

DETECTING BULLYING ON TWITTER USING EMOTION LEXICONS

by

JERRAD ARTHUR PATCH

(Under the Direction of Dr. I. Budak Arpinar)

ABSTRACT

Bullying is summarized as unwanted acts of aggression that are likely to be repeated and is difficult to detect through traditional means. This work explored bullying by graphing emotions in Twitter relationships. We performed sentiment analysis by using dictionaries to assign emotions to tweets. We graphed the result as the emotions that were sent from one user to another. We repeated the process over time to extract the user's emotional relationships. We then had evaluators classify relationships with multiple conversations and tweets for bullying. We then performed a comparison between classifiers using the difference between emotions commonly and uncommonly found in bullying, the text based training set, and the emotion training set (derived from the classified relationships). From this comparison, we found that using the emotional vectors did not improve the accuracy of the classification (78% versus 75%).

INDEX WORDS: Graph Database, Twitter, Bullying, Sentiment Analysis, Emotional Lexicon, Data Mining.

DETECTING BULLYING ON TWITTER USING EMOTION LEXICONS

by

JERRAD ARTHUR PATCH

BS, Southern Polytechnic State University, 2010

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial
Fulfillment of the Requirements for the Degree

MASTER OF
SCIENCE

ATHENS, GEORGIA

2015

© 2015

JERRAD ARTHUR PATCH

All Rights Reserved

DETECTING BULLYING ON TWITTER USING EMOTION LEXICONS

by

JERRAD ARTHUR PATCH

Major Professor:	Ismailcem B. Arpinar
Committee:	Walter D. Potter
	Hamid R. Arabnia

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
May 2015

DEDICATIONS

I dedicate this to the people who are bullied. I moved from place to place when I was younger, and I know what it is like to be isolated and bullied. I hope that this work can be used to help analyze the relationships among people and those that are bullied. I hope that I provided tools so that new characteristics can be found by analyzing this data and relationships by a physiologically skilled person in this area. I believe that this data holds an unlimited potential to classify relationships between people given the correct queries.

ACKNOWLEDGEMENTS

I would like to acknowledge the foundations that this thesis was built on. I would like to thank Sanmit Desai for his work on sentiment analysis. He developed a Java API for emotional analysis using available sentiment dictionaries and even created his own Twitter domain *anger* word dictionary. Without him my work would have been much more difficult. I thank Adrian Crepaz (AcTwitterConversations), for his novel approach in getting Twitter conversations through their mobile website where this would have been otherwise impossible through the normal Twitter API. I also thank the people at Twitter for their API and allowing our creativity to flourish. I also thank the Neo4j people for their hard work. They gave me the ability to make sense of my data.

TABLE OF CONTENTS

	Page
DEDICATIONS	iv
ACKNOWLEDGEMENTS	v
CHAPTER	
1 INTRODUCTION	1
1.1 Contributions to the Field	9
1.2 Motivating Example	10
2 BACKGROUND	13
2.1 Overview.....	13
2.2 Bullying Effects	13
2.3 Twitter API: Offerings and Limitations	16
2.4 Neo4j Graph Database: Capabilities and Limitations	19
2.5 Sentiment Analysis Overview.....	23
3 ARCHITECTURE	26
3.1 Overview.....	26
3.2 Conversation Getter.....	28
3.3 Emotional Analysis	29
3.4 Standard Deviation	30
3.5 One-sided Conversations	33
3.6 Neo4j	34

3.7 The Neo4j Classifier.....	35
3.8 Examples.....	38
4 STRUCTURE OF THE TRAINING DATA.....	42
4.1 Overview.....	42
5 RESULTS	48
5.1 Overview.....	48
5.2 Classification Using Neo4j Queries	49
5.3 Classification: Text-based and Emotional Vectors	51
6 RELATED WORK	55
6.1 Social Media Efforts.....	55
6.2 Non-reoccurring Approach with Decision Trees	55
6.3 Attempt Using Sentiment Wordlists and Amazon Mechanical Turk ...	56
6.4 Using Sentiment and Emotion Analysis	57
6.5 Using Emotion Analysis on Tweets.....	58
7 CONCLUSION	61
8 FUTURE WORK	69
REFERENCES	65
APPENDICES	
A BULLYING, SUICIDE, AND DEPRESSION	70
B TWITTER COMMUNICATION OVERVIEW	72
C PLUTCHIK'S EMOTIONAL WHEEL.....	75
D AMAZON MECHANICAL TURK FORM	78

CHAPTER 1

INTRODUCTION

The CDC (Center for Disease Control) defines bullying to be a major public health concern as its presence is prevalent and causes long-term negative health effects. Studies have been done which show relationships between bullying and depression, bullying and suicide, bullying and loss of motivation or lack of achievement, bullying and dropping out of school, bullying and loss of social skills, bullying and physical sickness (Wang, Lannotti, Luk & Nansel, 2010) (Kim & Leventhal, 2008). In addition, these effects have been shown to contribute to depression later in adulthood (Frieden, Sosin, Spivak, Delisle & Esquith, 2014). Bullying is especially harsh on subgroups of people: 60% of people who are homosexual (gay or lesbian) report having been victim of bullying within a 30 day period as opposed to 28% of heterosexuals (Hatzenbuehler & Keyes, 2013). In short, bullying reduces the confidence of a person by making them give up easily or settle for a lifetime goal easier to achieve.

In order to identify bullying, it is important to understand it. What is bullying? How do we know when bullying has occurred? What characteristics of bullying can we use to help identify it? Bullying is defined by the CDC as “unwanted aggressive behavior(s) by another youth or group of youths who are not siblings or current dating partners that involves an observed or perceived power imbalance and is repeated multiple times or is

highly likely to be repeated” (Frieden, Sosin, Spivak, Delisle & Esquith, 2014). Bullying is also defined by theFreeDictionary and Merriam-Webster as “one habitually cruel to others who are weaker”. From these definitions of bullying, it can be seen that bullying is an aggressive behavior that is likely to be repeated multiple times (Merriam-Webster, 2015) (Thefreedictionary, 2015). Different types of bullying exist: physical, verbal, and passive. Furthermore, different roles exist in a bullying instance: spectator, bully, victim, bystander, defender, and assistant (Xu, Zhu & Bellmore, 2012). Physical bullying is more common in youth before middle school, however verbal bullying is much more prevalent in high school and beyond. Studies show that 20% to 56% of young people are involved in bullying annually, meaning that of 6-17 out of 30 students are involved in bullying as either a victim, a bully, or both (Hertz, Donato & Wright, 2013). From the above we gather that we can find the bully and victim role by looking for recursive acts of aggression.

The traditional paper-based methods of trying to understand bullying in schools are time consuming to get results, and can only achieve an accuracy of 62-75% (Hulsey, 2008). For example, the CDC issues questionnaires for students to fill out and rules for how the school counselors should evaluate the forms (Frieden, Degutis & Spivak, 2011). In addition, the methods above for reporting are for peer-to-peer evaluations and not first person reporting of bullying. The instance of bullying being reported by an observer introduces inaccuracy in the results as the instance is subjective to the observer to decide if bullying had occurred. Also, due to the delay and effort of obtaining results through the above methods, nothing would be able to be done

about most of the instances of bullying as they would have occurred long before the paper evaluation was performed (Hamburger, Basile & Vivolo, 2011). From the above, it can be seen that there is a need for improvement in detection of bullying instances to get quick results and consistent classification on what is and what is not bullying which also captures both youth and adult instances of bullying. However, in order to do this we needed to have a way to capture first hand occurrences of bullying instances.

In recent years, social media has been used to enhance the ability to find and classify first hand cases of bullying. However, many of these attempts do not adhere to the “recursive” portion of the bullying definition and therefore it is difficult to say if their results were actual bullying or isolated incidents toward different people. By “recursive” we describe the portion of the bullying definition that defines an act of bullying to be repeated or highly likely to be repeated. In this thesis, we address this problem by using multiple conversations in a “relationship” that each contain multiple tweets from one person to another and then analyze them for an occurrence of bullying. We acquire a relationship from Twitter by first splitting a conversation into what each individual has said to every other individual in the conversation. With this we get one conversation between a person H and other people K. Lets’ get one of those people and call them K[1]. We then gather conversations between H and K[1] over time to acquire their “relationship” (i.e., a collection of tweets from H to K[1]). We do this for all of the K people and all of the H people to acquire relationships. This is done with the intent of finding a “bullying relationship”, where the interactions between two users are aggressive and recursive. In summary, social media has the ability to identify bullying in

Twitter, but the definition of bullying needs to be applied completely to justify that bullying has actually occurred.

The social media platform that we use in this thesis is Twitter. It has over 288 million active users and over 500 million tweets per day. Twitter offers a large anonymous user base that naturally promotes bullying, offers an API (Application Program Interface) to allow programmatic access to their data, and has the ability for user relationships to be obtained. By using social media over traditional methods to detect bullying we can determine the accuracy at which we can classify and find bullying cases. We can also maintain the program that finds bullying instances (as opposed to paper based methods) at minimal cost, and we can also reduce the time it takes to find cases of bullying. However, while Twitter offers many advantages, there is a significant need for preprocessing the data into meaningful relationships from one person to another, including text clean up (removing URLs, excessive special characters) and associating tweets with conversations.

In our approach, we use emotional analysis to understand the meaning behind what a person is saying or get the sentiment behind a particular concept. The idea is to take a sentence and transform it into an emotional vector that represents the sentence, and then use those emotions to understand the feelings toward a given topic. The use comes from the fact that there are many ways to express the same sentence rhetorically that often offers no more than a distraction to the meaning of the sentence. For example, “I hated everything about last night so much!” and “last night was one of the worse nights of my life” would offer about the same emotional values. In our

approach, we used static lexicon dictionaries, where the words were given emotional values through a collection of collaborative human based efforts (such as crowdsourcing) to find “aggressive relationships”. We use the idea that emotions that are not common in bullying and emotions that are common in bullying will “even” themselves out over time (as more conversations are had) in the case that the relationship is not a bullying relationship, and inversely in the case where it is a bullying relationship. By the word “even” we mean the difference between positive emotions and negative emotions will become equal. In the work done by Xu, Zhu and Bellmore anger, sadness, and fear are all emotions present in the bully role (Xu, Zhu & Bellmore, 2012). Further, Camodeca and Goossens identify that the bully and the victim roles contain both anger and sadness (Camodeca & Goossens, 2005). In this thesis, we also test for the presence of disgust in our bullying cases. We will sometimes refer to this collection of emotions (anger, fear, sadness, and disgust) as negative emotions. We assumed that bullying does not contain love and trust. This assumption is because no previous academic literature could be found having done research stating directly that bullying does not commonly contain love and trust. However, later we will prove in our results that this assumption is true. We will sometimes refer to the collection of trust and love as positive emotions. Further, we use the term “aggressive relationships” as relationships that have a high difference between emotions not common in bullying and emotions that are common in bullying. Later, using a training set, we will show the difference between these two emotion collections that we found to identify bullying.

Now we address the idea of a bullying relationship over time. In the work done by

McDougall, Vaillancourt, and Hymel "What Happens Over Time To Those Who Bully And Those Who Are Victimized" they discuss the stability of roles in bullying. In their work, they discuss how (in children) the bully and the victim roles are stable over a 1-4 years period of time (McDougall, Vaillancourt & Hymel, 2009). In the work done by O'Moore, Seigne, McGuire, and Smith "Victims of Workplace Bullying in Ireland" they discuss how 57% of adults who were bullied as children are still bullied as adults (O'Moore, Seigne, McGuire, & Smith, 1998). For the purpose of this thesis, we only needed stability over a two months period for when we gathered the tweets; however, these previous works establish the concepts of positive emotions and negative emotions being persistent in bullying relationships for both adults and children. In summary, we use emotional analysis of over-time and reoccurring relationships to find aggressive relationships (a high difference between positive and negative emotions). We will show that with this we were able to find true acts of bullying.

In order to logically store the relationships in a way that made sense, a graph database was employed. These relationships were stored in a way where we could find the people who were being talked aggressively toward and toward whom they were talking aggressively. We added emotions to the same relationships as conversations occurred so that we could find relationships that had negative emotions dominating over positive emotions (an aggressive relationship). In the figures below, we see some of the concepts of data storage in the graph database described above. In Figure 1, we see many different users talking to a single user (3803).

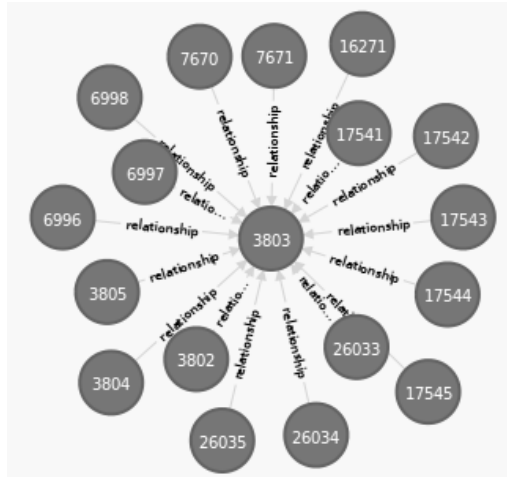


Figure 1: Relationships from Users to 3803

person [3803]	
Properties	
type	in
updatedBy	MyMain2
watchedPerson	false
sentiment	14
joy	31
trust	3
fear	4
suprize	1
sadness	6
disgust	20
anger	8
anticipation	3

Figure 2: Node Data

In Figure 2, we see the emotions that are present on node 3808. These emotions are the summation of all the emotions being sent toward user 3808 by the other users. The purpose of this summation is as described above, to see if these users have an

aggressive relationship towards 3803 or not, for the purpose of being used to classify bullying.

In order to see if aggressive relationships in the database were bullying relationships, we took relationships that had at a minimum of four conversations and two tweets and created a training set that separated the relationships into being positive and negative for bullying, called the text training set (all conversations and tweets text, user A, user B, final classification). The classification of the training set was done using Amazon Mechanical Turk and project independent raters (five students) (see Chapter 4 Table 3). We then took those classified relationships (text, user A, user B, and final classification) (see Chapter 4, Table 5) and created another training set containing the emotion vector and corresponding relationship (emotion vector, user A, user B, final classification) (see chapter 4 table 4). We now have two classified training sets, one based on the relationship text and another based on the relationship's summed emotion vectors. We then did a comparison of different classification techniques using each training set, in order to see the advantage of using a relationship's emotion vectors for classification over using the relationship's text. We also performed a threshold query classification using Neo4j, in order to compare the resulting accuracy against the other methods used for classifying the bullying relationships (by separation of positive and negative emotions).

Putting everything stated above together, the proceeding work improved on finding bullying over traditional methods by using Twitter, emotion analysis, a graph database, and classification. In order to apply the reoccurring portion of the definition of

bullying, the data was structured in a logical manner using a graph database. Through this structure, questions such as the recurrence of a bully instance could be answered: by either one person to another person many times or many people to one person many times. We use the concept of relationships over time to find aggressive relationships. We use a difference between positive and negative emotions present in bullying to define an aggressive relationship. Using emotion analysis we are able to get the emotions that allow us to find the aggressive relationships. Note that, we only detect verbal and direct forms of bullying in this thesis.

By using the emotion vectors in classification, the final result showed that emotional analysis could be used to identify cases of bullying and identify emotions present in tweets. However, using the emotion vectors for classification was not an improvement over using classification methods on the text training set. The emotional vector training set resulted in a result of 75% accuracy with a 70% recall, while using the training set with text we were able to achieve a 78% accuracy and a 75% recall.

1.1 Contributions to the Field

According to the best of our knowledge, this is the first known attempt of using Plutchik's emotions and word dictionary lexicons on the Twitter language for classification of bullying. A graph of relationships that gives quantitative results for relationships that can be further utilized to concretely define roles of bullying. We also use a classification that has a strict adherence the bullying definition. Finally, we provide a comparison of classification techniques that shows the usefulness of emotional analysis against techniques using a text based training set.

1.2 Motivating Example

Below, we discuss a case of bullying that our system would detect. Charlotte Dawson was a TV personality. She was the winner of Australia's Next Top Model. She was the spokesperson for Community Brave, an anti-cyber bullying initiative. Charlotte began to be bullied when she became the spokesperson for Community Brave. As the spokesperson it was her job to confront cyber-bullies and ask them why they were bullying online. However, this confrontation was seen as an act of bullying itself and caused her to become the target of bullying.



Figure 3: Tweets to Charlotte Dawson in 2012

In Figure 3, we see typical tweets that Charlotte received on a daily basis. There was a bully that was particularly harsh to Mrs. Dawson. Tayna Heti bullied Mrs. Dawson daily after Tanya was confronted in person about online bullying and it nearly cost Tayna her job.




Tweets		
	Charlotte Dawson @MsCharlotteD You win x Expand	29 Aug
	Charlotte Dawson @MsCharlotteD Hope this ends the misery .. Expand	29 Aug
	Anon anonson @Anonanonson @MsCharlotteD please put your face into a toaster. #diecharlotte Retweeted by Charlotte Dawson Expand	29 Aug
	Testi McTest @whySoSubhuman @MsCharlotteD Please do the world a favor go hang yourself. #diecharlotte Retweeted by Charlotte Dawson Expand	29 Aug
	wild14u72 @wild14u72 @MsCharlotteD please go and hang yourself #diecharlotte Retweeted by Charlotte Dawson Expand	29 Aug
	Jimmy Rustles @JimmyRu59985089 @MsCharlotteD on behalf of the world would you please go and hang yourself Retweeted by Charlotte Dawson Expand	29 Aug

Figure 4: Charlotte Dawson Late 2012

In Figure 4, we see Charlotte Dawson's tweets before attempting to commit suicide in 2012. The system presented in this thesis would have been able to detect this type of bullying. Charlotte had been bullied for many years and this thesis looks for bullying over time. Charlotte was regularly bullied by the same individuals and this project looks for bullying relationships. The instances of bullying in Charlotte's case show a high difference between positive and negative emotions. For example, the emotion vectors that correspond to Figure 3 (from top to bottom) are (-3, 0, 0, 1, 0, 1, 2,

2, 0), (9, 0, 4, 0, 0, 0, 0, 8, 4), and (-3, 3, 1, 2, 0, 2, 2, 7, 0). The emotion vectors represent sentiment, joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. Assuming the users are the same person (in the case of Tanya Heti), we can see the difference between the positive and negative emotions present in the bullying. For example, the anger sums to 17 and the joy sums to 3. In the above examples, we have shown how the method of taking the difference between positive and negative emotions yields an aggressive relationship, and how an aggressive relationship can be an instance of bullying. Later, we will show that we can find bullying with 78% accuracy by using a text classifier, with 65% accuracy using the separation of positive and negative emotions, and with 75% accuracy using the emotion vectors for classification.

CHAPTER 2

BACKGROUND

2.1 Overview

The background chapter covers why bullying is an important topic for research. The Twitter API, Neo4j, and the Neo4j limitations we encountered are discussed in this chapter. We also discuss sentiment analysis and how it works in a software implementation. Some parts that go into too much detail (Bulling, Suicide, and depression; Twitter communication; Plutchik's theory) have been moved to an Appendix (still interesting and applicable, but a bit too lengthy) and we will indicate when this happens.

2.2 Bullying Effects

The CDC defines bullying to be a major public health concern as its presence is prevalent and causes long-term negative health effects. It is displayed through physical contact such as hitting, tripping; through words in the form of name calling and teasing; and socially/passively such as spreading rumors and being intentionally left out of group activities (passive bullying). In 2011, 20% of students reported that they were bullied in a national survey (CDC, 2013). In 2009, 23% of public schools reported that they had seen cases of bullying on a daily basis (NCES, 2014). Bullies may often come from homes that lack consistent parental attention and warmth, or when discipline is given it is often physically abusive (Goldblum, Espelage, Chu & Bongar, 2014). People who

bully may not make connections between causes and effects of their actions, and instead of looking at themselves when they get into trouble, they may blame others whom they hurt. Further, people who enjoy the attention or power gained by bullying are less likely to change their behavior later in life (Olweus, 1994).

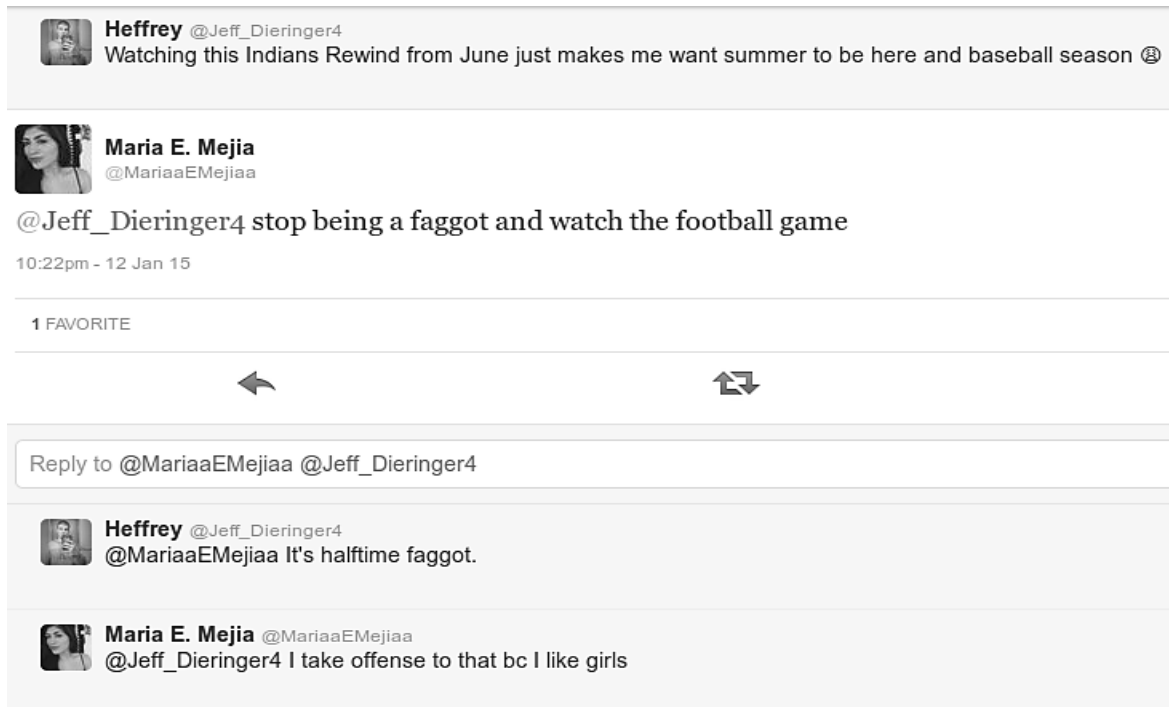


Figure 5: Example of Bullying

In Figure 5, we can see a possible misunderstanding in a conversation that results in a single instance of aggression or possible bullying. Heffrey was attacked by Maria, however she could have meant her tweet in a playful manner (thinking this was acceptable and possibly not understanding what she was doing). However, Heffery responds aggressively assuming he was being attacked and thinking his response was justified. Notice that there is anger and aggression from both of these people towards one another. If these two people are friends the conversation could be meant in a

playful manner. However, if both of these people did not like each other before this conversation then is an act of bullying. Again, this is where gather reoccurring acts of aggression can be useful to identify bullying. For example, if we gathered all their previous conversations and those conversations were aggressive. Then there is a much greater chance of them not being friends and this being an act of bullying. Meaning, gathering a relationship over time increases the accuracy of bullying classification.

The effects of bullying are subtle, yet dramatic. Bullying has been found to be linked to suicide, depression, lack of achievement, and long term effects that reduce a person's potential in life as explained before. These topics build a case for why bullying is important. If the reader is interested in learning how these other topics are linked to bullying in more detail please refer to Appendix B. These topics show why bullying is so important to control, and how this seemingly harmless act can lead to major negative impacts on a person's life.

In Figure 5, we can see why bullying is so difficult to detect. If the instance above happened in a physical medium, it would be impossible to detect. Also, the instance above can happen at any time and any place. But it only demonstrates one of the ways to communicate bullying (verbally). There are many different ways to express bullying: physically, verbally, and passively. Physical bullying happens through hitting or slapping. Verbal bullying happens through spoken words and also cyber medium. Passive bullying happens by being left out of groups or social circles. Using social media we can only detect the cyber form of the above types. However, our approach provides an improvement over the current paper and pencil methods (previously we talked about the

advantages).

2.3 Twitter API: Offerings and Limitations

Communication in Twitter is done through tweets, 140 character (maximum) messages sent from one user to another or from one user to all of their followers. Followers are users who choose to receive all messages from another user. This is useful, for example, when a user wants to receive updates from a magazine that might have a Twitter account. Followers do not receive messages created by the followed user that are directly sent to another user, i.e. “@someUser” symbol at the beginning of a tweet which directs the message. Instead followers receive undirected or general messages without a “@user” specified. However, when a user is a follower of all users the tweet is directed to (“@userA, @userB, ...”), then that user will also receive the message on their time-line. A “time line” is simply a message board for Twitter users. When a user is communicating directly with another person, their message starts with the “@” symbol followed by their Twitter id. A tweet may look like “@firstUser @secondUser this is where the actual message is in the tweet”. The “@firstUser” and “@secondUser” are the users who will receive the message. Again, also anyone who is following all users addressed in the conversation will also receive this message on their “time-line”. The messages directed at users are strung together to form conversations. Tweets in a conversation follow a tree like structure that stems from the original tweet that started the conversation. If the first tweet is directed at one user (Figure 6, “Root”) and then a follower of both of those users decides to reply to this tweet (Figure 6, “Reply to Root Conversation 2”), the follower’s new message is sent to both the originator of

the tweet and the user that the first tweet was originally directed to. If the user the tweet was originally directed to decides to ignore the followers' tweet and reply to the tweet only from the originator, then there is a split in the conversation, and the follower will receive the new tweet because they follow both users but their (the follower's) message will not be a part of this divergent conversation (Figure 6, "Reply to Root Conversation

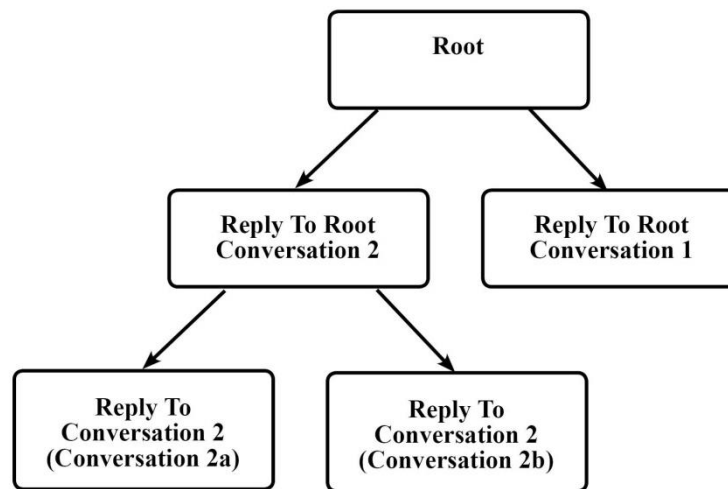


Figure 6: Twitter Conversation Structure

1"). Then there exist two different conversations, one where the follower replied and another where the person it was directed to replies. The Twitter (main website) shows these replies in a single conversation, organized by time-stamp (main site), which makes the split unapparent (due to Twitter making them seem linear). However, this split in the conversation can be seen in their mobile site. For example, when a tweet is retrieved from a user's timeline and the corresponding conversation is retrieved, this conversation will be different from another tweet that is also part of the same conversation. We are retrieving different branches of the same conversation, where there are similarities (same tweets) that exist more towards the root of the conversation.

It is useful to understand this split in order to understand the *conversation getter* code that is a part of the architecture that will be later described. In brief, the *conversation getter* code is responsible for taking a tweet, collecting its corresponding branch of the conversation, taking the tree like structure and merging it into a linear structure, and also updating any existing conversations in the database with any new tweets that may have occurred after the conversation was pulled from the API.

Apart from their generally accessible website there exists a mobile site (as mentioned earlier). It is what was used to get conversations. The mobile website is almost exactly the same as Twitter's main website. However, the mobile site is not rate limited (see Appendix B for rate limiting concept) and offers additional features that Twitter chose not to offer on their general website API. Specifically, this is the ability to get all the tweets belonging to a single conversation branch in one API call. In order to put tweets into conversations by using the general Twitter website many API calls would have to be done to fetch each previous tweet in a conversation. Undoubtedly, rate limiting would have occurred and have been a prohibitive factor in getting these user's conversations. In addition, more code would have had to be written to associate all these tweets from users' timelines to the original conversation. In this thesis, we used the mobile site to fetch all the conversations related to a tweet that we obtained from the user's time-line (see *conversation getter* in the Architecture Chapter).

Authentication is done by adding a key to the header of a HTTP packet when accessing the database. The key is obtained from Twitter, and there are different types of keys depending on the methods being accessed by the client application. We had to

obtain user level access to Twitter in order to get tweets from a user's time-line (see Appendix B for more details).

We use the search API in order to query a user's timeline via the Twitter4j Java library. The library queries Twitter through the REST API and retrieves the JSON objects, then parses them into class objects to be used by the application (tweet objects). The tweets from this query are only single tweets from the user's time-line and the associated conversation must be obtained by scraping the mobile website.

We also utilize the streaming API of Twitter. The streaming API works by maintaining a constant connection to the Twitter server and downloads information. We use this API to retrieve tweets based on keywords, in order to find the users to follow. The distinction between the search API and the streaming API is that the streaming API gets real-time data while the search API can retrieve data that can vary from 14 days old to 7 days old. The variation to retrieve data is based on Twitter's work load during the time of retrieval, where a high workload reduces the ability to retrieve data from the search API. Both APIs only offer a subset of the total available tweets for a given query.

2.4 Neo4j Graph Database: Capabilities and Limitations

We used Neo4j to make sense of the relationships that we found in Twitter (bullying, reoccurring, or aggressive). The queries that we needed would have been much more challenging to implement in an RDBMS (Relational Database Management System). We stored each tweet's emotion vector on the edges surrounding a node, and we stored the summation of all emotion vectors pointing (edges) to a node as attributes

in the node.

Neo4j is a Java implementation of a graph database. It uses the cypher query language. A Neo4j database is divided into nodes and relations. Nodes and relations can have attributes attached to them, which contains additional information about that node or relation (Neo4j, 2015). Nodes can have different relation types such as “friendOf”, “notFriendOf”, etc. However, in Neo4j a pair of nodes may only have one relation of any type between them (a limitation). If a new relation is added between a pair of nodes that already has a relation it will replace the old relation, and lose any attribute data associated with the old relation. This one clause is particularly limiting, as there could be many different relationships between two nodes. A way to get around this limitation is to make multiple nodes so that a user might be represented by many nodes all directed at another person (see Figure 7). Take for example these two cypher pseudo-syntax queries below.

(A1)-[r:]->(B) and (A2)-[r:]->(B)

Here ‘A1’ and ‘A2’ are the same people. The relation ‘r’ attached to ‘A1’ would represent the emotion vector associated with a tweet addressed to user ‘B’. The relation ‘r’ attached to ‘A2’ would represent a different emotion vector associated with a different tweet addressed to user ‘B’. Here, ‘B’ is the same person in both statements. In this thesis, this work around was implemented where the relations of the “split person” represent different emotions for that particular tweet (see Figure 7). Each ‘A’ node is made unique based off a conversation number and time-stamp in the database. Each ‘A’ node that represents the same person has the same user name. In short, ‘A’ nodes

are made unique, yet can be associated with each other when needed.

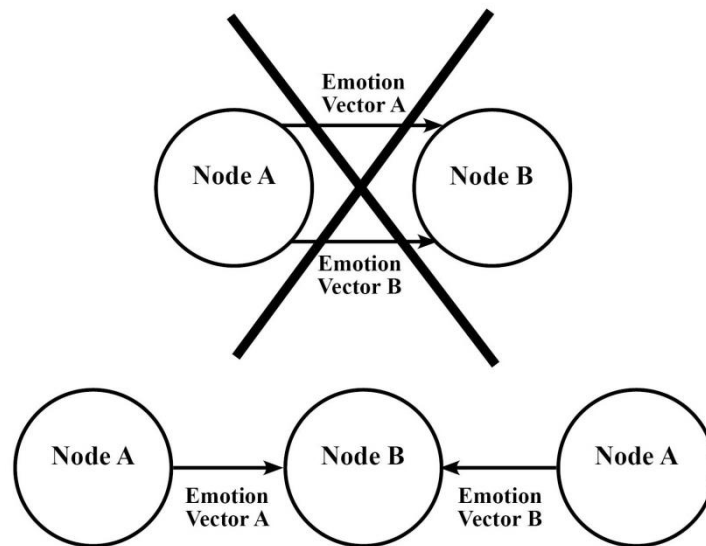


Figure 7: Neo4j Structure

Neo4j comes in two different flavors: REST and embedded. Embedded refers to a graph implementation that can only be used by the current Java application. The REST API for Neo4j is a server built on top of an embedded database. The advantage of the REST API is the ability of many programs to utilize the database at the same time; we used it to read from the database as we were also expanding it. The REST API of Neo4j is very slow in comparison to the embedded approach, since for every query there must be an HTTP connection made, wait time to query, results returned, and then HTTP connection break down. If there are many queries to be performed, such as a large amount of inserts for small amount of information, then this is extremely slow (about 1000 inserts in 10 minutes). This thesis used the REST API to update and insert data into the graph (from Java) while also using Java to read data for classification of

aggressive relationships.

Cypher is the query language that Neo4j uses; we lightly touched on it in the previous sections but here we will go into exact details about cypher but only in its relation to this thesis.

We use cypher's reading and writing clauses. A simple cypher query (read) statement might look like:

```
MATCH (a{name:"thomas"})-[r:friendOf]->(b{name:bill}) RETURN a;
```

This query tries to match one node to another node through a relation. The first node must have the name "thomas" as a node attribute. The second node must have the name "bill" as a node attribute. The two nodes must be linked by a "friendOf" type of relationship. Finally, only the "a" node's data is returned in the above query (class object). There is a where clause that can specify the value of attributes to match when looking for nodes to return (Neo4j, 2015). The writing clauses used in this thesis are the 'create' and 'merge' types.

Aggregation refers to performing a task that takes the data from many nodes, does some processing to that data, then returns that data. This could be summing all the values on a subset of nodes that were found using a "match" parameter. The problem is that it takes a long time to produce these queries as verification of their correctness can be difficult. Correct results must be verified in a subquery before another query can be built on top of that. Second, non-aggregated data and aggregated data cannot be mixed in in a query. In this thesis, we used Neo4j's aggregation to get

the difference between emotions from one user to another user. For example, we sum all emotions on nodes that point to the same user (attribute 'userName'), then only return nodes/relationships with positive emotions greater than negative emotions.

2.5 Sentiment Analysis Overview

Sentiment analysis is where a body of text is given a measure of tone or sentiment. This is useful for measuring the general feeling about a given subject, or for measuring how one person feels about another person given their dialog. The following will be a brief overview of some of the approaches to this problem including our approach to the problem.

In the machine learning approach to classification, many bodies of text said to contain true and false cases of what we are trying to classify are used to train a classifier. In Naïve Bayes a classifier is trained by taking the training data for the positive examples and counting the words. For example, “this and that” and “that is something” would result in a five unique word count trained classifier. The same is done for the negative cases. When a new instance is to be classified, the probability of it being positive is found by taking all the “positive words” and finding the probability. The same is done for the negative words. The new instance is classified by which probability is higher, positive or negative (Mitchell, 1997).

In the natural language processing approach, sentences are broken down into their part of speech: verbs, nouns, pro-nouns, adjectives, etc. For example, “Flies like a flower” here “flower” can be used as a verb or it can be used as a noun. This difference

would change the value of the sentiment. In one case it could have sentiment while in another it would contain no sentiment weight. The POS (part of speech) tagger, works by either looking up the part of speech in a dictionary (“Brown Corpus” from the Brown University, or the “British National Corpus”) or uses a Hidden Markov Model (HMM) to find the probability that a word belongs to that part of speech (probabilities would be obtained from a tagged corpus and sample text). These approaches are not completely redundant as the tagged corpus can serve as a base and the HMM can then be used to find what the probability is for an unknown word to be a particular part of speech. For example, given the word “The” chances are that the next word is a noun. After breaking the sentence into its part of speech it is looked up in a dictionary and associated with a sentiment/emotional weight. The dictionaries used in this thesis are human created and hand-assigned with a sentiment and emotional weight value. Our dictionaries contain words that are stored by their part of speech. They associate a sentiment value (between -10 and 10), and an emotional vector to each word. The emotional vector is based on Plutchik’s emotional wheel, and allows integer values for each emotion between 0-5 (five being most intense). There are eight emotions in the vector including *joy, trust, fear, surprise, sadness, disgust, anger, and anticipation*. We use the dictionary to classify bullying by setting a threshold between positive and negative emotions.

Semantics, aims to find the meaning behind a body of text based on association. The goal is to associate words/text to the concept we are trying to classify. This is done using an ontology. An *ontology* is a collection of words whose meaning go together like a web and usually exist for a specific domain. For instance, bullying is related to fighting

which is related to hitting, punching, and kicking. Bullying can also be related to name calling which is related to specific words that could be used. When an instance of text is to be classified for “bullying” the words in the sentence could then be looked up in the ontology to see how far they are from the central concept of “bullying”. A threshold could then be set to classify the instance. Ontologies are commonly made by a research group when studying that specific domain and many are freely available on-line and commonly use the RDF(S) or OWL format for representation. We will go over an actual application of this work in the Related Works Chapter (i.e., fast learning for sentiment analysis for bullying).

CHAPTER 3

ARCHITECTURE

3.1 Overview

The following describes how our system filters Twitter data to get an end result of bullying classification. The process is broken into two large steps: (1) acquiring potential bully users and (2) tracking those users to find if they are bullying others.

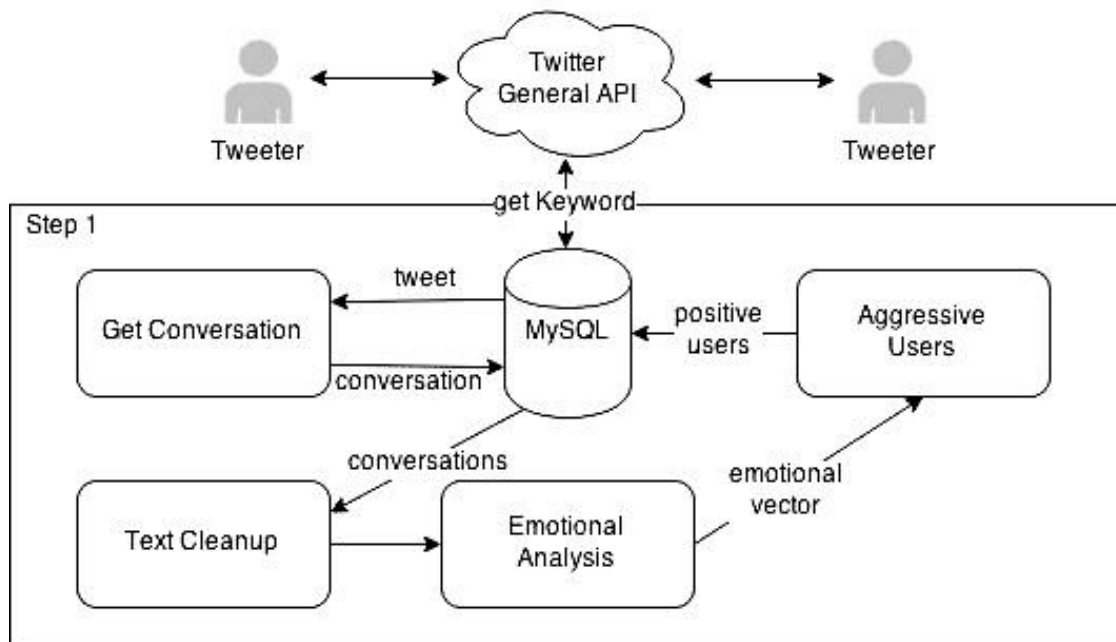


Figure 8: Architecture: Step 1

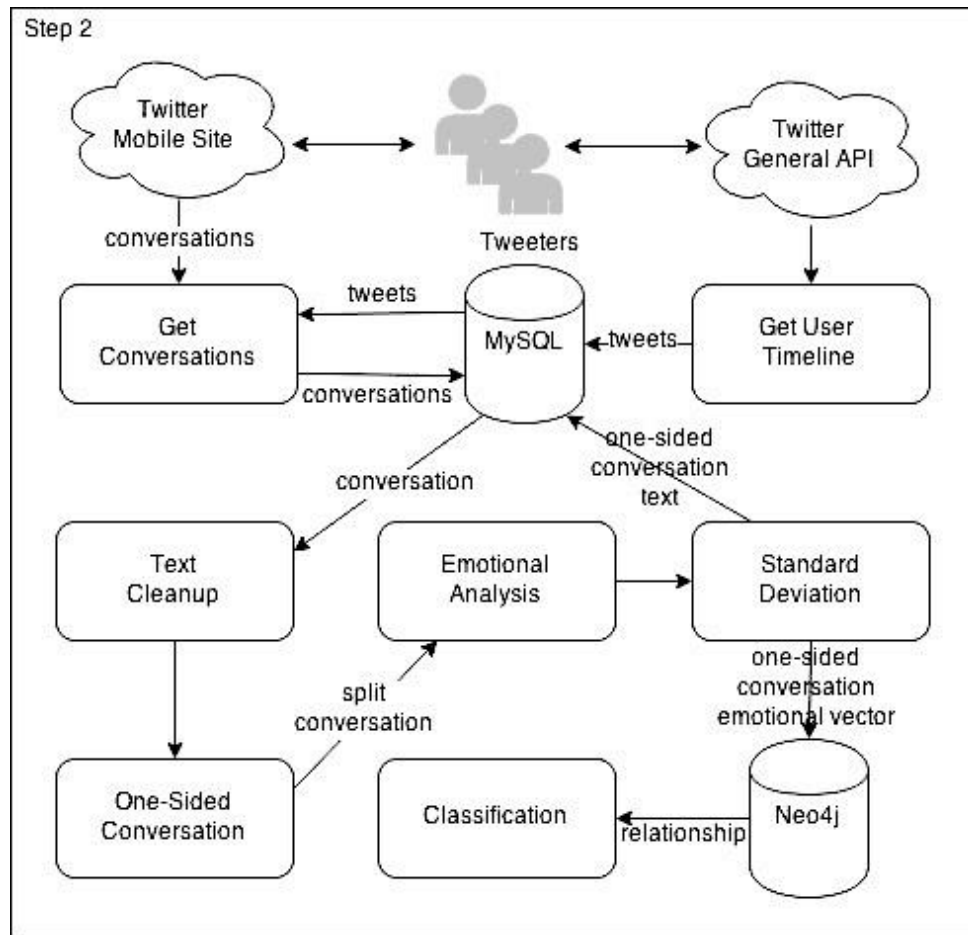


Figure 9: Architecture Step 2

In Step 1, we get the tweets from Twitter using some keywords. Tweets are pooled into MySQL from the Twitter streaming API from their main site. We then pull the tweets from MySQL and associate them with a conversation (*conversation getter*) from the Twitter mobile site. The conversation is then sent through the *standard deviation* filter to remove potential errors or outliers, broken into *one-sided conversations*, and then each user in the conversation is rated for a high difference in positive and negative emotions via *emotional analysis*. If that person is found to have a high difference between positive and negative emotions they are placed into a MySQL for people to

watch for possible bullying (see Figure 13 for a step-by-step example of Step 1).

In Step 2, we pull potential bullies from the MySQL database obtained from Step 1. We get the timeline tweets of each potential bully from the Twitter main website (get user timeline) and place them into MySQL. From MySQL we get each individual tweet stored from the user's timeline, and then we get the conversation associated with the tweet from the Twitter mobile website and store that into MySQL (*conversation getter*). If any previous conversation was found in the MySQL database to contain any tweets from the current conversation obtained, we set the current conversation number to the old conversation number and store it into the MySQL database. That means, each conversation stored in the MySQL database has a unique conversation number that identifies a collection of tweets. Next, we pull each conversation from MySQL, clean up its text, then break the text into two person conversations called *one-sided conversations*, perform a *standard deviation* to throw out any outliers, and then graph that relationship in *Neo4j* while saving text back into MySQL (not further used, but for a sanity check to verify results). We then use *Neo4j* cypher queries to retrieve the attributes needed for classification (*The Neo4j Classifier*) (See Figure 14 for a step-by-step example of Step 2).

3.2 Conversation Getter

Conversations were needed in order to see the relationship between two users. We use conversations to answer the following questions: is a person being bullied or what is the relationship between two given users? Without conversations these questions would have been impossible to be answered.

Twitter conversations are not linear and are usually very short with around 5 or 6 tweets. We started by getting a tweet from a user's time-line via the search API from Twitter's main website (see Figure 9: get conversations). We then took each individual tweet and got the branch of conversation that it was associated with from the Twitter mobile website using a PHP regex data scraper (see Figure 9: Twitter Mobile Site). After we got the new conversation, each tweet in that conversation had to have its tweet ID looked up to see if it was already part of an existing conversation. If any tweet in the conversation was found to be a part of an existing conversation then all of the tweets in the conversation were assigned the old conversation's number and reinserted into the database. Note that each conversation had a unique conversation number in the database. This was important, as we had to ensure that tweets were not counted more than once and thus skewing the data results. It was also important to have conversation numbers as these marked recurring events in which two people interacted (the bullying definition).

Conversation getter was used twice, once when getting the tweets from the keywords getter in Step 1 and then again after getting a tweet from the "person's being watched" Twitter time-line in Step 2.

3.3 Emotional Analysis

In order to understand the nature of two users' relationships emotional analysis was employed. In this thesis we use a lexical dictionary based approach to assign emotions and sentiment to tweets. We used the emotional analyzer developed by Sanmit Desai in his thesis "Smart Sentiment and Emotional Analysis" (Desai, 2015).

Here we make no attempt at improving the findings and overall accuracy of his emotional analysis tool. Instead, we focus on using his work in order to understand the relationships of users on Twitter, in order to find aggressive relationships.

His emotional analysis tool works by taking a sentence and assigning a sentiment and emotional value to each word. Then it totals all the values together for the final sentence value. Each word is located in the dictionary by its POS and its text, and after it is looked up it returns a vector. The vector contains eight emotions and one sentiment score for the word. The sentiment score can have a value between -10 to +10, and is the first value in the vector. The next 8 values in the vector are the emotions: *joy, trust, fear, surprise, sadness, disgust, anger, and anticipation*. Each emotion can have a value of zero to five, where five represents highest intensity of the emotion. Each word's emotional vector is then summed together to get the value of the sentence. For an example of emotion analysis on bullying tweets see Section 1.2. For more details on Desai's work see Section 6.5.

This tool was used once for finding conversations of people who have demonstrated high anger (Step 1), and again in Step 2 for adding emotions to relationships in the Neo4j graph. There was no analysis or filtering of emotions in Step 2 as opposed to Step 1 that used *standard deviation* for filtering.

3.4 Standard Deviation

Standard deviation is a measure of the amount of dispersion in a data set. It was noticed that the emotions and sentiment reported by the emotional analyzer was not

completely correct; however, these incorrect instances were only sporadic. These inconsistencies arrived due to some words not being in the dictionary (see Figure 10, below). For example, a sentence might appear to contain a lot of “love”, but it also contains a lot of “anger”; yet the “anger” words were not found in the dictionary. In the sentence “I would love to kiksumass” the emotional analyzer would find love but not anger which is an incorrect emotion analysis. In order to alleviate those inconsistencies standard deviation was used on each value in the emotional vector. In order for a *one-sided conversation* to pass the standard deviation test it would have to have consistent results for more than 70% of the sentences that were present in the conversation.

After the conversation was translated into emotional vectors we made sure that the conversation had a tweet count over 3 (otherwise, it would be difficult to get a good deviation measure). Every tweet in the conversation can be seen as a row and all

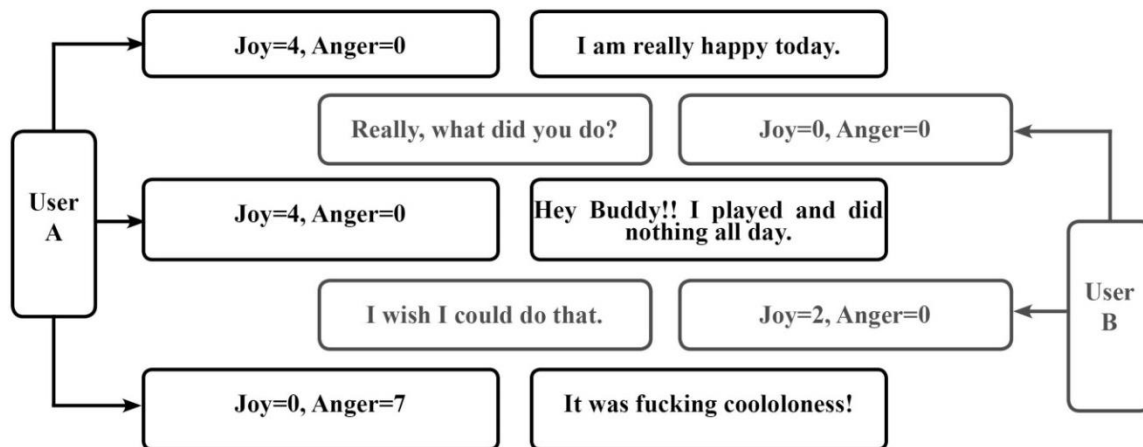


Figure 10: User A, One-sided Conversation that Fails the Standard Deviation Test

tweets together form a matrix. We then used standard deviation on each emotion and calculated the standard deviation across each column. We found by hand that allowing a variance of two times the standard deviation produced conversations that were consistent in emotion and thus a higher chance of accuracy. However, we admit this reduced the recall of the system to get all bullying cases but it was necessary to improve accuracy. That is, any tweet containing an emotion that was outside the deviation limit was thrown out of the conversation (typically called outliers). For example, the tweets that are part of the conversation in Figure 10 produce emotion vectors that are inconsistent. The third tweet “that was so fucking coolness” produces a joy of 0 and an anger of 7. The emotion vectors produced by the *one-sided* conversation that A is having with B are inconsistent with the remaining emotion vectors of the conversation, thus tweet 3 is thrown out. When tweet 3 is thrown out the percent of the conversation remaining is 66% which is below 70%. Because the conversation fell below the 70% limit the conversation is discarded. Thus, the conversation is thrown out in order to improve accuracy.

In Step 1, if a user’s conversation passes the deviation test (and emotional analysis for high anger) then that user (the one sided conversation from person) was added to a list of users that should be watched for bullying. If they already exist on that list then their priority count is increased by one. The priority count is used here because we are rate limited on Twitter data retrieval from the main site. The priority count is a method to focus on getting tweets from users that are likely to be bullies.

3.5 One-sided Conversations

This is the process of taking a Twitter conversation and breaking it into parts. Each part is what one person said to other people (one sided). For example, in Figure 10 user A is having a conversation with user B. User A's one-sided conversation with user B is shown in the figure by the arrows pointing to the bolded dark text. One-sided conversations were important because it allowed us to know what emotions any one person was sending to other people.

In a tweet, it is difficult to know exactly which person is being addressed by the single user sending the tweet. This is because names of users are listed at the beginning of a tweet (e.g., @user1, @user2, @user3, ..., tweets text here) and often there is no designation in the tweet text about which user they are addressing (unless of course the tweet was addressed to only one user). We solve this problem by addressing the tweet toward every person in the tweet. This is to say that there can be false emotions sent to a user that the user was not intended to receive by the sender. However, the definition of bullying states that it is a reoccurring act (or likely to reoccur) of aggression that will happen again perhaps with other users. We reason that a person bullied by another will have to be among those users being addressed (direct bullying) and the other people play other bullying roles. While the spectators will change in a recurrent bullying relationship the person being bullied should be constant. If a friend of a bully was in an instance of bullying as a spectator, it is reasonable to assume that this same friend will be in instances of nice or friendly conversations, where the emotions of love, joy, and trust are high (positive emotions). By having multiple conversations

between people, we can easily judge if the relationship is an aggressive relationship or not. However, in the case where some users consistently play other roles besides the victim and bully and are never in positive emotion conversations with the bullier, then our system would state that the bullier is also bullying a spectator. This is not an issue as we are still identifying the bully. We then can take the aggressive relationship and use our classifier on it to find if it is a bullying relationship or not. The concepts of the bully and victim roles being consistent over time were established in Chapter 1.

3.6 Neo4j

Using the Neo4j graph database we were able to model the relationships between users. We took the tweet's emotional vector from one user to another user, in a one-sided conversation, and inserted it into Neo4j (from 'A' person to 'B' person). For each 'A' person we inserted a local time, a type of "out", and the conversation number

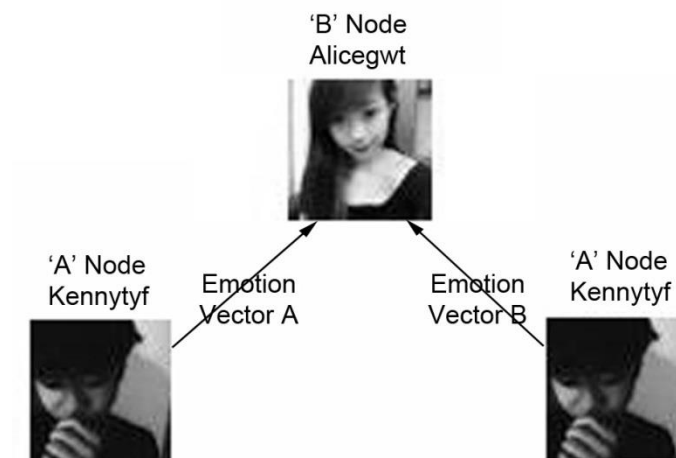


Figure 11: Neo4j Structure Example

that it was a part of. We appended local time to each 'A' node inserted to allow us to know relatively how old a tweet was. We used the type of "out" to allow us to know that

the 'A' node was directed toward another person. Here, the type of "out" means that person A is addressing person B (inversely with type of "in"). We assign the conversation number to each 'A' node to know how many different conversations these two people had. The 'B' person had a type of "in", no local time, and contains within it the summation of all the 'A' emotions/nodes around the 'B' node. For example, in Figure 11 all nodes around the center node are the 'A' nodes where the 'B' node is in the center. Through this process of data collection many interesting queries can be done: seeing a relationship change over time, seeing how one interaction causes a change in emotions of another person, and gauging the amount of bullies that exists around one person. These types of queries will be the concern of the Future Works chapter. As for this thesis, we were only concerned with finding when one person was directly bullying another person.

3.7 The Neo4j Classifier

For classification of bullying relationships, we attempted comparisons between using a simple Neo4j query with a threshold, against classification using the emotional vectors, and against text based methods. The purpose behind these comparisons was to see how well we could find bullying relationships with each of these different methods.

The Neo4j threshold method for finding bullies was done by using a simple cypher query. For example:

```
MATCH (B {type:"out"})-[]->(C {type:"in"})
```

```

WITH C.name as TO, collect(DISTINCT B.name) as FROM,
sum(toInt(B.sentiment)) as SENT,sum(toInt(B.joy)) as
SJ,sum(toInt(B.trust)) as ST,sum(toInt(B.fear)) as SF,sum(toInt(B.surprise))
as SP,sum(toInt(B.sadness)) as SS,sum(toInt(B.disgust)) as
SD,sum(toInt(B.anger)) as SA,sum(toInt(B.anticipation)) as SAA,
count(DISTINCT B.conversationNumber) as cCount

WHERE SF>(SJ+10) and SA>(SJ+10) and SD>(SJ+10) and cCount > 2

RETURN cCount, FROM,TO,SENT,SJ,ST,SF,SP,SS,SD,SA,SAA

ORDER BY TO

```

The above is an example query that could be used on the Neo4j database in order to find “reoccurring acts of bullying” for a single pair of users. By “reoccurring” we refer to the bullying definition of “reoccurring acts of aggression”. In the above “match” portion of the query says we are only looking for single relationships ‘B’ and ‘C’. We collect all nodes around the ‘C’ node into a single node and aggregate those emotional vectors. Recall from the discussion about Neo4j limitations, in the background section, that any two nodes can have only one relationship. Thus, we must aggregate all the ‘A’ user’s values that are around the user ‘B’. The idea of the difference in emotions was previously explained, but in short, by finding user relationships where we sum the emotions of all their sentences, then we take the difference between the summed positive emotions and the summed negative emotions. If that difference is above a certain threshold, this can be a method for classification. For instance, the ‘where’ clause of the above query is explained as, if anger is greater than joy and trust by a

difference of 10 amplitude ($SA > (SJ + 10)$), and if sadness and disgust is greater than joy and trust by a difference of 10 ($SD > (SJ + 10)$, $SF > (SJ + 10)$). When these are true then this instance is a possible case of bullying; here this was just an example, in the results

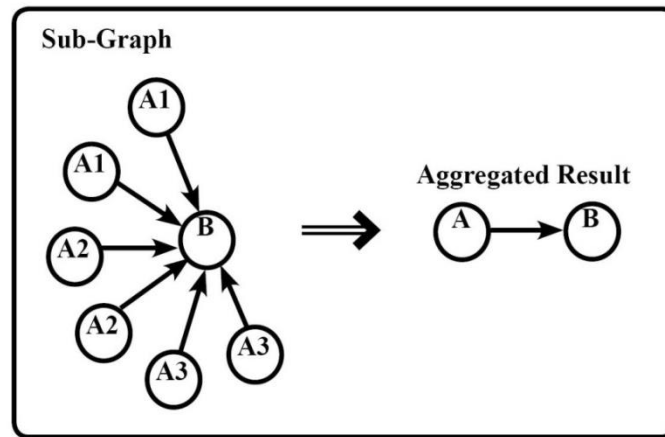


Figure 12: Queried Sub-graph and the Result

chapter we will show the values that we used in the above query to obtain a classification accuracy of 65% and how we found those values.

Figure 12 illustrates the previous query. Here user A has had 3 conversations with user B. The above query would aggregate the A users together (shown on the left) to form a single A user (shown on the right). The query would then test to see if the aggregated A user has a difference between positive and negative emotions by a factor of 10. This relationship is then classified as being positive for bullying if the difference is true. Again, this method has shown an accuracy of 65%.

3.8 Examples

In this section we will give examples of how the data was processed in order to illustrate the process of finding bullying relationships.

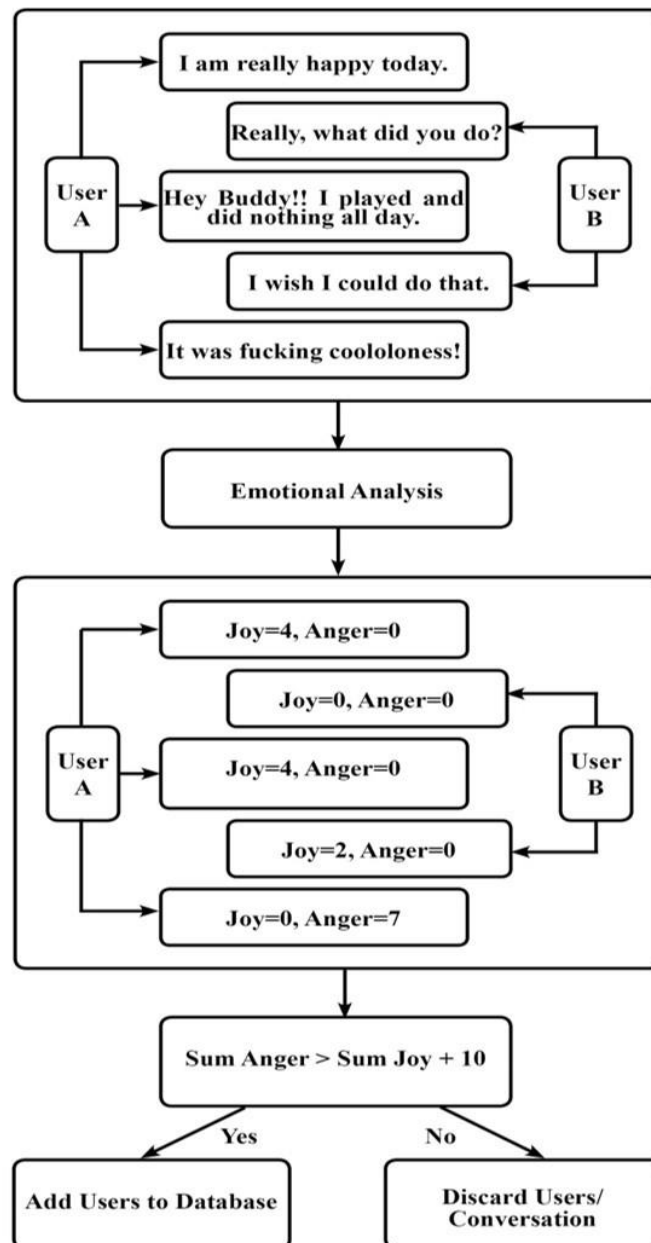


Figure 13: Step 1 in Processing the Data

Figure 13 shows getting a conversation. Then it shows taking that conversation and performing emotional analysis on it. It also depicts testing the result of the emotional analysis to see if the conversation contained a high difference in emotions

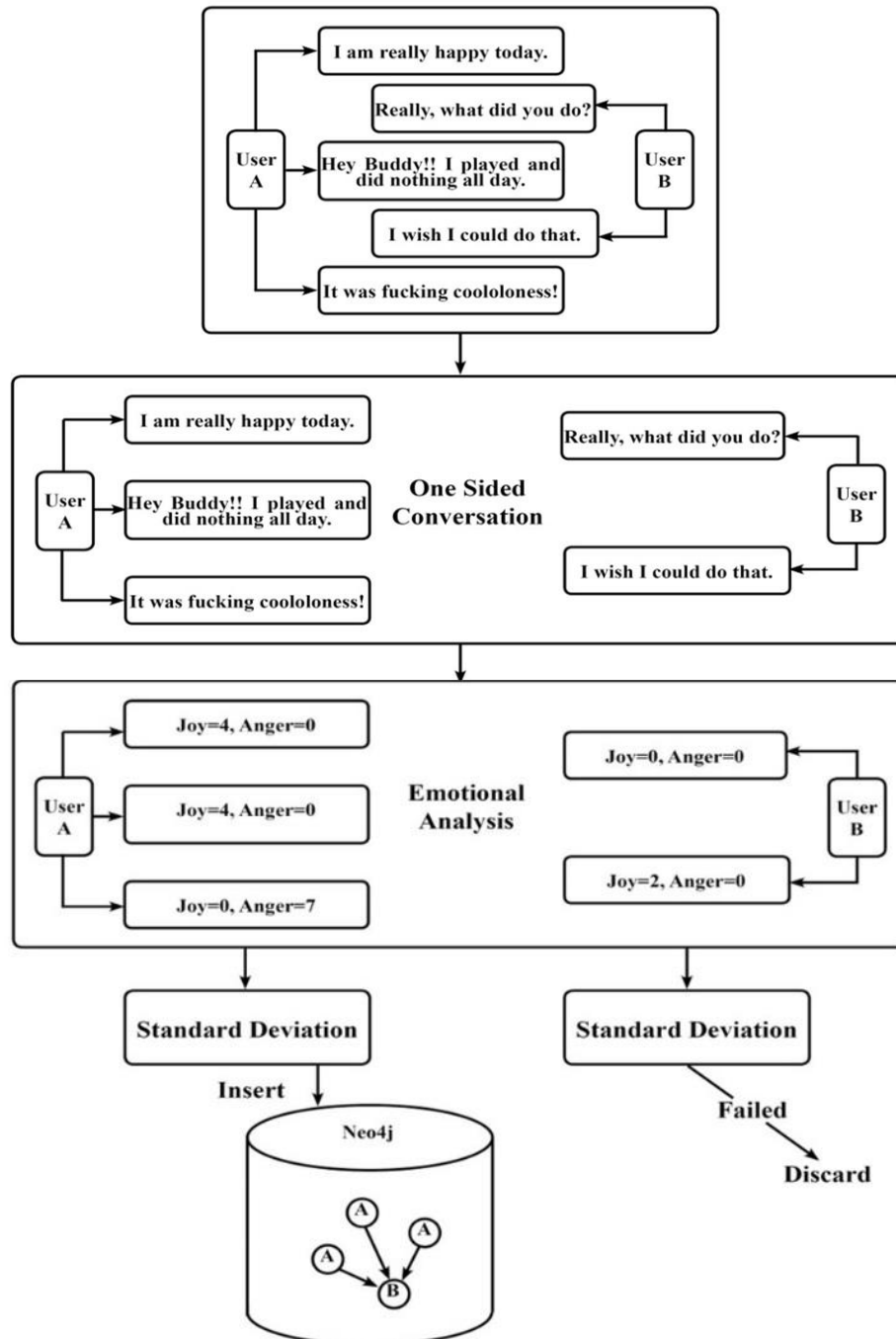


Figure 14: Step 2 in Processing the Data.

(between joy and anger). If it was found to be a true case, the users of the conversation were added as users to watch database for possible bullying. Here we are looking for possible bullies, so a true case means that this user might possibly be a bully.

Figure 14 illustrates Step 2 of the architecture. It shows the data as we get a conversation from a “user to watch” in the database, break the conversation into a one sided conversation, perform emotional analysis, and then perform the standard deviation test. If it passed the standard deviation test, we show that it is added to the database.

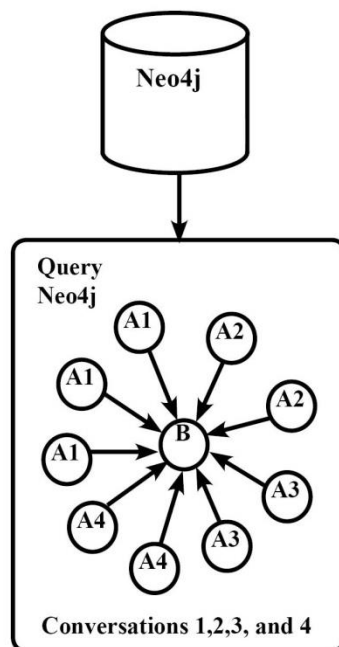


Figure 15: Querying the Database for Features

Figure 15 shows what a result from Neo4j would look like if we get one relationship. In this case user A has had four conversations with user B and each conversation has two or three tweets. We can then sum all the emotion vectors from

user A (all the A nodes) to user B and get sumA. Then we can use the separation of positive and negative emotions in sumA to classify the relationship as bullying with 65% accuracy. We could also use KNN on sumA (K-nearest neighbor) to classify the relationship as bullying with 75% accuracy.

CHAPTER 4

STRUCTURE OF THE TRAINING DATA

4.1 Overview

In the following sections, we discuss the relationships that we collected, and then how we obtained training data from the relationships that we used to build the classifiers in the results chapter.

After using keywords, standard deviation, and emotional filtering in Step 1, we had 316 users that were possible bullies in our MySQL database. We collected tweets from those users' time lines for a little under two months totaling to get 548,780 tweets. These tweets are not only from the 316 users they are also from re-tweets posted on those user's timelines, conversations from re-tweets posted on those users timeline's, and from users that these 316 people followed (see Section 2.3) who's tweets were posted to their timelines. In the following results we used all the tweets, since people commenting on tweets, friends of a possible bully, or other people who are in the same conversations as the bully also have a chance of being a bully, which makes them no less valid then the person's time-line that we were scanning. This follows the definition of bullying that states bullying can be a "group" activity, where a bully might be part of a conversation that has many other people bullying a user or have friends that often bully.

Before standard deviation, there were 548,780 tweets in the MySQL database. After the standard deviation filtering in stage two there were 144,124 tweets remaining.

In total, there were 42,385 distinct users in the Neo4j database. The below table shows how many relationships have one-sided conversations and how many tweets each conversation contained (in the Neo4j Database). Our goal here was to obtain the greatest amount of training data possible while having a high number of reoccurring conversations in the relationship. We needed many conversations and tweets per conversation in order to match the bullying definition. We created the table below in order to help choose the relationships for classification. For example, in the table below, there are 1,236 relationships in our Neo4j database that have at least three conversations and one tweet. We chose relationships that have had at least four conversations and two tweets for our training data (270 relationships). We chose these because they would better fit the bullying definition with a high conversation count.

Table 1: Number of One-sided Relationships having Y Tweets and X Conversations

	1 C	2 C	3 C	4 C	5 C
1 T	69939	4386	1236	538	121
2 T	38220	1800	434	270	84
3 T	17601	559	118	51	25
4 T	6258	137	32	10	4
5 T	1613	27	12	1	1

In the Table 1, notice the decrease in relationships from top to bottom and left to right. If we increase the amount of tweets required per conversation in a given relationship then there will be less relationships that meet those requirements. The same is also true for conversation count per relationship. Notice that most

conversations only have one tweet and one conversation present, which is something to be expected when someone is leaving a comment on a retweet.

For the training data we took the 270 relationships that had passed the standard deviation test, at a minimum of two tweets and four conversations per relationship ($2 \times 4 \times 270 = 2,160$ minimum tweets), and ran them through Amazon Mechanical Turk (AMT) and by independent raters (not belonging to the project's construction, four students from UGA and one from Georgia Perimeter College). See the Appendix D for an example of what the form looked like that the AMT workers filled out. In brief, AMT is a crowdsourcing platform that allows many people to participate in performing tasks. Here our task was to classify the relationships that we obtained from Twitter as bullying or not. So, there were many people on AMT that completed our two tasks. In both ratings we allowed for three types of classification: *positive*, *neutral*, and *negative*. A relationship was *positive* if it followed the definition for bullying (see Chapter 1 for the definition of a relationship), *neutral* if the one-sided conversation didn't make enough sense by itself (not enough information for classification), and *negative* if it did not follow the definition of bullying. We felt that the independent raters being graduates students made them more reliable for classification, so we gave more weight to their classification. Such that, if both AMT ratings (final classification values) disagreed with the independent rater's final classification then we agreed with the AMT rating, otherwise we agreed with the rating given by the independent raters. Where two out of three ratings agreed that the relationship was incomprehensible we discarded the relationships from the results (see Table 3 for an example). After we removed the

neutral relationships there were 198 remaining ($270 - 198 = 72$ instances removed). The ratings given by AMT and the independent raters are show in Table 2. After adjusting the results with the independent rater’s weights there were 88 positive cases and 110 negative cases in the final classification.

Table 2: Rating Results

	MT1 rating 1	MT2 rating 2	IRs
Positive	105	41	90
Neutral	41	51	45
Negative	124	178	135

We then used these relationship classifications to prepare our training sets that we used for each of the classifiers in the Results Chapter. For the classifier using text classification we took all the tweets composing a one-sided relationship and concatenated those together in a single text block that looked like a long paragraph (see Table 5 for a single instance example). At the end, this training set had the relationship (user, atUser), the text for the relationship, and the final classification as the attributes. For the training set using the emotional vectors, we took all the tweets that made up the relationship, applied emotional analysis to them, summed up all the emotions, and count the amount of conversations in the relationship. We then divided the relationships summed emotional values by their conversation count. We did this to allow relationships that had more than four conversations to have the same relative values as relationships with four conversations (the minimum for the training set) (see Table 4, for examples). The end result was that this training set had *sentiment*, *joy*, *trust*, *fear*, *surprise*,

sadness, disgust, anger, anticipation, user, atUser, and the final classification as the attributes.

Table 3: Example of Classified Relationships

user	atUser	AMT 1	AMT 2	IR	Final
_xBauty18	X_intimidator	Positive	Negative	Negative	Negative
xXxJ0DIExx	AdamCimmering	Negative	Negative	Negative	Negative
xXxJ0DIExx	ScarSnake	Negative	Negative	Negative	Negative
XXNoFleXX	Nicsdick	Positive	Negative	Positive	Positive

Table 4: Example of the Emotional Training Set

Sentiment	Joy	Trust	Fear	Surprise	Sadness	Disgust	Anger	Anticipation	User	atUser	Final
15	23	19	1	4	8	1	5	8	Walthizenberg	Nesta_carter	Negative
10	14	6	1	4	7	5	1	4	Sofiesassysofie	Leighaconnor	Negative
8	5	5	1	0	1	0	2	0	Junior_quantrel	Sureal187james	Positive

Table 5: Example of Text Training Set

Text	User	atUser	Final
@Grandes5sos @kxngmaraj @saturdaylrh @AZEALIABANKS youre life is a failure, just like the peopel you stan. and your parents are related. @Grandes5sos @kxngmaraj @saturdaylrh @AZEALIABANKS bitch your parents are fucking related shut the fuck up lmao. @TheMogulPrint @saturdaylrh @Grandes5sos @kxngmaraj @AZEALIABANKS omggg you found her!!! good job. @kxngmaraj @saturdaylrh @Grandes5sos @AZEALIABANKS LMAO BITCH READ BEFORE YOU START TALKING SHIT you ugly ass cunt. @Grandes5sos @saturdaylrh @darkskinwhyte @kxngmaraj @AZEALIABANKS you guys are so fucking racist its sad. @Grandes5sos @saturdaylrh @darkskinwhyte @kxngmaraj @AZEALIABANKS thats all u have to say... ugly emo... Youre so lame lmao.	SADBOYS_	AZEALIABANKS	Positive

CHAPTER 5

RESULTS

5.1 Overview

In the following sections we get to the most interesting part and what we have been building up to. We will first discuss the threshold classification method using Neo4j to classify the bullying instances (1), then classification of bullying instances with the text based training set using machine learning (2), and then classification of bullying instances with the emotion training set using machine learning (3). The purpose of this comparison is to see how well the difference between positive and negative emotions in a relationship can classify the act of bullying (done by using Neo4j), and compare it against other methods while also using the text based training set and the emotional vector training set to see the advantage of using emotional analysis. This will give a good comparison of the advantages of emotional analysis for classification purposes opposed to just simply comparing the results to another's work in the same area, because it eliminates a lot of the variables such as the domain the training set was obtained from (Twitter, Wikipedia, Formspring), the quality of that training set (how it was classified), or their adherence to the definition of bullying (arguments that what was classified was actually bullying or not).

5.2 Classification Using Neo4j Queries

In this section, we discuss the process in which we obtained the best values for the difference between positive and negative emotions to classify bullying relationships using Neo4j cypher queries.

“match (A:person{type:'out'})-[R]->(B:person{type:'in'})”

Using the above query on Neo4j to return all relationships in the training set, we summed all the values over those results matching the relationship from Neo4j to the relationship in the training set, i.e, summing all the emotional values for the sentences in that relationship. We then created a program that did a brute force search on the training set to find the differences that produced the best classification. For example in Figure 16, we would increase Ao by one and then take a relationship's *joy* value add it to Ao (call it sumJ). We would then see how many relationships had an *anger* value

```
for(relationship rel: table){
    if(rel.emotionalVector != null && !rel.emotionalVector.isEmpty()){
        int rjoy = rel.emotionalVector.get(joy);
        int rtrust = rel.emotionalVector.get(trust);
        int rfear = rel.emotionalVector.get(fear);
        int ranger = rel.emotionalVector.get(anger);
        int rdisgust = rel.emotionalVector.get(disgust);
        int rsadness = rel.emotionalVector.get(sadness);

        if((rjoy + Ao < ranger)
            && (rtrust + Bo < ranger)
            && (rjoy + Co < rsadness)
            && (rtrust + Do < rsadness)
            && (rjoy + Eo < rfear)
            && (rtrust + Fo < rfear)
            && (rjoy + Go < rdisgust)
            && (rtrust + Ho < rdisgust)){
            rel.classifierRes = "positive";
        }
    }
}
```

Figure 16: Code Snippet Illustrating Classification

greater than sumJ. The differences used to classify bullying were (*anger* – *joy*, Ao), (*anger* – *trust*, Bo), (*sadness* – *joy*, Co), (*sadness* – *trust*, Do), (*fear* – *joy*, Eo), (*fear* – *trust*, Fo), (*disgust* – *joy*, Go), and (*disgust* – *trust*, Ho). Here the idea was to use the difference between positive and negative emotions or emotions that were common bullying in order to see how well we could classify the data with a simple Neo4j query.

In the code snippet in Figure 16, we iterate over the training set until the best possible values were obtained. We then iterate over all possibilities of variables 'Ao' through 'Ho' for values of 0 – 10. During every iteration of 'Ao' through 'Ho', we looked to see how much of the training set these values classified correctly (by iterating through them with the emotional values and calculating the accuracy). At the end, we kept the result that had the highest accuracy. Through this we achieved a best accuracy hypothesis of 65.6% with 100% recall using values of Ao = 9, Bo = 9, Co = 3, Do = 3, Eo = 2, Fo = 2, Go = 5, and Ho = 3 that match Figure 16. From the values, the most dominant emotion found was *anger*, followed by *disgust*, then *sadness*, and then *fear*.

As an example of the previous process, given a relationship in the training set (relA) who's emotion vectors sum together to give a difference of 4 between *joy* and *anger* (Ao) and a final classification of *positive* for bullying. The code snippet in Figure 16 would initialize Ao to 0 and then increase the value of Ao by one every iteration. It would then test to see if that value of Ao could still properly classify relA by adding Ao to relA's *joy* and seeing if the result was less than relA's *anger*. By doing this we can find the difference between positive and negative emotions that best separates the training set. If relA could still be properly classified (ie, the new values still classify it as positive

which was what it originally was) then this increased the accuracy, otherwise the accuracy would decrease (for this iteration of the program). In short, the code does a brute force search on the hypothesis space to find the best separation of positive and negative emotions that classified bullying on the training set.

5.3 Classification: Text-based and Emotional Vectors

For the following classifiers we used the Weka software package. We will refer to the *trainText* as the text based training set and *trainEmot* as the emotional training set. For the *trainText* we tried a few different classifiers: KNN (K-nearest neighbor), Naive Bayes, and SVM (Support Vector Machines). Naïve Bayes is well known for its ability to classify text. There are no configurations present for this in Weka, since it's a very straight forward classifier. It works by taking a training set classified for positive and negative of the concept to be found. Then we take all the text in the positive examples and find how frequent each word appears in that classification. This process is repeated for all words. This is also done for the negative training examples. It is common to remove all words not appearing more than a threshold amount (we tried this at various values with no better results). It is also common to normalize the remaining data, so that all data (for its respective classification) appears a percentage out of 100. When a new instance is to be classified, each word in that new instance is associated with an attribute, and for the attributes it contains their percentages are summed up. The new instance is classified according to the class with the highest percentage. Support Vector Machines find a plane in multidimensional space that can separate the data perfectly between the classes. When a new instance is to be classified it just simply looks to see

which side of the line it is on (for binary classification). K-nearest neighbor works by just plotting points in multi-dimensional space. When classifying a new instance it finds the K nearest neighbors and takes a vote, which ever class gets the most votes classifies the new instance. All three of the classifiers were chosen for their ability to perform well when there are many attributes present.

We found that KNN provided the best classification by a wide margin. We normalized vectors to make the values between zero and one, the stop words were removed, and stemming was performed. The data set would be too large to use a classifier that finds the global optimum hypothesis. With this in mind we tried a few different adjustments with KNN to get the best values. Finally, we arrived at 78% accuracy and 75% recall. This was obtained using $K = 3$ and using the Manhattan Distance between the points. Also note that these accuracies were obtained using 10-fold cross validation. We used 10-fold cross validation since our training set was small and we did not have a testing set.

For trainEmot we again tried Naive Bays, KNN, and SVM. We found that SVM produced the best results. We did some preprocessing on the training data before doing these classifications, by dividing all the emotional vector values for a relationship by the amount of conversations that the relationship had. We needed to put all the relationships on the same scale, so it didn't matter if a person had a few conversations or many conversations. This was because the training set contained users that had a minimum of four conversations and two tweets. There could have been many users with many more conversations and tweets per conversation then the minimum. We also

removed some attributes that were not used for the classification including sentiment, user, and atUser. With the SVM we used a $\gamma = 0.1$ and a RBF kernel. We obtained best accuracy of 75.25% and a 70% recall using 10-fold cross validation.

The above results show that actually using emotional vectors for classification reduces the accuracy of the classification (78% vs 75.25%). This is believed to be because the dictionaries for the emotion analyzer was human annotated and thus inherently contain a small amount of errors. However, the result also shows that the emotional analysis is consistent. If the emotional analysis was not consistent a much lower accuracy would have been obtained as the emotion vectors would have not been able to classify the bullying cases. Thus, anger (and other emotions) would not have been consistent for different examples of bullying (we see that this is not the case 75% of the time). For example the sentence “I hate you” should contain anger, and the sentence “I love you” should contain love, and all the sentences that could form hate or love actually produce hate or love when changed into an emotion vector. This means that the emotional analysis can be used accurately in other manners such as identifying the emotions present in each bullying role (see Future Work Chapter). In short, the other roles need to display certain emotions in order to be classified as that role, and this shows that emotional analysis can reliably be used to show if those emotions are present or not (75% of the time) (Xu, Zhu & Bellmore, 2012).

In the above results, we have seen that standard methods have a far greater accuracy than the Neo4j query threshold classification. We also have seen that text classification methods work better than classification with the emotional vectors, but not

by much. It was also shown that the emotional vectors could reliably play a supporting role to identify roles in bullying, while the final classification should be done using the text of the relationship.

CHAPTER 6

RELATED WORK

6.1 Social Media Efforts

Emotional analysis has become the trend for analyzing cases of bullying. Facebook has teamed with Yale: Center for Emotional Intelligence in order to take an active approach to stopping bullying (Balkam, 2015). Twitter seems to be relying on crowdsourcing in order to recognize instances of bullying and deal with them via a report abuse button (Twitter, 2014). However, in both cases the popular social media platforms are relying on human based emotion analysis in order to find instances of bullying.

6.2 Non-reoccurring Approach with Decision Trees

In the paper, “Using Machine Learning to Detect Cyber-bullying” by Kelly Reynolds (Reynolds, 2012), they parsed testing data from a website (Formspring.me). The website allows users to anonymously answer questions posted by other users. For example, “what do you think of me?” question is inherently high in bullying due to anonymity of responders. Their data was in the form of a question by a user and a response from a user, and they collected 2,696 posts for training and 1,219 posts for testing. A single question and answer was considered to be a single post. The training and testing data was sent to Amazon Mechanical Turk three times for classification, where two out of three classifications decided the final classification of the instance.

They then made a list of features by hand around swear words since they noticed that these were of high frequency in bullying posts. Using these features they used Weka to test the classification ability of different algorithms: J48, JRIP, IBK1, IBK3 and SMO. They achieved the highest accuracy with J48 with a classification of 67% and a true positive rate of 81% (recall).

The definition of bullying refers to “reoccurring acts”, but in their work they make no attempts to find if either one user was being bullied by many different users or to find users that are recursively bullied by a single user. Instead, they just focused on if the single post was a bullying or not. This is to say they don't focus on the relationship of the two individuals. They also limit themselves to a subset of bullying by considering only attributes that contain swear words.

6.3 Attempt Using Sentiment Wordlists and Amazon Mechanical Turk

In the paper “Twitter Bullying Detection” (Sanchez & Kumar, 2012), they used sentiment analysis and Naive Bayes to classify tweets found between potential bullies and a person that is potentially being bullied. The overall idea was to use machine learning on sentiment vectors. They used Amazon Mechanical Turk to add sentiment to a word-list (later used as attributes), which was then used on a testing set of tweets that were from the potential bullies that mentioned the user. The testing set was classified by hand and contained 460 tweets. Their results had an accuracy of 67% using 10 fold cross validation.

They made an attempt to classify relationships between users, but did not show

the “reoccurring” portion of the bullying definition. They also attempted to use sentiment analysis, but their results did not give any improvement over other classification methods.

6.4 Using Sentiment and Emotion Analysis

The following is an interesting article that tries to capture the entire definition of bullying using sentiment analysis and emotion analysis but does so in a different method than this thesis does. In “Fast Learning for Sentiment Analysis on Bullying” (Xu, Zhu & Bellmore, 2012), they started by finding emotions that were common in bullying tweets: *anger*, *embarrassment*, *empathy*, *fear*, *pride*, *relief*, and *sadness*. They found their machine learning features by first looking up synonyms for each emotion from WordNet and an on-line dictionary of synonyms (Miller, 1995). They obtained tweets that contained these words from the Twitter API. They broke the tweets into unigrams and bigrams and counted the occurrences. Finally, they combined the resulting words from the tweets and synonym lists to obtain the features for classification. When finished, they had seven different vectors (words) of attributes, one for each emotion. Using Wikipedia pages that were automatically categorized to each emotion, they did feature/attribute extraction on each page for training data. They compared kernels and found that a seven class SVM was better using the RBF kernel (radial based function) producing an accuracy of 85% (from Wikipedia pages) (10-fold cross validation). The seven class SVM worked by a one vs many approach. This means that each instance was classified using a binary classification seven times (one for each emotion). Every binary classification returned a real number for how well it classified the data. The

binary classification returning the smallest number was what classified the instance. If the values returned were all greater than a threshold then the instance would be classified as “other”. They then applied the process to 1,500,000 tweets finding that 6% of the tweets were parts of bullying traces (single conversations with many tweets). They did not specify, in the paper but it is assumed that they identified these bullying traces based on the emotions shown by the participants. In a prior work, they identify the emotions that are present in every role in bullying. They found that 50% of tweets in the bullying traces were classified to contain fear, then anger of 18%, and then sadness of 11%.

The most obvious critique is that they used Wikipedia pages as their training data. Twitter data often has missing words, misspelled words, and abbreviated words, which are not in structured pages like Wikipedia. Second, they make an attempt to use emotions on each tweet found in a trace, however this is a rough classification and truncation of all other emotions in the tweet, and the other emotions could be used to give that user a different role in the trace if there was a magnitude assigned to each emotion and the emotions were then summed over the user’s tweets in the trace.

6.5 Using Emotion Analysis on Tweets

The final related work we identified, “Smart Sentiment and Emotional Analysis” by Sanmit Tatoba, is the methodology that this thesis is based on (Desai, 2015). In his thesis he used emotional dictionaries/lexicons to assign sentiment and emotion to each word in a sentence. The dictionaries used part of speech in order to better assign sentiment to words. He used SentiWordNet 3.0 for the sentiment portion of a given word

(Baccianella, Esuli, & Sebastiani, 2010). He then used another dictionary that assigned emotional amplitude (0-5) to each POS distinguished word: NRC dictionary (Mohammad & Turney, 2013). He also created his own anger-oriented dictionary to rate the Twitter language for emotion and sentiment where the previous dictionaries were lacking: profanity, emoticons (☺, :/, :>, ☹), etc. In Sanmit's modified dictionary he used another dictionary to assign the sentiment portion of the words, where Sanmit assigned the emotion values (Davies & Ghahramani, 2011). In his dictionary he used both uni-grams and bi-grams. The emotions chosen were based on Plutchik's work: *joy, trust, fear, surprise, sadness, disgust, anger, and anticipation* (Plutchik, 2001). Where a zero assignment means no emotion of that type and a +5 assignment is the most intense of that emotion. When a sentence is analyzed for sentiment and emotional value, it is broken down into parts of speech, then it is looked up in the hand built dictionary. If not found there it is looked up in the NRC and SentiWordNet. If the word is still not found then the word is assigned a zero value to both the sentiment and emotional vector. When combining the words for a final emotional outcome of a tweet, he simply added the vector of each word together (see Section 1.2 for an example). He also used valence shifters, intensifier, stop word removal, and stemming lists that might modify the meaning of a sentence and thus the output. An example of a valence shifter is "not good"; if "not" was encountered it would invert the sentiment and emotional measures of the next word. An intensifier is similar: if the word "very" was encountered following a word its emotion and sentiment values were increased by one where values are not zero. A stemmer reduces words to their root word, for example "cats" or "catty" would be reduced to the word "cat". This is useful when the words "cats" or "catty" are not in the

dictionary. The stemmer that he used was the Porter Stemmer (Porter, 1980).

He evaluated his results using three human testers that were each given a tweet evaluated by the system. They were asked to give the ranking of the emotions present and a rating of sentiment. The most prominent emotion was identified correctly by the system 63% of the time and the correct sentiment analysis was found 66% of the time.

Using this approach above, this thesis attempts to classify recursive acts of bullying by finding relationships that consist of high amounts of anger, sadness, and disgust over other emotions such as joy and trust.

CHAPTER 7

CONCLUSION

In conclusion, using emotional analysis was effective at finding relationships that contained aggressive behavior and thus could be used to identify bullying relationships. However, not having the complete Twitter language prevented us from gathering all the instances of bullying. Even after we filtered out all the unclassifiable cases using standard deviation, we were only able to obtain 75% accuracy, marginally greater than the accuracy of the underlying emotional dictionary/lexicon of sentiment and emotions. From the difference in accuracy between text-based and emotional based classification, there must be some error in representing the text as emotions. This error is likely due to the dictionaries human element. The dictionaries were created by AMT workers and researchers perspective on the emotions that belonged to each word. However, the difference is not significant enough that this method of emotional analysis cannot play a secondary and reliable role for classifying emotions in sentences, but a machine based dictionary expansion approach should be used over the human element.

Taking into account others' works that have accomplished similar accuracies in bullying classification, it can be concluded that the language of Twitter is the issue at hand. Where the variance in the amount of words that has the same meaning is the real issue and therefore is difficult to classify. This could explain how the related work by

Jun-Ming Xu, Xiaojin Zhu, and Amy Bellmore (Xu, Zhu & Bellmore, 2012) was able to obtain an 85% accuracy using structured Wikipedia pages over our Twitter training set.

Also, in the related work of Ming Xu, Xiaojin Zhu, and Amy Bellmore they found that fear was the most prevalent emotion while we found it was anger. This could have been due to their usage of the Wikipedia pages for a training set, but this could have been also due to our usage of dictionaries that were not extensive enough in emotions, besides anger, for the Twitter domain. This lack of a comprehensive dictionary could have led to its lack of ability to give a good classification. This also leads to the need for a machine based dictionary learning approach.

CHAPTER 8

FUTURE WORK

Below we discuss a few ideas that we could implement with the current tools developed but just simply did not have time to implement, such as classifying the other roles in the bullying relationship. We also discuss an idea that would improve quality for quantifying emotional vectors that would improve the performance of this work.

We produced a database of relations that could be used to further classify other types of bullying relationships based on the emotions that were displayed by each participant during one conversation or multiple conversations: friend of the bully, observer of the bullying instance, the victim's defender, someone who reports the bullying instance, and someone who accuses the bully of being a bully. All these roles are possible to identify with the correct cypher query. These roles could be even identified to a greater degree than was done in the similar related work, because we have the ability to see a relationship over time, and thus we know who is actually a friend or a victim of the bully. In addition, the graph database could be used to study bullying relationships and how they progress over time. A relationship could become positive and then negative again, and additional information about bullying relationships can be gathered through these observations.

Another concept that would improve this work would be a better way to understand the language of Twitter, such as an ontology of all Twitter words, to replace

the current dictionaries (that was used to change the sentences into emotional vectors) in this thesis. One such idea is to use a feedback mechanism built from seed sentences and a single word variable. The feedback is how a person reacts (with a return tweet) to a tweet. The seed sentence is a sentence that has been rated by hand for known emotional value. The feedback tweet would also be of known value. We could then look at tweets with one word variance from our original sentence and continue gathering known tweet responses to see how the single word made their emotional response change. This could potentially offer a more reliable and consistent emotional vector of Plutchik's emotions, and with better base results we could perform a better classification of bullying.

REFERENCES

- Baccianella, S., Esuli, A., & Sebastiani, F. (2010, January 01). SentiWordNet. Retrieved from <http://sentiwordnet.isti.cnr.it/>
- Balkam, S. (2015, January 01). Put a Stop To Bullying. Retrieved from <https://www.facebook.com/safety/bullying/>
- Berant, J. (2014, October 01). The Stanford NLP Group. Retrieved from <http://nlp.stanford.edu/>
- Camodeca, M., & Goossens, F. (2005, February). Aggression, Social Cognitions, Anger and Sadness in Bullies and Victims. *Journal of Child Psychology, Psychiatry*, 46(2), 186-97 Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/15679527>
- CDC (2013). "Understanding Bullying: Fact Sheet". Retrieved from http://www.cdc.gov/violenceprevention/pdf/bullying_factsheet.pdf
- Copeland, W., Wolke, D., Lereya, S., Shanahan, L., Worthman, C., & Costello, J. (2014, May 27). Childhood Bullying Involvement Predicts Low-grade Systemic Inflammation into Adulthood. *Proceedings of National Academy of Sciences*, vol. 111, no. 21.
Retrieved from <http://www.pnas.org/content/111/21/7570.full.pdf+html>
- Davies, A., & Ghahramani, Z. (2011, January 01). *Language-independent Bayesian Sentiment Mining of Twitter*. University of Cambridge.

- Desai, S. T. (2015). Smart Sentiment and Emotion Analysis. Master's Thesis, University of Georgia.
- Frieden, T. R., Degutis, L. C., & Spivak, H. (2011). Measuring Bullying Victimization, Perpetration, and Bystander Experiences. Retrieved from <http://www.cdc.gov/violenceprevention/pdf/bullycompendium-a.pdf>
- Frieden, T. R. Sosin, D. M., Spivak, H. R., Delisle, D. S., & Esquith, D. G. (2014). Bullying Surveillance Among Youths. pp. 4-7. Retrieved from <http://www.cdc.gov/violenceprevention/pdf/bullying-definitions-final-a.pdf>
- Goldblum, P., Espelage, D., Chu, J., & Bongar, B. (2014, October 29). Youth Suicide and Bullying: Challenges and Strategies for Prevention and Intervention. pp. 70-71.
- Goldweber, A., Waasdorp, T., & Johnson, S. (2013). Bullies, Gangs, Drugs, and School: Understanding the Overlap and the Role of Ethnicity and Urbanicity. *Journal of Youth & Adolescence*, 42(2), pp. 220-234.
- Hamburger, M. E., Basile, K. C., & Vivolo, A. M. (Eds.). (2011). Measuring Bullying Victimization, Perpetration, and Bystander Experiences: A Compendium of Assessment Tools. Centers for Disease Control and Prevention, National Center for Injury Prevention and Control, Division of Violence Prevention.
- Hatzenbuehler, M. L., & Keyes, K. M. (2013). Inclusive Anti-bullying Policies and Reduced Risk of Suicide Attempts in Lesbian and Gay Youth. *Journal of Adolescent Health*, 53(1), pp. S21-S26. Retrieved from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3696185/>

- Hertz, M., Donato, I., & Wright, J. (2013, July 01). Bullying and Suicide: A Public Health Approach.
- Hulsey, C. (2008). Examining the Psychometric Properties of Self-report Measures of Bullying: Reliability of the Peer Relations Questionnaire. Master's Thesis, Wichita State University. Retrieved from <http://soar.wichita.edu/xmlui/bitstream/handle/10057/2050/t08022.pdf>
- Juvonen, J., Wang, Y., & Espinoza, G. (2010). Bullying Experiences and Compromised Academic Performance Across Middle School Grades. *Journal of Early Adolescence*.
- Kim, Y. S., & Leventhal, B. (2008). Bullying and Suicide. A Review. *International Journal of Adolescent Medicine and Health*, 20(2), pp. 133-154. Retrieved from http://equalitytexas.org/app_themes/images/site/10/pages/10/Bullying-Suicide.pdf
- Merriam-Webster (2015) Bullying definition. Retrieved from <http://www.merriam-webster.com/dictionary/bully>
- McDougall, P., Vaillancourt, T., & Hymel, S. (2009). What Happens Over Time to Those Who Bully and Those Who are Victimized. *Bullying at School and Online*. Retrieved from http://www.education.com/reference/article/Ref_What_Happens_Over/
- Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11), pp. 39-41
- Mitchell, T. M. (1997). Bayesian Learning. In *Machine Learning*, pp. 177-179. New York: McGraw-Hill

Mohammad, S. M., & Turnkey, P. D. (2013). NRC Emotional Lexicon Publications. Computational Intelligence, 29 (3), pp. 436-465

Retrieved from
<http://www.saifmohammad.com/WebPages/ResearchInterests.html>

NCES: "Indicator 7: Discipline Problems Reported by Public Schools". (2014, June).

Retrieved from
https://nces.ed.gov/programs/crimeindicators/crimeindicators2013/ind_07.asp

Neo4j (2015, April 03). "The Neo4j Manual v2.2.0-RC01". Retrieved from
<http://neo4j.com/docs/milestone/>

Olweus, D. (1994). Bullying at School: What We Know and What We Can Do. British Journal of Educational Studies, 42(4), pp. 403-406

O'Moore, M., Seigne, E., McGuire, L., & Smith, M. (1998). Victims of workplace bullying in Ireland. The Irish Journal of Psychology, 19, pp. 345-357

Plutchik, R. (2001). Integration, Differentiation, and Derivatives of Emotion. Evolution and Cognition, 7(2), pp. 114-126

Porter, M. F. (1980, January 01). The Porter Stemming Algorithm. Retrieved from
<http://tartarus.org/martin/PorterStemmer/index-old.html>

Reynolds, K. (2012, January). Using Machine Learning to Detect Cyberbullying. Submitted to the faculty of Ursinus College in fulfillment of the requirements for Distinguished Honors in Computer Science

Retrieved from
<http://webpages.ursinus.edu/akontostathis/ReynoldsHonors.pdf>

- Sanchez, H., & Kumar, S. (2012). Twitter Bullying Detection, ser. NSDI'12. Berkeley, CA, USA: USENIX Association. Retrieved from <https://users.soe.ucsc.edu/~shreyask/ism245-rpt.pdf>
- Thefreedictionary (2015). Bullying Definition. Retrieved from <http://www.thefreedictionary.com/bully>
- Twitter (2014, January 22). Taking Action Against Abuse. Retrieved from <https://blog.Twitter.com/en-gb/2014/taking-action-against-abuse>
- Twitter (2015). Connecting to a Streaming Endpoint. Retrieved from <https://dev.Twitter.com/streaming/overview/connecting>
- Twitter (2015). The Search API. Retrieved from <https://dev.Twitter.com/rest/public/search>
- Twitter (2015). Streaming Message Types. Retrieved from <https://dev.Twitter.com/streaming/overview/messages-types>
- Twitter (2015). Tweets. Retrieved from <https://dev.Twitter.com/overview/api/tweets>
- Wang, J., Lannotti, R. J., Luk, J. W., & Nansel, T. R. (2010). Co-occurrence of Victimization From Five Subtypes of Bullying: Physical, Verbal, Social Exclusion, Spreading Rumors, and Cyber. *Journal of Pediatric Psychology*, 35(10), 1103-1112.
- Retrieved from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2980945/>
- Xu, J., Zhu, X., & Bellmore, A. (2012, August 12). Fast Learning for Sentiment Analysis on Bullying.
- Retrieved from <http://pages.cs.wisc.edu/~jerryzhu/pub/wisdom12.pdf>

APPENDIX A

BULLYING, SUICIDE, AND DEPRESSION

A.1 Bullying, Suicide, and Depression

In the past few years, reports about suicides and bullying have become common. While it is good to raise awareness about bullying and suicide, it must be controlled since it could lead to negative impacts. It should be understood that suicide is not a natural response to bullying, and that blaming and punishing may mislead from the cause of the problem rather than getting help for bullying. There are other important risk factors that lead to bullying that should be addressed in coordination with bullying treatment. This said there are no direct links found between bullying and suicide.

A.2 Bullying and Lack of Achievement

Bullying happens to everyone, whether you stand out due to excellence, or stand out due to lack of achievement. An UCLA study appearing in the *Journal of Early Adolescence*, showed a significant difference in performance between bullied and non-bullied students. The study was conducted on 2,300 students in 11 different Los Angeles public middle schools. The study was conducted by asking all students on a 1 to 4 scale if they were bullied, then asking their peers which other students they felt were bullied, and finally asking the teacher to rate the students' performance. While bullying was shown to correspond to a lower GPA, in the area of math it was most

evident where 1 point on a 1-4 rating scale corresponded to a 1.5 loss in GPA on a 1 – 4 scale (Juvonen, Wang & Espinoza, 2010). These findings show the significance of bullying and its effects on students.

A.3 Bullying and Long Term Effects

There are many long term effects that go beyond the classroom. These effects are subtle and impact the bullied and bullies dramatically as they mature.

One study, done at the John Hopkins School of Medicine, shows how youth that are either bullied or bullies are at a higher risk of violence including carrying weapons and become members of gangs (Goldweber, Waasdorp & Johnson, 2013).

Yet another study shows positive effects of bullying for the bullier. In a study done at the University of North Carolina at Chapel Hill, they show how bullies show no change in stress levels over time as compared to the bullied that show increased stress levels. This was done by using CRP (C-reactive protein) measurements at different time intervals with 1,420 participants. Bullied and bullies were asked if they were being bullied and a blood sample was taken. Over time all participants of the study had increased levels of CRP, however, the people who were bullied showed a much larger increase in CRP levels. These long terms effects can lead to health problems and social difficulties (Copeland, Wolke, Lereya, Shanahan, Worthman & Costello, 2014).

APPENDIX B

TWITTER COMMUNICATION OVERVIEW

Twitter offers a REST API on their general accessible website that allows their tweets to be consumed. The REST API consists of a few different parts: authentication, rate-limiting, streaming, and search. Authentication refers to the ability to access the API resources through a program by using keys that identify who you are. Streaming and search refers to services offered by the API. The streaming API provides real-time tweets that users are sending. The search API data is a small portion of previous tweets that were once part of the steaming API. In all cases of data consumption, streaming and searching, there exists rate-limiting. Current rate limits average at 15 calls per 10 minutes for search (including profile information fetching), and is much lower for streaming. Although, this changes as Twitter becomes more burdened with users and the program will begin to download information at a slower rate.

The streaming API is broken down into three different types: The public stream, user stream, and site streams. The real-time streams work by the application performing an HTTP get request for a file, and constantly receiving data from a file of infinite size. The file being received is all the streaming data from the Twitter API that was requested by the client. Each account can only have one streaming connection at one time, if an account attempts to make more than one connection then they run the risk of having

their IPAddress banned from the API (Twitter, 2014). This thesis uses the public stream to get real-time bully biased tweets just based off key word searches, in the early stages of the project. The streaming API is picky where rate-limiting, reconnection, missing data fields, and error codes are all issues that need to be addressed when utilizing this interface. When the data is returned it is returned starting with “\r\n” characters, to state this is the start of a new tweet. The “\r\n” will be followed by a JSON formatted message (JSON formatting is not covered in background information in this thesis). The types of JSON formatted responses from the Twitter API are standard tweet formats and additional message types which are usually meta information about the tweets (status has been deleted or user has unfollowed another user) or meta information about the current connection (e.g., your connection is falling behind) (Twitter: Tweets, 2015) (Twitter: Streaming Message Types, 2015).

The public stream is real-time publicly available data including public tweets, hashtag filters, etc. This information can be obtained with the application keys. This type of connection receives general tweet messages and meta data about the connection in a JSON format.

The user stream type of the streaming API allows an application to track a specific user’s tweet time-line in real-time. The type of streaming connection requires user authentication, thus will require user keys. The types of messages are friends’ lists, direct messages, event messages, and too many follower warning messages in JSON format. This type connection is meant for when a single user is utilizing an application, ie not one application for many users (Twitter: Streaming, 2015). In this thesis, we use this

API to get the tweets from each user's time-line that we are following.

The last type of stream is the "type streams". Type streams are a special stream type that requires twitter's prior approval, where it monitors the tweet time-line for many users. The types of messages here are the same as the user streams, but on a much larger scale. Each user has to grant permission to the application, thus the application must have user keys for every user they are gathering information from.

The search API offered by Twitter returns a sample set of tweets based off the given query parameters through the GET method of HTTP. The tweets are based on popularity rather than relevance to the search parameters. For searching there are different query parameters that will allow AND, OR, exclusions between keywords, and also allow specifying a geo-location.

APPENDIX C

PLUTCHIK'S EMOTIONAL WHEEL

Robert Plutchik originally came up with the emotional wheel in the 1980's (Plutchik, 2001). He talks about the difficulties of coming up with a set of emotions that would be representative of every emotion, the basic emotions from which others can be derived. He eventually derived these emotions from previous attempts and from observations of animals. He stated that animals have emotions and even bacteria perform a mating dance. Based off these observations, he hypothesized the emotions an animal was expressing, and these eventually lead to his list of primal emotions. He states that emotions come from evolutionary responses in an effort to adapt and survive in the environment. The below figures are examples of how he found primal emotions by observation.

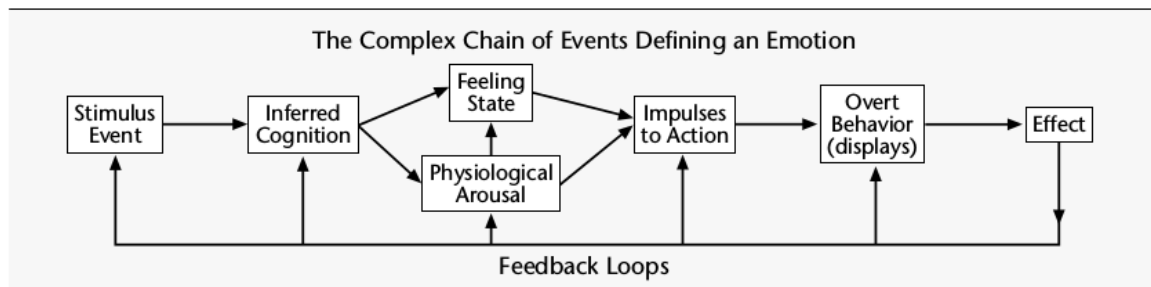


Figure 17: Emotion Cause and Effect (Plutchik, 2001)

Figure 18, is an example of the color wheel and how multiple emotions can be combined to form other emotions.

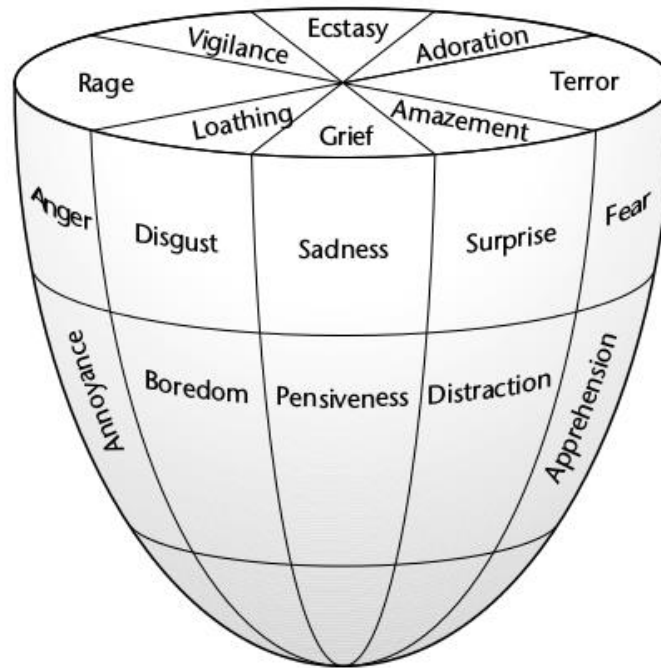


Figure 18: Emotion Amplitude Chart (Plutchik, 2001)

The link between feelings and cognitive thought and that there are feelings evident in all creatures is what Plutchik was attempting to convey. Through his work we can see an ontology forming around emotions, where it is possible to have two emotions combined to form different emotions, as shown in Figure 19.

Primary Emotion Components	Labels for Mixed Emotions
Joy + Acceptance	= Love, Friendliness
Fear + Surprise	= Alarm, Awe
Sadness + Disgust	= Remorse
Disgust + Anger	= Contempt, Hatred, Hostility
Joy + Fear	= Guilt
Anger + Joy	= Pride
Fear + Disgust	= Shame, Prudishness
Anticipation + Fear	= Anxiety, Caution

Figure 19: Emotion Combination Chart (Plutchik, 2001)

APPENDIX D

AMAZON MECHANICAL TURK FORM

Figure 20 is an example of a form that was used by an Amazon Mechanical Turk rater in order to classify a relationship. In this picture it can be seen that the rater classified the relationship as being negative.

[–] Instructions (Open full instructions in a separate window)

Caution: I will be validating these results by posting 3 batches of the same stuff and comparing with randomly selected results. If you dont understand what I am asking for or how to do it, then dont do it, you will be wasting your time. However, this will be quick once you have read the intructions one time. Thanks for your time and good work.

The following are conversations from one peron to another, these are only one side of the conversation: person A to B. The goal is to find out if person A is bullying person B using the official definition provided to evaluate the "text" portion of the information below. Here a recursive act is multiplue conversations, different "CN" values of the information below.

POSITIVE = is case of bullying

CN = conversation number, same number = sentences are part of the same conversation

fromUser and toUser = self explanatory feilds

text = sentences to look at and see if bullying is occuring

Official Definition: "unwanted aggressive behavior(s) by another youth or group of youths who are not siblings or current dating partners that involves an observed or perceived power imbalance and is repeated multiple times or is highly likely to be repeated"

CN: 9236 fromUser: MattyMcCalahan toUser: talking_2_crazy text: Also-notice how amp always appear together use the same stupid language then disappear 2gether

CN: 9236 fromUser: MattyMcCalahan toUser: talking_2_crazy text: Jesusbruce is a little moron thinks he actually insulting some1 LOL

CN: 8341 fromUser: MattyMcCalahan toUser: talking_2_crazy text: So jesus just happened to appear at ur whim How fricken stupid r you Liar

CN: 8341 fromUser: MattyMcCalahan toUser: talking_2_crazy text: Bruce is a little loser that that spends all his time creating multiple accounts harassing

CN: 8341 fromUser: MattyMcCalahan toUser: talking_2_crazy text: That was your Jesus account not this one duh u drunk

CN: 8340 fromUser: MattyMcCalahan toUser: talking_2_crazy text: This little moron couldnt terrorize a gnat LOL

CN: 8340 fromUser: MattyMcCalahan toUser: talking_2_crazy text: I blocked YOU cause u used racial slurs that were sickening About as low as possible

CN: 8340 fromUser: MattyMcCalahan toUser: talking_2_crazy text: Its 1 person same person Hostile little moron who prides himself on trying to bully


Category	Worker 1
NEGATIVE	

Figure 20: Example of an AMT Form for Rating Relationships