AUTOMATED DETECTION OF IMPROPER BEHAVIOUR ON ONLINE SOCIAL NETWORKS

Nathaniel D. Brown

Cape Peninsula University of Technology

B.Tech Information Technology

2017

Abstract

The purpose of this study was to research the capability of using sentiment analysis, a form of natural language processing, to detect improper behaviour on Online Social Networks through automated computer software. The need for this specific research is evident of the fact that as social media content has been growing, so has improper behaviour on Online Social Networks. The improper behaviour tested in this study includes hate speech, cyber-bullying and cyber-stalking, with sentiment analysis used to evaluate the sentiment of each of the text.

Data in this study was collected using public social media posts and messages anonymously. The data was inserted into a software artefact which incorporated the sentiment analysis algorithm. The software artefact produced the experimental data and the data was analysed using statistical methods. It was determined that the sentiment analysis algorithm used in this study could accurately detect cyber-bullying, but hate speech and cyber-stalking with decreasing accuracy.

Keywords

Sentiment analysis, natural language processing, social networks, hate speech, cyber-bullying, cyber-stalking

Table of content

1 Introduction

This paper examines the complexities of automatically detecting improper behaviour on online Social Networks. Improper behaviour such as cyber-bullying, cyber-stalking and hate speech detection will be explored. The paper considers how technological innovations such as sentiment analysis can restrict the harm caused by improper behaviour and content on online Social Networks.

People have the natural ability to identify whether text in a paragraph has a positive or negative impression or opinion of the matter at hand. However, automating this ability is a challenge as computers have no concept of natural language.

A technology known as sentiment analysis seeks to solve this problem and gives computers the ability of automatically processing human opinions in text (Pang & Lee, 2004). Sentiment analysis is also known as opinion mining, a form of natural language processing (NLP), which mines text to understand people's subjective expressions (Liu, 2010). In 2011, Kouloumpis, Wilson & Moore (2011) used sentiment analysis to identify negative, positive and neutral tweets on Twitter of how people feel about certain products or services. Additionally, Schmidt & Wiegand (2017) recently did a survey using sentiment analysis to automatically detect hate speech on the Internet. Their primary aim of the survey was to reach researchers specialising in natural language processing who were new to hate speech detection. Their work is used as a foundation in this paper to not only detect hate speech, but cyber-bullying and cyber-stalking as well.

2 Background to the research problem

The online Social Networks have dramatically changed the way people express their views and opinions (Liu, 2010). Unfortunately, this has also opened door for people to express improper behaviour such as cyber-bullying, cyber-stalking, hate speech, etc. online. The Internet provides anonymity for users which also allow offenders to pretend to be someone completely different. This gives the offenders the opportunity to be in contact with anyone with Internet access with no regard of the repercussions or fear of being caught (Ellison & Akdeniz, 1998). In recent years, improper behaviour has been increasing on online Social Networks in South Africa (Lujabe, 2017).

Before appropriate action can be taken, most of the time improper behaviour on online Social Networks must be reported manually or flagged by users who witnessed or was a victim of the improper content (Crawford & Gillespie, 2016). The online Social Networks have dedicated teams to moderate inappropriate content, but mostly rely on users to report it as it is infeasible for them to check each post posted by millions of users (Facebook, n.d.). Online Social

Networks also put much weight behind the user report and flag mechanism, that sometimes it removes posts incorrectly on the basis on the users own definition of morality (Crawford & Gillespie, 2016). Therefore, this must be done in line with the online Social Network's terms of service and local laws.

Researchers consider indirect aggression and relational/social aggression as forms of bullying (Smith, 2006). Cyber-bullying is a recent phenomenon that only emerged a few years ago, in which the aggression occurs through technological devices, especially mobile phones and the Internet. On schools, research as shown that the identity of the cyberbully is known by the victim (Smith, 2006), but cyber-bullying on online Social Networks may span geographical distances.

YouGov and Vodafone conducted a global online study that surveyed almost 5 000 teenagers for cyber-bullying (YouGov, 2015). The teenagers were aged between 13 and 18 and the survey was done across 11 countries. In South Africa, it found an average of 24% of teenagers had been cyber-bullied, whereas the average across the countries was 18% of teenagers (YouGov, 2015). In October 2016, a video showing a teenage girl bullying another girl went viral on social media (Johns, 2016). The teenage girl subsequently apologised for her behaviour (Tswanya, 2016).

Additionally, cyber-stalking is distinguished from traditional stalking as it relies on the Internet to perpetrate the offence, but may or may not be illegal under law in certain jurisdictions (Ellison & Akdeniz, 1998). Traditional stalking on the hand is a criminal offence.

Hate speech online alone has increased dramatically in South Africa and has been widely published by media (Lujabe. 2017). On January 3 2016, a woman posted remarks on Facebook which was constituted as hate speech by the Equality Court magistrate. Another woman, in February 2016 faced charges of crimen injuria following a viral video speaking improperly about black police officers. In May 2016, a political activist posted on social media that he made a waitress cry at a restaurant due to white ownership of land. Also in May, a young man posted a racist remark on Facebook and agreed to complete community service facilitated by the Human Rights Commission. In January 2017, another man posted a racist Facebook status update which was caught by media.

As social media content has been growing, along with hate speech and the like, interest in improper behaviour detection has been increasing. Unfortunately, basic word filters do not provide a sufficient way to address the issue. Nevertheless, the use of sentiment analysis and natural language processing may prove to automatically detect improper behaviour in the light of cultural, race and language differences (Schmidt & Wiegand, 2017).

3 Statement of research problem

Online Social Networks is a breeding ground for improper behaviour as many of the inappropriate content goes undetected (Banks, 2010). As social media content has been growing, improper behaviour such as hate speech has been increasing as well (Lujabe. 2017). The online Social Networks, such as Facebook rely mostly on the users who witnessed or was a victim to report the improper behaviour (Facebook, n.d.; Crawford & Gillespie, 2016). However, the harm has already been inflicted (Crawford & Gillespie, 2016). In addition, Twitter increased the size of their teams to identify improper behaviour but it was not enough, revealing that manual identification alone is not sufficient (Panzarino, 2017).

4 Research questions and sub-questions

- How can we use sentiment analysis to detect improper behaviour on online Social Networks?

4.1 Aim and Objectives of the research

Aim:

- Determine whether using sentiment analysis will automatically detect improper behaviour on online Social Networks. Additionally, develop a strategy or policy that can be considered by online Social Network providers.

Objectives:

- Gather public Facebook and Twitter data from their respective application program interfaces (API).
- Investigate sentiment analysis, and determine if it can detect improper behaviour.
- Develop an application artefact using the sentiment analysis technique.
- Develop a strategy or policy that can be considered by online Social Network providers.

5 Literature review

5.1 Introduction

In this chapter, the published literature of sentiment analysis and related research topics will be analysed and summarised. This will include a report of the fundamental background of this research topic, its latest findings, challenges and other related work.

5.2 Sentiment analysis

Facts and opinions are the two categorisation types of textual information in the world, with facts being objective and opinions being subjective (Liu, 2010; Yu & Hatzivassiloglou, 2003). Analysing opinions computationally in text is rather challenging, requiring question-answering systems that can extract information with the ability to distinguish between facts and opinions (Pang & Lee, 2004). Pang and Lee (2004) explored that sentiment analysis can identify people's viewpoints in text, including the ability to differentiate between positive and negative statements. Analysing opinions computationally, its subjectivity and sentiment have attracted a great deal of attention due to the great value for practical applications (Liu, 2010; Pang & Lee, 2004; Wiebe, Wilson & Cardie, 2005). According to the research done by Liu (2010), the companies that provide sentiment analysis services at least number 20-30 in the USA alone.

5.3 Analysing opinions

Sentiment analysis begins with analysing opinions. Opinions typically constitute the feelings, appraisals or sentiments of people toward entities, events and its properties (Liu, 2010). The term "sentiment", according to the Cambridge (n.d.) dictionary, is "a thought, opinion, or idea based on a feeling about a situation, or a way of thinking about something". In general, this is what sentiment analysis attempts to identify from textual information. Pang and Lee (2008) stated that sentiment is context-sensitive and domain dependent, as different domains can indicate different sentiment for the same expression. For example, for book reviews the sentence "go read the book" indicates a positive sentiment, but a negative sentiment for movie reviews (Pang & Lee, 2008). Analysing opinions are important because people tend to make decisions based other people's opinions, which is also true for organisations (Liu, 2010).

Earlier research into sentiment analysis and NLP by separating opinions from facts divided the textual identification into two sub-topics; document-level and sentence-level identification (Yu & Hatzivassiloglou, 2003; Wiebe et al., 2005, Liu, 2010). This is the most researched area in academia, but sentiment analysis is treated only as a text classification problem (Liu, 2010). Document-level identification of opinionated text classifies entire documents or paragraphs as either fact or opinion (Yu & Hatzivassiloglou, 2003). For example, classifying customer reviews as positive or negative and distinguishing news articles from editorials (Pang, Lee, & Vaithyanathan, 2002; Yu & Hatzivassiloglou, 2003). Sentence-level identification limits classifying sentences as either fact or opinion (Yu & Hatzivassiloglou, 2003). Subsequently, sentence-level identification is closely associated with document-level identification, because documents that have mostly opinion sentences tend to be classified as opinionated documents (Yu & Hatzivassiloglou, 2003). Yu and Hatzivassiloglou (2003) used the Naive Bayes classifier to calculate the likelihood of whether a document is fact or opinion, portrayed in Figure 1.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Figure 1. Naive Bayes classifier theorem (Yu & Hatzivassiloglou, 2003; Murphy, 2006). Where A is a class, B is a document and single words are used as a feature.

Researchers Wiebe, Wilson and Cardie (2005) later argued that identifying and classifying opinionated text at document-level or sentence-level is not sufficient. A sentence may contain two or more opinions, or include both opinions and facts. Wiebe et al. (2005) continue to suggest that not only should the system be able to identify whether documents or sentences are opinionated, but also identify the strength of the opinion. This is important for identifying the emotions and attitudes of the people stating the opinions (Liu, 2010). For example, the NLP system should be able to identify strong rants or rhetoric of violent persons or groups of interest (Wiebe et al., 2005). Wiebe et al. (2005) investigated the identification of opinion and emotion in text using a corpus annotation study, instead of document-level or sentence-level. The corpus annotation study used was a fine-grained annotation scheme, annotating text at the phrase and word level (Wiebe et al., 2005). The goal was to represent internal mental and emotional states of people. Wiebe et al. (2005) termed these states as a private state, saying that a private state is "a general term that covers opinions, beliefs, thoughts, feelings, emotions, goals, evaluations, and judgments.". The private state is divided into functional components – experiencers, attitudes and targets (Wiebe et al., 2005). For example, in the text "John hates Jane", John is the experiencer, hate is the attitude and Jane is the target. This aids the identification of emotions and strength of opinions needed to identify improper behaviour on online Social Networks.

5.4 Analysing emotions

As stated previously, dealing with improper behaviour includes identifying the feelings and emotions portrayed by the individual's text. This is a key area of sentiment analysis still being researched (Liu, 2010). Parrot (2001) states that people have 6 types of primary emotions which are joy, love, surprise, sadness, fear and anger. There are different intensities for each of these primary emotions and the strength of an opinion is closely related to it, however emotions and opinions are not equivalent (Liu, 2010). This presents a problem of analysing the subjective feelings of emotions and opinions in text and the type of behaviour the person is portraying.

People have mental states (feelings) and language expressions used to describe their feelings (Liu, 2010). A language expression is "the act of saying what you think or showing how you feel using words or actions" (Cambridge, n.d.). Liu (2010) states that "Sentiment analysis essentially tries to infer people's sentiments based on their language expressions". Unfortunately, many language expressions can be used to express the 6 types of primary

emotions. Additionally, describing positive or negative sentiments has an unlimited number of opinion expressions (Liu, 2010).

Pang and Lee (2014) proposed using machine-learning with sentiment analysis to contend with the time and complexity of analysing the large number of language expressions. Wiebe et al. (2005) agree, stating that the use of machine learning has become the method of choice for NLP systems, including statistical approaches. However, these methods require training the NLP system with test data that must be manually configured to be affective (Wiebe et al., 2005).

5.5 Sentiment analysis models

There are different types and models of sentiment analysis.

Before it determines if the opinion is positive, negative or neutral, the feature-based model first identifies the targets on which opinions have been expressed (Liu, 2010). The targets are the objects, and their attributes and features. For example, "The camera quality of the phone is bad", the comment is on "camera quality" of the phone object and the opinion is negative. An important thing to note here is that document-level and sentence-level identification does not discover this kind of information (Liu, 2010).

Another sentiment analysis model is analysis of comparative sentences. This model identifies opinions comparing similar objects such as competing products (Liu, 2010). For example, "The camera quality of this phone is better than phone-x".

Opinion search is a model that empowers search engines to retrieve results in accordance with the user's opinion of the search query (Liu, 2010; Pang & Lee, 2008). For example, a user dislikes a certain car manufacturer and when searching for that car manufacturer, the search engine will retrieve negative opinionated articles. This model is a combination of sentiment analysis and information retrieval (Liu, 2010).

5.6 Challenges facing sentiment analysis

According to Liu (2010), NLP is generally a difficult task especially studying the text with arbitrary level of detail. In many sentiment identification tasks opinion words are employed to identify opinions in text (Liu, 2010). Opinion words are the words used to identify the states of opinions in text. For example, positive opinion words include the words "beautiful" and "wonderful", and negative opinion words include "bad" and "terrible" (Liu, 2010). These opinion words are also known as sentiment words. Additionally, there are also opinion phrases and idioms. Opinion words, phrases and idioms are collectively known as opinion lexicon and are instrumental for sentiment analysis (Liu, 2010). However, the the quality of the opinion lexicon

is related to the language being analysed, as most sentiment analysis research has been done in English (Pang & Lee, 2008).

Researchers typically use two types of approaches to analysing text using opinion lexicons (Liu, 2010). These include the dictionary based approach and the corpus-based approach. Yu and Hatzivassiloglou (2003) used the dictionary based approach with the Naive Bayes classifier, and Wiebe et al. (2005) used the corpus-based approach which they argued was better. However, both approaches have advantages and disadvantages (Liu, 2010). The dictionary based approach is the simplest technique, containing a collection of opinion words to compare against. This approach is however unable to adequately analyse text in context and within specific domains (Liu, 2010). For example, the term "quiet" may be a negative opinion for a person but a positive opinion for a car. The corpus-based approach relies on co-occurrence patterns in a large corpus of words (Liu, 2010). The weakness of this approach is that when used alone, it cannot identify all opinion words as it is hard to prepare a huge corpus to cover all words in a specific language (Liu, 2010). This approach however can identify domain specific opinion words in context due to the large corpus containing patterns to compare against (Liu, 2010).

Pang & Lee (2008) stated that it is easy to produce text that machines find difficult to analyse. Figure 2 gives an example of movie reviews that are difficult to analyse (Pang & Lee, 2008), where the first review is ambiguous and the second contains linguistic errors.

Figure 2. Example of movie reviews (Pang & Lee, 2008)

5.7 Related work

Kouloumpis et al. (2011) investigated the use of sentiment analysis for detecting the sentiment of Twitter messages. Due to the character limitations on tweets, identifying the sentiment was mostly like sentence-level sentiment analysis. The researchers used the Edinburgh Twitter corpus which is a collection of 97 million tweets to carry out their research (Kouloumpis et al.,2011). Additionally, they added features to each tweet which counts the nouns, verbs, adverbs and adjectives (Kouloumpis et al., 2011). Kouloumpis et al. (2011) concluded that their experiments showed that these features that were added were not useful

for sentiment analysis and will require further investigation. However, the researchers found that using hashtagged data to train the NLP system provided the best results for determining the sentiment of Twitter messages (Kouloumpis et al., 2011).

Schmidt and Wiegand (2017) conducted a survey on automated hate speech detection using NLP. This survey is a foundation for this study. Schmidt and Wiegand (2017) stated that sentiment analysis and hate speech are closely related, as a hate speech message typically pertains to a negative sentiment. Polarity classifiers can predict the polar intensity of an utterance, whether negative or positive, with hate speech resting on negative polar utterances (Schmidt & Wiegand, 2017). Therefore, polarity classifiers are also employed in conjunction with setting the polarity (positive/negative) to be identified. Schmidt and Wiegand (2017) stated that many researchers used the general assumption that hate speech contains specific negative words and used these lexicons to aid hate speech detection. One researcher was noted to have manually compiled a lexicon, named the Insulting and Abusing Language Dictionary which additionally assigns weights to each entry representing the degree of the hate speech impact (Schmidt & Wiegand, 2017). Another researcher devised a set of linguistic features to detect hate speech, detecting imperative statements such as "Get lost!" and using regular expressions to detect predefined good words (Schmidt & Wiegand, 2017). However, the method of just using lexicons faces the challenge of identifying variations of spelling in text. For example, the hate speech "ki11 yrslef a$$hole" poses a problem for using lexicons (Schmidt & Wiegand, 2017). Therefore, hate speech detection cannot be solved by merely looking at keywords, even if researchers use linguistic features (Schmidt & Wiegand, 2017). As Pang and Lee (2008) noted that sentiment is context-sensitive, Schmidt and Wiegand (2017) stated that hate speech cannot be read in isolation but additionally the target of the statement must be identified. For example, the sentence "Put on lipstick and be who you really are" may not be regarded as hate speech if the target of the statement is a woman. Schmidt and Wiegand (2017) continued to note that hate speech exhibits most types of stereotypes, but requires the help of a knowledge base with domain specific assertions.

5.8 Conclusion

The published literature of sentiment analysis portrays that the technology has the ability to identify the sentiment of opinionated text. However, there are some challenges with sentiment analysis being context sensitive which is a requirement of analysing improper behaviour on Online Social Networks.

6 Research design and methodology

6.1 Introduction

This chapter describes the research methodology used in this study. This includes the research design, strategies, the instruments used to collect the data and the research method.

6.2 Research method and approach

This study takes a realistic view, using objective methods to collect data which can support or refute the hypothesis of this study.

According to Creswell (2003), the practice of research is not limited to philosophical assumptions, but philosophical ideas should be combined with strategies and implemented with specific methods. Figure 3 displays how the elements of inquiry combine to form different research approaches (Creswell, 2003).

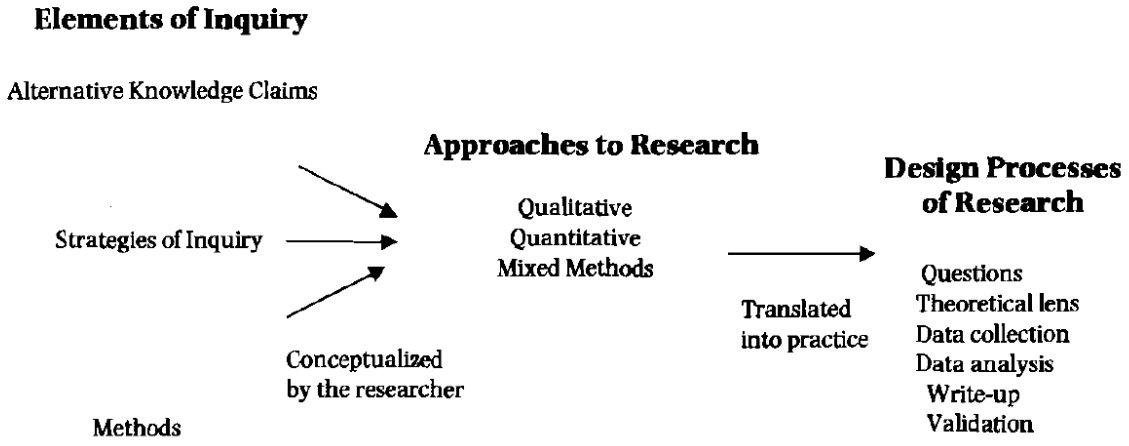

Figure 3: Methods Leading to Approaches and the Design Process (Creswell, 2003)

The three approaches used in research include quantitative, qualitative and mixed methods (Creswell, 2003). A key element of all research, be it quantitative or qualitative is explaining phenomena (Muijs, 2010). In the quantitative research approach scenario, the researcher tests a theory by firstly producing a hypothesis from that theory and collecting data to support or refute the hypothesis (Creswell, 2003). Statistical procedures and hypothesis testing is used to analyse the data collected by this approach, with the data being quantifiable (Creswell, 2003). Aliaga and Gunderson (2002) describe quantitative research as "Explaining phenomena by collecting numerical data that are analysed using mathematically based methods (in particular statistics)." With the qualitative research approach, the researcher collects open-ended data which are typically subjective (Creswell, 2003; Muijs, 2010). Qualitative research results are mostly not quantifiable and measurable and do not limit the scope of the study and the participant's responses (Flyvbjerg, 2006). The mixed methods research approach is flexible; with quantitative and qualitative components can have equal status (Muijs, 2010).

The quantitative research approach is used to satisfy the aim of this study. Muijs (2010) stated that with quantitative research, researchers are far less limited than typically perceived compared to qualitative research. This is because some data that naturally don't appear in quantitative form can be collected in a quantitative method. Additionally, quantitative research is flexible and able to study almost an unlimited number of phenomena (Muijs, 2010). This is important for objectively studying the data and results produced by subjective opinions and emotions in sentiment analysis. Sentiment analysis uses large sample sizes, measuring ranges and frequencies of opinions, and analyses the numerical data (Pang & Lee, 2002; Wiebe et al., 2005; Liu, 2010; Kouloumpis et al., 2011; Schmidt & Wiegand, 2017). The researcher needs to be detached from the research as much as possible which is a key attribute of the quantitative research approach (Muijs, 2010). Testing theories and hypotheses is also a key specialty of the quantitative research approach (Muijs, 2010).

6.3 Research strategy

Table 1: Strategies of Inquiry (Creswell, 2003)

Quantitative	**Qualitative**	**Mixed Methods**
Experimental designs Non-experimental designs, such as surveys	Narratives Phenomenologies Ethnographies Grounded theory Case studies	Sequential Concurrent Transformative

In Table 1, Creswell (2003) focused on two types of strategies of inquiry in the quantitative research approach: experiments and surveys. Experiments are sometimes known as "the scientific method" and happen in controlled environments (Muijs, 2010). Muijs (2010) defined an experiment as "a test under controlled conditions that is made to demonstrate a known truth or examine the validity of a hypothesis". Surveys use questionnaires or interviews for data collection in cross-sectional and longitudinal studies from a sample or population (Creswell, 2003). For this study, an experiment strategy of inquiry will be used.

6.4 Data sample

To detect improper behaviour such as cyber-bullying, the range of data collected from public text data will begin from the age of 13. The public text data is collected from comments and messages from Online Social Networks.

6.5 Data collection method

Public text data such as comments and messages from Facebook and Twitter were collected. All-caps text will remain as is and repeated characters be replaced by their single

character. Personal identifiable information was removed. The text data was confined to positive, negative, hate speech, cyber-bullying and cyber-stalking comments and messages.

Thereafter, the text was categorised into five groups of opinionated sentence lists. These five groups of opinionated sentence lists are: positive, negative, hate speech, cyber-bullying and cyber-stalking. The sentences contained in these lists are the dependent variables of the hypothesis of this study. The positive and negative sentience lists are the controls of the experiment. The Social Sentiment Analysis algorithm should accurately identify the sentiment of these lists using the software artefact.

6.6 Data collection instruments

The data collection instruments that will be used in this study includes; Application Programmable Interfaces (APIs), opinion lexicons, sentiment classifiers and algorithms. These instruments are integrated into a single software artefact to produce the results needed for this study.

6.6.1 Social Sentiment Analysis algorithm

The Social Sentiment Analysis algorithm is used as part of the Algorithmia API (Algorithmia, n.d.). The algorithm is based on the sentiment analysis research of Hutto and Gilbert (2014). Hutto and Gilbert (2014) built the sentiment analysis algorithm with the aim to work well with social media styled text, handle multiple domains and requires no training data. Most importantly for this study, it is fast enough to be used with the streaming data of Social Online Networks (Hutto & Gilbert, 2014).

The algorithm consumes a single English sentence or multiple sentences and returns sentiment ratings of positive, negative, neutral and compound in the JavaScript Object Notation (JSON) format. The first three sentiments ratings scale from 0 to 1. Compound sentiment is the overall sentiment, where it scales between -1 to 1, negative to positive respectively.

6.6.2 Opinion lexicons

For accurate sentence-level or document-level sentiment analysis, a comprehensive and high-quality opinion lexicon is essential (Hutto & Gilbert, 2014). The Linguistic Inquiry and Word Count (LIWC) opinion lexicon is widely used in social media, as it has been extensively validated and reliable to extract emotion or sentiment from text (Hutto & Gilbert, 2014). However, Hutto and Gilbert (2014) argued that LIWC has little regard for the suitability of the domain. Therefore, they developed the Valence Aware Dictionary for Sentiment Reasoning (VADER) opinion lexicon to produce a gold-standard opinion lexicon for microblog-like domains such as social media.

The Social Sentiment Analysis algorithm used in this study uses the VADER opinion lexicon. According to the research of Hutto and Gilbert (2014) VADER performs exceptionally well in the social media domain, performing close to human raters using the Pearson Product Moment Correlation Coefficient and similar to competing machine learning approaches.

6.6.3 Software artefact

In order to produce the experimental data for this study, a software artefact was built using agile methodology to incorporate the APIs and algorithms in an easy to use interface. See Appendix A Figure 2 for a view of the artefact's dashboard.

A live demo of the artefact can be located at http://project4.natebrown.co.za. In addition, the website contains links to the project planning and record of version control of the artefact.

6.7 Hypothesis

6.7.1 Variables

Independent variable:

- An opinionated/emotional English text grouped into one of these five categories: positive, negative, hate speech, cyber-bullying or cyber-stalking.

Dependent variables:

- The sentiment polarity of the English text which are ratios of positivity, negativity, neutrality and the compounded polarity.

6.7.2 The hypothesis

The null hypothesis of this study (H_o):

- Sentiment analysis cannot accurately detect improper behaviour such as hate speech, cyber-bullying and cyber-stalking on Online Social Networks.

The alternative/experimental hypothesis of this study (H_a):

- Sentiment analysis can detect improper behaviour such as hate speech, cyber-bullying and cyber-stalking on Online Social Networks.

7 Data collection process, analysis and findings

7.1 Data collection procedure

The test of the Social Sentiment Analysis algorithm used five groups of opinionated sentence lists based on the dependant variables of the hypothesis of this study. The lists include

positive, negative, hate speech, cyber-bullying and cyber-stalking sentence lists. The opinionated sentences in each list are the independent variable of the hypothesis. The positive and negative opinion list contained general examples of opinions people posted on the Online Social Networks. The hate speech and cyber-bullying sentence lists contained examples pulled from Online Social Networks. However, the cyber-stalking sentence list contained a conversation of the perpetrator and victim as cyber-stalking is context sensitive as seen in Appendix A.

For the first four groups of opinionated sentence lists ten sentences were captured per group. The last group of cyber-stalking captured a conversion of a perpetrator and the victim with an additional control cyber-stalking sentence. See Appendix A for the opinionated sentence lists.

The five opinionated sentence lists were captured into the software artefact in code and posted to the Social Sentiment Analysis API, with tests running on each list. The tests were run several times on each list, with each test returning the same results per list.

7.2 Data analysis and findings

The results were captured an analysed to determine if the compound sentiment result accurately identified the correct sentiment of the opinionated sentence (Appendix A).

As stated earlier in the Data collection instruments, the first three sentiments ratings for positive, negative and neutral scale from 0 to 1. Compound sentiment is the overall sentiment, where it scales between -1 to 1, negative to positive respectively.

7.2.1 Positive opinionated sentence list results

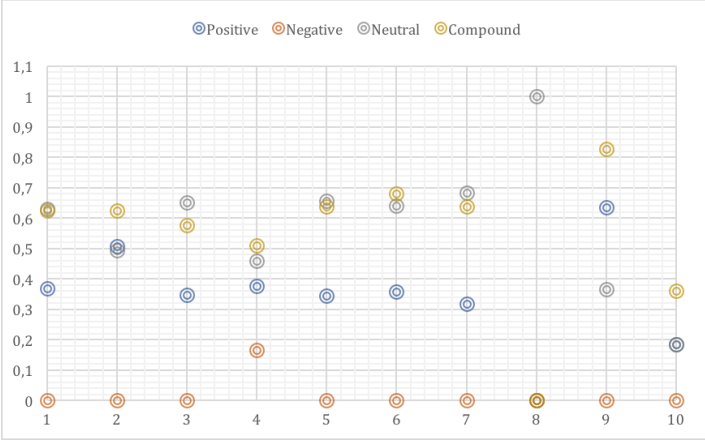

Figure 4: Positive opinions sentiment analysis results

7.2.2 Negative opinionated sentence list results

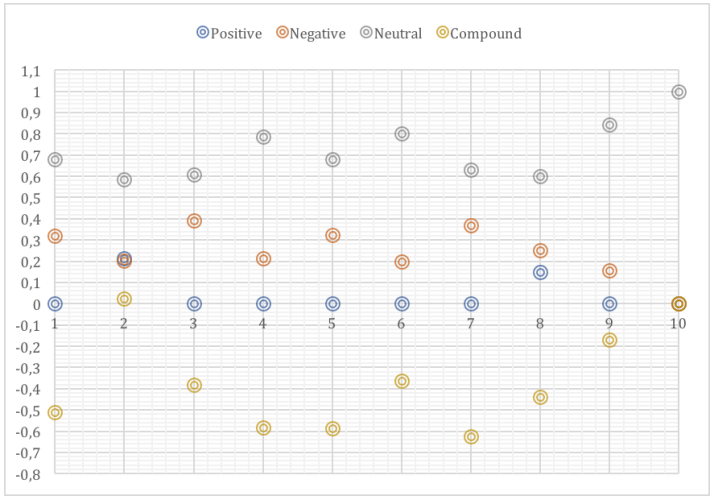

Figure 5: Negative opinions sentiment analysis results

7.2.3 Hate speech opinionated sentence list results

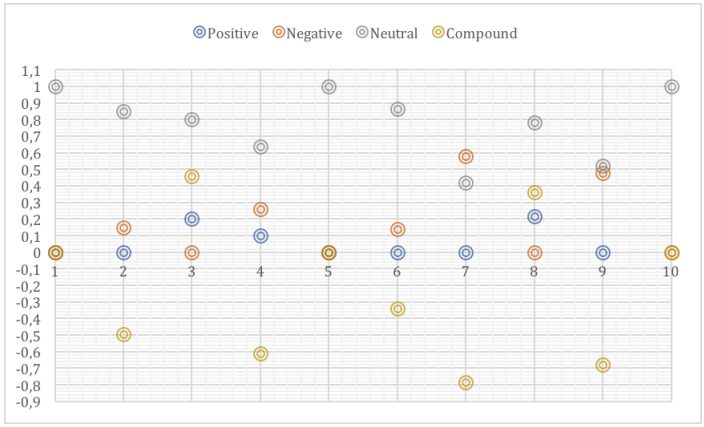

Figure 6: Hate speech sentiment analysis results

7.2.4 Cyber-bullying opinionated sentence list results

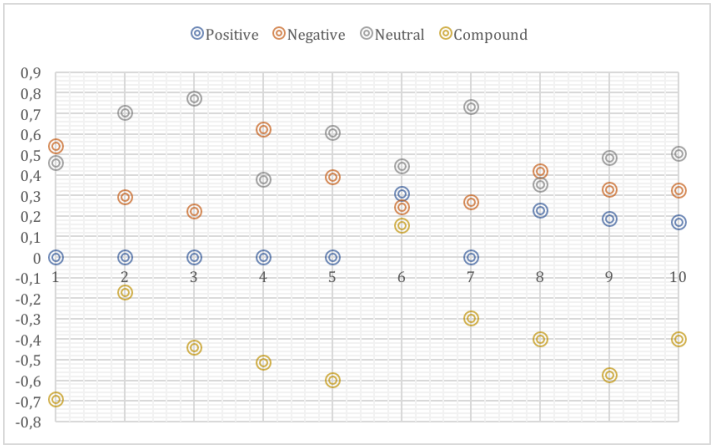

Figure 7: Cyber-bullying sentiment analysis results.

7.2.5 Cyber-stalking opinionated sentence list results

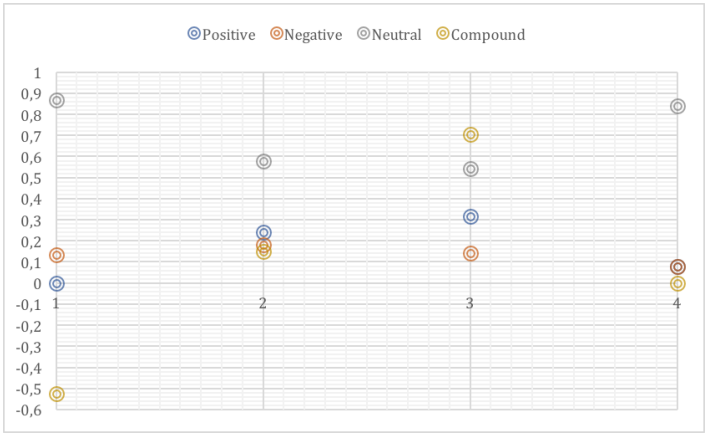

Figure 8: Cyber-stalking sentiment analysis results (perpetrator only)

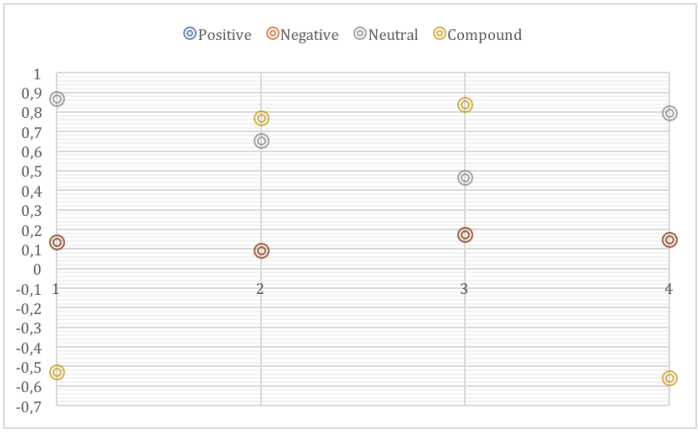

Figure 9: Cyber-stalking sentiment analysis results (perpetrator and victim messages)

7.2.6 Statistical test

The dependent variables for this study are the sentiment polarity of the results which are ratios of positivity, negativity, neutrality and the compounded polarity. The primary dependent variable to be tested is the compounded polarity, as this is the overall sentiment ratio for each opinionated sentence.

As noted earlier in the study, the VADER opinion lexicon that the Social Sentiment Analysis algorithm uses has a Pearson Product Moment Correlation Coefficient of (r = 0.881) (Hutto & Gilbert, 2014). Compare this to individual human raters who obtained a Pearson Product Moment Correlation Coefficient of (r = 0.888) (Hutto & Gilbert, 2014).

The compounded polarity was tested using ANOVA, see Table 2, 3, 4 and 5.

Table 2: Summary of Analysis of Variance in compound sentiment for hate speech and cyber-bullying opinionated list results

	Count	Sum	Average	Variance
Hate Speech	10	-2,0906	-0,20906	0,190686456
Cyber-bullying	10	-3,9227	-0,39227	0,059518142

Table 3: Analysis of Variance in compound sentiment for hate speech and cyber-bullying opinionated list results

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0,167829521	1	0,167829521	1,341538258	0,261900222	0,47384473
Within Groups	2,251841385	18	0,125102299			
Total	2,419670906	19				

Table 4: Summary of Analysis of Variance in compound sentiment for the cyber-stalking opinionated list results

Groups	Count	Sum	Average	Variance
Perpetrator only	4	0,3278	0,08195	0,255977203
Perpetrator and victim	4	0,5256	0,1314	0,604597907

Table 5: Analysis of Variance in compound sentiment for the cyber-stalking opinionated list results

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0,004890605	1	0,004890605	0,011365899	0,918572869	0,514889765
Within Groups	2,58172533	6	0,430287555			
Total	2,586615935	7				

7.2.7 Summary of findings

The Social Sentiment Analysis algorithm using the VADER opinion lexicon was able to identify the negative sentiment of hate speech in text. Unfortunately, problems were found with accurately identifying the sentiment of cyber-bullying and cyber-stalking in text with increasing difficulty.

Therefore, we do not reject the null hypothesis (H_o) of this study.

8 Discussion and interpretation of findings

8.1 Introduction

The interpretation of the compound sentiment polarity results will be discussed in the chapter. The compound sentiment results scales between -1 to 1, negative to positive respectively for each opinionated sentence tested. A compound sentiment result of zero would be an indication of neutral compound sentiment.

8.2 Positive and negative sentence list results

The Social Sentiment Analysis algorithm identified the correct compound sentiment for nine out of ten sentences in the positive and negative sentence lists as seen in Figure 4 and Figure 5. Both sentences that were incorrectly identified as neither positive or negative were identified to have compound sentiment result of zero and a neutral sentiment of 1. The positive sentence that was identified as neutral was "Nelson Mandela is a role model to many of us, and we should follow in his footsteps". While the opinion in this sentence may be a positive for many, the Social Sentiment Analysis algorithm was correct in identifying it as neutral compound sentiment, as the opinion is strongly subjective. However, the negative sentence "Your application has too many bugs, uninstalled" was incorrectly identified as neutral compound sentiment. Nevertheless, the results of the positive and negative control opinionated lists are in line with the VADER opinion lexicon's Pearson Product Moment Correlation Coefficient of ($r = 0.881$) used by the Social Sentiment Analysis algorithm (Hutto & Gilbert, 2014).

8.3 Hate speech sentence list results

As stated earlier by Schmidt and Wiegand (2017) in the literature review, sentiment analysis and hate speech are closely related. Hate speech messages typically pertain to a negative sentiment as seen in Figure 6. However, the results in Figure 6 also reveals the issues that Schmidt and Wiegand's (2017) argued that hate speech detection cannot be solved by merely looking at keywords only. Therefore, four out of the ten sentences were identified incorrectly, with three of the four having compound sentiment result of zero and a neutral sentiment of 1. The hate speech sentence: "Women are like grass, they need to be beaten/cut regularly" scored an incorrect compound value of 0.3612 which indicates that its sentiment is positive. The mean result of the compound sentiment was -0,20906, below the point of neutrality and negative.

8.4 Cyber-bullying sentence list results

Regarding the cyber-bullying opinionated sentence list, the correct sentiment of negative compound polarity has been identified nine out of the ten sentences as seen in Figure 7. No one

likes you, you know…" was incorrectly identified with positive compound sentiment of 0.1531. The mean result of the compound sentiment was -0,39227, below the point of neutrality and negative.

8.5 Cyber-stalking sentence list results

Identifying cyber-staling in text proved more difficult for the Social Sentiment Analysis algorithm to identity the correct compound sentiment polarity as seen in Figure 8. The first control sentence was correctly identified, but the perpetrator's messages in the next three sentences were incorrectly identified as positive or neutral sentiment. The mean result of the compound sentiment was 0,08195, above the point of neutrality and positive.

Due to the three cyber-stalking sentences being incorrectly identified, the victim's messages were then also added to the sentence list and the list reran against the Social Sentiment Analysis algorithm. The reason for this is that in most cases, cyber-stalking in text is normally a conversation between the perpetrator and victim. The control sentence remained correctly identified. However, two of the three previously incorrectly identified sentences were still incorrectly identified with positive compound sentiment, as seen in Figure 9. The mean result of the compound sentiment was 0,1314, and remained above the point of neutrality and positive.

9 Conclusion and recommendation

The research question asked how can we use sentiment analysis to detect improper behaviour on online Social Networks. Whilst, the primary aim of this study was to investigate and experiment if sentiment analysis could detect improper behaviour on Online Social Networks using computer software.

The results of this study reveal that cyber-bullying can be identified by sentiment analysis, which is in accordance with related work by Schmidt and Wiegand (2017) and others. However, sentiment analysis continually loses accuracy as we move from trying to detect hate speech to cyber-stalking in text. Therefore, automated sentiment analysis can be used to detect some forms of improper behaviour, but not all using computer software.

The extent of this study determined the compound sentiment polarity of the text, with improper behaviour identified as negative compound sentiment. Future research could investigate whether sentiment analysis could determine different types of improper behaviour and identifying a specific type in text. In addition, future research in sentiment analysis could investigate conversational based text to improve the identification of cyber-stalking.

During the course of this study, Twitter CEO Jack Dorsey posted tweets stating that the company was bolstering their effort against improper behaviour on their network (Panzarino, 2017). The CEO stated that, "We [Twitter] updated our policies and increased the size of our teams. It was not enough" revealing that manual identification alone is not enough (Panzarino, 2017). As research into natural language processing and specifically sentiment analysis, one day improper behaviour can be automatically identified with speed and accuracy using computer software. The behaviour will be flagged for immediate moderation and restrict the harm caused by hate speech, cyber-bullying and cyber-stalking.

10 Limitation and delineation of the research study

There were a few constraints encountered in this study. Firstly, finding public data of improper behaviour on Online Social Networks to test. Due to the size of major Online Social Networks tested in this study, the amount of posts and messages to filter through was vast. I addition, improper behaviour such as cyber-stalking typically did not transpire in public posts and messages, but in private.

Secondly, the Social Sentiment Algorithm used in this study use a predefined non-configurable opinion lexicon, the VADER opinion lexicon (Hutto & Gilbert, 2014). Even though Hutto and Gilbert (2014) augured against the LIWC opinion lexicon, the ability to configure different opinion lexicons used in the Social Sentiment Algorithm would have been advantageous to this study. This constraint did not allow this study to investigate whether sentiment analysis could determine the different types of improper behaviour in the same text.

11 References

Aliaga, M. and Gunderson, B. 2002. Interactive Statistics. Sage.

Algorithmia, n.d, Social Sentiment Analysis - Algorithm. https://algorithmia.com/algorithms/nlp/SocialSentimentAnalysis [Accessed 10 September 2017].

Banks, J., 2010. Regulating hate speech online. International Review of Law, Computers & Technology, 24(3), pp.233-239.

Cambridge, n.d., Expression Meaning. http://dictionary.cambridge.org/dictionary/english/expression [Accessed 15 June 2017].

Cambridge, n.d., Sentiment Meaning. http://dictionary.cambridge.org/dictionary/english/sentiment [Accessed 15 June 2017].

Crawford, K. and Gillespie, T., 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. New Media & Society, 18(3), pp.410-428.

Creswell, J.W., 2003. Research design: Qualitative, quantitative, and mixed methods approaches. Sage Publications.

Ellison, L. and Akdeniz, Y., 1998. Cyber-stalking: the Regulation of Harassment on the Internet. Criminal Law Review, 29, pp.29-48.

Facebook. n.d.. Community standards. https://www.facebook.com/communitystandards [Accessed 27 April 2017].

Flyvbjerg, B., 2006. Five Misunderstandings About Case-Study Research. Qualitative Inquiry, 12(2):219-245.

Hutto, C.J and Gilbert, E.E, 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14)

Kouloumpis, E., Wilson, T. and Moore, J.D., 2011. Twitter sentiment analysis: The good the bad and the omg!. Icwsm, 11(538-541), p.164.

Liu, B., 2010. Sentiment analysis and subjectivity. In Handbook of Natural Language Processing, Second Edition (pp. 627-666). Chapman and Hall/CRC.

Lynette Johns. 2016. Bullies busted as ugly videos go viral on FB. http://www.iol.co.za/capetimes/news/bullies-busted-as-ugly-videos-go-viral-on-fb-2078775 [Accessed 28 April 2017].

Muijs, D., 2010. Doing quantitative research in education with SPSS. Sage.

Murphy, K.P., 2006. Naive bayes classifiers. University of British Columbia.

Ndileka Lujabe. 2017. 2017 started just like 2016, with racist rants on social media. http://city-press.news24.com/News/2017-started-just-like-2016-with-racist-rants-on-social-media-20170106 [Accessed 28 April 2017].

Pang, B., L. Lee, and S. Vaithyanathan, 2002. 'Thumbs up? Sentiment Classification Using Machine Learning Techniques'. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002). Philadelphia, Pennsylvania, pp. 79–86.

Pang, B. and Lee, L., 2004, July. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271). Association for Computational Linguistics.

Pang, B. and Lee, L., 2008. Opinion Mining and Sentiment Analysis. Information Retrieval, 2(1-2), pp.1-135.

Panzarino, M. 2017. Twitter CEO promises to crack down on hate, violence and harassment with "more aggressive" rules. https://techcrunch.com/2017/10/13/twitter-ceo-promises-to-crack-down-on-hate-violence-and-harassment-with-more-aggressive-rules. [Accessed 30 October 2017].

Parrott, W.G., 2001. Emotions in social psychology: Essential readings. Psychology Press.

Yolisa Tswanya. 2016. Cape teen apologises for bullying video. http://www.iol.co.za/news/south-africa/western-cape/cape-teen-apologises-for-bullying-video-2078813 [Accessed 28 April 2017].

YouGov. Cyberbullying. 2015. Web. 26 Apr. 2017.

Yu, H. and Hatzivassiloglou, V., 2003, July. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the 2003 conference on Empirical methods in natural language processing (pp. 129-136). Association for Computational Linguistics.

Schmidt, A. and Wiegand, M., 2017. A Survey on Hate Speech Detection using Natural Language Processing. SocialNLP 2017, p.1.

Smith, P.K., Mahdavi, J., Carvalho, M. and Tippett, N., 2006. An investigation into cyberbullying, its forms, awareness and impact, and the relationship between age and gender in cyberbullying. Research Brief No. RBX03-06. London: DfES.

Wiebe, J., Wilson, T. and Cardie, C., 2005. Annotating expressions of opinions and emotions in language. Language resources and evaluation, 39(2), pp.165-210.

Appendix A

Opinionated sentence lists and sentiment analysis results

Table 6: Positive opinionated sentence list

Sentence
The camera on my phone takes great pictures
That was an awesome event
The book I read was so good and spine-chilling
The roads were empty, it was great
I think blue is the best colour in the world
The company provides exceptional customer service. I would recommend them to my friends
I think Romans pizza is the best pizza in South Africa
Nelson Mandela is a role model to many of us, and we should follow in his foot steps
The movie was awesome, definitely worth the money
The application on my phone does everything I need it to, thank you

Table 7: Negative opinionated sentence list

Sentence
My phones camera does not take great pictures
The event could have been better, what a waste of time
The book was completely boring
The traffic really was bad, the city has to do something about it really soon
Black is a really bad colour for the new-borns room
That company's customer service sucks. I won't buy anything from them again
Débonnaires makes the worst pizza in South Africa
Adolf Hitler is something we should definitely not role model, he was a murderer
The movies are too expensive nowadays, not worth my money
Your application has too many bugs, uninstalled

Table 8: Hate speech sentence list

Sentence
These monkeys that are allowed to be released on New years Eve And new years day on to public beaches towns etc
From now on I shall address the blacks of south Africa as monkeys as I see the cut little wild monkeys do the same pick drop and litter.
I want to clean this country of all white people. we must act as Hitler did to the Jews.
Queers are an abomination and need to be helped to go straight to Hell!

The Palestinians are beasts walking on two legs.
Now that Trump is president, I'm going to shoot you and all the blacks I can find
The Jews are criminals and our enemies.
Women are like grass, they need to be beaten/cut regularly.
You just need to be raped.
Your Muslim headscarf isn't allowed anymore. Why don't you tie it around your neck & hang yourself with it

Table 9: Cyber-bullying sentence list

Sentence
I'm going to kill you
You are not worth anything
Your are worthless. You are just on Facebook to get attention
You are ugly
I can't believe I have seen anyone this ugly
No one likes you, you know…
There's no reason for you to live
No one wants to be your friend, loser!
You are ugly and fat. You have no friends and no one will ever love you. Why do you even bother coming to school anymore freak
She is ugly and fat!, We all know you have no friends

Table 10: Cyber-stalking sentence list with perpetrator messaging only

Sentence
Let's make this simple. You have until noon. I am not bluffing. Don't be stupid. Once I send pics of you they cannot be unsent.
Well what can I say, I do what I want. But, no worries since you made it very easy to see why you single
Really, yeah I just went through ever single one of your pictures… sure your married lol and I am not a creep, just curious and stubborn thanks
Privacy settings don't stop people who know how to get around those things on facebook, yeah and i still made it into your pics to see

Table 11: Cyber-stalking sentence list of with perpetrator and victim messaging

Sentence
Let's make this simple. You have until noon. I am not bluffing. Don't be stupid. Once I send pics of you they cannot be unsent.

Next time a woman politely declines an unwarranted advance from you, please take that with grace and don't try to convince her otherwise. Well what can I say, I do what I want. But, no worries since you made it very easy to see why you single
HAHAHA IM MARRIED YOU IDIOT. But nice try. Really, yeah I just went through ever single one of your pictures… sure your married lol and I am not a creep, just curious and stubborn thanks
My album is set to private you f*****g idiot. Privacy settings don't stop people who know how to get around those things on facebook, yeah and i still made it into your pics to see

Table 12: Positive opinions sentiment analysis results

Sentence	Positive	Negative	Neutral	Compound
The camera on my phone takes great pictures	0.369	0	0.631	0.6249
That was an awesome event	0.506	0	0.494	0.6249
The book I read was so good and spine-chilling	0.348	0	0.652	0.5777
The roads were empty, it was great	0.376	0.165	0.459	0.5106
I think blue is the best colour in the world	0.344	0	0.656	0.6369
The company provides exceptional customer service. I would recommend them to my friends	0.359	0	0.641	0.6808
I think Romans pizza is the best pizza in South Africa	0.318	0	0.682	0.6369
Nelson Mandela is a role model to many of us, and we should follow in his foot steps	0	0	1	0
The movie was awesome, definitely worth the money	0.635	0	0.365	0.8271
The application on my phone does everything I need it to, thank you	0.185	0	0.185	0.3612

Table 13: Negative opinions sentiment analysis results

Sentence	Positive	Negative	Neutral	Compound
My phones camera does not take great pictures	0	0.32	0.68	-0.5096
The event could have been better, what a waste of time	0.212	0.204	0.584	0.0258
The book was completely boring	0	0.393	0.607	-0.3804
The traffic really was bad, the city has to do something about it really soon	0	0.213	0.787	-0.5829

Black is a really bad colour for the new-borns room	0	0.322	0.678	-0.5849
That company's customer service sucks. I won't buy anything from them again	0	0.2	0.8	-0.3612
Débonnaires makes the worst pizza in South Africa	0	0.369	0.631	-0.6249
Adolf Hitler is something we should definitely not role model, he was a murderer	0.148	0.251	0.601	-0.4404
The movies are too expensive nowadays, not worth my money	0	0.156	0.844	-0.1695
Your application has too many bugs, uninstalled	0	0	1	0

Table 14: Hate speech sentiment analysis results

Sentence	Positive	Negative	Neutral	Compound
These monkeys that are allowed to be released on New years Eve And new years day on to public beaches towns etc	0	0	1	0
From now on I shall address the blacks of south Africa as monkeys as I see the cut little wild monkeys do the same pick drop and litter.	0	0.149	0.851	-0.4939
I want to clean this country of all white people. we must act as Hitler did to the Jews.	0.2	0	0.8	0.4588
Queers are an abomination and need to be helped to go straight to Hell!	0.101	0.26	0.639	-0.6114
The Palestinians are beasts walking on two legs.	0	0	1	0
Now that Trump is president, I'm going to shoot you and all the blacks I can find	0	0.138	0.862	-0.34
The Jews are criminals and our enemies.	0	0.58	0.42	-0.7845
Women are like grass, they need to be beaten/cut regularly.	0.217	0	0.783	0.3612
You just need to be raped.	0	0.479	0.521	-0.6808
Your Muslim headscarf isn't allowed anymore. Why don't you tie it around your neck & hang yourself with it	0	0	1	0

Table 15: Cyber-bullying sentiment analysis results

Sentence	Positive	Negative	Neutral	Compound
I'm going to kill you	0	0.54	0.46	-0.6908
You are not worth anything	0	0.294	0.706	-0.1695
Your are worthless. You are just on Facebook to get attention	0	0.225	0.775	-0.4404
You are ugly	0	0.623	0.377	-0.5106
I can't believe I have seen anyone this ugly	0	0.392	0.608	-0.596
No one likes you, you know…	0.311	0.244	0.444	0.1531
There's no reason for you to live	0	0.268	0.732	-0.296
No one wants to be your friend, loser!	0.227	0.418	0.355	-0.4003
You are ugly and fat. You have no friends and no one will ever love you. Why do you even bother coming to school anymore freak	0.186	0.331	0.483	-0.5719
She is ugly and fat!, We all know you have no friends	0.173	0.324	0.503	-0.4003

Table 16: Cyber-stalking sentiment analysis results of the perpetrator's messages

Sentence	Positive	Negative	Neutral	Compound
Let's make this simple. You have until noon. I am not bluffing. Don't be stupid. Once I send pics of you they cannot be unsent.	0	0.134	0.866	-0.5267
Well what can I say, I do what I want. But, no worries since you made it very easy to see why you single	0.239	0.181	0.58	0.1513
Really, yeah I just went through ever single one of your pictures… sure your married lol and I am not a creep, just curious and stubborn thanks	0.316	0.141	0.543	0.7032
Privacy settings don't stop people who know how to get around those things on facebook, yeah and i still made it into your pics to see	0.08	0.08	0.839	0

Table 17: Cyber-stalking sentiment results of combination of the perpetrator and victim's messages

Sentence	Positive	Negative	Neutral	Compound
Let's make this simple. You have until noon. I am not bluffing. Don't be stupid.	0	0.134	0.866	-0.5267

Once I send pics of you they cannot be unsent.				
Next time a woman politely declines an unwarranted advance from you, please take that with grace and don't try to convince her otherwise. Well what can I say, I do what I want. But, no worries since you made it very easy to see why you single	0.255	0.093	0.652	0.7713
HAHAHA IM MARRIED YOU IDIOT. But nice try. Really, yeah I just went through ever single one of your pictures… sure your married lol and I am not a creep, just curious and stubborn thanks	0.357	0.176	0.466	0.8373
My album is set to private you f*****g idiot. Privacy settings don't stop people who know how to get around those things on facebook, yeah and i still made it into your pics to see	0.056	0.149	0.795	-0.5563

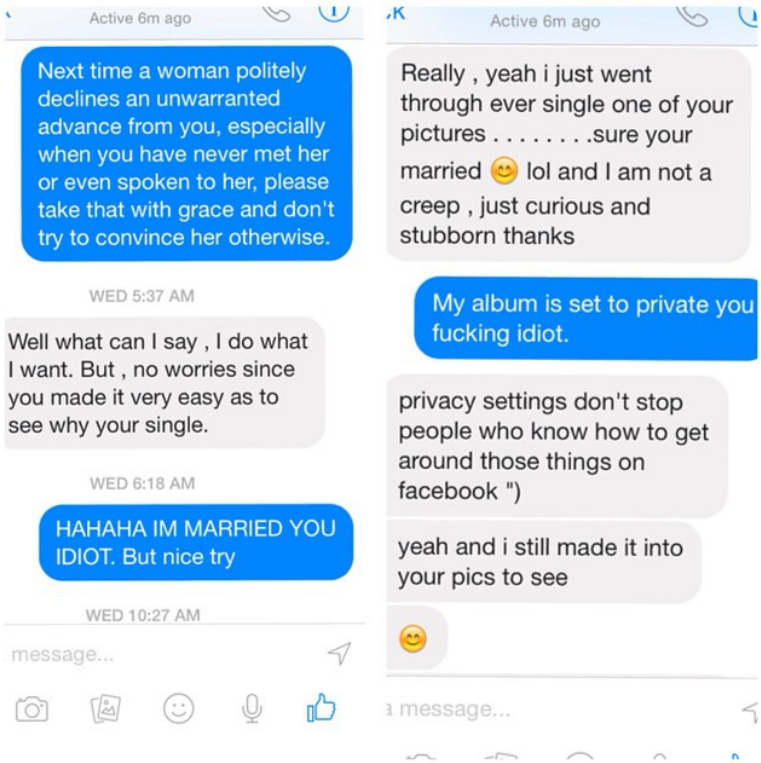

Figure 10: Cyber-stalking conversation between perpetrator and victim

Table 18: Artefact development tools

Category	Tool	Reason
IDE:	JetBrains WebStorm	
Languages:	TypeScript 2.2.3	Strongly typed JavaScript.
	HTML 5 & CSS 3	Web development standard.
Frameworks:	Angular 4.2.4	
	Karma 1.7	Unit test JavaScript frameworks.
	Node.js 6	JavaScript framework package manager.
Platform:	Microsoft or Linux Web Server	
APIs:	Algorithmia: Social Sentiment Analyses API	Sentence-level sentiment analysis algorithm. Can detect more than one sentiment in a sentence.
	Twitter API	
	Facebook Graph API	
Version Control	GitHub	Free and open source.

The version control of this project can be found at https://github.com/nate-brown/automated-online-behaviour-detection. It includes the build instructions to build the project and is located in the README.md file situated in the parent directory.

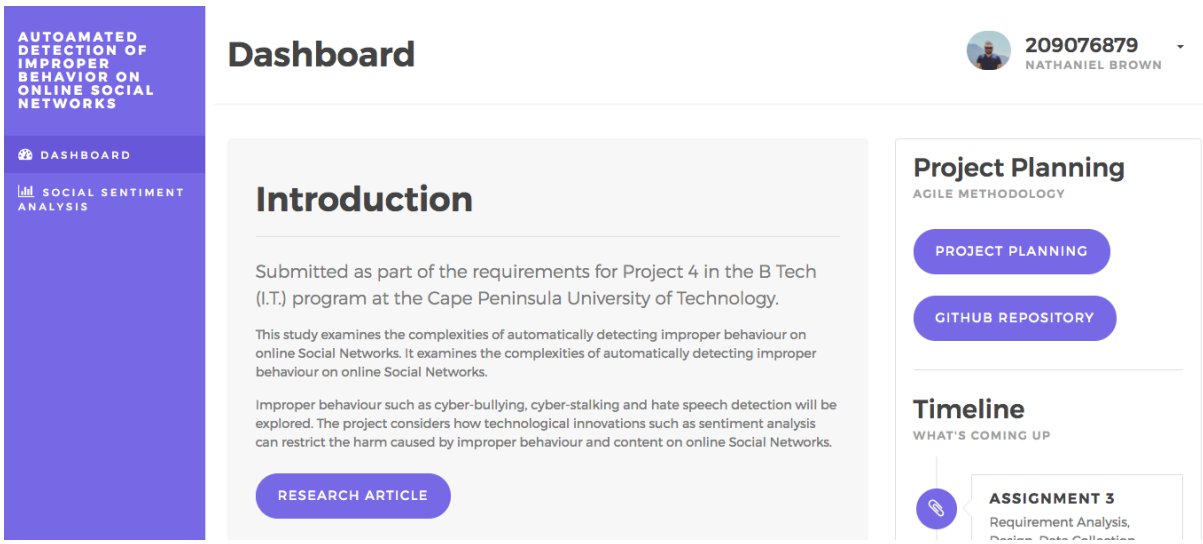

Figure 11: Live demo of software artefact