ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

A Study on Geographical Prediction of Cruelty Attacks Using Data from Forums

Kiran Kalshetty, Khadija Sitabkhan, Akshay Yelpale, Dhananjay Chavan

B.E. Student, Department of Computer Engineering, PES Modern college of Engineering, Shivajinagar, Pune, India

ABSTRACT: In today's day, the Internet provides infinite amount of data. A lot of this data is useless. But the useful data can be mined and used for further analysis. We have tried to use the data provided from the twitter forum to predict the occurrence of any mishap. We are using sentiment analysis to find out the peoples' reaction to certain events. Based on their reaction is many people with different views are present in close locations there is a possibility of a crime. The web provides volumes of text-based data which are stored in online chatting websites like Twitter, Facebook, Blog and Forum etc. Cyberbullying is a socially aggressive and has powerful negative effects for individuals, specifically adolescents and youngsters. In the recent times many methods for automatic thoughts of mining in the online data are becoming increasingly important, to increase the safety parameter of the people. This framework is proposed to extract Cyberbully polarity from the Forum using Fuzzy logic technique. At first, the given input is pre-processed and the useful content is gathered. Subsequently, the pre-processed data will be sent to the features extraction method. Probabilities of the words are calculated by using Fuzzy Decision Tree Method. Fuzzy rules can be applied in all these features to extract the certain set of cyberbully words like bad words, insulting words, threatening words and terrorism words from the given input, hence we use text mining here. Finally this method will return the reduced and accurate cyberbully words. This method is performed by human annotation using the existing methods like Mamdani Fuzzy System and Naïve Bayes classifier. Extensive experiments are performed by using fuzzy logic on crime debate forum and the results show that this proposed approach is better than the traditional one.

KEYWORDS: Cyberbully, text mining, Forum, fuzzy-logic, fuzzy decision tree, Naive Bayes classifier, tweets, feature extraction, R language, Geographical prediction. Twitter, Python, Sentiment analysis.

I. INTRODUCTION

The industrial revolution changed the face of the world. It not only became easier but also more technology dependent. People became more expressive as they received more exposure. Social media became a very useful platform to express ourselves. The expressions have adverse reactions as well. We intend to take data from these platforms and make use of it to improve on the safety parameter. For the development of the system we take the data available on Twitter and filter all the useful contents. This content will depict the mood and will also give the location information in response to a particular event (death of a famous personality or a new motion passed by the Govt.). Using sentiment analysis we classify the good and the harsh words. Then we locate them on the map (of India for now) in 2 different colours showing the happy and the sad. Then calculating their frequency we will alert the user if that area is in danger. It will also save previous history for the purpose of analysis.

II. RELATED WORK

J.I.Sheeba introduced detection of online social cruelty attack from forums using fuzzy logic technique.[1] MaralDadvar proposed Support Machine model to train a gender-specific text classifier and it will detect the written language used by a harasser varies with the author's features including gender [2].KarthikDinakar proposed a new method to detect the textual Cyberbullying which can be tackled by building individual topic- sensitive classifiers .In this methods, binary classifiers for individual labels outperform multiclass classifiers[3].Kelly Reynolds introduced a new method for detecting language patterns used detecting language patterns used by bullies and their victims .To develop the rules for automatically detect cyberbullying content using machine learning algorithm [4].Saberi Nasr introduced a newly devised Mamdani Fuzzy inference system to assess groundwater quality in Yazd province

DOI: 10.15680/IJIRCCE.2015. 0310118

[5].Jennifer Bayzick proposed a new method to detect the presence of cyberbullying in the online chat conversations by using computer software [6].Michael J. Moore proposed a new method for identifying aggressiveness Forum posts including both attacks and defends .Anonymity of forum posts was identified as a criterion to distinguish attackers from defenders [7].Marine Cambodia discussed about the bullies, victims and how cyberbullying were involved in the process of social information [8].Julian J. Dooley discussed about the similarity , differences between cyberbullying and face-to-face bullying [9].Michael Hay and Gerome Miklau introduced a novel anonymization technique based on perturbing the network and demonstrate empirically that it leads to substantial reduction of the privacy threat[10].Sameer Hinduja and Justin W. Patchin discussed about bullying, cyberbullying online and suicide concepts and they suggested suicide prevention and intervention component was essential within comprehensive bullying response programs implemented in schools[11].Robin M. Kowalski proposed a new method about cyberbullying ,electronic bullying , online social cruelty and this phenomenon includes bullying through chat room, web site,e-mail,digital messages and images sent to a cell phone[12].Qing Li investigated the nature and the extent of adolescences experience of cyberbullying [13].QiangShen and TossapanBoongoen proposed a new approach based on order-of-magnitude reasoning with which the theory of fuzzy set was blended to provide quantitative semantics of descriptors and their unambiguous mathematical manipulation[14].Selva Kumar introduced a new technique mixed C-means clustering. This method was mainly used to test against a brain tumour gene expression [15].

III. PROPOSED TECHNOLOGY

1. **Data collection:** The first step in building this project was to gather data to compare the users. We had to analyse the moods of the people that are depicted through their tweets, whether angry words, foul language or jargons. Whether people responded positively or negatively to the occurrence of a particular event.
   Twitter's API provides a straightforward way to query for users and returns results in a JSON format which makes it easy to parse in a Python script, this makes our data reliable.

2. **Pre-processing:** Today, more than 80% of the data is unstructured – it is either present in data silos or scattered around the digital archives. In order to produce any meaningful actionable insight from data, it was important to know how to work with it in its unstructured form. One of the first steps in working with text data is to pre-process it. It is an essential step before the data is ready for analysis.

   Cleaning & structuring: One approach is to directly remove useless html characters by the use of specific regular expressions. Another approach is to use appropriate packages and modules (for example htmlparser of Python), which can convert these entities to standard html tags. It is necessary to keep the complete data in standard encoding format. UTF-8 encoding is widely accepted and is recommended to use.
   1. **Removal of Stop-words:** When data analysis needs to be data driven at the word level, the commonly occurring words (stop-words) should be removed. One can either create a long list of stop-words or one can use predefined language specific libraries**.**
   2. **Removal of Punctuations**: All the punctuation marks according to the priorities should be dealt with. For example: “.”, “,”,”?” are important punctuations that should be retained while others need to be removed.
   3. **Removal of Expressions**: Textual data (usually speech transcripts) may contain human expressions like [laughing], [Crying], [Audience paused]. These expressions are usually non relevant to content of the speech and hence need to be removed. Simple regular expression can be useful in this case.

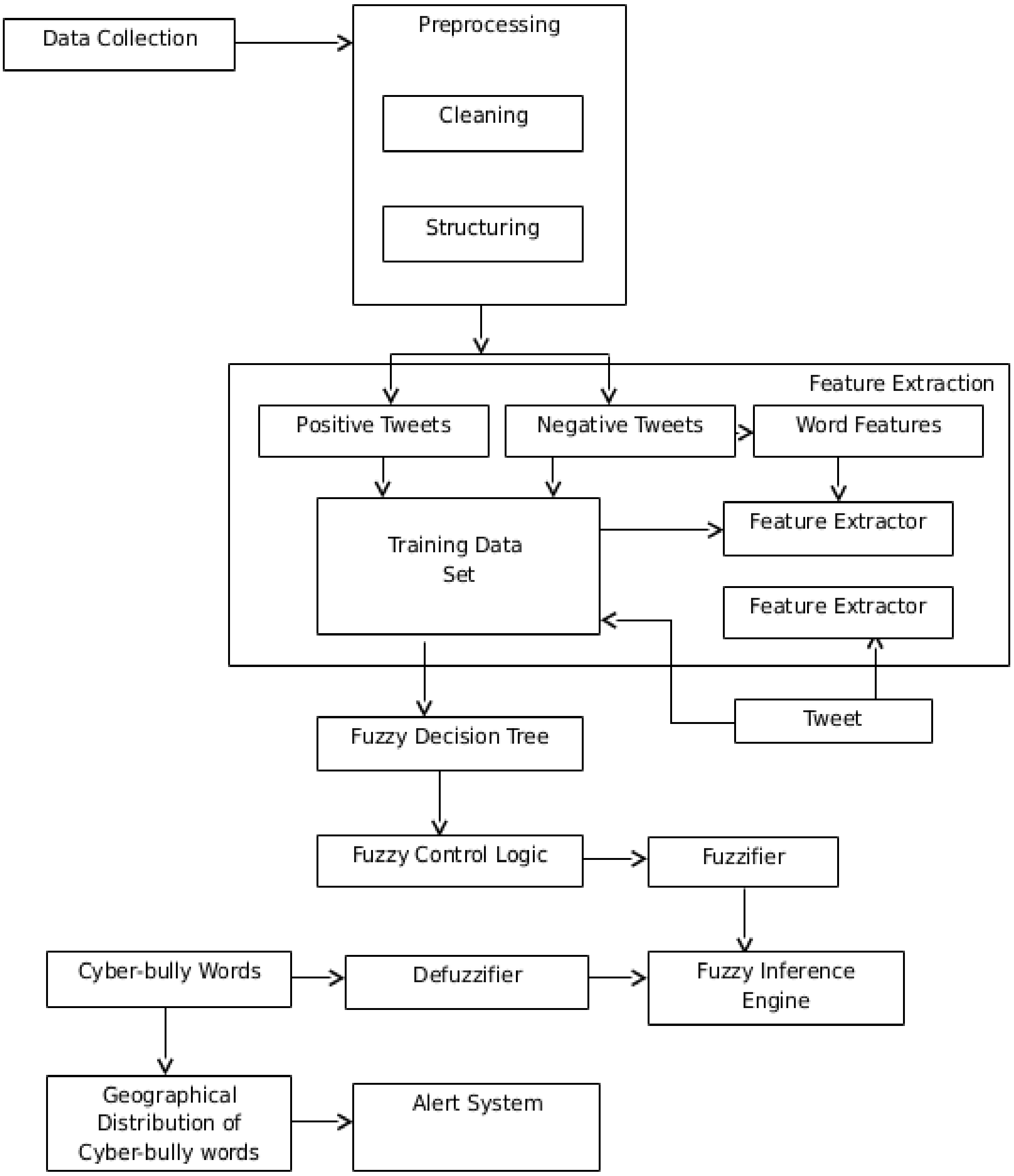

Fig 1: System architecture diagram

4. **Split Attached Words**: We humans in the social forums generate text data, which is completely informal in nature. Most of the tweets are accompanied with multiple attached words like RainyDay, PlayingInTheCold etc. These entities can be split into their normal forms using simple rules and regex.

5. **Slangs lookup**: Social media comprises of a majority of slang words. These words should be transformed into standard words to make free text. The words like luv will be converted to love,

Heloto Hello. These words are extremely useful for our system and hence they will be collected for further analysis.

3. **Feature Extraction:** The feature extraction techniques are used to obtain the important words in the text. After pre-processing, each word from the document was represented by a vector of features. Using the training data set we are designing our system.

4. **Fuzzy decision tree:** Fuzzy logic for solving the introduced uncertainties were called Fuzzy Decision Trees.(FDTs)In this a method, selecting a word split at every new node of the tree and a rule for determining when a node should be considered terminal. The of frequency calculation words were given input to this method. It will find a probability of all the words ,based on probability values the words were assigned into three ranges like Low,Medium,High.Finally results were obtained and it was concluded whether the particular forums contains cyberbully words or not and assigning ranges were categorized as Low or Medium or High.[1]

5. **Fuzzy control logic:** Fuzzy control logic was forming a knowledge-based consisting of IF-THEN fuzzy rules. These rules were obtained from human experts based on their respective domain of knowledge and observations. Fuzzy control logic contains 3 steps like Evaluation, Fuzzification and Defuzzification.Fuzzification transforms the real-valued input variables into a fuzzy set and using membership functions graphically to describe a situation. Rule Evaluation would evaluate the fuzzy rules.Defuzzification transforms back these fuzzy sets into real-valued variable outputs. The output of Feature extraction and Fuzzy decision tree system were given input to this method and assign fuzzy rules. Based on the fuzzy rules, the output of cyberbully words were categorized as Low or Medium or High [1].

6. **Fuzzifier:** Fuzzifer can be defined as crisp inputs which were translated into linguistic values by using a membership function. In this system, three word features were extracted where each word was associated with the vector of three features. These three features were operated as input to the fuzzifier separately and to manage the vector set, each feature is categorized into three sets like Low, Medium and High [1].

7. **Fuzzy Inference Engine:**After fuzzification, the inference engine was activated .Fuzzy rules were implemented to this process in order to assign linguistic values to all the three features. The inference engine was linked with all the category values and they were later converted into a single group that was eventually categorized as Low, Medium and High. The vital role of the Fuzzy inference engine was to assign the fuzzy rules. The important sentences were extracted from these rules according to the criteria of the features. In fuzzy rules, the set of fuzzy propositions associated with the if-condition rule was known as the premise or the antecedent. In the rule, if x is Low and y is High then z is Small .The premise consists of the the two fuzzy propositions x is Low and y id High connected by the and (&&)operator [1].

8. **Defuzzifier:** The output of the linguistic variables from the inference engine will be converted into the final crisp values by the defuzzifier using membership function for representing the final sentence score. In this step, Defuzzification utilizes the output membership function which can be classified as output (Low, Medium, and High) that converts the fuzzy results from the inference engine into a crisp output to derive the final evaluation of each sentence. The suitable words were represented as High and it must be considered as the principal cyberbully words [1].

9. **Cyberbully words:** Slang term used to describe online harassment, which can be in the form of flames, comments made in chat rooms. We classify the words we get, as positive and negative then study their frequency. We discard the neutral words.

10. **Geographical distribution of cyberbully words:** The maps library for R language is a powerful tool for creating maps of countries and regions of the world. The coordinate system of the graph is latitude and longitude, so it's easy to overlay other spatial data on this map. We can plot all the tweets on the map and study their density.

11. **Alert system:** The alert system includes all the tweets on the graph of India. The positive responses will be depicted in blue & negative in red. Based on the density of their occurrence we will warn the user if there is a chance of a tragedy. We can also depict statistical analysis with the help of bar graphs, etc.

IV. CONCLUSION AND FUTURE WORKS

While studying about the crime scenes in the city we found out that these crimes can be divided on the basis of their cause. We found a really interesting and growing trend in those crimes caused by social media. Since social media can be inspected in a way better than any other possibility, then why not strive to reduce it. While doing more research we found that there is no new requirement of data to do this surveillance, the already available data is very efficient. We have considered only twitter, but during the commercial development of the system we can make use of more data from other forums and make the system even more efficient. We have used a very simple logic in this system which is if different minded people are together there is danger. Using this logic, this system is being developed for the first time. Our papers gave us details about the alternative ways in which this system can be implemented, they helped us to study the opportunities and obstacles of each method and select the best one. We hope that this project turns out to be of some help and it won't constrain itself to our university.

REFERENCES

[1] J.I.Sheeba, K.Vivekanandan "Detection of Online Social Attacks from Forums" International Journal Of data Mining and Emerging Technologies. Volume 4, Number 2. November, 2014

[2] Dadvar M, de Jong FMG ,Ordelman RJF and Trischnigg RB.Improved cyberbullying detection using gender information .DIR'2012.2012.

[3] DinakarK,REichart and Lieberman H. Modeling the detection of textual cyberbullying .Associated for all the advancement of artificial intelligence.2011;11-27

[4] Reynolds K,kontostathisA and Edwards L.Using machine learning to detect cyberbullying. In machine learning and application and workshops(ICMLA).2011 10TH international conference on,2011;Vol.2:241-244

[5] Nasr AS,Rezaei M and BarmakiMD.Analysis of groundwater quality using Mamdani Fuzzy Inference System in Yazd Province,Iran Intern JComput Appl. Analusis,2012;59 (7):45-53

[6] BayzickJ,Kontosthathis A and Edwards Detecting the presence of cyberbullying using computer software.WebSci'11,Koblenz,Germany;2011

[7] Moore MJ,NakanoT,Enomoto A and Anonymity and roles associated with aggressive posts in an online forum.Computers in Human Behavior,2012;(3);861-867

[8] CamodecaM,Goossens FA<Schuengel,C and Terwogt MM . Links between social information processing in middle childhood and involvement in bullying. Aggressive behavior 2003;29(2);116-127

[9] Dooley JJ,Pyalski J and Cross D.Cyberbullying versus face to face bullying .Zeitschrift fur Psychologies/Journal of Psychology 2009;217(4);182-188

[10] Hay M,Miklau G. Anonymizing social networks. Computer Science Department Faculty Publication Series 180,2007

[11] Hinduja S and PatchinJW.Bullying, cyberbullying and suicide.Archives of Suicide Research 2010; 14(3):206-221.

[12] Kowalski RM and Limber SP.Electronic bullying among middle school students .Journals of adolescent health 2007;41(6):S22-S30

[13] Li Q. New bottle but old wine:A research of cyberbullying in school.Computers in human behavior.2007;23(4):1777-1791

[14] Shen Q and BoongoenT.Fuzzy Order-of-Magnitude-Based Link Analysis for Qualitative Alias Detection.IEEE Transactions on knowledge and data engineering 2012;24(4):649-664

[15] Kumar S and InbaraniH.Analysis of mixed C-means clustering approach for brain tumor gene expression data.International Journal of Data Analysis Techniquies and Strategies,2013;5(2):214-228

BIOGRAPHY

Kiran Kalshetty is pursuing BE computer from PES Modern college of Engineering, Shivajinagar Pune(SavitribhaiPhule Pune University)

Khadija Sitabkhan is pursuing BE computer from PES Modern college of Engineering, Shivajinagar Pune(SavitribhaiPhule Pune University)

Dhananjay Chavan is pursuing BE computer from PES Modern college of Engineering, Shivajinagar Pune(SavitribhaiPhule Pune University)

Akshay Yelpale is pursuing BE computer from PES Modern college of Engineering, Shivajinagar Pune(SavitribhaiPhule Pune University)