Lameck Onyango Oyare
MSc. Software Engineering

Reg.SCT313 – 3506/2017

ICS 3101: Research Methodology

Continuous Assessment

Implementing an empathic agent to detect textual Cyberbullying

Table of Contents

1.0 Introduction

The social nature of Web 2.0 applications is increasingly affecting communication and collaboration in today's society. It was argued that message boards, blogs and social networking platforms like Facebook, Twitter, telegram, Instagram or WhatsApp have become an important means of communication and collaboration, especially among teenagers (Van Hee, C). According to Soko Directory (2017), a total of 12 million Kenyans were on WhatsApp with 7.1 million of them on Facebook. Although most of the time, youth's Internet use is thought to be perfectly safe and enjoyable, there are risks involved with the Internet on social media websites. Like offline communities, online communities can be very harmful. Youngsters can be confronted with threatening situations, such as cyberbullying, suicidal behavior or grooming by pedophiles.

As a response to those threats, governments have come up with Cybercrime laws as preventive initiatives. For instance, the Kenyan government enacted the Computer Misuse and Cybercrime Act to increase online youth safety. In spite of these efforts, much undesirable or even hurtful content remains online.

According to nderi, s. (2017, September 27) Cyberbullying can be described as the use of cell phones, instant messaging, email or social networking sites such as Facebook and Twitter to harass, threaten and intimidate someone. Back in the day, bullying used to be on the playgrounds of schools or in a case of high school in the dorms where the older boys used to pick on the formative. However, with the advent of technology, bullying has gone digital whereby someone can bully you anonymously and from anywhere in the world. All one needs is access to communication technology. Given the gravity of the problem and its rapid spread among the youth, there is an immediate and pressing need for research to understand how textual cyberbullying occurs today so that techniques can be rapidly developed to accurately detect, prevent, and mitigate textual cyberbullying.

Sentiment analysis is generally denoted as techniques used to determine the predisposition of text, usually expressed in free text form. In this research, I used sentiment analysis for text classification and analyze incoming messages and tell

whether the underlying sentiment is positive or negative.

Subjective information in source materials is recognized and extracted by the means of natural language processing, text analysis, and computational linguistics. It is used to determine an author's attitude, with respect to a particular topic or the overall contextual polarity in the text. This is a promising technology, which has resulted in remarkable interest among academics. The development of research in the field of text analytics has allowed researchers to formulate algorithms and techniques to discover sentiments from free text more efficiently than ever. This work covers all features from mining texts from social media particularly twitter, applying sentiment analysis based on people's opinions expressed on social media to finally assigning the polarity to them as positive, negative or neutral.

1.2 Problem Statement

The word cyberbullying did not even exist a decade ago, yet the problem has become a pervasive one today. Cyber bullies do not have to be physically strong or fast; they just require access to gadgets such as a cell phone or computer and a desire to terrorize. Anyone can be a cyberbully, and such persons usually have few worries about having face-to-face confrontation with their victims.

In this paper, we focus on the detection of textual cyberbullying with an empathic agent, which is one of the main forms of cyberbullying. We use a corpus of comments from twitter tweets involving sensitive topics related to race & culture, sexuality and intelligence i.e., topics involving aspects that people cannot change about themselves and hence become both personal and sensitive. We pre-process the data, subjecting it to standard operations of removal of stop words and stemming, before annotating it to assigning respective labels to each comment.

According to the survey by Digital Trends, more youths experienced cyberbullying on Instagram than any other platform at 42%, with Facebook following close behind at 37%. Snapchat ranked third at 31%. Seventy-one percent of the survey participants said that social media platforms do not do enough to prevent cyberbullying.

The survey also considered the other side of the story, asking the same age group how often they were the bullies, instead of being on the receiving end. Nearly 70% of those surveyed said they were abusive online toward another user, compared to just 12% that admitted to bullying in general. Despite the prevalence of youth initiating the bullying, more than 60% disagreed with the idea that "saying something nasty" is less hurtful online than in person.

2.2 Literature review

So many other researchers have worked mostly with finding out crime pattern from social media or Internet blog using sentiment analysis. Very few researches have been conducted on the context of textual cyberbullying. The approaches they took differ from mine in various ways; I will summarize briefly their research and results.

To curb cyberbullying Kenyan government recently enacted a bill the Computer and Cybercrimes Bill (2016) whereby harassing and stalking someone on Facebook or Twitter can now earn you a 10-year prison sentence or a Sh20 million fine or both. This follows Cabinet's approval of the Computer and Cybercrimes Bill (2016) that spells out stiff penalties for digital crimes including illegal breach of systems and networks, cyber-bullying and stalking among others. "A person who, individually or with other persons, willfully and repeatedly communicates, either directly or indirectly, with another person or anyone known to that person, commits an offence, if they know or ought to know that their conduct is likely to cause those persons apprehension or fear of violence to them or damage or loss on that persons' property detrimentally affects that person," reads Section 14 of the Bill in part.
The Computer and Cybercrime Bill 2016 is part of a raft of legislation mooted by the Government by various agencies in the last three years to combat rising cases of cybercrime.

Pang and Lee have labeled sentences in the document as subjective or objective (B. Pang and L. Lee, A sentimental education, 2004). They have applied machine learning classifier to the subjective group, which prevents polarity classification from considering useless, and misleading data. They have explored extraction of methods on the basis of minimum cut formulation.

Wang and Wang have proposed a variance mean based feature-filtering method that reduces the feature for representational phrase of text classification. The final performance of the method was observed to be better as it only considered the best feature and also the computation time got decreased as incoming text classified automatically.

The Research on opinion mining on YouTube performed for discussing how social media can be utilized to radicalize a person (Etzioni, O., Cafarella, M., "Unsupervised named-entity extraction 94 from the web: An experimental study). The research idea, which illustrates in Crawling, a global social networking platform, such as YouTube, has the potential to unearth content and interaction aimed at radicalization of those with little or no apparent prior interest in violent Jihadism. Their work examines an approach is indeed fruitful. They got together a large dataset from a collection within YouTube that was recognized as potentially having a radicalizing agenda. The data is analyzed using social network analysis and sentiment analysis tools. It also examines the topics discussed and what the sentiment polarity (positive or negative) is towards these topics. Particularly, they focused on gender differences in this group of users, suggesting most extreme and less tolerant views among female users. With (automatically) labeled data collected from the online websites, the researchers approached the related task of detecting a sentiment polarity in reviews via supervised learning approaches. Interestingly, our baseline experiments on this task show that humans may not always have the best intuition for choosing discriminating words. While they did experiment with a set of different features in the previous research, their essential focus was not on feature engineering.

The research in sentiment analysis trend is not limited yet. In order to improve accuracy and performance of the proposed techniques, applications, or algorithms. It enables them to more compatible with understanding meaning and features. But still there are some problems and challenges in text analysis of reviews/documents and evaluate sentiment scores.

2.0 Background of sentiment Analysis

2.1 What is Sentiment Analysis

Sentiment analysis: is also called opinion mining which is a computational study of reviews, sentiments, opinions, evaluations, attitudes, subjective, views, emotions, etc., expressed in a textual form.

People who speak a language can easily read through a paragraph and quickly identify whether the writer had an overall positive or negative impression of the topic at hand. However, for a computer, which has no concept of natural spoken language, this problem must be reduced to mathematics. Without any context of what words actually mean, it cannot simply deduce whether a piece of text conveys joy, anger, frustration, or otherwise. Sentiment analysis seeks to solve this problem by using natural language processing to recognize keywords within a document and thus classify the emotional status of the piece.

In this research I used Sentiment analysis to classify text as positive, neutral, or negative.

Natural Language Processing (NLP): branch of data science that consists of systematic processes for analyzing, understanding, and deriving information from the text data in a smart and efficient manner. By utilizing NLP and its components, one can organize the massive chunks of text data, perform numerous automated tasks and solve a wide range of problems such as – automatic summarization, machine translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation etc.

2.2 How Does Sentiment Analysis work?

There is an approach to use sentiment analysis is with constructing a lexicon with information about which words and phrases are positive and which are negative. For example, SentiWordNet is an overtly obtainable lexical resource in which each WordNet. Synset is ascribed three numerical scores describing how objective, positive, and negative the terms in the synset. This lexicon can either compile

manually or be acquired automatically. The annotation of lexical or corpora is usually done by hand, and classifiers are then trained with large sets of features to classify a new batch of words or phrases. There are other approaches to analyze sentiments focus on the mining of sentences or entire documents, rather than to depend on the parity of words. This approach usually works with corpora of text documents. The essential problem with document classification (polarity classification) which is that it has to determine the overall sentiment characteristics of an entire document, while the expressed sentiment can be included in just one sentence or word. In other cases, the sentiment can be expressed implicitly, which makes it even more difficult to detect and classify. However, the context surrounding these 'hidden' sentiments can provide very beneficial information for classifying it. Based on this division of the field of sentiment analysis, we often speak of word-level, sentence-level and document-level sentiment classification.

By a synopsis of Sentiment analysis defection (also called as opinion mining) that refers to the use of natural language processing (NLP), text analysis (TA) and computational linguistics (CL) to identify and extract subjective information in source materials. Sentiment analysis is widely utilized for online reviews and social media for a variety of applications, ranging from marketing to customer service.

2.3 Why sentiment analysis is Important?

There are millions online users, who write and read online and Internet usage around the world. Online daily sentiments become the most significant issue in making a decision. According to a new survey conducted by Dimensional Research, the survey discusses the percentage of trust online customer reviews as much as personal recommendations. According to 2011 Study: 74% of customer's confidence is based on online personal recommendation reviews, 60% in 2012 study, and 57% in 2013 Study. But this percentage increases with respect to 2014 Study: 94% of customer's trust on online sentiment reviews

3.0. System prototype Design

3.1 System Design

Figure 3.1 shows the basic architectural diagram of the implemented system. Basically it consists three modules, they are:

- Twitter application and twitter database
- Sentiment analysis process
- Web application

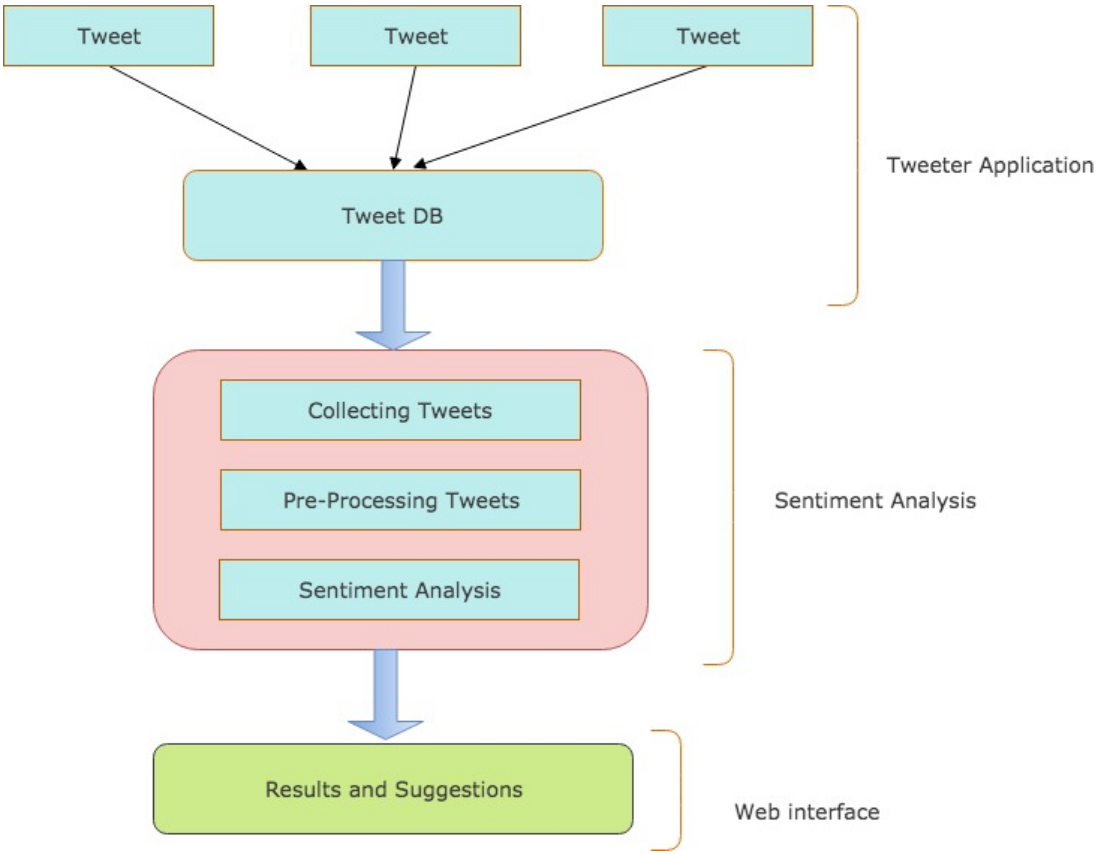

Figure 3.1 architectural designs

Figure 3.2 shows the dataflow diagram of the implemented system. At first, a twitter application is created and tweets are collected from the twitter database. Collected tweets are stored as data set and is pre-processed and parsed by removing common unwanted words, symbols, characters, numbers and converts the upper case letters to lower case letters. After pre-processing, the sentiments will be analyzed by using Natural language processing tool. Each sentence is provided with sentiment value, based on this sentiment value the data is cataloged as positive or negative. We will also be able to detect cyber bullying on the text that has been classified as negative by Natural processing Language. Both positive and negative data are analyzed and similar data are identified. Then by using a web application,

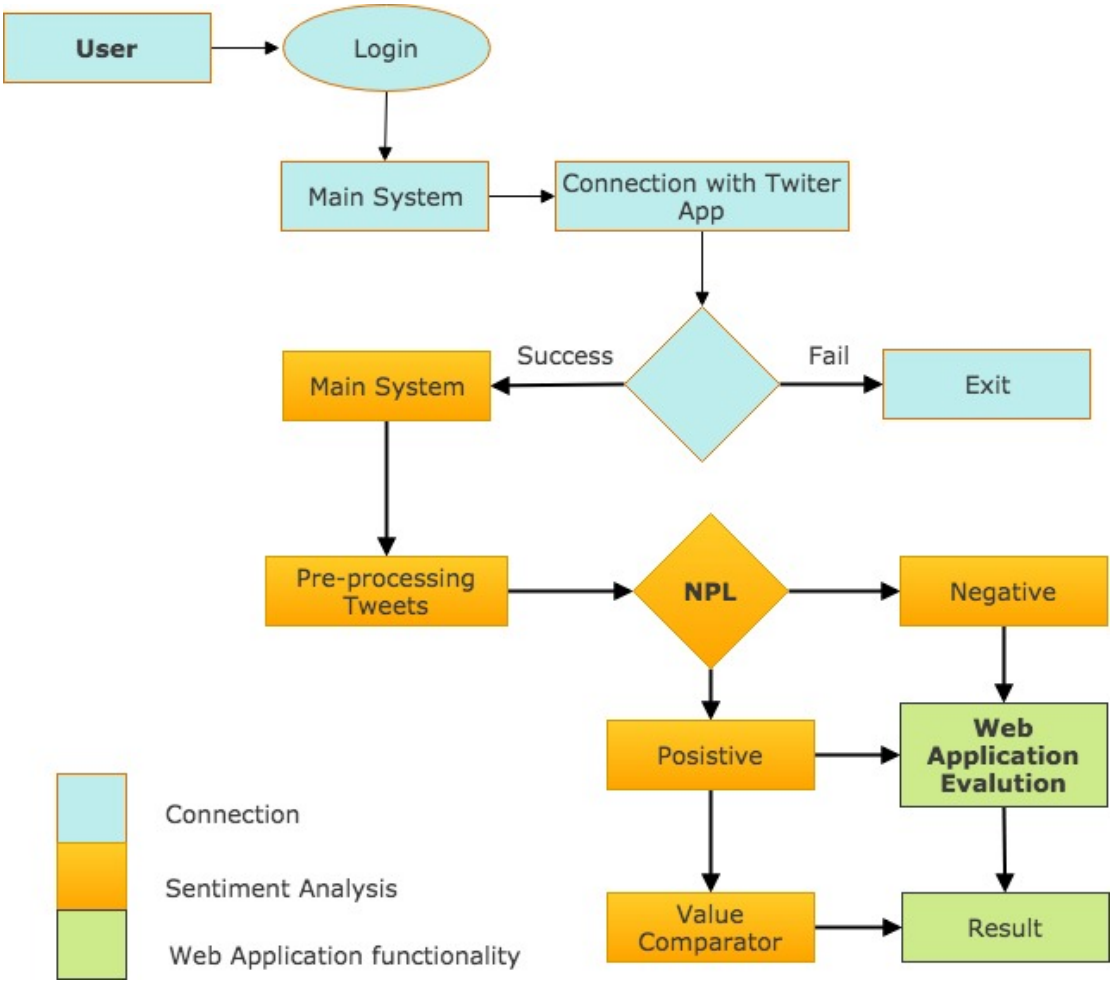

Figure 3.2 Data Flow Diagram

3.2 Data Collection:

This is happens on the connection part. I used twitter to collect relevant data since it is

one of major social networking website. To get access and gather I developed an API with python that integrates with twitter to pool all the tweets. You first needed to login then the API establishes a connection with the twitter app to gather all the data required then move it to the database within the main system.

3.3 Pre- Processing:

Initially for pre-processing I corrected the spelling mistakes in the tweets as many people tend to right English words in short form i.e., “u r nt pretty” is converted to “you are not pretty”. Then we converted the uppercase letters of the tweets to lowercase order. Then we removed all the usernames, URLs and unnecessary white spaces from the tweets. Stop words are words that are generally considered useless. Most search engines ignore these words because they are so common that including them would greatly increase the size of the index without improving precision or recall. Any group of words can be chosen as the stop words for a given purpose.
For some search engines, these are some of the most common, short function words, such as the, is, at, which, and on. In this case, stop words can cause problems when searching for phrases that include them, particularly in names such as " The Who ", " The The ", or " Take That ".

3.4 Natural Language Processing

Using natural language sentiment analysis requires a set of tweets labeled positive, negative or objective, it can be created by hand, by labeling tweets manually. I opted for a more automatic approach by collecting hundreds of thousands of tweets and running an algorithm through those tweets, which compares each individual word with positive list and negative list of words.

3.5 Evaluation

The tweets that are positive of cyber bullying are flagged and passed onto value comparator for further screening. While the negative ones will be displayed on the web interface as a normal tweet.

4.0 Methodology

In order to detect cyberbullying I Implemented Machine Learning techniques to find out which was the most efficient and accurate one. The training dataset I used for the experiment was obtained from Sentiment140 dataset (A., Bhayani, R. and Huang, L., 2009) with 1.6 million tweets. I only selected a random sample of 1,000 tweets labeled positive and another set of 1,000 tweets labeled negative. The sample size selected was sufficient to carry out this experiment and draw a conclusion.

Sentiment classification has two approaches where the tweets are converted into binary for easy detection of whether they are either positive or negative. Binary classification is important when you want to compare two products. In this study implementation was done with respect to binary classification.

The dataset on sentiment140 tweets has been studied and the repository is stored in unstructured textual format. The data needs to be converted into meaningful order for the machine learning algorithms to be applied. The processing will involve the removal of unnecessary information, blank spaces etc. The processed data will be converted to numerical vectors, where each of the vectors corresponds to a review and entries of each vector represent the presence of feature in that particular review.

The conversion of textual data into vector is done using the following methodologies;

- Counter Vectorizer:
- Term Frequency – Inverse Document frequency (TF-IDF)

When you already have the labeled dataset we will used the Supervised machine-learning algorithm. The two different algorithms are; Native Bayes classifier and support vector machine classifier. The two classifiers were the best for analyzing the textual tweeter messages and detecting cyberbullying. They proved to be efficient and accurate in terms of handling massive tweets from different users.

4.1 Naive Bayes (NB) Classifier:

It is the simplest probability classifier; it computes the posterior probability of a class based on distribution of words. The classifier can be advantageous since it can handle

a small or large amount of training data to calculate the parameters for prediction. Instead of calculating the complete covariance matrix, only variance of the feature is computed because of independence of features.

For a given textual review 'd' and for a class 'c' (positive, negative), the conditional probability for each class given a review is P(c|d). According to Bayes theorem this quantity can be computed using the following equation:

$$P(c \mid d) = \frac{P(d \mid c) * P(c)}{P(d)}$$

To further compute the term P(d|c), it is decomposed by assuming that fi's are conditionally independent given d's class. This decomposition of P(d|c) is expressed in following equation:

$$P_{NB}(c \mid d) = \frac{P(c)\left(\prod_{i=1}^{m} P(f_i)|c\right)^{n_i(d)}}{P(d)}$$

4.2 Support Vector Machine (SVM) as a classifier:

SVM is a non-probabilistic binary linear classifier. The main principle of this method is to determine linear separators in the search space that can be used to separate different classes. In this study, SVM Model represents each review in vectorized form as a data point in the space. This method is used to analyze the complete vectorized data and the key idea behind the training of model is to find a hyperplane represented by $\vec{w}$. The set of textual data vectors are said to be optimally separated by hyperplane only when it is separated without error and the distance between closest points of each class and hyperplane is maximum. After training of the model, the testing reviews are mapped in-to same space and predicted to belong to a class based on which side of the hyperplane they fall on.

Let be the class (positive, negative) for a textual message dj, the equation for $\vec{w}$ is given by

$$\vec{w} = \sum_j \alpha_j c_j \vec{d}_j, \alpha_j \geq 0$$

Dual optimization problem gives the values for αj's. All the dj such that αj is greater than zero are termed as Support vectors as they are the only document vectors which are contributing to $\vec{w}$.

Confusion matrix is generated to tabulate the performance of any classifier. This matrix shows the relation between correctly and wrongly predicted reviews. In the confusion matrix, TP (True Positive) represents the number of positive movie reviews that are correctly predicted whereas FP (False positive) gives the value for number of positive movie reviews that are predicted as negative by the classifier. Similarly, TN (True Negative) is number of negative reviews correctly predicted and FN (False Negative) is number of negative reviews predicted as positive by the classifier.

Correct Labels		
	Positives	**Negatives**
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

Table 4.1: Confusion Matrix

From this confusion matrix, different Performance evaluation parameter like precision, recall, F-measure and accuracy are calculated. The table of confusion matrix formation is shown in table 4.1. Precision: It gives the exactness of the classifier. It is the ratio of number of correctly predicted positive reviews to the total number of reviews predicted as positive.

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

Recall: It measures the completeness of the classifier. It is the ratio of number of correctly predicted positive reviews to the actual number of positive reviews present in the corpus.

$$Recall = \frac{TP}{TP+FN}$$

F-measure: It is the harmonic mean of precision and recall. F-measure can have best value as 1 and worst value as 0. The formula for calculating F-measure is presented as:

$$F - Measure = \frac{2*Precision*Recall}{Precision+Recall}$$

Accuracy: It is one of the most common performance evaluation parameter and it is calculated as the ratio of number of correctly predicted reviews to the number of total number of reviews present in the corpus. The formula for calculating accuracy is given as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The dataset, which was considered in this study, is the Sentiment140 dataset (A., Bhayani, R. and Huang, L., 2009) with 1.6 million tweets extracted using the twitter API. The tweets have been annotated (0 = negative, 2 = neutral, 4 = positive) and they can be used to detect sentiment. It contains the following 6 fields:

- **target**: the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
- **ids**: The id of the tweet
- **date**: the date of the tweet
- **flag**: The query (lyx). If there is no query, then this value is NO_QUERY.
- **user**: the user that tweeted
- **text**: the text of the tweet

I only selected a sample of 1000 positively, labeled and 1000 negative labeled tweets (A., Bhayani, R. and Huang, L., 2009).

5.0 Proposed Approach

Labeled Twitter sentiment classification has been taken in the consideration, which consist of 1000 positive and 1000 negative reviews (A., Bhayani, R. and Huang, L., 2009). The tweets will go preprocessing step, where all the vague information is removed. From the cleaned dataset, potential features are extracted. The vectorization techniques are used to convert textual data to numerical format. Using vectorization, a matrix is created where each column represents a feature and each row represents an individual review. This matrix is used as input to classification algorithm and cross validation technique is applied to choose the training and testing set for each fold. Step-wise presentation of proposed approach is shown in the figure 5.1.

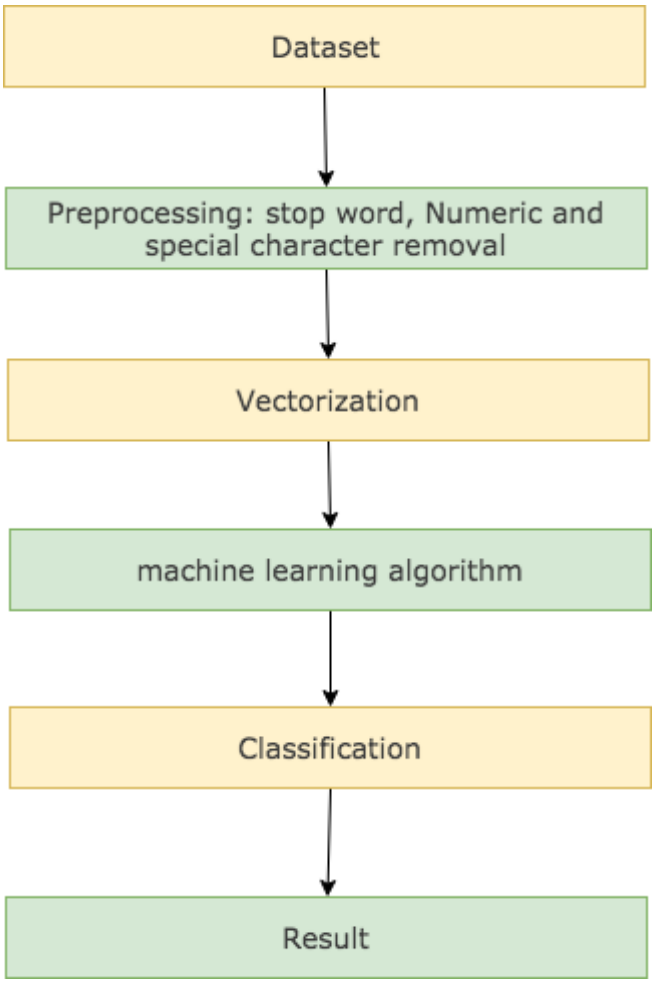

Figure 5.1: Process flow of proposed approach

5.1 Steps followed for classification

Step 1. The Twitter sentiment classification dataset is considered for analysis, which consists of 1000 positive, and 1000 negative labeled reviews. For each review a separate text file is maintained.

Step 2. For the preprocessing I corrected the spelling errors in the tweets and eliminated vague information. English words written in shorthand such as "u r nt beautiful" is converted to "you are not beautiful". The uppercase letter tweets are converted to letters lowercase. It is observed that reviewers often repeat a particular character of a word to give more emphasis to an expression or to make the review trendy. Words like yesssssssss, heheehhehehehe falls in this category. The repetition of characters is also eliminated in this step. Most of the words that do not contribute to any sentiment used in English language are termed as stop words. So, second step in preprocessing involves the removal of all the stop words of English language.

Step 3. In this step the features are tokenized word of a review. These words need to be converted to numerical vectors so that each review can be represented in the form of numerical data. The vectorization of features is done using the following two methods.

- **CountVectorizer**: It transforms the review to token count matrix. First, it tokenizes the review and according to number of occurrence of each token, a sparse matrix is created.

- **TF-IDF**: Its value represents the importance of a word to a document in a corpus. TF-IDF value is proportional to the frequency of a word in a document.
    - Calculation of TF-IDF value: suppose sentiment140 dataset contain 100 words wherein the word Awesome appears 5 times. The term frequency (i.e., TF) for Awesome then (5 / 100) = 0.05. Again, suppose there are 1 million reviews in the corpus and the word Awesome appears 1000 times in whole corpus Then, the inverse

document frequency (i.e., IDF) is calculated as log (1,000,000 / 1,000) = 3. Thus, the TF-IDF value is calculated as: 0.05 * 3 = 0.15.

Step 4. The numeric vectors can be given as input to the classification algorithm. The different classification algorithm used is as follows:

- Naive Bayes (NB) algorithm: Using probabilistic analysis, features are extracted from numeric vectors.
  These features help in training of the Naive Bayes classifier model.

- Support vector machine (SVM) algorithm: SVM plots all the numeric vectors in space and defines decision boundaries by hyperplanes. This hyperplane separates the vectors in two categories such that, the distance from each category to the hyperplane is maximum.

  Initially, the dataset was not divided between testing and training subsets. So, k-fold cross validation technique is used, the number of folds used are 10.

Step 5. After training of model, confusion matrix is generated which shows the number of positive and negative reviews that are correctly predicted and number of positive and negative reviews that are wrongly predicted. For each fold, prediction accuracy is calculated based on this confusion matrix and final accuracy is given by taking the mean of all the individual accuracies of 10 folds. However, individual accuracy of a particular fold can be much higher than the mean of all accuracies.

Step 6. For each model, values of precision recall and F-measure as performance evaluation parameters are found out. The confusion matrix and a table containing performance evaluation parameter are generated. Finally, these obtained results are compared with the values obtained by other authors in literature.

6.0. Implementation

The implementation of above mentioned algorithms are carried out on Sentiment140 dataset (A., Bhayani, R. and Huang, L., 2009). After implementing the two machine learning approaches to determine which of the classifiers used for detecting cyberbullying gave the most accurate results and is preferable over the other, we determined the accuracy of the classifier, precision, recall and f-score of the positive, negative and neutral tweets.

6.1 Naive Bayes Algorithm:

The confusion matrix obtained after implementation of Naive Bayes classification algorithm is shown in table 6.1.

Table 6.1: confusion matrix for Naïve Bayes classifier

Correct Labels		
	Positives	**Negatives**
Positive	11107	1393
Negative	2384	9666

The performance evaluation parameters obtained for Naive Bayes classifier is shown in table 6.2.

Table 6.2: Evaluation parameters for Naïve Bayes classifier

	Precision	**Recall**	**F-Measure**
Positive	0.80	0.89	0.84
Negative	0.87	0.77	0.82

Maximum accuracy achieved after the cross validation analysis of Naive Bayes classifier is **0.8953**.

6.2 Support Vector Machine Algorithm:

The confusion matrix obtained after implementation of Support Vector Machine algorithm is shown in table 6.3.

Table 6.3: Confusion matrix for Support Vector classifier

Correct Labels		
	Positives	**Negatives**
Positive	11102	1393
Negative	1688	10812

The performance evaluation parameters obtained for Support Vector Machine classifier is shown in table 6.4.

Table 6.4: Evaluation parameters for Support Vector classifier

	Precision	**Recall**	**F-Measure**
Positive	0.87	0.89	0.88
Negative	0.89	0.86	0.88

Maximum accuracy achieved after the cross validation analysis of Support Vector Machine classifier is **0.9406**.

7.0 Comparison Analysis

In this section we will compare the output obtained from the two proposed machine-learning approaches with the output from other manuscripts. The manuscripts that were taken into consideration are by Pang Lee and another by read. The two manuscripts used the same polarity dataset with 1000 positives and 1000 negatives reviews.

Table 7.0: Comparison of proposed work with existing literatures

	(Pang and Lee)	**(Read)**	**Proposed approach**
Naïve Bayes	0.864	0.789	0.895
Negative	0.8615	0.815	0.940

From the table 7.0 it is clear that the accuracy obtained in the proposed approach is more accurate than those from both manuscripts. All the other research methods had a high precision but a low recall meaning that most of the predicted labels are correct, the precision and recall are normally used to determine the quality of classifier output.

8.0 Conclusions

In this thesis we have investigated the possibilities of building a system capable of automatically identifying cyberbullying on social media using an empathic agent. This chapter summarizes what we have learned, discusses limitations of the achievements and concludes with a brief look at future work.

The goal with the project was to generate knowledge of how an automatic system for detecting bullying on social media could be constructed. During the project we have learned how to employ state of the art methods in Natural Processing Language (NLP) for bullying classification. Which is the core in any bullying detection system. We have learned the limitations of this method in the form of differentiating between common bad language and cyberbullying. We have learned about some of the challenges that we are facing when trying to automatically scan a social media platform. Most importantly we have learned that an automatic system for bullying detection is possible to some extent as shown by the implemented prototype. The greatest limiting factor is how well classification can be performed. With the knowledge gained during this project it is the author's opinion that the step to a real world application is not very far.

8.1 Future Scope

As of now this system is implemented only on textual data. In the future we plan on extending the scope of our system by incorporating cross-media detection in the form of audio, video and images too. We also plan to try to make our system be context-aware with the help of deep learning in the future.

9.0 References

1. Ronan Collobert et al. (2011). Natural Language Processing (Almost) from Scratch". [Online]. Available: http://jmlr.org/papers/volume12/collobert11a/collobert11a.pdf

2. Schukla, A. (2011). Sentiment analysis of document based on annotation, CORR Journal, Vol. abs/1111.1648.

3. Love Engman. (2016). Automatic Detection of Cyberbullying on Social Media. [online] available on http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1052691&dswid=991

4. Fabian Fagerholm, Scientific writing , (2016) online . Available on https://www.cs.helsinki.fi/group/ese/ScientificWritingGuide.pdf

5. soko directory, (February 15, 2018) , online https://sokodirectory.com/2018/02/number-kenyans-social-media-platforms-increases-bake/

6. nderi, s. (2017, September 27). Cyber bullying as a form of violence against women & how to tackle it. Retrieved from https://hapakenya.com/2017/09/27/cyber-bullying-as-a-form-of-violence-against-woman-and-how-to-tackle-it/

7. Kenya Law Reports Kenya gazette supplement, 2017, Kenya cyber crime act http://kenyalaw.org/kl/fileadmin/pdfdownloads/bills/2017/ComputerandCybercrimesBill_2017.pdf

8. GitBook, Sentiment Analysis, online: https://lizrush.gitbooks.io/algorithms-for-webdevs-ebook/content/chapters/sentiment-analysis.html

9. Sentiment analysis algorithms and applications: A survey by Walaa Medhata, Ahmed Hassan, Hoda Korashyb https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0015

10. UML diagrams, online: https://www.draw.io/

11. 2004 Kothari_ Research Methodology~Methods and Techniques (2)

12. Mifta Sintaha, Shahed Bin Sattter, Niamat Zawad, Ahanaf Hassan "Cyber bullying detection using sentiment analysis in social media", 2016. [Online] http://dspace.bracu.ac.bd/xmlui/bitstream/handle/10361/6420/13101123%2C%2013101258%2C%2013101283%2C%2013101290%20%26%2013101002_CSE.pdf?sequence=1&isAllowed=y

13. Doaa Mohey El-Din Mohamed Hussein, "Analyzing Scientific Papers Based on Sentiment Analysis", 2016 [Online], https://www.researchgate.net/profile/Doaa_Mohey_El-Din/publication/301649777_Analyzing_Scientific_Papers_Based_on_Sentiment_Analysis_First_Draft/links/571fb75208aeaced788ac760/Analyzing-Scientific-Papers-Based-on-Sentiment-Analysis-First-Draft.pdf

14. Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. Detection of harassment on web 2.0. Proceedings of the Content Analysis in the WEB, 2:1-7, 2009.

15. Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. In Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on, volume 2, pages 241-244. IEEE, 2011.

16. A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(2009), p.12. Online. Available: https://www.kaggle.com/kazanova/sentiment140

17. Van Hee, C. Automatic Detection and Prevention of Cyberbullying. Retrieved from https://www.clips.uantwerpen.be/sites/default/files/automatic_detection_and_prevention_of_cyberbullying.pdf

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

19. J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in Proceedings of the ACL Student Research Workshop. Association for Computational Linguistics, 2005.

20. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S., Weld, D. and Yates, A., "Unsupervised named-entity extraction 94 from the web: An experimental study", Artificial Intelligence, 165(1): 91–134, 2005.

21. Y. Wang and X.-J. Wang, "A new approach to feature selection in text classification," in Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on, vol. 6. IEEE, 2005, pp. 3814–3819.