$$= I(C) - H \quad (2.2)$$

From equation 2.2, quality is the mutual information between adjacent clusters, $I(C)$, minus the entropy of the discovered word distribution, $H$ [45]. The counts, $n(\cdot)$, derived empirically from the text, are used to calculate the probabilities in Equation 2.1. $P(w) = \frac{n(w)}{n}$, $P(c) = \frac{n(c)}{n}$, $P(c, c') = \frac{n(c, c')}{n}$, where $n(w)$ is the number of times word *w* occurs in the text, $n(c)$ is the number of times a word *w* in cluster *c* appears in the text, and $n(c, c')$ is the number of times a bigram[3] $(w, w')$ with $w$ in cluster $c$ and $w'$ in cluster $c'$ occurs in the text.

Using Brown clusters as features, each leaf is assigned a unique bitstring which is related to the bitstring of its parent. Taking only the first *n* bits in a bit string will provide groups of words that occur in the same cluster. A Boolean vector with one column for each possible cluster can be created, just like in a term-document matrix. A column will contain a true value if a word in the given document occurs in the particular cluster assigned to that column. These Brown cluster features have been shown to improve performance in a variety of standard NLP tasks like dependency parsing

[3] A sequence of 2 words, with $w$ before $w'$.

[68].

Another type of word representation is called a *word embedding* or *word vector*. These vectors are low-dimensional representations (relative to the dimensionality of the known vocabulary) of the tokens. Surprising results from the word vectors trained by Mikolov et al. [50, 51] revealed that simple algebraic operations on these vectors can result in semantically similar words. For example, the vectors *King* - *Man* + *Woman* resulted in the closest vector being *Queen*. Not only was the understanding of gendered words included, but also the relative royal titles *King* and *Queen*.

The word vectors in use in this work are derived from an extension of the skip-gram model in Mikolov et al. [50, 51], introduced in Levy and Goldberg [44]. In the original skip-gram model, each word $w \in W$ has a vector $v_w \in \mathbb{R}^d$ and a surrounding context $c \in C$ with a vector $v_c \in \mathbb{R}^d$, where $W$ is the vocabulary of words, $C$ is the vocabulary of contexts, and $d$ is the vector dimensionality. The parameters to be learned, $v_w$ and $v_c$, are determined using a network model shown in Figure 2.2.

As can be seen in Figure 2.2, the learning objective is to take a word as input, pass it into a projection layer (which is implemented as a log-linear classifier), and predict the surrounding context words. In Levy and Goldberg [44], the word contexts based on position from the original word are dropped in favor of contexts derived from dependency parsing. For an input word $w$, gold-standard contexts are the combination of that word's modifiers

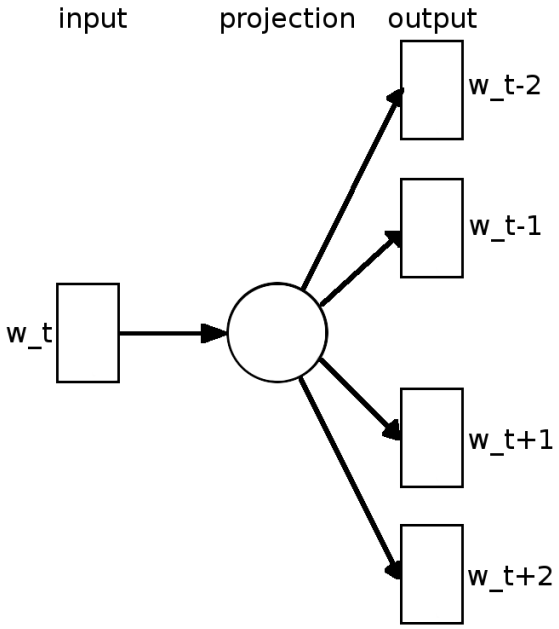



Figure 2.2: The original skip-gram network architecture to learn model parameters $v_w$ and $v_c$. Figure modified from Mikolov et al. [51].

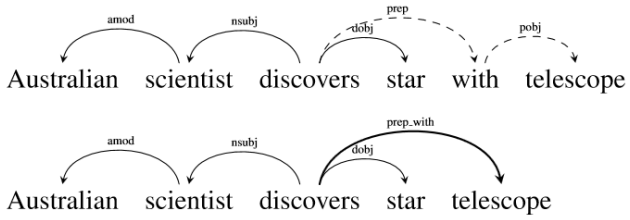
$m_1, ..., m_k$, head $h$, and dependency relations. An example of the derived contexts is shown in Figure 2.3.

The classifier is trained using stochastic gradient descent with the objective function in Equation 2.3:

$$\arg\max_{v_w, v_c}( \sum_{(w,c)\in D} \log \sigma(v_c \cdot v_w) + \sum_{(w,c)\in D'} \log \sigma(-v_c \cdot v_w)) \tag{2.3}$$

where $D$ is the dataset of $(w, c)$ pairs, $D'$ is a randomly generated dataset of $(w, c)$ pairs never seen in the dataset, and $\sigma(x) = \dfrac{1}{1+e^x}$.

By making this modification, empirical results show that the word vectors




| WORD | CONTEXTS |
|---|---|
| australian | scientist/amod$^{-1}$ |
| scientist | australian/amod, discovers/nsubj$^{-1}$ |
| discovers | scientist/nsubj, star/dobj, telescope/prep_with |
| star | discovers/dobj$^{-1}$ |
| telescope | discovers/prep_with$^{-1}$ |

Figure 2.3: Contexts derived from dependency parsing. Preposition relations are collapsed by taking the object of the preposition as the relation (e.g., *telescope* becomes a direct modifer of *discovers*). A -1 indicates a relationship going from head to dependent where the dependent is the word and the head is the context. Figure from Levy and Goldberg [44].

capture more semantic relations rather than domain relations. For example in the original model of Mikolov et al. [50, 51], the most similar words to *turing* are domain words related to Turing like *nondeterministic*, *computability*, and *finite-state*. In the dependency model, the most similar words to *turing* are other famous scientists related to Turing like *pauling*, *hotelling*, and *hamming* [44]. In this work, the word vectors of Levy and Goldberg [44] (trained on an English Wikipedia corpus) as included in spaCy [35] are used to compute cosine similarity between forum submissions and the comments within those submissions (see Section 5.1.2).

### 2.3.7 Feature Vectors

In text mining contexts, the data is usually represented as a document-term matrix, where each row represents a document, $d$ (a single text instance),

and each column represents a term, *t* (a token), that has been seen at least once in the corpus. This matrix is often extremely high-dimensional and sparse. The entries in this matrix can be Boolean, representing whether that term occurs in a specific document, frequency counts, or normalized scores calculated by a method called term frequency - inverse document frequency (TF*IDF). In Equation 2.4 $f(t, d)$ is the number of times a term *t* occurs in a specific document *d*. The calculation for TF*IDF is in Equation 2.6.

$$tf(t, d) = f(t, d) \tag{2.4}$$

$$idf(t, D) = \log \frac{N}{n_t} \tag{2.5}$$

$$tf * idf(t, d, D) = tf(t, d) \times idf(t, D) \tag{2.6}$$

where *D* is the total set of documents, *N* is the total number of documents in *D*, and $n_t$ is the number of documents in which term *t* appears. Intuitively, if a term *t* occurs many times in a document *d*, the term's corresponding concept is probably important in that document, and therefore the term should have a high value in that document – unless *t* is simply a common term across all documents. This is what $idf(t, D)$ corrects for. Some variations of TF*IDF log normalize $tf(t, d)$ such that it equals $1 + \log f(t, d)$. This

technique is called *sublinear* TF*IDF or *log-normalized* TF*IDF. Log normalization replaces the absolute, linearly increasing effect of $tf$ with a relative effect, increasing with powers of $e$. Non-normalized and sublinear TF*IDF were both used in experiments in this thesis. Some experiments benefited slightly from log normalization while others did not. When sublinear TF*IDF is used, it is noted in the experiment.

## 2.4 Machine Learning Algorithms

### 2.4.1 Perceptron

It has been shown in literature that the perceptron learning algorithm, while quite simple, is a powerful machine learning algorithm for natural language processing tasks like POS tagging [14]. It works efficiently with the sparse, high-dimensional vectors typical of NLP datasets. Additionally, it can be used in multi-class problems. A multi-class perceptron is shown in Figure 2.4.

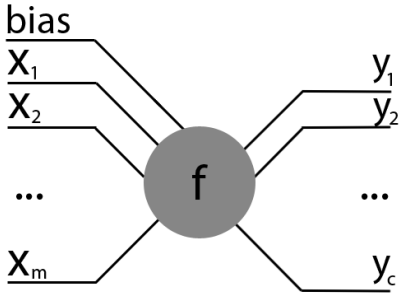



Figure 2.4: A multi-class perceptron.

The perceptron predicts labels given an input instance by taking the corresponding input vector of length *m*, $x \in \mathbb{R}^m$, with feature values (Boolean or a weighted representation like TF*IDF), operating on them with the function $f$, and taking the $\arg\max$. The function $f$ outputs a prediction $y \in \mathbb{R}^c$ for each of *c* classes, where $f = w \cdot x$ is a dot product of the input vector, $x$, and a weight vector, $w \in \mathbb{R}^m$, per class.

In training, the weight vector, $w$, is initialized to 0. A set of features extracted from training documents are passed to the perceptron in an online fashion, where it will use them to predict an output. The perceptron is error-driven: if the prediction is not the same as the ground-truth label for that document, the weights of the perceptron are penalized for the guessed class, and boosted for the ground-truth class. The penalty works by iterating through each feature, and for the guessed class decrementing the weight (usually by 1), while for the ground-truth class boosting the weight. Typically this process will be repeated for a certain number of iterations, randomly shuffling the training set on each iteration.

A problem with a regular perceptron is that it does not generalize well to different inputs. When training on varying training sets, wildly different weight models will be learned. Additionally, since it is error-driven, the weights only update on an incorrect guess, potentially ruining the weights that guessed correctly for previous inputs. In order to help prevent these issues, an averaged perceptron is used. Averaged perceptrons simply use

the average of the weight across all training iterations for a feature and class pair, rather than the final weight at the end of training [37].

### 2.4.2 Support Vector Machines

Support Vector Machines (SVMs) are powerful machine learning classifiers that are capable of accurately discriminating between two classes. When used in NLP, SVMs are often applied to classify data, due to their ability to work with sparse, high-dimensional vectors, their tuning capabilities, and their general performance.

SVMs seek to map input vectors to higher-dimensional spaces, such that a separating hyper-plane can split the two classes from each other with an optimal amount of distance between the support vectors and the hyperplane. Support vectors are the data points that touch the separating hyperplane. The distance between a support vector and the hyperplane is called the margin, which is defined to be equal to 1. For linear SVMs the separating boundary is linear. Figure 2.5 is an example of an SVM splitting between two classes in 2-dimensional space [16].

Kernel modifications allow for nonlinear separating hyperplanes by modifying the kernel function, $k(\cdot)$. In addition, a parameter, C, allows for tolerable error in the number of support vectors and incorporation of a soft margin distance. This parameter emulates regularization, and is tunable through cross-validation. Generally, higher C values decrease tolerance for incorrect

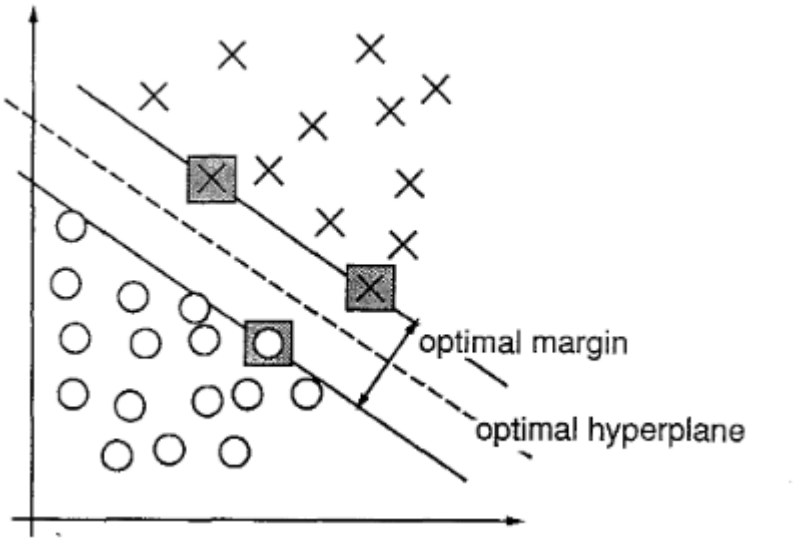



Figure 2.5: An example of an SVM decision in 2 dimensional space. The support vectors are in squares. Figure from Cortes and Vapnik [16].

classification, at the risk of overfitting to the training data [16].

In order to work, the SVM needs to solve for a parameter $\alpha$ in its Lagrangian dual objective function:

$$\sum_{i}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \tag{2.7}$$

where $y_i$ is the ground truth vector of a training example, $x_i$ is its input feature vector, and *n* is the number of training examples. After solving for $\alpha$ in an optimization problem, the prediction can be calculated with:

$$f(x) = \sum_{i}^{n} \alpha_i y_i k(x, x_i) \tag{2.8}$$

A nice property of SVMs is that the values in $\alpha$ will all be zero except for the

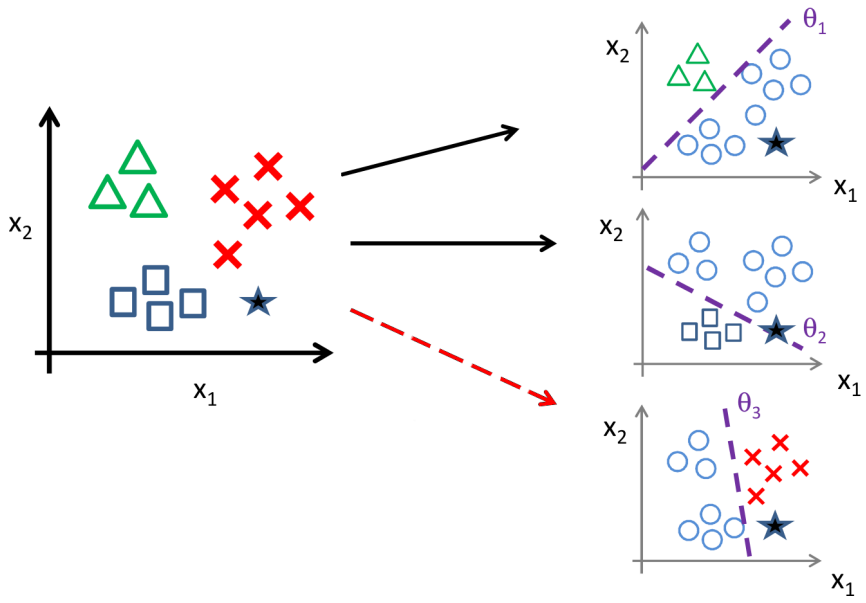



Figure 2.6: An example of a 3-class *one-versus-rest* problem. Each plot on the right is an individual SVM hyperplane, separating one class from the rest. The dotted arrow indicates the class label that is ultimately chosen for the never-before-seen instance indicated by the star. Figure modified from Ptucha [62].

support vectors, meaning only a small number of inner products between $x$ and the support vectors needs to be computed [53]. A limitation to SVMs is that a single SVM can only discriminate between 2 classes. The traditional way to work around this problem, and the way it is done in this thesis (per the implementation in Scikit-learn [57]), is to use a strategy called *one-versus-rest* (OVR) a.k.a *one-versus-all* (OVA). In OVR with $C$ classes to predict, $C$ individual SVMs are trained. Each SVM has one of the $C$ classes as its positive class and the rest as its negative class. When prediction is done, each of the $C$ SVMs are given the input and the SVM that places that input the largest positive distance away from its hyperplane is chosen. An illustration of OVR is in Figure 2.6.

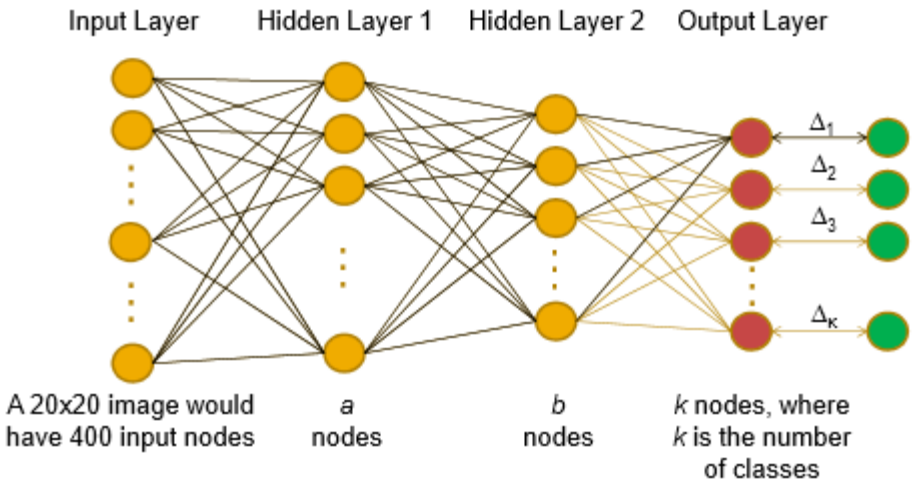



Figure 2.7: An example of 4-layer neural network. 1 input layer, 2 hidden layers, and an output layer. Figure from Ptucha [62].

### 2.4.3 Neural Networks

Like SVMs, neural networks are powerful classifiers. The basic idea behind a neural network is that layers of nodes are connected to one another (in the simplest case, all nodes in each layer are fully connected to the following layer's nodes), each node computes a summation between its inputs multiplied by its edge weights, and outputs a value, usually restricted by an activation function like a sigmoid. Figure 2.7 is an example of a 4-layer neural network. Each hidden layer tends to operate as a feature detector from the original inputs, becoming more abstract as the layers proceed into deeper levels. Currently, deep neural networks (those consisting of multiple hidden layers and multiple non-linear feature transformations) are dominating several machine-learning competitions.

In order to train, two processes are followed: *feed-forward* and *back propagation*. The network is initialized with random edge weights. In the feed-forward step, a training instance is chosen and propagated through the layers until the final output layer is reached. The output is compared to a ground truth label, and the differences between the prediction and the ground truth are sent backwards through the network in *back propagation*. This step updates the weights from the last layer to the first to make the network learn. A new instance is then chosen and the process repeats. Once every instance has been seen, the network is said to have completed an *epoch*. Training consists of multiple epochs, each time randomly shuffling the input data.

Neural networks usually take much longer to train than SVMs, and have many more parameters to tune. In addition, too many input nodes will lead to a prohibitively long training time. However, their results are often state-of-the-art. In order to work well on natural language processing data, with extremely high dimensions, dimensionality reduction techniques should be used on the data before training.

### 2.4.4 Long Short-Term Memory

Long Short-Term Memory (LSTM) models are gated recurrent neural network models that are designed to learn sequences from input. They are effective even with long-term dependencies between units, and have been

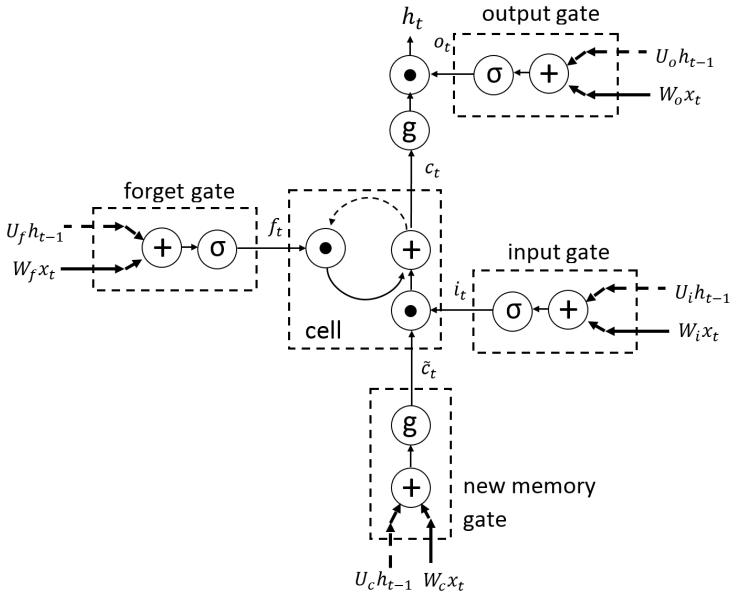



Figure 2.8: A single LSTM memory block. Dotted arrows represent time-delayed input, solid arrows represent current-time input. The function $g$ is the *tanh* function, the function $\sigma$ is the *sigmoid* function, and the function $\circ$ is the Hadamard product (element-wise multiplication). Figure modified from Greff et al. [28].

shown to perform at state of the art levels for many tasks, including handwriting recognition and generation, language modeling, and machine translation [28]. LSTMs modify the standard design of neural networks in several ways: they eliminate the strict requirement that neurons only connect to other neurons in succeeding layers (adding recurrence), convert the standard neuron into a more complex *memory cell*, and add non-linear gating units which serve to govern the information flowing out of and recursively flowing back into the cell [28]. The memory cell differentiates itself from a simple neuron by including the ability to remember its state over time; this coupled with gating units gives the LSTM the ability to recognize important long-term dependencies while simultaneously forgetting unimportant

collocations.

The original LSTM design was introduced by Hochreiter and Schmidhuber [31] in 1997, but it was not until 2005 that the most common design for LSTMs was described by Graves and Schmidhuber [27]. An excellent illustration of this design can be seen in Figure 1 of the recent largescale LSTM analysis paper by Greff et al. [28]. The LSTM in use in this thesis, as implemented by Karpathy [38] and taught by Socher's Stanford course *Deep Learning for Natural Language Processing* [66], modifies the original architecture by removing *peephole connections*. This architecture can be seen in Figure 2.8.

The equations defining an LSTM memory block are as follows:

$$i_t = \sigma(W_i x_t + U_i h_{t-1}) \tag{2.9}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1}) \tag{2.10}$$

$$\tilde{c}_t = tanh(W_c x_t + U_c h_{t-1}) \tag{2.11}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1}) \tag{2.12}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_{t-1} \tag{2.13}$$

$$h_t = o_t \circ tanh(c_t) \tag{2.14}$$

where $x_t$ is the input vector, $i_t$ is the output of the input gate, $f_t$ is the output of the forget gate, $\tilde{c_t}$ is the output of the new memory gate, $o_t$ is the output of the output gate, $c_t$ is the output of the cell (this becomes the memory of the cell in the next timestep), and $h_t$ is the output of the LSTM memory block (the hidden state). The matrices $W_i$, $W_f$, $W_c$, $W_o$, $U_i$, $U_f$, $U_c$, and $U_o$ are weight matrices to be learned for the input gate, forget gate, cell, and output gate respectively, where the $W$ matrices are rectangular matrices and the $U$ matrices are square recurrent weight matrices [28].

The intuitive understanding of the components in an LSTM memory block are described by Socher [66]:

1. **Input Gate**: Takes the input and the past hidden state to determine the importance of the current input as it effects the cell.

2. **Forget Gate**: Takes the input and the past hidden state and determines the usefulness of the previous cell output on the current cell.

3. **New Memory Gate**: Takes the input and the past hidden state to summarize the new input in light of the past context from $h_t$. Does not care about the importance of this new input – this is what the input gate

concerns itself with.

4. **Output Gate**: Determines what parts of the cell output $c_t$ need to be present in the new hidden state $h_t$ for the next timestep.

5. **Cell Output**: Takes advice from the forget gate to determine the usefulness of the previous memory ($c_{t-1}$) and advice from the input gate to determine the usefulness of the new memory ($\tilde{c_t}$) to produce a summation of the two, equaling the new memory ($c_t$).

The functionality above describes only how a *single* LSTM memory block works, analogous to a single neuron in a regular neural network. To create an LSTM which learns, hundreds of these blocks are combined in a single layer, with the hidden output, $h_t$, of one block feeding into the input of another. Further complexity (and learning power) is added by including further layers of LSTM memory blocks. The final output of LSTM memory blocks (or inputs from one layer to the next) are provided by calculating $y_t = W_y f(h_t)$, where $W_y$ is an output weight matrix to learn and $f(\cdot)$ is an activation function which can vary depending on use case.

The input, $x_t$, to an LSTM memory block differs depending on implementation and use-case. LSTMs can be word or character-based if using LSTMs for NLP. The LSTM used in this work, from Karpathy [38], is a character-level LSTM, which means that it takes as input a vector representing an individual character and predicts the most probable character given the current character and the LSTM's previous states. Training, therefore, is done

by taking an example sequence of characters, predicting the next character using the current weights in $W$ and $U$, calculating the difference between what was predicted and what should have been predicted, and backpropagating this difference to update the weights. Language generation can be performed after training, in which the LSTM is given a starting sequence of characters (or it calculates the most probable sequence to start with), and then generates new characters based on its own predictions in previous timesteps.

## 2.5 Dimensionality Reduction

Dimensionality reduction is an important step for many machine learning algorithms. Essentially, any dimensionality reduction algorithm seeks to find a representation of the original data $x_i \in \mathbb{R}^D$ such that the new representation $y_i \in \mathbb{R}^d$ has $d < D$. Principal component analysis (PCA) [61] and linear discriminant analysis (LDA[4]) [61] are two such common dimensionality reduction techniques, however they suffer from problems. PCA optimizes reconstruction error between the two representations and LDA optimizes for linear classification by separating distinct classes as much as possible. Both of these are desirable, but only one or the other can be applied, and both assume that the features lie in simple linear manifolds. Using a variant of PCA on sparse high-dimensional NLP data is called Latent Semantic

[4]Not to be confused with Latent Dirichlet Allocation [4].

Analysis (LSA) [41].

### 2.5.1 Latent Semantic Analysis

LSA, like PCA, seeks to find a lower-dimensional representation of the original data which minimizes reconstruction error. The returned lower-dimensional data can be thought of as meaning derived from linear combinations of different tokens.[5] Therefore, not only is the dimensionality reduced, but higher-order understanding of the association of tokens with documents is captured [23].

LSA works by applying Singular Value Decomposition (SVD) to the document-term matrix.[6] The key difference from PCA is that the original data is not converted to a covariance matrix. Instead, LSA decomposes the original document-term matrix, $X$, into three matrices, $U$, $\Sigma$, and $V$ such that $X = U\Sigma V^T$. $\Sigma$ contains the singular values in its diagonal, allowing the top $d$ largest singular values to be chosen, reducing dimensionality [18].

### 2.5.2 Supervised Locality Preserving Projections

Supervised Locality Preserving Projections (SLPP) is an alternative to PCA, LSA, LDA, and other techniques [61]. For SLPP, a fully-connected graph of all input points is constructed, with edge weights $0 \leq w_{ij} \leq 1$. The

[5] Tokens here refers to the tokens (i.e., terms) derived from a tokenizer as well as any sequence of these tokens (higher order ngrams) used.

[6] See Section 2.3.7 for a description of document-term matrices.

edge weight $w_{ij}$ is set to 1 when $x_i$ is a near neighbor in Euclidean distance to $x_j$, and 0 when it is far away. SLPP's goal is to find an alternative low dimensional representation of the data while preserving the neighborhood structure of the high-dimensional space. SLPP attempts to minimize the function:

$$\sum_{i,j} \|y_i - y_j\|^2 w_{ij} \tag{2.15}$$

where $y_i$ and $y_j$ are points in the new feature embedding. SLPP defines a neighbor as those points that share similar class labels. In this way, supervision can be added to the process. This method borrows concepts from PCA to combine supervised results with unsupervised results, avoiding overzealous dimensionality reduction [60]. Usually, $d \ll D$. Additionally, from Ptucha [59], SLPP generalizes to new points, and therefore usually works better than other manifold methods like Isomap [67] and locally linear embedding (LLE) [64]. Figure 2.9, from an experiment on facial expression recognition [61], shows SLPP versus PCA, clearly showing a better separation of the various facial expression classes in the SLPP-reduced space.

## 2.6 Model Evaluation and Error Metrics

Standard procedures for training and evaluating the performance of classifiers in this thesis are followed. When tuning model parameters, *k*-fold

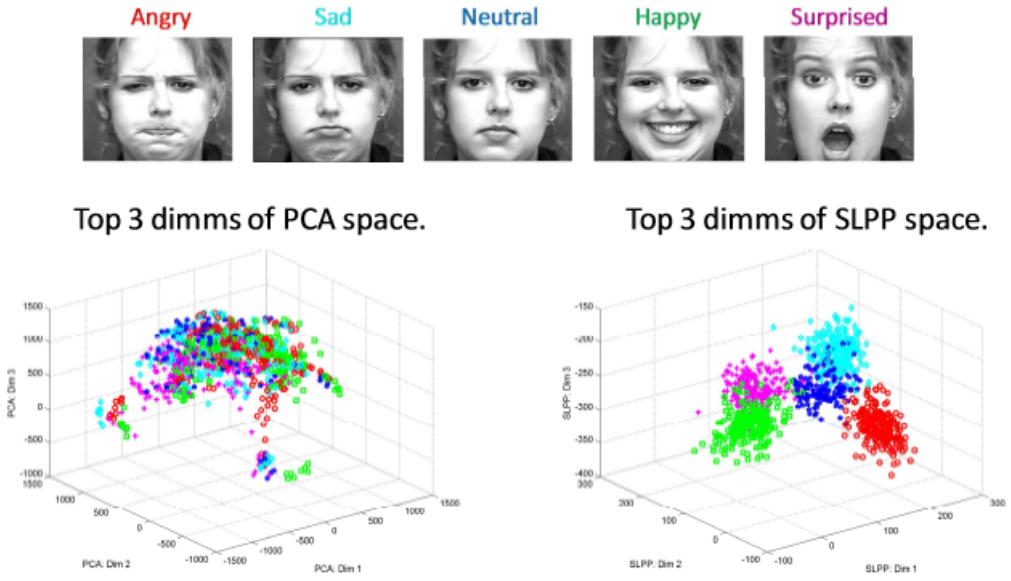



Figure 2.9: A comparison of the first 3 dimensions as computed using PCA and SLPP on a high-dimensional facial expression recognition task. Figure from Ptucha [61].

cross-validation is used, and the average metric (accuracy or F1) with standard deviation is reported. When enough data is available, a final held out testset is used to report final evaluation metrics with the chosen model parameters from cross-validation. This held out testset is evaluated by using all the folds to train a classifier with the best model parameters from the cross-validation set.[7] Following this procedure ensures that the model parameters are not unfairly tuned to the final testset, providing confidence in the ability of the classifier to generalize to unseen data points in the wild.

[7] In this work, when multiple experiments use the same set of data, they always have the same exact cross-validation devset and final held out testset so that their results can be compared. When datasets are expanded or modified, a new random split is used to create a different cross-validation set and testset.

### 2.6.1 *K*-fold Cross-validation

*K*-fold cross-validation is a method used to tune model parameters and evaluate the performance of machine learning classifiers. This method splits the dataset into $k$ partitions, called folds (usually $k = 5$ or $k = 10$), where $k - 1$ folds are used for training and 1 fold is used for testing. This training and testing step is done $k$ times, each time changing which fold is used for testing such that each fold is used for testing only once. The average metric (accuracy or F1) and standard deviation across all $k$ train/test splits is reported.

### 2.6.2 Accuracy, Precision, Recall, and F1 Metrics

In standard two-class problems, the model must take an unseen instance and predict whether it belongs to class 1 (the 'positive' class, e.g., *abuse*) or class 2 (the 'negative' class, e.g., *non-abuse*). If there are an even number of examples across both classes, then it suffices to report only the accuracy:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{N} \tag{2.16}$$

where $TP$ is the total number of *true positives*, $TN$ is the total number of *true negatives*, $FP$ is the total number of *false positives*, $FN$ is the total number of *false negatives*, and $N$ is the total number of instances. A true positive occurs when the classifier correctly predicts an instance as class 1

which in truth is class 1. Similarly, a true negative occurs when the classifier predicts that an instance which is labeled as class 2 in truth is class 2. A false positive is characterized by the classifier labeling as class 1 an instance which in truth is class 2, and a false negative is when the classifier labels an instance as class 2 when in truth it is class 1. These values can be easily visualized using a *confusion matrix*. For example, if the classifier is predicting whether a text instance is about abuse or not, the confusion matrix would look like Table 2.2.

Table 2.2: An example confusion matrix for the Abuse/Non-Abuse classifier.

| | | **Predicted Class** | |
|---|---|---|---|
| | | **Abuse (Positive)** | **Non-Abuse (Negative)** |
| **Actual Class** | **Abuse (Positive)** | TP | FN |
| | **Non-Abuse (Negative)** | FP | TN |

Ideally, the TP and TN values on the diagonal will be high, and the FN and FP values will be low (or 0 in a perfect case). If an imbalance exists between the two classes, then further metrics should be reported to avoid the problem of accuracy being skewed by the class with more instances. For example, if class 1 has only 10 instances while class 2 has 90 instances, and the classifier predicts all 90 class 2 instances correctly while misclassifying all 10 class 1 instances, the accuracy will still be $\frac{TP+TN}{N} = \frac{0+90}{100} = 0.9$.[8] Metrics that help avoid this bias are *precision* and *recall*.

[8]This is also why a strong baseline metric should be used to compare against. The simplest baseline in an even set of data is 50% accuracy – achieved by randomly guessing each instance.

$$precision = \frac{TP}{TP + FP} \tag{2.17}$$

$$recall = \frac{TP}{TP + FN} \tag{2.18}$$

Intuitively, precision is the ability of the classifier to not label as 'positive' an instance that is 'negative' (achieved with low $FP$), while recall is the ability of the classifier to retrieve all 'positive' instances (achieved with a low $FN$). Depending on the problem, high precision or high recall may be desired over the other. For example, in a hypothetical test used for detecting cancer in patients, high recall may be desired over precision so that every patient with cancer is found, at the expense of potentially more false positives. If this test decides that the patient has cancer, then a second test with high precision may be used, to determine if the original test was a false positive. If, however, both metrics are required to be high, a single metric which takes their harmonic mean, called *F1 Score*, can be used instead:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{2.19}$$

# Chapter 3

# Datasets and Data Analytics

Datasets are available from http://nicschrading.com/data/.

## 3.1 Twitter

In September of 2014, the fear of discussing abusive relationships broke down in response to the Ray Rice Assault Scandal[1], as thousands of Twitter users participated in a viral discussion of domestic abuse. The hashtags #WhyIStayed and #WhyILeft were utilized to denote reasons for staying in or leaving abusive relationships. A new large corpus of tweets with the hashtags #WhyIStayed or #WhyILeft was collected for this thesis.

### 3.1.1 Preprocessing

Spam tweets based on the usernames of prevalent spammers, as well as key spam hashtags[2] were removed. Additionally, tweets related to a key controversy, in which the Twitter account for DiGiorno Pizza (ignorant of

[1] See http://www.sbnation.com/nfl/2014/5/23/5744964/ray-rice-arrest-assault-statement-apology-ravens.
[2] Such as *#MTVEMA, #AppleWatch, #CMWorld.*

the trend's meaning) tweeted *#WhyIStayed You had pizza*[3] were removed. This resulted in over 57,000 unique tweets in the corpus.

Many tweets in the dataset were reflections on the trend itself or contained messages of support to the users sharing their stories, for example, *Not usually a fan of hashtag trends, but #WhyIStayed is incredibly powerful. #NFL #RayRice*.[4] These tweets, here denoted *meta-tweets*, were often retweeted, but they rarely contained reasons for staying or leaving (the interest of the study), so they were filtered out by keyword.[5] In section 3.1.3 the remaining instances are empirically explored. For a generated example of what this set of data looks like, see Table E.1 in Appendix E.

### 3.1.2 Extracting Gold Standard Labels

Typically, users provided reasons for staying and leaving, with the reasons prefixed by or appended with the hashtags #WhyIStayed or #WhyILeft as in this example: *#WhyIStayed because he told me no one else would love me. #WhyILeft because I gained the courage to love myself.* Regular expressions matched these structures and for tweets marked by both tags, split them into multiple instances, labeled with their respective tag. If the tweet contained only one of the target hashtags, the instance was labeled with that hashtag. If the tweet contained both hashtags but did not match with any of the regular

[3]Removed by keywords *pizza* and *digiorno*.

[4]Illustrative tweet examples were anonymized and sensitive content was purposefully attempted to be minimized.

[5]Including *janay/ray rice, football, tweets, trend, video, etc.*

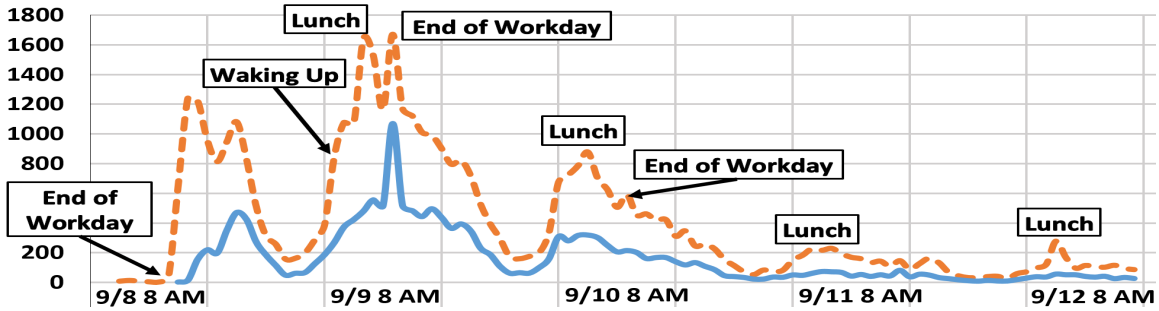



Figure 3.1: Tweet count per hour with #WhyIStayed (dotted) or #WhyILeft (solid) from 9/8 to 9/12. Times in EST, vertical lines mark 12 hour periods, with label corresponding to its left line. Spam removed and includes meta tweets.

expressions, it was excluded to ensure data quality.

The resulting corpus comprised 24,861 #WhyIStayed and 8,767 #WhyILeft labeled datapoints. The class imbalance may be a result of the origins of the trend rather than an indicator that more victims stay than leave. The tweet that started the trend contained only the hashtag #WhyIStayed, and media reporting on the trend tended to refer to it as the "#WhyIStayed phenomenon." As Figure 3.1 shows, the first #WhyILeft tweet occurred hours after the #WhyIStayed trend had taken off, and never gained as much use. By this reasoning, it was concluded that an even set of data would be appropriate, and enable the ratio metric (see Equation 3.1) in experiments, as well be used to compare themes in the two sets. By random sampling of #WhyIStayed, a balanced set of 8,767 examples per class was obtained. From this set, 15% were held out as a final testset, to be considered after a tuning procedure with the remaining 85% devset. All data analytics to follow in this chapter utilize the even 85% devset.

### 3.1.3 Annotation Study

Four people (co-authors of Schrading et al. [65]) annotated a random sample of 1000 instances from the devset, to further characterize the filtered corpus and to assess the automated extraction of gold standard labels. This random subset is composed of 47% #WhyIStayed and 53% #WhyILeft gold standard instances. Overall agreement overlap was 77% and Randolph's free-marginal multirater kappa[6] [71] score was 0.72. By the annotations of the four annotators, on average 36% of the instances are reasons for staying (S), 44% are reasons for leaving (L), 12% are meta comments (M), 2% are jokes (J), 2% are ads (A), and 4% do not match prior categories (O). Table 3.1 shows that most related directly to S or L, with annotators identifying more clearly L. Of interest are examples in which annotators did not agree, as these are indicative of complexities in the data, and are instances that a classifier may likely label incorrectly. The tweet *because i was slowly dying anyway* was marked by two annotators as S and two annotators as L. Did the victim have no hope left and decide to stay? Or did the victim decide that since they were "slowly dying anyway" they could attempt to leave despite the possibility of potentially being killed in the attempt? The ground truth label is #WhyILeft. Another example with two annotators labeling as S and two as L is *two years of bliss, followed by uncertainty and fear*. This tweet's

[6]This multirater kappa was chosen because it allows any distribution of the class labels that annotators assign (it is free-marginal), unlike Fleiss' multirater kappa which assumes a fixed distribution.

label is #WhyIStayed. The limited context from these instances makes it difficult to interpret fully, and causes human annotators to disagree; however, most cases contain clear enough reasoning to interpret correctly.

Table 3.1: Confusion matrices of all 4 annotators, compared to the gold standard. Annotators mostly identified reasons for staying or leaving, and only a small fraction were unrelated. #L=#WhyILeft, #S=#WhyIStayed.

| | | A | J | L | M | O | S |
|---|---|---|---|---|---|---|---|
| A1 | #L | .01 | .01 | .78 | .11 | .03 | .07 |
| | #S | .01 | .03 | .10 | .21 | .02 | .63 |
| A2 | #L | .02 | .01 | .72 | .06 | .09 | .10 |
| | #S | .03 | .01 | .07 | .16 | .10 | .63 |
| A3 | #L | .00 | .02 | .77 | .09 | 0 | .11 |
| | #S | .01 | .04 | .06 | .21 | 0 | .68 |
| A4 | #L | .02 | .01 | .75 | .05 | .04 | .14 |
| | #S | .03 | .01 | .16 | .12 | .05 | .63 |

### 3.1.4 Lexical Usage

Both tweet sets have unique lexical structures explaining reasons for leaving or staying. Basic lexical statistics in the even devset *before* lowercasing, stoplisting, and lemmatizing are shown in Table 3.2.

Table 3.2: Basic lexical statistics on the tokens and types in the two sets. *Types* are unique tokens while *hapax legomena* are those tokens that only occur once in the dataset.

| | **#WhyIStayed** | **#WhyILeft** |
|---|---|---|
| **Total number of tokens** | 130545 | 118768 |
| **Total number of types** | 7094 | 6269 |
| **Type:token ratio** | .054 | .053 |
| **Number of hapax legomena** | 3871 | 3340 |

The lexical diversity is approximately equal in both sets. This means that the users explained their reasons for staying and leaving using approximately

the same ratio of different words.

The top 10 most frequent unigrams, bigrams, and trigrams in the even dataset *after* lowercasing, stoplisting, and lemmatizing are shown in Tables 3.3, 3.4, and 3.5 respectively. Stoplisting, lemmatizing and not including tokens for *start of sentence* and *end of sentence*[7] was done before extracting and examining the ngrams in this study in order to generally discard *function* words and focus on the *content* words. In doing so, the reasoning of users is more interpretable from the data.[8]

Table 3.3: Top 10 most frequent unigrams after preprocessing with their respective frequencies in the Twitter dataset.

| **Unigrams** | |
|---|---|
| **#WhyIStayed** | **#WhyILeft** |
| think, 1061 | love, 930 |
| love, 971 | realize, 888 |
| leave, 872 | want, 702 |
| abuse, 754 | leave, 613 |
| believe, 578 | know, 594 |
| tell, 550 | better, 570 |
| want, 540 | deserve, 558 |
| say, 529 | abuse, 507 |
| know, 518 | life, 497 |

From these frequently occurring phrases, initial ideas about the dynamics between staying and leaving emerge. The ngrams *think*, *believe*, *tell*, *feel*

[7]Indicators for *start of sentence* and *end of sentence* are often used in ngram experiments, but in this work are considered *functional* unlike *content* words.

[8]The special tokens *@mention* and *url* were not included in the stoplist, but for unigrams and bigrams they were not considered in the top 10, as they are similar to *function* words without context.

Table 3.4: Top 10 most frequent bigrams after preprocessing with their respective frequencies in the Twitter dataset.

| **Bigrams** | |
|---|---|
| **#WhyIStayed** | **#WhyILeft** |
| think love, 127 | deserve better, 298 |
| abusive relationship, 112 | finally realize, 103 |
| feel like, 95 | realize deserve, 80 |
| make feel, 89 | realize love, 67 |
| try leave, 78 | want live, 66 |
| emotional abuse, 72 | learn love, 61 |
| think deserve, 67 | want daughter, 59 |
| make believe, 64 | year old, 56 |
| kill leave, 57 | know deserve, 55 |

Table 3.5: Top 10 most frequent trigrams after preprocessing with their respective frequencies in the Twitter dataset.

| **Trigrams** | |
|---|---|
| **#WhyIStayed** | **#WhyILeft** |
| make feel like, 37 | realize deserve better, 56 |
| pregnant hit url, 25 | know deserve better, 40 |
| stay abusive relationship, 25 | finally realize deserve, 19 |
| change conversation url, 22 | son deserve better, 18 |
| leave man yell, 21 | true love hurt, 18 |
| abusive relationship url, 20 | daughter deserve better, 17 |
| man yell url, 20 | want daughter think, 15 |
| say kill leave, 20 | want daughter grow, 15 |
| church support spousal, 19 | daughter grow think, 15 |

*like*, *make feel*, *think deserve*, *make believe*, and *make feel like* in the #WhyIStayed class indicate cognitive manipulation in the victim – the abuser may have made them feel or believe that they deserve their abuse. In higher order ngrams violent aspects of abuse emerge, including *kill leave*, *pregnant hit url*, *leave man yell*, *man yell url*, and *say kill leave*. Indications of threats on the victim's life appear as critical reasons for staying. Conversely, in the

#WhyILeft class, the victims indicate that they had an awakening – a moment of clarity – which allowed them to leave (*realize*, *know*, *finally realize*, *realize deserve*, etc.). Apparently linked to this realization is the desire for a better life, either for themselves or their children, as indicated by ngrams like *deserve better*, *realize deserve*, *realize love*, *want live*, *want daughter*, *realize deserve better*, *know deserve better*, *finally realize deserve*, *son deserve better*, *daughter deserve better*, *want daughter think*, and *want daughter grow*.

### 3.1.5 Analysis of Subject-Verb-Object Structures

Data inspection suggested that many users explained their reasons using a Subject-Verb-Object (SVO) structure, in which the abuser is doing something to the victim, or the victim is explaining something about the abuser or oneself.[9] Here, unlike in Schrading et al. [65] which used TurboParser [49], the open-source tool spaCy [35] was used to heuristically extract syntactic dependencies, constrained by pronomial and restricted lexical usage. This parser performed well since it is trained to handle social media data and many instances in the corpus had standard English. While tweets are known for non-standard forms, the seriousness of the discourse domain may have encouraged more standard writing conventions.

An analysis was conducted for both male and female genders acting as the

[9]Example: *He hurt my child* S: *He*, V: *hurt*, O: *child*.

abuser in the subject position. Starting at the lemmatized predicate verb in each dependency parse, if the predicate verb followed an abuser subject word[10] per the dependency links, and preceded a victim object word,[11] it was added to a conditional frequency distribution, with the two classes as conditions. These structures are here denoted *abuser onto victim*. Similar methods were used to extract structures in which the victim is the subject. Improvements of note from Schrading et al. [65] include adding negations to predicate verbs and considering neutral-gender abusers deemed relevant for analysis. A negation indicator (an exclamation point) was added to the front of a predicate verb if a negation token[12] occurred in its direct left or right dependencies.

Instances with female abusers were rare (approximately 230 instances), and statistical gender differences could not be pursued. Accordingly, both genders' frequency counts were combined. Discriminative predicates from these conditional frequency distributions were determined by Equation 3.1. Table 3.6 reports on those where the ratio is greater than 0.70 and the total count exceeds a threshold to avoid bias towards lower frequency verbs.[13]

$$ratio = \frac{count_{largerOfCounts}}{count_{left} + count_{stayed}} \tag{3.1}$$

[10]Male abuser: *he, bf, boyfriend, father, dad, husband, brother, man*. Female: *she, gf, girlfriend, mother, mom, wife, sister, woman*. Neutral: *pastor, abuser, offender, ex, x, lover, church, they*.

[11]Victim object words: *me, sister, brother, child, kid, baby, friend, her, him, man, woman*.

[12]Negation tokens: *no, not, n't, never, none*.

[13]This threshold was defined to be .5% of the total number of instances. In the case of *abuser onto victim* this came to be a frequency threshold of 11, and in the case of *victim as subject* this was a threshold of 68.

Table 3.6: Discriminative verbs for *abuser onto victim* and *victim as subject* structures. An exclamation point (!) before a verb indicates negation, e.g., the phrase *he did not love me* would give the verb *!love*.

| **Legend** |
|---|
| Stayed |
| Left |

| Most discriminative *abuser onto victim* verbs | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| convince | need | isolate | promise | love | !love | !hit | have | leave | tell | be | find | choke | kill |
| 0.95 | 0.94 | 0.94 | 0.92 | 0.90 | 0.89 | 0.89 | 0.87 | 0.80 | 0.80 | 0.78 | 0.76 | 0.75 | 0.74 |

| Most discriminative *victim as subject* verbs | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| realize | think | !think | find | learn | believe | !know | try | felt | know | tell | get |
| 0.98 | 0.91 | 0.91 | 0.88 | 0.88 | 0.86 | 0.84 | 0.80 | 0.73 | 0.71 | 0.71 | 0.70 |

From Table 3.6, agreements with clinical literature on the reasons for staying and leaving can be seen. Heise et al. [30] suggested that victims of abuse leave after an increase in violence triggers a realization. The narrative of Table 3.6 suggests exactly this; physical *abuser onto victim* verbs like *choke* and *kill* are indicative of #WhyILeft, while the *victim as subject* verb *realize* appears as the most discriminative verb in the data, along with *find* and *learn*. Additionally, a predominance of verbs indicative of cognitive manipulation appear for #WhyIStayed, such as *convince*, *promise*, *believe*, *think*, *!think* (e.g., *I didn't think he would...*) and *tell*. Heise et al. [30] suggested that emotional dependence and an optimistic hope for change are reasons for staying, and these manipulative verbs seem to corroborate this finding. Other interesting findings in this data are the equal and opposite effects of *love* and *!love*, and the verb *!hit* suggesting that perhaps because the abuse was not physical, the victims stayed. This could be due to a number of factors, including the victim considering physical abuse as the only form of abuse, or confusion in the general populace of how to define verbal abuse.

Heated arguments can be a component of many relationships, and this may give victims of verbal abuse (repeated patterns of belittling and threats) the idea that they are not being abused, and instead that their relationship just has normal, healthy verbal disputes.

## 3.2 Reddit

Reddit[14] has a wide range of forums dedicated to various topics, called *sub-reddits*, each of which are moderated by community volunteers. For sub-reddits dedicated to sensitive topics such as depression, domestic abuse, and suicide, the moderators tend to ensure that the anonymous submitter has access to local help hotlines if a life-threatening situation is described. They also enforce respectful behavior and ensure that the submissions are on topic by deleting disrespectful or off-topic posts. Finally, they ensure that all site rules are followed, including the strict disallowal of *doxing*, the practice of using submission details to reveal user identities.

Reddit allows lengthy submissions, unlike Twitter, and therefore the use of standard English is more common. Additionally, Reddit's informal list of rules called *reddiquette*[15] includes the rule *use proper grammar and spelling*, which has lead to a more widespread use of standard English. This allows natural language processing tools like semantic role labelers trained

[14] See www.reddit.com.
[15] See https://www.reddit.com/wiki/reddiquette.

on standard English to function better. Finally, Reddit allows users to comment on submissions, providing them with the ability to ask questions, give advice, and provide support. This makes its data ideal for studies of sensitive subjects not typically discussed in social media.[16]

Following the procedure in Balani and De Choudhury [2] for subreddit discovery, several subreddits that focus on domestic abuse were manually identified. Additionally, several subreddits unrelated to domestic abuse were identified to be used as a control set. Table 3.7 shows the subreddits, the total number of unique posts (called *submissions*) and total number of replies in those submissions (called *comments*) collected, and the number of active users of these subreddits (called *subscribers*).

Table 3.7: The domestic abuse subreddits and control subreddits with the total number of submissions and comments collected, along with their number of subscribers.

| **Domestic Abuse** | **# Submissions** | **# Comments** | **# Subscribers** |
|---|---|---|---|
| abuseinterrupted | 1653 | 1069 | 1344 |
| domesticviolence | 749 | 2145 | 1184 |
| survivorsofabuse | 512 | 2172 | 2039 |
| **Control** | **# Submissions** | **# Comments** | **# Subscribers** |
| casualconversation | 7286 | 285575 | 93525 |
| advice | 5913 | 31323 | 24485 |
| anxiety | 4183 | 23300 | 64743 |
| anger | 837 | 3693 | 4033 |

The *anger* and *anxiety* subreddits were chosen as control subreddits in order to help the classifier discriminate between the dynamics of abusive relationships and the potential effects of abuse on victims. For example, anxiety

[16]The Twitter phenomena discussed in Section 3.1 was a rather rare event spurred on by the Ray Rice scandal. This is unlike the subreddits on Reddit that have discussions of sensitive subjects, which are permanent message boards for users to discuss in a safe environment.

and anger may be affect caused by domestic abuse, but they are also caused by a wide variety of other factors. By including these subreddits in the control set, a classifier should utilize the situations, causes, and stakeholders in abusive relationships as features, not the affect particularly associated with abusive relationships. Similarly, the *advice* subreddit was chosen as a way to help the classifier understand that advice-seeking behavior is not indicative of abuse. The *casualconversation* subreddit allows discussion of anything, providing an excellent sample of general written discourse.

The domestic abuse subreddits have far fewer active users, called subscribers, than the others – with the exception of the *anger* subreddit. Low activity subreddits have far fewer submissions and comments in total.

### 3.2.1 Preprocessing

All experiments used the same preprocessing steps. From the collected subreddits, only submissions with at least 1 comment were chosen to be included for study. This was done to ensure that each submission could have comment data appended to its text in classification experiments. The title and selftext[17] (concatenated together) of each submission were processed with the Illinois Curator [13] to obtain semantic role labels. A total of 552 domestic abuse submissions were parsed, and an even distribution of the

[17] An optional text body of a submission, for elaboration on the title.

control subreddits (138 each) were randomly chosen, yielding a total sample size of 1104. All data instances were lowercased, lemmatized, and stoplisted. External links and URLs were replaced with *url* and references to subreddits, e.g., */r/domesticviolence*, were replaced with *subreddit_link*.

### 3.2.2 Corpus Characteristics

Basic descriptive statistics on the set of 552 abuse submissions and 552 non-abuse submissions *before* lowercasing, stoplisting, and lemmatizing are presented in Table 3.8.

Table 3.8: Basic descriptive statistics. The score is provided by users voting on submissions/comments they feel are informative. Users are given the option to *upvote* or *downvote* a submission or comment. If they appreciate the content, they *upvote*, increasing the total score of the content, while *downvotes* decrease the score. The depth of a comment indicates where in a reply chain it falls. A depth of 0 means it is in reply to the submission, a depth of 1 means it is in reply to a depth 0 comment, etc. The $\pm$ values are standard deviation metrics.

| | **Abuse** | **Non-Abuse** |
|---|---|---|
| Average number of comments per submission | $5.4 \pm 6.1$ | $13.2 \pm 25.3$ |
| Average number of tokens per submission | $278 \pm 170$ | $208 \pm 164$ |
| Average submission score | $6.1 \pm 5.1$ | $7.5 \pm 16.4$ |
| Average submission sentiment | $-0.08 \pm 0.18$ | $-0.02 \pm 0.20$ |
| Percent negative sentiment | 69.7 | 54.7 |
| Number of unique submitters | 482 | 535 |
| Average number of tokens per comment | $107 \pm 128$ | $53.4 \pm 79.9$ |
| Average comment score | $2.2 \pm 2.7$ | $2.0 \pm 2.9$ |
| Average comment sentiment | $0.08 \pm 0.28$ | $0.13 \pm 0.28$ |
| Average comment depth | $0.96 \pm 1.5$ | $1.5 \pm 1.9$ |
| Percent comments with negative sentiment | 34.2 | 22.5 |
| Number of unique commenters | 1022 | 2519 |
| Number of comments | 2989 | 6964 |

Sentiment scores were provided by VADER: a rule-based sentiment analyzer designed for social media texts [36]. A negative score means negative sentiment, while a positive score means positive sentiment, with a range constrained between -1 and 1. A comment or submission is considered to have negative sentiment if its sentiment score is less than 0.

In general, abuse subreddits have more negative sentiment submissions and comments than non-abuse subreddits, and their average sentiments are slightly lower than non-abuse. Additionally, the non-abuse subreddits have more discourse between commenters, as indicated by a larger comment depth, however, the abuse subreddits tend to have longer submissions and replies. The abuse subreddits also have a smaller, perhaps more tight-knit, community as indicated by fewer numbers of unique submitters and commenters.

### 3.2.3 Lexical Usage

Basic lexical statistics in the even dataset *before* lowercasing, stoplisting, and lemmatizing are shown in Table 3.9. Past and present tense verbs were determined using the same method in Lamb et al. [40]: POS tagging was performed, and all verbs were examined for suffix matches (*ed* for past-tense and *ing* for present tense) or existence in a hand-crafted set of verbs.[18]

Higher lexical diversity in the non-abuse class for both submissions and

[18]Past tense: *was, did, had, got, were*. Present tense: *is, am, are, have, has*.

Table 3.9: Basic lexical statistics on the tokens and types in the two sets. *Types* are unique tokens while *hapax legomena* are those tokens that only occur once in the dataset.

| | **Abuse** | **Non-Abuse** |
|---|---|---|
| Total number of tokens, submissions | 153644 | 114542 |
| Total number of types, submissions | 8565 | 8319 |
| Type:token ratio, submissions | 0.056 | 0.072 |
| Number of hapax legomena, submissions | 4275 | 4373 |
| Ratio of present:past tense verbs, submissions | 1.12 | 1.65 |
| Total number of tokens, comments | 319345 | 372024 |
| Total number of types, comments | 12760 | 20076 |
| Type:token ratio, comments | 0.040 | 0.054 |
| Number of hapax legomena, comments | 6378 | 10673 |
| Ratio of present:past tense verbs, comments | 1.68 | 1.84 |

comments is observed. This makes sense, as the non-abuse class is derived from subreddits on various topics. The ratio of present tense verbs to past tense verbs differs between the abuse and non-abuse submissions. Often, submissions in the abuse class recount their abuse in the past tense, leading to higher overall past-tense usage. In non-abuse submissions, and all comments, the discourse tends to be in the present-tense, with slightly higher past-tense usage in abuse comments compared to non-abuse comments. Tentatively, higher present tense usage in comments is due to users discussing topics and events that are currently happening to them or that they are interested in, as it relates to the submission. The slight drop in present tense usage in the abuse comments with respect to the non-abuse comments is probably from users sympathizing with the submitter by recounting past stories of their own.

To get a sense of the lexical content between the two sets of subreddits, the

most frequent uni- bi- and trigrams were examined. In Tables 3.10, 3.11, and 3.12 are the top 10 unigrams, bigrams, and trigrams of the submission and comment data combined (after lowercasing, stoplisting, and lemmatizing). As in Section 3.1.4, these ngrams were examined after stoplisting and lemmatizing in order to analyze only the *function* words.

Table 3.10: Top 10 unigrams after preprocessing with their frequencies in the two sets of data.

| **Unigrams** | |
|---|---|
| **Abuse** | **Non-Abuse** |
| know, 1987 | like, 2620 |
| like, 1802 | feel, 1586 |
| feel, 1686 | know, 1557 |
| help, 1624 | make, 1505 |
| abuse, 1595 | time, 1473 |
| time, 1391 | think, 1451 |
| want, 1335 | really, 1433 |
| make, 1333 | thing, 1371 |
| thing, 1310 | people, 1280 |
| think, 1252 | want, 1238 |

Table 3.11: Top 10 bigrams after preprocessing with their frequencies in the two sets of data.

| **Bigrams** | |
|---|---|
| **Abuse** | **Non-Abuse** |
| feel like, 389 | feel like, 423 |
| domestic violence, 202 | sound like, 134 |
| sound like, 170 | make feel, 121 |
| abusive relationship, 166 | high school, 118 |
| make feel, 131 | panic attack, 107 |
| good luck, 121 | good luck, 86 |
| let know, 121 | year old, 83 |
| year old, 112 | year ago, 80 |
| sexual abuse, 103 | feel better, 79 |
| year ago, 97 | really like, 76 |

Table 3.12: Top 10 trigrams after preprocessing with their frequencies in the two sets of data.

| **Trigrams** | |
|---|---|
| **Abuse** | **Non-Abuse** |
| feel free pm, 27 | play video game, 27 |
| make feel like, 26 | make feel better, 22 |
| domestic violence hotline, 23 | make feel like, 20 |
| local domestic violence, 17 | time feel like, 11 |
| domestic violence shelter, 17 | url url url, 11 |
| long story short, 15 | feel like need, 9 |
| year old male, 14 | meet new people, 8 |
| local dv agency, 12 | feel like want, 8 |
| feel like need, 12 | spend lot time, 8 |
| make feel bad, 11 | feel like talk, 8 |

While there are many common and overlapping ngrams in the two sets (e.g., *like, feel*, and *sound like*), each set does have distinct ngrams. In the abuse set, distinct ngrams include the obvious *abuse*, *domestic violence*, *abusive relationship*, and *sexual abuse*. Additionally, unique trigrams related to the agents and situations in abusive relationships like *local dv agency* and *make feel bad* appear. Also included are unique empathetic and helping discourse from comments, including *let know*, and *feel free pm*[19]. This indicates that comment data has potential to improve classification results, as the unique desire to help and sympathize in the abuse subreddits may be more prevalent than in the control subreddits.

[19]The abbreviation *pm* stands for private message.

### 3.2.4 Semantic Role Attributes

Using the Semantic Role Labeler (SRL) in the Illinois Curator [13], the dataset was tagged with various arguments of predicates, along with the particular sense numbers of the predicates. This data is particularly useful to study, as the semantic agents, actions, and patients within an abusive relationship are desired to be examined. As noted in Section 2.3.5, the SRL tool gives a given argument number, predicate, and sense. Next, these are used to perform a lookup in Proposition Bank (PropBank) [47] to retrieve the unique role labels for each tagged argument.

The top 100 most frequent roles and predicates in the two sets were determined. Significant overlap between the two sets exists, so only the top 10 unique roles and predicates within each top 100 set were taken. This provides the frequently occurring but unique roles and predicates within the abuse and control set. Tables 3.13 and 3.14 contain this data.

Tables 3.13 and 3.14 mark important distinctions between the two groups. The SRL processed abuse data contains agents, actions, and patients that are powerful indicators of an abusive relationship, including *hitter*, *thing hit*, *abuser*, and *entity experiencing hurt/damage*. The role label *benefactive* in the abuse set is interesting because it may mark submissions where victims of abuse are getting help. In one submission, a father posts *I just found out that my 24 yr old daughter is a victim of domestic violence...What can I do to support and help her?* The SRL tool marked *her* as the benefactive

Table 3.13: Top 10 unique role labels for the abuse and non-abuse sets.

| **Role Labels** | |
|---|---|
| **Abuse** | **Non-Abuse** |
| *caller*, 175 | *maker*, 175 |
| *thing hit*, 174 | *attributive*, 79 |
| *agent, hitter - animate only!*, 164 | *target or hatred*, 78 |
| *abuser, agent*, 162 | *thing lost*, 75 |
| *entity abused*, 139 | *thing disappearing*, 71 |
| *utterance*, 115 | *extent*, 69 |
| *patient, entity experiencing hurt/damage*, 113 | *agent, setter*, 67 |
| *utterance, sound*, 104 | *entity losing something*, 65 |
| *belief*, 104 | *thing set*, 64 |
| *benefactive*, 103 | *dealer*, 57 |

Table 3.14: Top 10 unique predicates of the abuse and non-abuse classes with their respective frequencies. An exclamation point on a predicate indicates negation. The dotted number attached to the predicate is its sense number.

| **Predicates** | |
|---|---|
| **Abuse** | **Non-Abuse** |
| abuse.01, 433 | pay.01, 127 |
| share.01, 167 | sit.01, 108 |
| believe.01, 164 | !help.01, 107 |
| call.02, 151 | play.01, 105 |
| remember.01, 149 | enjoy.01, 104 |
| cry.01, 147 | spend.01, 100 |
| !tell.01, 142 | go.06, 90 |
| send.01, 127 | watch.01, 87 |
| thank.01, 127 | mean.01, 86 |
| realize.01, 124 | decide.01, 83 |

and linked it to the predicate *help*. Another example is from a victim of abuse recounting when they left: *It took awhile but I realized just how NOT-NORMAL and destructive this behavior was and I was able to make the safest and healthiest choice for me to end my relationship with him.* Here the SRL tool marked *me* as the benefactive and linked it to the predicate *make*.

Importantly, in the predicate data, several abuse predicates that appeared also occurred in the Twitter data, including *believe* and *realize*.

### 3.2.5 Analysis of Subject-Verb-Object Structures

Following the same procedures in Section 3.1.5, the *abuser onto victim* SVO structures and *victim as subject* SV structures of the abuse and non-abuse sets were examined. Because only the lexicon in use was needed, rather than semantic role labels, a larger set of data (1336 submissions per abuse and non-abuse class) was used.[20]

As the SVO structures are still constrained by abuser subject and object words, only a single discriminative verb appears for the non-abuse class: *like*, with a ratio of 0.71. This makes sense, as users often post about boys or girls "liking" them in the *casualconversation* subreddit. This also gives strength to the validity of the chosen control data. If further discriminative abuse SVO structures appeared in the control data, it would indicate that submissions there were also about abuse. Since this does not happen, it is reasonable to assume that noise of this type is not too prevalent. The discriminative verbs and their ratios for the abuse class are in Table 3.16.

The discriminative *victim as subject* structures were also examined.[21] The top verbs for the abuse class are in Table 3.15. There were no discriminative non-abuse *victim as subject* verbs.

[20]Resulting in 2740 *abuser onto victim* SVO structures and a frequency threshold of 13.
[21]Resulting in 27762 *victim as subject* SV structures and a frequency threshold of 83.

Table 3.15: Discriminative *victim as subject* verbs in the abuse class.

| **Discriminative Verb** | **Ratio** |
|---|---|
| abuse | 1 |
| remember | 0.83 |
| call | 0.80 |
| leave | 0.79 |
| tell | 0.73 |

The verbs in these experiments are useful for determining the actions that occur within abusive relationships. There are several physical assault words like *throw*, *slap*, and *punch*, but also included are sexual assault words like *rape*, stalking words like *follow* and *contact*, and verbal abuse words like *call* and *yell*. The verb *abuse* appearing in the *victim as subject* table is interesting. Most examples where this structure occurs appear as a result of the passive voice phrase *i was abused*, however a few indicate that they were the *abuser* or they participated in a mutually abusive relationship.

Table 3.16: Discriminative *abuser onto victim* verbs in the abuse class.

| **Discriminative Verb** | **Ratio** |
|---|---|
| throw | 1 |
| slap | 1 |
| rape | 1 |
| !hit | 1 |
| pull | 1 |
| hit | 0.99 |
| beat | 0.98 |
| hurt | 0.97 |
| push | 0.96 |
| abuse | 0.96 |
| punch | 0.96 |
| do | 0.96 |
| kill | 0.95 |
| contact | 0.94 |
| follow | 0.93 |
| love | 0.93 |
| leave | 0.86 |
| take | 0.85 |
| treat | 0.83 |
| kick | 0.83 |
| have | 0.82 |
| say | 0.81 |
| send | 0.76 |
| call | 0.76 |
| yell | 0.76 |
| !tell | 0.76 |
| ask | 0.75 |
| tell | 0.74 |
| send | 0.74 |
| be | 0.71 |

# Chapter 4

# Twitter Data Experiments

## 4.1 Classification Experiments

The following sections contain machine learning classification experiments on the Twitter dataset outlined in Section 3.1. A table describing all of the properties of these experiments, including the classifier used, devset and testset size, metrics, etc. are in Table D.1 in Appendix D.

### 4.1.1 SVO Features Only

The usefulness of the *abuser onto victim* SVO structures were examined, using subsets of the devset and testset having *abuser onto victim* structures. In total, 14% of these instances had these structures. While 14% is not a large proportion overall, given the massive number of possible dependency structures, it is a pattern worth examining – not only for corpus analytics but also classification, particularly as these SVO structures provide insight into the abuser-victim relationship. A linear SVM using boolean SVO features performed best (C=1), obtaining $70\% \pm 3\%$ accuracy on the devset and 72%

accuracy on the testset.

The weights assigned to features by a Linear SVM are indicative of their importance [29]. The top SVO structures are presented in Table 4.1. Some interesting structures, separate from the most discriminative verbs in Section 3.1.5, appear as features in this table. For example, in the #WhyILeft set, indications of intervention from non-abusers (*sister tell me*) appear as important features. Taking a closer look at the tweets that these structures originated from, e.g., *because my sorority sisters and roommates told me nothing about how he treated me was okay.*, suggests that these are features originating from external support structures that the victim gained access to. An interesting structure in the #WhyIStayed class is *church tell me*. Several tweets indicated that their church condoned abuse as a means of avoiding embarrasment and divorce, e.g., *because the church told me that it was my responsibility as a godly wife to not embarrass him and just pray.*

Table 4.1: Top 10 SVO features for #WhyIStayed and #WhyILeft with their SVM weights. An exclamation point (!) in front of a predicate verb indicates negation.

| **#WhyIStayed** | **#WhyILeft** |
|---|---|
| he hunt me, 1.1 | he tell him, 1.3 |
| they !remember him, 1.1 | he !protect me, 1.2 |
| he need me, 1.1 | he !tell me, 1.0 |
| he convince me, 1.1 | he lie me, 1.0 |
| she convince me, 1.1 | he stab me, 1.0 |
| he give child, 1.0 | he do kid, 0.9 |
| he remind me, 1.0 | sister tell me, 0.89 |
| he wear me, 1.0 | she have baby, 0.89 |
| he !abuse kid, 1.0 | he strangle me, 0.78 |
| church tell me, 0.99 | he attack me, 0.77 |

The SVO structures capture meaning related to staying and leaving, and are

useful for analyzing the phenomenon, but are limited in their data coverage. Another experiment explored an extended feature set including uni-, bi-, and trigrams in sublinear $tf \times idf$ vectors, tweet instance character length, its retweet count[1], and SVO structures.

### 4.1.2 Full Feature Set

Naïve bayes, logistic regression, linear SVM, and RBF SVM classifiers from the Scikit-learn package [57] were compared. The RBF SVM performed slightly better than the others, achieving a maximum accuracy of $78\% \pm 1\%$ on the devset and 78% on the testset usinga subset of features.[2] In Schrading et al. [65], dimensionality reduction with supervised locality preserving projections (SLPP) [59] was attempted using a slightly different pipeline.[3] This reduced the feature set from the extremely high dimensional, sparse vector space of 197,176 features to a dense matrix of 134 features; however, accuracy on both the cross-validation and testset was reduced by $1\%$. Ablation, following the procedure in Fraser et al. [24], was utilized to determine the most important features and preprocessing steps for the classifier, the results of which can be seen in Table 4.2.

Interestingly, the SVO features combined with n-grams worsened performance slightly, perhaps due to trigrams capturing the majority of SVO cases.

[1]The number of times a particular instance was *retweeted* (shared) by other Twitter users

[2]Tuned parameters: max df = 12%, C=10, gamma=1.

[3]In this thesis, the tokenizer was changed from Scikit-learn's [57] to spaCy's [35], the stoplist was expanded, SVO extraction was improved, and many small changes were implemented.

Table 4.2: Feature and preprocessing parameter ablation study with an RBF SVM and no dimensionality reduction. NG = ngrams, E = emoticon replacement, IR = informal register replacement, TL = tweet length, RT = retweet count, SVO = subject-verb-object structures. % Acc is accuracy on the testset.

| **Removed** | **Remaining Features** | **% Acc** |
|---|---|---|
| | NG+E+IR+TL+RT+SVO | 77.07 |
| SVO | NG+E+IR+TL+RT | 77.60 |
| TL | NG+E+IR+RT | 77.91 |
| E | NG+IR+RT | 77.95 |
| RT | NG+IR | 77.83 |
| IR | NG | 76.65 |

The highest accuracy, 78% on the testset, could be achieved with a combination of ngrams and retweet count for features and informal register replacement in the preprocessing step. However, the vast majority of cases can be classified accurately with ngrams alone. Emoticon replacement may not have contributed to performance since they were rare in the corpus. Standardizing non-standard forms presumably helped the SVM slightly by boosting the frequency counts of ngrams while removing non-standard ngrams. Tweet length reduced accuracy slightly, while the number of retweets helped. Retweets appear to help due to the distribution of retweets between the two classes. An approximately equal proportion of tweets in both classes get a low number of retweets (0-10) or a very high number of retweets (>100), but an unequal proportion of #WhyIStayed tweets have a retweet count between 10 and 100. For those tweets with a retweet count between 10 and 100, 63% are #WhyIStayed while only 37% are #WhyILeft. This is probably due to the exposure in the media of the #WhyIStayed hashtag, leading to a larger number of Twitter users seeing and retweeting #WhyIStayed tweets.

The top features from a Linear SVM trained using ngrams and retweet count as features, and informal register replacement in the preprocessing, are shown in Table 4.3.

Table 4.3: Top 10 features with their linear SVM weights using ngrams and retweet counts as features, and informal register replacement in the preprocessing. The top features are all ngrams.

| **#WhyIStayed** | **#WhyILeft** |
|---|---|
| think, 3.0 | realize, 3.3 |
| believe, 1.6 | finally, 2.4 |
| convince, 1.6 | tired, 1.7 |
| tell, 1.5 | realise, 1.4 |
| say, 1.3 | daughter, 1.4 |
| try leave, 1.1 | son 1.4 |
| money, 1.0 | die, 1.3 |
| abuser, 0.9 | strong, 1.3 |
| feel, 0.9 | kill, 1.2 |
| young, 0.9 | anymore, 1.2 |

The SVM picked up on many of the key reasons for leaving and staying that have been discussed. For leaving, a realization (*realize*, *realise*) after an escalation of violence or threats of violence (*kill*, *die*) and concern for children (*son*, *daughter*). New reasons that appear are the words *tired* and *strong*. These may come from victims explaining that they became worn down, sick, and *tired* of the abuse or that they gained courage and strength to leave - either through their own fortitude or external support structures. For staying, again cognitive and verbal manipulation is key (*think*, *believe*, *convince*, *tell*, *say*, and *feel*). Several new reasons also appear: *try leave*, *money*, and *young*. The phrase *try leave* backs up claims in clinical literature that it is often difficult to gain external support to leave, and that

victims of abuse frequently go through cycles of abuse that involve leaving and coming back multiple times [30]. Financial distress is also a key factor for staying [6, 30], so it is no surprise that *money* appears as a top feature for the SVM. Without financial independence it is extremely difficult for victims of abuse to leave. The word *young* is interesting. It suggests that Twitter users explained that they were too young to be able to leave, or that their naivety (related to being young) may have made them think the abuse was normal.

## 4.2 Long Short-term Memory Experiments

Using a trained LSTM, novel character sequences can be generated.[4] This experiment was done to see if interesting language patterns, longer than the restricted ngrams in previous sections, emerge from a generated set of data that resembles the training corpus of tweets. In order to test, two Twitter sets were created to train the LSTM. The first was generated from the entire set of tweets after removing spam and meta-commentary, but before cleaning urls, hashtags, emoticons, and informal register. The character sequences generated from an LSTM trained on this corpus should resemble the tweets as they existed on Twitter, before being split and cleaned into their #WhyIStayed and #WhyILeft ground truth instances. The second corpus was generated from the cleaned and split tweet instances. An LSTM trained on this

[4]Character-level generation can also lead to nonsense words.

set should generate character sequences resembling the reasons for staying and leaving. These experiments used Karpathy's Char-RNN project [38].

### 4.2.1 LSTM Training Set 1

An LSTM with a size of 300 nodes per layer and 3 layers was trained on the entire set of tweets after removing spam and meta commentary, but before preprocessing and splitting. A dropout factor of 0.5 was used, achieving a cross-validation loss of 1.0938 in its $50^{th}$ (final) epoch. Given the starting sequence *#WhyIStayed*, 2000 characters were generated, resulting in the output in Table E.1 in Appendix E.

This generated text looks similar to the real set of tweets, with the exception of ungrammatical structures and made up words, hashtags, and urls. For example, the words *musly* and *dolfar* are nonsense words generated due to the LSTM using character-level units rather than word-level (a limitation of the implementation in Char-RNN [38]). Additionally, the phrase *my mom should an game to punished by the best* is clearly ungrammatical and nonsensical, due to the LSTM's inability to understand complete grammatical structure. However, key words and phrases that have been identified in previous experiments appear in this generated set as well: *He never hit me*, *financial*, *Love isn't enough*, *Because I was scared*, *he would tell me he would change*, *realized*, *I believed him*. These phrases are both longer and

provide more context than the decontextualized ngrams examined earlier.[5]

### 4.2.2 LSTM Training Set 2

An LSTM with a size of 300 nodes per layer and 1 layer was trained using the cleaned and split tweet instances using a dropout factor of 0.5, achieving a cross-validation loss of 1.0924 in its $41^{st}$ epoch. Priming the LSTM generator with a starting phrase (*primed text*) and different random number generator seeds gives various generated reasons for staying and leaving. The following results are the primed texts and various generated reasons, with their seed number in parentheses.[6]

**because he ...**

(2) ... deserved better i thought i would heal!

(10) ... had hit me

(12) ... made me crazy

(13) ... said this was real.

(16) ... had called me through my clienting and choked me.

(23) ... tried to kill me.

[5]Generated text is useful to avoid releasing real tweets against Twitter's data sharing policy (see https://dev.twitter.com/overview/terms/agreement-and-policy). According to Section 6, part b, providing datasets to third parties is only allowed if the tweet IDs are the only part of the tweet released, or less than 50,000 public tweets are shared. In this thesis, less than 50,000 tweets could be shared in full, but they may contain sensitive and private information that should not be provided publicly.

[6]The training text was provided such that each new line was an instance, therefore only a single line is shown per seed number. Additionally, instances were selected when complete ideas were expressed in legible phrases. Entire generated instances are shown.

(25) ... told me he would kill himself if i left.

**because i ...**

(2) ... deserved better like my kids were hurting me again.

(5) ... couldnt believe you told me he was sorry.

(6) ... didnt love myself for him

(10) ... felt strong enough to save myself in the corsen to leave

(12) ... know my friends. because so he started to say he was the only reasons he would hurt my family which helped me leave!

(22) ... thought things would get better.

**i stayed because ...**

(6) ... verbal and financial abuse reusons url

(5) ... i loved him or change. ”for do to put him through a fide ome

(14) ... they all want to control.

(15) ... of a gun. . the day.

(34) ... he convinced me that kind of treatment she told me. i am been too.

(37) ... he didn’t believe it was

**i left because ...**

(103) ... i learned to love myself and my family. i had to change him, things had no one would love me, he would take anyone else, fere

complecetel n

(119) ... he destroyed me.

(120) ... i was pregnant! because i am worth more after i was stronger than this.

(123) ... no one should have killed my own car. i found my family within her family. url

(300) ... he killed me.

These examples in (semi) complete sentences outline the various reasons for staying and leaving identified in previous experiments. Physical threats or escalation of violence (e.g., *i stayed because of a gun* or *because he tried to kill me*), cognitive manipulation (*i stayed because he convinced me that kind of treatment...*), and financial control (*i stayed because verbal and financial abuse...*) dominate reasons for staying, while gaining security through family, friends, and improved personal courage appear, along with concern for children (*i left because i learned to love myself and my family...* or *i left because i was pregnant! because i am worth more...*) appear as reasons for leaving.

# Chapter 5

# Reddit Data Experiments

## 5.1 Classification Experiments

A classifier for detecting text discussing domestic abuse was desired to further examine the semantic and lexical features in detecting abusive relationships. The subreddit to which a submission was posted (see Table 3.7) was used as a way to map the instance into the the gold standard label *abuse* or *non-abuse*. Confidence that these labels are appropriate is gained by examining the top ngrams, roles, and predicates in Section 3.2, and by taking into account that these subreddits are moderated for on-topic content. Several experiments to determine the optimal classifier, best combination of features, and the effect of comments on prediction accuracy were run. A table describing properties of these experiments, including the classifier used, devset and testset size, metrics, etc. are in Table D.1 in Appendix D.

### 5.1.1 Combinations of Features

The uni-, bi-, and trigrams in the submission title and selftext, hereafter denoted *submission text*, the verb predicates from the SRL tool, and the semantic role labels (see Section 3.2.4) were used as features after TF*IDF vectorization.[1] Perceptron, naïve Bayes, logistic regression, random forest with chi-squared feature selection, radial basis function SVM, and linear SVM classifiers were parameter optimized using 10-fold cross-validation. Table 5.1 contains the results for the optimized classifiers. The best features are the ngrams, achieving the highest performance alone. Predicate and role features perform admirably, and usually give a 1% increase in accuracy when combined, but bring the classifier accuracies down slightly when combined with text features. The top performing classifier, Scikit-learn's [57] Linear SVM with C=0.1, had its weights examined to determine the top features for prediction [29]. These features along with their weights are shown in Table 5.2.

Table 5.1: Classification accuracies of all attempted classifiers. NG=Ngrams, P=Predicates, R=Roles. The best result is bolded.

| Classifer | NG | P | R | NG+P | NG+R | P+R | NG+P+R |
|---|---|---|---|---|---|---|---|
| Linear SVM | **90 ± 3** | 72 ± 5 | 73 ± 4 | 89 ± 3 | 88 ± 3 | 73 ± 4 | 87 ± 3 |
| RBF SVM | 90 ± 3 | 72 ± 5 | 73 ± 4 | 89 ± 3 | 89 ± 4 | 74 ± 4 | 86 ± 3 |
| Logistic Regression | 90 ± 3 | 72 ± 5 | 73 ± 3 | 88 ± 3 | 88 ± 3 | 73 ± 5 | 86 ± 3 |
| Naïve Bayes | 88 ± 3 | 71 ± 3 | 72 ± 3 | 86 ± 3 | 87 ± 3 | 73 ± 3 | 84 ± 4 |
| Random Forest | 88 ± 5 | 71 ± 4 | 71 ± 6 | 87 ± 4 | 86 ± 5 | 71 ± 6 | 86 ± 4 |
| Perceptron | 86 ± 3 | 68 ± 4 | 69 ± 3 | 86 ± 3 | 87 ± 4 | 69 ± 4 | 85 ± 3 |

[1] Binary features and only unigrams were tried but these did not improve results.

Table 5.2: Top 10 features based on Linear SVM weights for each class, using only ngrams from submission titles and selftext. The classifier may be relying heavily on the anxiety and anger subreddits to discriminate between abuse and non-abuse, as indicated by the sharp drop in SVM weight from *anger* to *job* in the *non-abuse* class. *Abuse* word weights are more evenly distributed.

| **Abuse** | **Non-Abuse** |
|---|---|
| abusive, 1.3 | anxiety, 1.1 |
| child, 0.93 | anger, 1.1 |
| abuser, 0.86 | job, 0.52 |
| relationship, 0.84 | school, 0.46 |
| therapy, 0.83 | hour, 0.45 |
| survivor, 0.83 | week, 0.45 |
| domestic, 0.73 | fuck, 0.44 |
| happen, 0.72 | class, 0.42 |
| violence, 0.68 | college, 0.41 |
| father, 0.67 | fun, 0.40 |

### 5.1.2 Comment Data Only

Comment data alone was experimented with to determine if the discussions within abuse subreddits differed from those in non-abuse subreddits. Taking all comments individually, the task was to predict if they were posted in an abuse or non-abuse subreddit. Because ngram features performed best in the previous experiment, only ngrams were used from a larger set of data (1336 submissions per class). A final held out testset was created from 10% of these submissions giving 1202 submissions per class for the devset and 134 per class for the testset. Taking the comments from these submissions yielded 4712 abuse and 19349 non-abuse comments for the devset and 642 abuse and 2264 non-abuse comments for the testset. 10-fold cross-validation was used on the devset to tune the classifier. Using a Linear

SVM[2] with C=1 achieved an F1 score of $0.70 \pm .02$ on the devset. On the held out testset, it achieved a precision of 0.68, recall of 0.62, and F1 score of 0.65. Examining its weights gives features similar to those in Table 5.2.

Because comments can be completely off-topic or in reply to other comments, only the top-scoring[3] comments and those most similar to the submission text were also examined. To compute similarity, a sum of the word vector representations of each word in the submission and comment, respectively, was used. Word vectors were taken from Levy and Goldberg [44] as included in spaCy [35] and cosine similarity was used to determine the similarity score between submission and comment.

$$similarity = \frac{A \cdot B}{||A||||B||} \tag{5.1}$$

In Equation 5.1 $A$ and $B$ are both vectors with 300 dimensions (as created by Levy and Goldberg [44]). Taking only the top $90^{th}$ percentile for both submission score and similarity from the same devset/testset split above yielded 2381 abuse and 6928 non-abuse comments for the devset and 307 abuse and 924 non-abuse comments for the testset. Again, 10-fold cross-validation was used on the devset for tuning. A Linear SVM[4] with C=1 achieved an F1 score of $0.75 \pm 0.03$ on the devset. On the held out testset, it achieved a precision of 0.72, recall of 0.72, and F1 score of 0.72. The confusion matrix for the testset is in Table 5.3.

[2]max df = 5%, min df = 1
[3]See *score* in Table 3.8.
[4]With max df = 8%, min df = 1.

Table 5.3: Confusion matrix for the Abuse/Non-Abuse comment text classifier trained on an even set of data, using the testset after removing noisy comments.

| | | **Predicted Class** | |
|---|---|---|---|
| | | **Abuse** | **Non-Abuse** |
| **Actual Class** | **Abuse** | 221 | 86 |
| | **Non-Abuse** | 86 | 838 |

Examining the SVM's weights gives features similar to those in Table 5.2, with additional empathetic discourse like *leave*, *hug*, and *help* in the abuse class and casual discourse like *probably* and *haha* in the non-abuse class. This indicates that comments should be useful in predicting if the original submission is about abuse or not, and that the method to filter comments by similarity and score helps to remove noisy data.

### 5.1.3 Comment and Submission Predictors Cascaded

A cascade of the best classifiers for submission text data and comment text data was examined to determine if this improves accuracy over just the submission text classifier. The task is to predict whether a submission from the larger set of data (1336 submissions per class, using only ngrams) is *abuse* or *non-abuse*, given predictions from both a classifier trained on submission text and a classifier trained on comment text. The classifier trained only on this submission text data achieves an accuracy of $90\% \pm 2\%$ on the devset and $86\%$ on the testset, and this can be treated as the baseline to compare against.[5]

[5]With a Linear SVM with C = 100, min df = 1 and max df = 28%.

Using the $90^{th}$ percentile method to filter comments and the same devset/testset split as above, the trained comment classifier (trained on the comments in the training fold) is passed only comments that pass the filter for an individual submission. Using the signed distance to the hyperplane in the linear SVM as a confidence score, confidences for each comment and the confidence score of the submission text are determined.[6] Adding the confidences together provides a final score, where a negative score is the *abuse* class and a positive score is the *non-abuse* class. Theoretically this method should improve accuracies, since even if the submission classifier classifies the submission as one class (perhaps the submitter does not think they are being abused), but the comments are highly confident that it is the other class (the commenters are persuading the submitter that they are experiencing abuse), then the final prediction will be for the correct class.

In fact, it improves accuracies by about 2% overall. Using this method the accuracy on the devset improves to $92\% \pm 2\%$ and on the testset improves to $88\%$.

### 5.1.4 Comment and Submission Text Combined

Concatenating the comments within a submission to its title and selftext may also improve results. Using the $90^{th}$ percentile method as above, the comment text within a submission was concatenated to the submission text.

[6] Each confidence score is treated equally, because the number of comments per submission can change and the relative signed distance from the hyperplane differs in the submission classifier and the comment classifier. Tuning weights for these scores has the potential to improve results.

Again, the same devset/testset split of the larger dataset (using only ngrams) is used with 10-fold cross-validation to tune the classifier on the devset. This method achieves extremely high accuracy of $94\% \pm 2\%$ on the devset and $92\%$ on the testset using a Linear SVM with C=1,[7] and reduces the complexity by using a single classifier rather than multiple cascaded together. The top features are similar to those in Table 5.2.

### 5.1.5 Uneven Set of Submissions

Using the method in Section 5.1.4 to train the classifier, a much larger, but uneven set, of data was examined (still using only ngrams). This set contained all collected submissions with at least 1 comment, leading to 1336 *abuse* and 17020 *non-abuse* instances. From this set, 15% were held out for final examination as a testset and the rest was used as a devset with 5-fold cross-validation. On the devset, an F1 score of $0.81 \pm 0.01$ was achieved[8] while on the testset a precision of 0.84, recall of 0.74, and F1 score of 0.79 was achieved. The best classifier was a Linear SVM with C=100.[9] The confusion matrix of the testset is in Table 5.4.

This classifier has an excellent precision for the positive class (*abuse*), and decent recall, meaning that there can be high confidence that submissions flagged as *abuse* are indeed about *abuse*. By applying this classifier to a

[7] With max df = 31%, min df = 2.

[8] POS tags were added to ngrams in an additional experiment. This addition had no major effect on the results of all Reddit classification experiments, leading to only slight differences in performance metrics.

[9] With max df = 35%, min df = 20.

Table 5.4: Confusion matrix for the Abuse/Non-Abuse classifier trained on an uneven set of data, on the testset.

| | | **Predicted Class** | |
|---|---|---|---|
| | | **Abuse** | **Non-Abuse** |
| **Actual Class** | **Abuse** | 152 | 53 |
| | **Non-Abuse** | 29 | 2520 |

large held out set of data, these results suggest that many submissions should be flagged for examination, and they should mostly be about *abuse*.

### 5.1.6 Testing on Completely Held Out Subreddits

To get a sense of efficacy in the wild in detecting submissions that are discussing abuse, the best classifier from Section 5.1.5 was taken (trained on the devset data) and run on a large set of submissions from the *relationships* and *relationship_advice* subreddits. Statistics on this held out set is in Table 5.5.

Table 5.5: Held-out subreddits with the total number of submissions and comments collected as well as number of subscribers.

| **Held Out** | **# Collected** | **# Comments** | **# Subscribers** |
|---|---|---|---|
| relationships | 8201 | 192977 | 339807 |
| relationship_advice | 5874 | 55275 | 108090 |

These subreddits are general forums for discussion and advice on any relationship (not necessarily intimate). Their submissions tend to be long, descriptive, and extremely personal. Additionally, the moderators of these subreddits require that users include the age and gender of the major actors within the relationship. By running the abuse classifier on these subreddits,

not only can precision statistics be determined for a completely different dataset, but also interesting census data can be gathered about abusive relationships discussed online.

After running the abuse classifier on the submissions from these subreddits with at least 1 comment (13623 in total, with their $90^{th}$ percentile comments concatenated), 423 submissions were flagged as being about abuse. 101 of these 423 were annotated by 3 annotators, using the labels *A*, *M*, *N*, and *O*. Guidelines defining these categories are below:

- **A**: This submission is about abuse. It does not have to be abuse affecting the submitter - the submitter could be posting on behalf of someone else, could be the abuser, or could be posting asking for advice about a relationship affected by abuse in some other way (e.g., their girlfriend was abused by a relative and this is affecting their relationship). If the submitter is asking for advice about a problem that would not exist without the abuse, then it should be labeled *A*. Abuse in general is defined in Section 2.2. If any of these factors of abuse are present then it should be considered abuse.

- **M**: This submission has a mention of abuse, but is not related to the abuse. For example, if the poster mentions in passing that their friend was abused, but they are asking for advice about an unrelated topic, it is a mention.

- **N**: This submission is not about abuse. It has no mention of abuse and

you have no idea why it was flagged as abuse, but it is still on-topic for the subreddit, i.e., it is asking for advice or talking about a relationship (not necessarily an intimate relationship).

- **O**: Off-topic submissions/ads/jokes/other. This submission has no mention of abuse or is joking about abuse in some way, or it is a submission completely unrelated to relationships or relationship advice.

From the three annotator's annotations, on average 59% are *A*, 16% are *M*, 23% are *N*, and 2% are *O*. The percentage of overall agreement was 72% and Randolph's free-marginal multirater kappa[10] [71] score was 0.63.

Annotators occasionally had a hard time distinguishing between *A* and *M*, as context may have been missing, or the definitions between the two options were too vague. Combining the two by considering all *M* as *A*, the average percent of *A* increases to 75%, the percentage of overall agreement improves to 86% and Randolph's free-marginal multirater kappa improves to 0.79. Taking the statistic that on average 75% of the flagged submissions in the annotated subset are about abuse or have a mention of abuse indicates that this classifier should hopefully have a precision of around 0.75 on unseen Reddit data at large. Understandably, the precision drops by about .1 compared to its use on the subreddits it was trained and tested on. A precision of 0.75 on this set of data would mean that any statistics from this

[10]This multirater kappa was chosen because it allows any distribution of the class labels that annotators assign (it is free-marginal), unlike Fleiss' multirater kappa which assumes a fixed distribution.

set may include some noise, but overall, the trends should reveal important results about abuse.

By using regular expressions to capture the ages and genders mentioned in the titles of these 423 flagged submissions, a small census was conducted on the agents involved in these abusive relationships.[11] In these submissions, 345 agents were of the ages 18 to 25, 285 were 26 to 35, 81 were 36 to 54, 40 were 13 to 17, 27 were under the age of 13, and 9 were 55+. Additionally, 424 were female and 363 were male. The ages align with expected values, as the prevalence of abuse is greatest amongst the ages 18-35 (see Section 1.1) and the active users of Reddit tend to be in this age group as well. It is difficult to analyze these gender statistics further, since it is not immediately known which gender is the abuser and which is the victim.

### 5.1.7 Dense Features Experiment

Using the same uneven dataset in Section 5.1.5, dense features rather than sparse TF*IDF vectors were created and analyzed to see if performance could be increased, and to gain insight into the relative importance of different aspects of abuse discourse. Feature engineering using the devset data results from Sections 3.1.4, 3.2.2, 4.1.2, 5.1.1, and 5.1.5 was applied in order to create lists of features that may be indicative of abuse. The following features were created:

[11] It is standard in these titles to include age and gender in square brackets (e.g., [23F] means a 23 year old female). Most titles contain 2 actors (one abuser, one victim), however some have only 1 (usually the submitter of the post), and others have more than 2 (the ages and genders of all people involved).

- **Actors (ACTR)**: This set includes commonly seen stakeholders involved in abusive relationships, from data inspection.

- **Acts (ACTS)**: This set includes commonly seen actions (verbs) involved in abusive relationships, from data inspection, and expanded upon using Levin verb classes [43].

- **Sympathy (SMP)**: This set includes common sympathetic and helping discourse, usually from comments within Reddit submissions.

- **Abuser onto victim verbs (AOV)**: This set includes indicative verbs appearing in the *abuser onto victim* structures from Sections 3.1.5 and 3.2.5.

- **Victim as subject verbs (VAS)**: This set includes indicative verbs appearing in the *victim as subject* structures from Sections 3.1.5 and 3.2.5.[12]

- **Top Features (TF)**: This set includes the top ngrams based on Linear SVM weights of the Reddit abuse versus non-abuse classifier from Section 5.1.5 for the abuse class only.[13]

[12]The *abuser onto victim* and *victim as subject* sets are built from structures that occur more often for the abuse class than the non-abuse class, but not necessarily in a large ratio as in the experiments leading to Table 3.16.

[13]Feature weights with an absolute value of 0.9 or above were considered. The set of features was then modified by manually removing dataset-specific features that may not help in general classification tasks, e.g., by removing names, and by removing ngrams that already existed in the other feature sets (actors, acts, etc.).

A full list of the ngrams in these sets are in Appendix C. The feature vector for each training instance was constructed by incrementing a count for the above feature sets if a token in those sets occurred in the instance. Several additional features used in this experiment were derived from the text, as explained below:

- **Sentiment (SNT)**: The sentiment score derived from VADER [36].
- **Number of Tokens (NT)**: The total number of tokens in the instance.
- **Number of Present Tense Verbs (PRT)**: The total number of present tense verbs in the instance.[14]
- **Number of Past Tense Verbs (PST)**: The total number of past tense verbs in the instance.[15]

It is important to note that none of these features were used in any of the previously discussed experiments; only ngrams or SRL features were used in previous experiments. This was done for several reasons. First, Reddit-specific features were avoided in order to make the classifier applicable to any domain. Second, experience from the Twitter experiments indicated that adding additional features to TF*IDF vectors only serves to introduce code complexity, reduce accuracy or barely improve it, and slow down training. Finally, using only ngrams, an understanding of the usefulness of the lexical features without influence from other features can be gained.

[14] See Section 3.2.3 for a description of how present tense verbs were determined.
[15] See Section 3.2.3 for a description of how past tense verbs were determined.

In this experiment, the training data was scaled to have a mean of 0 and a standard deviation of 1. Using the same devset in Section 5.1.5, several classifiers were tested and parameter-optimized. Using all features, an AdaBoost classifier performs best with 275 estimators, achieving an F1 score of $0.71 \pm 0.04$ on the devset and $0.70$ on the testset.

The results of a feature ablation experiment, following the procedure in Fraser et al. [24], are shown in Table 5.6.

Table 5.6: Feature ablation study with an AdaBoost classifier. ACTR = Actors, ACTS = acts, SMP = Sympathy, AOV = Abuser onto victim, VAS = Victim as subject, TF = Top features, SNT = Sentiment score, NT = Number of tokens, PRT = Number of present tense verbs, PST = Number of past tense verbs. F1 here is the F1 score on the testset.

| **Removed** | **Remaining Features** | **F1** |
|---|---|---|
| | ACTR+ACTS+SMP+AOV+VAS+TF+SNT+NT+PRT+PST | 0.70 |
| AOV | ACTR+ACTS+SMP+VAS+TF+SNT+NT+PRT+PST | 0.71 |
| VAS | ACTR+ACTS+SMP+TF+SNT+NT+PRT+PST | 0.69 |
| SMP | ACTR+ACTS+TF+SNT+NT+PRT+PST | 0.68 |
| PRT | ACTR+ACTS+TF+SNT+NT+PST | 0.70 |
| ACTS | ACTR+TF+SNT+NT+PST | 0.70 |
| PST | ACTR+TF+SNT+NT | 0.68 |
| SNT | ACTR+TF+NT | 0.64 |
| ACTR | TF+NT | 0.61 |
| NT | TF | 0.17 |

From the feature ablation experiment, it can be seen that many of the features play important roles in increasing F1 score. The only feature that worsens performance slightly is the *abuser onto victim* set, which was also observed to reduce accuracy in the Twitter classifier experiment of Section 4.1.2. Again, this may be due to trigrams capturing the important AOV features, although not many trigrams are included in the TF set. Another

possibility, then, is that errors introduced by incorrect parses caused these features to introduce confusion in the classifier. The TF set is left as the last feature, which indicates its importance relative to the other feature sets, but it is not a good feature alone. This makes sense since this is a reduced set of features identified in earlier experiments to be useful in classification. Examining the TF set in Appendix C shows that it could be called ngrams *describing* or *related to* domestic abuse. New ngrams like *aggressive*, *alcohol*, *attorney*, *brainwash*, *childhood*, *counselling*, *crisis*, *depressive*, *flashback*, *heal*, *human interaction*, *interpersonal*, *oh god*, and many more appear to describe some of the causes, situations, and consequences of domestic abuse as described by victims or knowledgeable commenters. Other important features were (in order): NT, ACTR, SNT, and PST. The number of tokens makes sense, since it was shown in Section 3.2.2 that many abuse submissions are longer than non-abuse submissions. TF alone has a poor-performing F1 score;[16] however, TF along with NT provides the classifier with enough information to give an F1 score of 0.61. Adding ACTR and SNT further improve the F1 score by 0.04 each, and finally adding PST improves the F1 score by another 0.02. Taking the ablation experiment into consideration and looking at the confusion matrix in Table 5.7 reveals that the presence of a few ngrams related to abuse (TF), a relatively long submission (NT), the presence of some stakeholders involved in abuse (ACTR), a

[16]Note that if TF is removed and NT is left as the only feature, F1 score is only 0.09! This indicates that each feature is relatively weak alone, but powerful when combined.

relatively low sentiment score (SNT), and a high number of past tense verbs (PST) indicates abuse submissions with high precision.

Table 5.7: Confusion matrix for the Abuse/Non-Abuse classifier trained on an uneven set of dense data (ACTR+ACTS+SMP+VAS+TF+SNT+NT+PRT+PST), on the testset.

| | | **Predicted Class** | |
|---|---|---|---|
| | | **Abuse** | **Non-Abuse** |
| **Actual Class** | **Abuse** | 131 | 74 |
| | **Non-Abuse** | 33 | 2516 |

This experiment suffers from lower precision and recall scores relative to the sparse TF*IDF experiment in Section 5.1.5, but it drastically reduces the number of features (from the order of hundreds of thousands to 10). Removing Reddit-specific features (NT and possibly SMP due to the absence of comments in other social media domains) and applying this same experiment to other domains, e.g., Twitter, may perform better than the sparse TF*IDF SVM classifier, which may suffer from inherent over-fitting to the domain (caused by the number of features).[17]

## 5.2 Long Short-term Memory Experiments

As in Section 4.2, an LSTM was used to see if interesting language patterns emerge from generated Reddit submissions. All submissions labeled as abuse and with at least 1 comment, along with their top $90^{th}$ percentile comments, were included as training data. An LSTM with 400 nodes per

[17] An experiment was run on a large Twitter dataset using the TF*IDF SVM classifier, however the number of false positives was too high to warrant examining precision through an annotation study.

layer, 3 layers, and a dropout factor of 0.5 was trained, leading to a cross-validation loss of 1.0393. A generated sample submission, with interesting sections bolded and the primed text *Help*, is shown in Table E.2.

Again, ungrammatical structures e.g., *that makes sure that didnt know most of a good psychologist from that affection movie* and nonsense words e.g., *Dimas* and *teecance* appear, but many phrases contain meaning, and the structure follows a typical submission in the *abuse* subreddits.

The very beginning of the submission is the title. It appears the LSTM took the primed text as the only word for the title (*Help*) and generated a link to go along with it.[18] Following the title is the *selftext* and then the LSTM transfers into comments in the second paragraph. This can be seen with references to *you*, *help*, and pieces of advice. This is an advantage of LSTMs over other language models; they do well at determining long-term content dependencies.

The bolded texts in the sample have important discourse discussing the dynamics of domestic abuse. References to family members (*family*, *sister*, *abusive father*, *parents*, *child*), loving oneself (*I still love myself...*), thinking the abuse was their fault (*I just thought it was my fault.*), threats (*Your father is threatening...* or *scared me*), secrets (*working with my secret*), and PTSD are all involved in discussions of abuse.

[18] On Reddit, when words are wrapped in square brackets and followed by a url in parantheses, only the words in the square brackets appear, but they link to the url. For example [This is google](http://www.google.com) would appear only as *This is google* to users but would link to google.com.

# Chapter 6

# Conclusion and Future Work

Across two distinct datasets derived from different social media websites, meaningful structural and semantic, linguistic, and textual characteristics, including actions, stakeholders, and situations involved in abusive relationships are uncovered. Analyses of Twitter data reveal micro-narratives in tweeted reasons for staying versus leaving abusive relationships, and Reddit data is helpful in uncovering the dynamics of abusive relationships, the thoughts and motivations of the stakeholders within these relationships, and the lexical features used in discussing abuse online. A classifier to discriminate between tweeted reasons for staying versus leaving abusive relationships achieves an accuracy of 78% while a classifier to detect general text discussing abuse achieves an F1 score of 0.79 on a final held out testset. Additionally, from an annotation study, this classifier performs well on a large held out set derived from subreddits unused in training. Data analytics and various experiments reveal lexical items important for discovering abuse-related text, and the power of ngrams for text classification is confirmed. Importantly, many findings in this thesis overlap with an ecological model

proposed by Heise et al. and expanded on by the World Health Organization. All four levels that increase the likelihood that a man will abuse his partner are found in these data:

1. **Individual**: Ngrams like *alcohol* (alcoholism), *hit* and *choke* (acceptance of violence as a means of solving issues), *childhood* (experiencing or witnessing abuse as a child) and *want daughter, want son, son deserve better* (trying to prevent their children from experiencing or witnessing abuse).

2. **Relationship**: Ngrams like *money* and *financial* (control of finances, economic stress) and the *abuser onto victim* verb *!love* (marital conflict).

3. **Community**: Ngrams like *try leave* and the *abuser onto victim* verb *isolate* (women's isolation), and *church*, *church support spousal*, and *church tell me* (social groups that condone abuse).

4. **Societal**: The *abuser onto victim* structure *he need me* and the LSTM generated text *they all want to control* (concept of control/ownership of women).

Findings are consistent with different methods and datasets, correspond to observations in the clinical literature, and affirm the relevance of natural language processing techniques for exploring issues of social importance in social media.

## 6.1 Limitations

Several limitations are important to be noted and understood in this work.

- **Underrepresented Groups**: As discussed in Section 1.1, the age groups 0-17 and 55+ are significantly underrepresented on the websites used in this thesis [21, 22]. This means that unique aspects of domestic abuse affecting these age groups could be missed. In particular, adult-dependent abuse is rarely discussed in the datasets used (child abuse is occasionally discussed, with older submitters reflecting on their childhood).

- **Bias Towards Female-victim Abuse**: As noted in Section 2.2, males have significant inhibitions in reporting their abuse [3]. This may bias the results to the aspects of abuse in which the victim is female.

- **Unique and/or Rare Forms of Abuse Missing**: The properties of abuse and reasons for staying and leaving discovered in these data are affected by their relative frequency of occurrence. Unique and/or rare reasons for staying and leaving, and rare aspects of abusive relationships, may not be discovered using the methods in this thesis. For these to be uncovered, individual submissions would have to be examined by hand, or template-matching would have to be implemented. This may eliminate the speed and cost advantages over surveys.

- **Noise**: As with most data from the internet, it is important to know

that these datasets contain noise. These data can include accidental submissions to incorrect subreddits, submissions by spam bots, lies by the users, or jokes that were missed by moderators and filters.

- **Handcrafted Pronouns and Lexical Items**: The pronouns and lexical items used to convert the SVO features to *abuser onto victim* structures were handcrafted, potentially restricting the discriminative verbs that appear in sections 3.1.5 and 3.2.5.

- **Preprocessing**: Lowercasing, stoplisting, and lemmatizing helps to reduce dimensionality, but case, tense, and certain ngrams that appear in the stoplist may be important features that were missed due to these preprocessing steps.

- **Single Devset/Testset Split**: In all experiments, a single random devset and testset split was created, rather than creating multiple random devset/testset splits and averaging over their results. This means that the devset or testset split in each experiment could potentially be easier or more challenging to classify than compared to an average split. This is unlikely to effect the results by more than a few percentage points, and all testset results are near or within the standard deviation of the devset split.

## 6.2 Future Work

Significant amounts of future work are possible with the collected datasets.[1]

- **Domestic Abuse Communication Frames**: These data could be examined to study the communication frames involved in discussing domestic abuse to further qualitatively analyze the patterns of abuse and compare to theories of domestic abuse in clinical literature.
- **Demographics**: The Reddit data (especially data collected from the *relationships* and *relationship_advice* subreddits) could be used to study user demographics for those submissions related to domestic abuse, taking into account normal Reddit demographics, which may provide an estimate for the prevalence of domestic abuse and the ages and genders most affected by abuse and most likely to be the abusers.
- **Geotag Study**: The Twitter data with geotags could be used to study whether reasons for staying and leaving differ across different geological locations. This could be used to study how varying characteristics of those locations (e.g., poverty level, population density, education levels, etc.) affect domestic abuse victims.
- **Large Dataset Generation**: Using the currently trained abuse classifier, an extremely large Reddit-specific dataset could be developed to help improve domestic abuse research in the future. Additionally, the

[1] Datasets are available from http://nicschrading.com/data/.

#WhyIStayed and #WhyILeft tweets could continue to be collected for a larger study of these instances.

- **Insight Into Rare Forms of Abuse**: To gain insight into rarer forms of abuse, the classifier could be applied to different subreddits not examined in this thesis. For example, male-victim abuse is occasionally discussed in the subreddits *MensRights* and *AskMen*. The classifier should be able to find these submissions, and then analysis of these specific posts could help reveal the differences and similarities of male-victim and female-victim abuse. Similarly, to obtain specifically female-victim abuse, the subreddits *WomensRights* and *TwoXChromosomes* may be useful. Other abuse cases that could be examined are abuse within same-sex relationships or between other gender and sexual minorities.

- **Disjoint Domain Study**: Efforts could also focus on developing an abuse classifier that works on multiple online sites. This could be useful in developing machine learning and natural language processing techniques that work on disjoint domains. It could also be used to collect data from varied sources, improving the quality of the research data. Analysis of the features and patterns of online abuse discourse across varied forums will strengthen the present findings if they overlap, and perhaps reveal undiscovered features of abuse. Using forums

focused more on child or adult-dependent neglect may help to add further lexical items, and remove bias towards intimate partner violence observed in this thesis.

- **Comparison of LSTMs with other Language Models**: A comparison between the sequences generated by LSTMs and those generated by more traditional models may help to gain a thorough understanding of the trade-offs between them. Additionally, an LSTM implemented to generate word-level, rather than character-level, sequences could be studied.

- **General Improvement of Methods**: The experiments in this thesis could be performed again, making changes in the methods and classifiers in order to attempt to improve upon the reported metrics.

# Bibliography

[1] Olga Babko-Malaya. Propbank annotation guidelines. `http://clear.colorado.edu/compsem/documents/propbank_guidelines.pdf`, 2010. Accessed: 2015-06-16.

[2] Sairam Balani and Munmun De Choudhury. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual Association for Computing Machinery Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, pages 1373–1378, New York, NY, USA, 2015. ACM.

[3] Michele C Black, Kathleen C Basile, Matthew J Breiding, Sharon G Smith, Mikel L Walters, Melissa T Merrick, and MR Stevens. National intimate partner and sexual violence survey. *Atlanta, GA: Centers for Disease Control and Prevention*, 75, 2011.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

[6] Sarah M. Buel. Fifty obstacles to leaving, a.k.a, why abuse victims stay. *The Colorado Lawyer*, 28(10):19–28, Oct 1999.

[7] Sandra K. Burge, Johanna Becho, Robert L. Ferrer, Robert C. Wood, Melissa Talamantes, and David A. Katerndahl. Safely examining complex dynamics of intimate partner violence. *Families, Systems, & Health*, 32(3):259 – 270, 2014.

[8] Shannan Catalano. Stalking victims in the United States-revised. Technical Report NCJ 224527, United States Department of Justice, Washington, DC, 2012.

[9] Jinho D Choi and Andrew McCallum. Transition-based dependency parsing with selectional branching. In *Proceedings of the Association of Computational Linguistics*, pages 1052–1062, 2013.

[10] Jinho D Choi and Andrew McCallum. ClearNLP dependency labels. `https://github.com/clir/clearnlp-guidelines/blob/master/md/dependency/dependency_guidelines.md`, 2015. Accessed: 2015-06-17.

[11] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual Association for Computing Machinery Web Science Conference*, WebSci '13, pages 47–56, New York, NY, USA, 2013. ACM.

[12] Munmun De Choudhury, Scott Counts, Eric Horvitz, and Michael Gamon. Predicting depression via social media. In *Proceedings of the Seventh Annual International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media*. AAAI, July 2013.

[13] James Clarke, Vivek Srikumar, Mark Sammons, and Dan Roth. An

NLP curator (or: How I learned to stop worrying and love NLP pipelines). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3276–3283, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[14] Michael Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[15] Glen Coppersmith, Craig Harman, and Mark Dredze. Measuring post traumatic stress disorder in Twitter. In *Proceedings of the 8th Annual International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media*, 2014.

[16] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[17] Carole Craft. Costs of intimate partner violence against women in the United States. Technical report, National Center for Injury Prevention and Control, Centers for Disease Control and Prevention, Atlanta, GA, 2003.

[18] Danilo Croce and Daniele Previtali. Manifold learning for the semi-supervised induction of framenet predicates: An empirical investigation. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 7–16. Association for Computational Linguistics, 2010.

[19] Kristin Denham and Anne Lobeck. *Linguistics for everyone: An introduction*. Cengage Learning, Boston, MA, second edition, 2012.

[20] K.M. Devries, Joelle Y.T. Mak, C. García-Moreno, M. Petzold, J.C. Child, G. Falder, S. Lim, L.J. Bacchus, R.E. Engell, L. Rosenfeld, C. Pallitto, T. Voss, and C.H. Watts. The global prevalence of intimate partner violence against women. *Science*, 340(6140):1527–1528, 2013.

[21] Maeve Duggan and Aaron Smith. 6% of online adults are Reddit users. *Pew Internet & American Life Project*, 3, 2013.

[22] Maeve Duggan and Aaron Smith. Social media update 2014. *Pew Internet and American Life Project*, 2014.

[23] Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Scott Deerwester, and Richard Harshman. Using latent semantic analysis to improve access to textual information. In *Proceedings of the Special Interest Group on Computer-Human Interaction (SIGCHI) conference on Human factors in Computing Systems*, pages 281–285. ACM, 1988.

[24] Kathleen C. Fraser, Graeme Hirst, Naida L. Graham, Jed A. Meltzer, Sandra E. Black, and Elizabeth Rochon. Comparison of different feature sets for identification of variants in progressive aphasia. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 17–26, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[25] Claudia Garcia-Moreno, C Watts, and L Heise. Putting women first: Ethical and safety recommendations for research on domestic violence

against women. *Department of Gender and Womens Health, World Health Organization. Geneva, Switzerland*, 2001.

[26] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

[27] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005.

[28] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *arXiv preprint arXiv:1503.04069*, 2015.

[29] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, March 2002.

[30] Lori Heise, Mary Ellsberg, and Megan Gottemoeller. Ending violence against women. *Population Reports. Series L: Issues in World Health*, (11):1–43, 1999.

[31] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[32] Christopher Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilia Ovesdotter Alm. Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 107–117, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[33] Matthew Honnibal. A good POS tagger in about 200 lines of Python. https://honnibal.wordpress.com/2013/09/11/a-good-part-of-speechpos-tagger-in-about-200-lines-of-python/#comments, 2013. Accessed: 2015-06-16.

[34] Matthew Honnibal. Parsing English with 500 lines of Python. https://honnibal.wordpress.com/2013/12/18/a-simple-fast-algorithm-for-natural-language-dependency-parsing/, 2013. Accessed: 2015-06-17.

[35] Matthew Honnibal. spacy: Industrial strength NLP with Python and Cython. https://github.com/honnibal/spaCy, 2015.

[36] CJ Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth Annual International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media*. AAAI, 2014.

[37] Hal Daumé III. A course in machine learning. http://ciml.info/dl/v0_8/ciml-v0_8-ch03.pdf, 2012. Accessed: 2015-06-16.

[38] Andrej Karpathy. Char-RNN: Multi-layer recurrent neural networks (LSTM, GRU, RNN) for character-level language models in torch. https://github.com/karpathy/char-rnn, 2015.

[39] Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. A dependency parser for tweets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, WebSci '13, 2014.

[40] Alex Lamb, Michael J Paul, and Mark Dredze. Separating fact from fear: Tracking flu infections on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795, June 2013.

[41] Thomas K Landauer and Susan T Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[42] Rebecca T Leeb. *Child maltreatment surveillance: Uniform definitions for public health and recommended data elements*. Centers for Disease Control and Prevention, National Center for Injury Prevention and Control, 2008.

[43] Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.

[44] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308, 2014.

[45] Percy Liang. *Semi-supervised learning for natural language*. PhD thesis, Massachusetts Institute of Technology, 2005.

[46] Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. Technical report, University of Pennsylvania.

[47] Palmer Martha, Gildea Dan, and Kingsbury Paul. The Proposition Bank: A corpus annotated with semantic roles. *Computational Linguistics Journal*, 31:1, 2005.

[48] James H Martin and Daniel Jurafsky. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Second edition, 2008.

[49] Andre Martins, Miguel Almeida, and Noah A. Smith. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 617–622, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[50] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[51] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[52] Christopher JL Murray, Jerry Abraham, Mohammed K Ali, Miriam Alvarado, Charles Atkinson, Larry M Baddour, David H Bartels, Emelia J Benjamin, Kavi Bhalla, Gretchen Birbeck, et al. The state of US health, 1990-2010: Burden of diseases, injuries, and risk factors. *JAMA*, 310(6):591–606, 2013.

[53] Andrew Ng. CS229 lecture notes: Part V support vector machines. `http://cs229.stanford.edu/notes/cs229-notes3.pdf`, 2014. Accessed: 2015-06-19.

[54] Joakim Nivre. Dependency parsing. *Language and Linguistics Compass*, 4(3):138–152, 2010.

[55] World Health Organization. Understanding and addressing violence against women: intimate partner violence. 2012.

[56] Umashanthi Pavalanathan and Munmun De Choudhury. Identity management and mental health discourse in social media. In *Proceedings of WWW'15 Companion: 24th International World Wide Web Conference, Web Science Track*, Florence, Italy, May 2015. WWW'15 Companion.

[57] Fabian Pedregosa, Gaël Varoquaux., Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[58] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*, 2011.

[59] R. Ptucha and A.E. Savakis. LGE-KSVD: Robust sparse representation classification. *IEEE Transactions on Image Processing*, 23(4):1737–1750, April 2014.

[60] R. Ptucha, G. Tsagkatakis, and A. Savakis. Manifold based sparse representation for robust expression recognition without neutral subtraction. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2136–2143, Nov 2011.

[61] Raymond Ptucha. *Joint Optimization of Manifold Learning and Sparse Representations for Face and Gesture Analysis*. PhD thesis, Rochester Institute of Technology, 2013.

[62] Raymond Ptucha. Cmpe-789 machine intelligence lecture slides. CMPE-789 Machine Intelligence, Fall 2014.

[63] Vasin Punyakanok, Dan Roth, and Wen-tau Yih. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287, 2008.

[64] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[65] Nicolas Schrading, Cecilia Ovesdotter Alm, Raymond Ptucha, and Christopher Homan. #WhyIStayed, #WhyILeft: Microblogging to make sense of domestic abuse. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1281–1286, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[66] Richard Socher, Milad Mohammadi, and Rohit Mundra. Cs 224d: Deep learning for NLP. http://cs224d.stanford.edu/lecture_notes/LectureNotes4.pdf, Spring 2015.

[67] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[68] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In

*Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[69] Paola Velardi, Giovanni Stilo, Alberto E. Tozzi, and Francesco Gesualdo. Twitter mining for fine-grained syndromic surveillance. *Artificial intelligence in medicine*, 61(3):153–163, Jul 2014.

[70] Joseph Walther. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research*, 23(1):3–43, Feb 1996.

[71] Matthijs J. Warrens. Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification*, 4(4):271–286, 2010.

[72] Jun-Ming Xu, Benjamin Burchfiel, Xiaojin Zhu, and Amy Bellmore. An examination of regret in bullying tweets. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 697–702, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[73] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666. Association for Computational Linguistics, 2012.

[74] Nianwen Xue and Martha Palmer. Calibrating features for semantic role labeling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 88–94, 2004.

# Appendix A

# Lemmatization Rules

Table A.1: Rules for lemmatizing tokens. Each part of speech is broken by table section.

| **Noun Rules** | |
|---|---|
| **Ends With?** | **Becomes** |
| s | |
| ses | s |
| ves | f |
| xes | x |
| zes | z |
| ches | ch |
| shes | sh |
| men | man |
| ies | y |
| **Verb Rules** | |
| **Ends With?** | **Becomes** |
| s | |
| ies | y |
| es | e |
| es | |
| ed | e |
| ed | |
| ing | e |
| ing | |
| **Adjective Rules** | |
| **Ends With?** | **Becomes** |
| er | |
| est | |
| er | e |
| est | e |

# Appendix B

# Dependency Relations

Table B.1: Dependency relation types and their descriptions. A full description can be found in Choi and McCallum [10].

| Dependency Relation | Description |
|---|---|
| ACL | Clausal modifier of noun. |
| ACOMP | Adjectival complement. |
| ADVCL | Adverbial clause modifier. |
| ADVMOD | Adverbial modifier. |
| AGENT | Agent. |
| AMOD | Adjectival modifier. |
| APPOS | Appositional modifier. |
| ATTR | Attribute. |
| AUX | Auxiliary. |
| AUXPASS | Passive auxiliary. |
| CASE | Case marker. |
| CC | Coordinating conjunction. |
| CCOMP | Clausal complement. |
| COMPOUND | Compound modifier. |
| CONJ | Conjunct. |
| CSUBJ | Clausal subject. |
| CSUBJPASS | Clausal subject (passive). |
| DATIVE | Dative. |
| DEP | Unclassified dependent. |
| DET | Determiner. |
| DOBJ | Direct Object. |
| EXPL | Expletive. |
| INTJ | Interjection. |
| MARK | Marker. |
| META | Meta modifier. |
| NEG | Negation modifier. |
| NOUNMOD | Modifier of nominal. |
| NPMOD | Noun phrase as adverbial modifier. |
| NSUBJ | Nominal subject. |
| NSUBJPASS | Nominal subject (passive). |
| NUMMOD | Number modifier. |
| OPRD | Object predicate. |
| PARATAXIS | Parataxis. |
| PCOMP | Complement of preposition. |
| POBJ | Object of preposition. |
| POSS | Possession modifier. |
| PRECONJ | Pre-correlative conjunction. |
| PREDET | Pre-determiner. |
| PREP | Prepositional modifier. |
| PRT | Particle. |
| PUNCT | Punctuation. |
| QUANTMOD | Modifier of quantifier. |
| RELCL | Relative clause modifier. |
| ROOT | Artificial root. |
| XCOMP | Open clausal complement. |

# Appendix C

# Dense Feature Sets

## C.1 Actors

abuser, abusers, aunt, bf, boss, boyfriend, brother, brothers, child, children, church, cousin, cousins, coworker, dad, daughter, daughters, ex, family, fiance, fiancee, fiancé, fiancée, friend, friend family, friends, gf, girlfriend, husband, infant, kid, kids, mother, parent, parents, partner, pastor, sibling, siblings, sister, sisters, son, sons, stalker, survivor, survivors, teen, toddler, uncle, victim, victims, youth

## C.2 Acts

abuse, afflict, agonize, alienate, antagonize, asphyxiate, bash, batter, beat, beguile, belittle, bite, brawl, break, bruise, burn, butcher, castigate, chastise, choke, claw, coerce, control, convince, cower, criticize, cut, demean, demoralize, deprecate, depress, deride, devastate, disappoint, discourage, disgust, dishearten, disparage, distress, disturb, divorce, drown, embarrass, enrage, exhaust, fault, fight, flinch, frighten, grope, hate, hit, horrify, humiliate, hunt, hurl, hurt, injure, insult, intimidate, isolate, kick, kill, knock, lash, loathe, love, malign, mock, molest, monitor, mortify, murder, pity, plow, poison, punch, punish, push, pushed, rape, recoil, reprimand, ridicule, sadden, scare, scratch, scrutinize, shame, shock, shoot, shove, sicken, slam, slap, smack, smash, smother, spank, stab, strike, suffer, suffocate, tear, tease, terrify, terrorize, thrash, threaten, throw, tire, torment, torture, track, victimize, weary, weep, worry, yell

## C.3 Sympathy

awesome :), call 911, feel for you, feel free pm, find safety, get better, get help, get to safety, good luck, hug, hugs, i feel, i understand, leave, life leave, love, press charge, share, share story, sorry, thank you, thanks

## C.4 Abuser onto Victim Verbs

!abandon, !abuse, !beat, !consider, !face, !grab, !harm, !hit, !marry, !occur, !protect, !rape, !recognize, !release, !remind, !remove, !scar, !shout, !smack, !strike, !survive, !trivialize, abort, abuse, alienate, anchor, appear, appreciate, asaulted, assault, attack, babysit, bar, bash, beat, believe, belong, bribe, call, charm, choke, cleanse, coach, coerce, collaborate, contact, convince, cost, court, cross, cuff, damage, deem, defend, defriended, demand, demean, discover, disregard, do, downgrade, drown, ejaculate, excommunicate, explode, fail, fear, finish, flare, fling, follow, forward, frame, groom, guide, guilt, guilts, handcuff, harm, heal, hit, hoist, humiliate, hurt, injure, insert, isolate, kick, kill, lunge, mirror, molest, overhear, perform, photograph, pin, place, promise, prosecute, pull, punch, push, rape, release, rescue, sacrifice, separate, shape, shout, shove, slam, slap, sling, spank, spit, spread, straighten, strangle, strip, subject, survive, tazed, tear, throw, touch, tower, treat, twist, withdraw, yank, yell

## C.5 Victim as Subject Verbs

!assault, !block, !cease, !confirm, !consent, !cooperate, !disapprove, !disassociate, !discount, !formulate, !gather, !hint, !inconvenience, !induce, !know, !nod, !overstate, !protest, !rehash, !report, !retaliate, !shed, !slip, !spank, !summon, !think, !title, !transfer, !trap, !unpack, abuse, alibi, antogonising, appeal, arm, backpedal, barricade, believe, blanket, brainwash, cage, call, characterize, coerce, compel, convulse, cower, delt, demonize, disdain, dispose, disprove, dissociate, dredge, educate, elect, endure, eschew, estrange, felt, find, flail, flinch, friendless, grapple, huff, hypothesise, idolise, incapacitate, insert, interfere, interrogate, kidnap, kowtow, learn, leave, limp, login, mangle, misbehave, model, molest, mourn, obligate, outlive, possible, pout, protect, quiver, rape, ration, realize, recount, regurgitate, reign, relive, remember, reopen, scrub, scrunch, sentence, sin, strangle, tell, testify, think, try, unleash, victimize, writhe

## C.6 Top Features

abuse, abusive, abusive relationship, accept, aggressive, alcohol, anxiety feel like, attorney, belief, bias, boggle, boundary, brainwash, bruise, care love, challenge, childhood, clarity, compassion, confide, counsel, counselling, crisis, cycle, danger, decision make, depressive, domestic, domestic abuse, domestic violence, dv, emotion, emotional abuse, extreme, fault, finally, flashback, forward, grumpy, heal, healing, healthy, healthy relationship, hotline, human interaction, improve, independent, infamous, interaction, interpersonal, intimacy, ipv, leash, loving, make peace, manipulative, mindset, narcissist, narcissistic, new life, nightmare, obligate, oh god, painful, parenting, paycheck, people believe, people care, permission, personality disorder, perspective, pick phone, post, prisoner, protect, recovery, relationship, relive, remember, resource, restraining, restraining order, rethink, safe, sexual, shelter, situation help, social service, sociopath, spank, spiral, strong, stuck, success, support, survive, tactic, terrible person, thinking, tight, touch, toxic, trauma, trust, truth, upsetting, validation, verbal, verbal abuse, vice, violence, violent, want end

# Appendix D

# Experiment Summary

Table D.1: List of experiments with their properties and results. Refer to the sections noted in the *Experiment* column for a full description of the experimental setup and features used. The results displayed here are using the *best* parameters and features found after tuning on the devset.

| **Experiment** | **Features Used** | **Vectorization Used** | **Classifier Used** | **Devset Size** | **Devset Results** | **Testset Size** | **Testset Results** |
|---|---|---|---|---|---|---|---|
| 4.1.1 | AOV | Boolean | Linear SVM C=1 | 848 per class | $70\% \pm 3\%$ Acc | 142 per class | 72% Acc |
| 4.1.2 | NG+IR +RT | TF*IDF | RBF SVM C=10 gamma=1 | 7451 per class | $78\% \pm 1\%$ Acc | 1315 per class | 78% Acc |
| 5.1.1 | NG | TF*IDF | Linear SVM C=0.1 | 552 per class | $90\% \pm 3\%$ Acc | N/A | N/A |
| 5.1.2 with all comments | NG | TF*IDF | Linear SVM C=1 | 4712 abuse 19349 non-abuse | $0.70 \pm 0.02$ F1 | 642 abuse 2264 non-abuse | 0.65 F1 |
| 5.1.2 with $90^{th}$ percentile method | NG | TF*IDF | Linear SVM C=1 | 2381 abuse 6928 non-abuse | $0.75 \pm 0.03$ F1 | 307 abuse 924 non-abuse | 0.72 F1 |
| 5.1.3 Submission Text Only | NG | TF*IDF | Linear SVM C=100 | 1202 per class | $90\% \pm 2\%$ Acc | 134 per class | 86% Acc |
| 5.1.3 Comment and Submission Predictors Cascaded | NG | TF*IDF | Submissions: Linear SVM C=100 Comments: Linear SVM C=1 | 1202 per class | $92\% \pm 2\%$ Acc | 134 per class | 88% Acc |
| 5.1.4 | NG | TF*IDF | Linear SVM C=1 | 1202 per class | $94\% \pm 2\%$ Acc | 134 per class | 92% Acc |
| 5.1.5 | NG | TF*IDF | Linear SVM C=100 | 1131 abuse 14471 non-abuse | $0.81 \pm 0.01$ F1 | 205 abuse 2549 non-abuse | 0.79 F1 |
| 5.1.6 | NG | TF*IDF | Linear SVM C=100 | N/A | N/A | N/A | N/A |
| 5.1.7 | ACTR +ACTS +SMP +VAS +TF +SNT +NT +PRT +PST | Frequency Counts | AdaBoost num estimators =275 | 1131 abuse 14471 non-abuse | $0.71 \pm 0.04$ F1 | 205 abuse 2549 non-abuse | 0.71 F1 |

# Appendix E

# LSTM Output

Table E.1: LSTM generated text of the Twitter dataset before splitting into reasons based on hashtags. Interesting sections are bolded.

#WhyIStayed because everyone got sleeping on care. **He never hit me** at the abuse, but this is a **financial** way out to ever escape.
After my life to be killed too
#WhyIStayed **Love isn't enough to walk away.**
#WhyIStayed **Because I was scared** that he wanted me for those who had a musly a hand to a from 6.
I thought I was full he was just my brother until love. **I was afraid to be greater**
Forgiveness on #WhyIStayed. "#RayRice needs was abuse. #MereYesnoole when the ones high PSN #whyileft." For Relonic Alls I often helped my mother #WhyIStayed
rT @cennoiler
realized I was victims.
I found myself enough. Maybe lie. #IfEGEDNotCiplicth @Jackol relationship: no car to trapped up out of it. #WhyIStayed #WhyILeft http://t.co/Ij8J5XaZM4
He told me **he would tell me he would change.** I thought if I left him. **I believed him**, do I stay if I left. #WhyIStayed
Because I was hope of a dolfar #whyileft I was a broken me was forced
#WhyIStayed **I thought no one would love me.**,#WhyILeft is NEVER. #whyileft
#WHYIStayed
**I didn't want to love myself** to live. #WhyIStayed
#WhyIStayed,my mom should an game to punished by the best he would go any tleening, **being hit** this is. **I was worthless.**,**I believed the reasons** to alive
I was in the father because I finally was mentally alone going to have my mom & my getcountallyy later."
#WhyIStayed: For Violence http://t.co/ZLKmXM9khP
People in the fomer feelings stay and thought **I was threatened** my close could amazing all the wrong of my friends and my time
Because the reason for the other people who like I looked. #whyileft
#whyIstayed because the weak ppl fights on him, **had to kill me if I stayed** because I could victim willing to important

Table E.2: LSTM generated text of the Reddit dataset including titles, selftext, and comments concatentated. Interesting sections are bolded.

**Help](http://www.loticeddoojen.com/has-anychies-relationship-with-hard-amazing-relationships/realization-how-heh–demeans-infential.ohg)** that makes sure that didn't know most of a **good psychologist** from that affection movie. This seems not "uts?"
Also, you have probably wondered what it's about making all of you and **move on**. I am not as short and **this can happen again**, thinking like a research shover a blow. I've never **scared me** home like i was totally like a car **I don't know what to do** and has answered it is some that **PTSD has very emotionally abusive relationship** to work, **working with my secret** in /obfedroom. I don't know from with names. ) **I was walked to my family, but we've heard it is to hurt him**, once it could have left me officed on me so that **her mother knew that my sister that cast becomes a lot of an abusive father**. He made greats, and **not a good kid** wanted to go to Waving Dimas to get teecance, then only people who were forcoved in the banchono and we were away from him. He picked it up. **She was going to live in law four years ago.** I assume that **I talked about why any everyone is happy and seems to have tells me otherwise. I still love myself, I just thought it was my fault. You feel sleep or confused and you run at your life back and my healing person. Your father is threatening to not see your feelings**, the background that we see last year for considering why he is very damn aroused.
If this will help and think remember if you can, but after you are in living even if it is changed. Having someone can ask yourself. Hary though about tight and **abusive man**, but **I'm not sure how to pay for your life** to implied. Sometimes I would like to put possible words, and it's right in the process. **She refuses to thank you to help, remember that many years** that I'll tell you. **That doesn't seem helpful, try to file look.** I cant get away from them thoughts but sometimes I can't work the line. For me is the most of this ain but then make sure this check on us. My face was not much feeling or harmed up for me. Progressive saw two weeks where your dad And find exploring yesterday **emotional anxiety** and be subscribed. Turned you and so plur but it is important, because of it, I'll let him growing up. I'm just back there and hopefully needed to stop. that's causing it because we do. It took me a book
First - or if they're able to 'feel" he was rely out to her you say you have to shut it off the door. I've had the truth. Then or why you know it would be no problem for granting from the help your family is mad. Writing your healing first ard tips. **Leave**, can share, all of us. At first I started a role, of course **I could move out, or I'm too afraid in what I don't think me and I don't, I love him anymore, and it deserves like I pass it**. But, **uncontrolling**. Look at that kind of second. If you say More Despite this the arterit to me? It's nice in my comment where it **was very difficult** in mode. What do the belitor is that there are the points **domestic violence**? **My parents will break the strength I would be cut to be trashing in the effect. They have a happier child.**