BEHAVIORAL ANALYSIS OF INTERNET MESSAGING AND MALICIOUS ACTIVITY DETECTION

Authors Name: K.Srividya
Computer Science and Engineering
GMR Institute of Technology
Rajam, India
srividya.kotagiri@gmail.com

Authors Name: A.Mary Sowjanya Computer Science and Systems Engineering Andhra University
Visakapatnam, India
sowmaa@yahoo.com

1. Abstract

With the outburst in the growth of the social networking, it is alarming how the youth could become a victim of social trapping and psychological depressions. Nowadays internet messaging service plays a vital role in the social networking paradigm.Internet messaging which previously was meant for communication however turned out to have adverse effects on youth and country. Increased exposure and misuse of "chatting" has become a recent, common phenomenon. Internets messaging in social networking sites like Facebook, WhatsApp have also led to disastrous scenario where teenagers communicate through bawdy or explicit messages, and prowlers trap opposite genders.

In this paper the methodology for dealing with such type of messages is discussed. First, text messages are processed through the algorithm which would predict the behavior of both the communicating parties.Eventually it would alarm if there is a malicious activity detected. This process majorly uses analysis of content using LSA (Latent Semantic Analysis) and other original algorithms for determining the context of the chat.

Keywords-Social networking Analysis, Internet messaging, computer mediated communication, Keyword Matching, Latent Semantic analysis, Frequency Counting

2. Introduction:

Instant messaging (IM) has become a necessary evil in the form of communication in the present era thus instant Messaging Clients like AIM, MSN and more recently, GTalk, Facebook and WhatsApp have triggered a wave of instant Messaging based communication in particular and Computer Mediated Communication (CMC) in general. Since IM is the affective or emotional content of the information involved in a CMC ability to identify affective content and classify the nature of the affect has given rise lot of practical applications. Applications range from administration point of view and moderation of communication (Holzman and Pottenger, 2003),area of Affective User Interfaces (AUI) which include improved chat clients with real time feedback on the users' emotional state inferred from the conversation text. Another emerging applicationis automating facial expression of Avatars in online games, especially in Massively Multiplayer Online Role Playing Games (MMORPGs). But most of the work has involved emotion analysis of text from domains (Strapparava and Mihalcea, 2008; Wiebe and Cardie, 2005), rather than proper attention to similar analysis of instant messages. This could be because the text in instant messages is less structured than news headlines and blog posts, is less grammatical, has more unintended typographical errors, has special morphological conventions like vowel elongation and, also is loaded with a colloquialism of its own, resulting in a large number of out-of-vocabulary (OOV) words. ReferTagliamonte and Derek (2008) for a study of IM-speak from a linguistic point of view. As search it is difficult to identify emotion in instant messages as compared to more structured data from sources of formal text. For a detailed analysis of these difficulties refer[2] Schmidt and Stone.

3. Related work:

3.1 Social Networking

[3]A social networking service (also social networking site or SNS) is a platform to build social networks or social relations among people who share similar interests, activities, backgrounds or real-life connections. It consists of a representation of each user (often a profile), his or her social links, and a variety of additional services. Most social network services are web based, allowing users to create a public profile to interact over the internet with e-mail and instant messaging, create a list of users with whom to interact, view and share information, ideas, pictures, posts, events, activities and interests. They also incorporate communication tools like mobile connectivity, photo/video/sharing, blogging, online community services which could be individual-centred or group-centred.

The main types of social networking services are those that contain category places (such as former school year or classmates), means to connect with friends (usually with self-description pages), and a recommendation system linked to trust.

3.2 Semantic Analysis :

With the data already processed through the first phase of Keyword Matching. There is already an associated score with the particular message. The message now undergoes the second phase of processing in terms of deriving the meaning that the message intends to carry. The second phase uses an algorithm called as Latent Semantic Analysis, which is widely used for the semantic analysis of data all through the internet.

4. Motivation

Anoutburst of Social networking pages, without proper monitoring highly endangers personal security especially of teens and youth who are highly exposed to many inhumanely activities. They get addicted to internet Messaging over social networking sites presenting serious threat of getting beguiled into extreme and even horrendous acts like Trapping, girl-Trafficking, motivation to become extremists and other evils.

This paper proposes usage of behavioral analysis on messaged that are exchanged while people chat, followed by semantic analysis, scoring algorithm and a data mining technique to detect a malicious are an abusive activity so that the user can be intimated of what is the actual intent of the message. i.e., behavioral analysis is made on internet messages so as to detect any malicious activity.

Extensive work done to analyze the behavior of the data to predict the positive or negative feedback using classification, which eventually can be used to predict the quality and response of people towards that particular thing or product. But, rather negligible work towards addressing this kind of social media vulnerabilities.

5. Methodology

The entire process of behavioral analysis of internet/chat messages for malicious activity consists of the following phases:

Phase1:

Keyword Matching:

The most important and also tedious part of this process is the correction of all the set of abusive/flirty words. Then these words are separated into files categorized as flagged, non-flagged and highly-flagged based upon their general usage and the frequency with which they can be used to signify an abusive and explicit meaning. WordNet is used to retrieve all the synsets or synonyms of the words and are stored in the database.

Now keyword matching has to be performed by matching the keywords from the given text to the already prepared list of words in the database. A few stop words are also added to the list so as to increase the efficiency. In this phase, the important functions are input extraction from the internet/chat messaging tokenization of the extracted data, keyword matching and frequency counting (which is explained in phase 3).

Algorithm for Keyword Matching:

```
        1. PreCursorFunction();
        2. ReadData();
        3. FilterStopWords();
4.  SplitWords(); For
    Each Word
        5.CheckInDB();
        6.if(found) 7.addScore(corresponding
        Score)
        8.  modifyFrequencyMatrix        ();
        9.break;
else
continue;
        10.preCursorFunction()
11.Read from Flagged File
12.Read from Non-Flagged File
13.Read from High-Flagged File
14.For Each word in Each File
15.getSynonyms(eachWord)
16.insertIntoCorrespondingDBTable
```

Phase 2:

Semantic Analysis:

The aim of this phase is to derive the meaning that the message intends to carry. LSA (Latent Semantic Analysis) algorithm is used for the purpose of semantic analysis of data. Generally, all behavioral detection and analysis algorithms use LSA and define a custom score for the data so as to classify the documents as positive or negative.

But in our approach semantic analysis is used to process the data in terms of meaning and to extract different emotions like joy, anger, sadness etc from the given text. Then, this extracted meaning is further used to score the message in accordance of maliciousness and explicit content.

Phase 3:

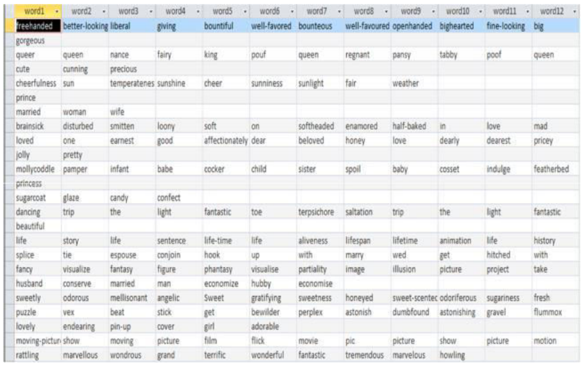
Frequency Counting:

This phase is used to predict the actual score of the message after being processed in the above two phases. Actually, frequency counting starts implicitly in the keyword matching phase where the individual usages of every single keyword counted in the background and tabulated in terms of matrix.

The keywords are scored from the matrix and a final score of every single sentence is computed. So the sentenced are marked either as malicious or not malicious along with their cumulative scores. An average cumulative score is eventually computed for the entire message and then compared to the threshold value that has been already predefined.If the average score for that particular message exceeds the threshold value,an alert is fired.

6. Results & Conclusions:

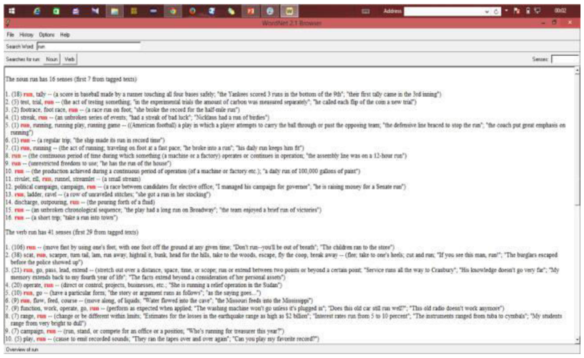
The database collection of all abusive/flirty/bawdy/explicit words looks as follows.



word1	word2	word3	word4	word5	word6	word7	word8	word9	word10	word11	word12
freehanded	better-looking	liberal	giving	bountiful	well-favored	bounteous	well-favoured	openhanded	bighearted	fine-looking	big
gorgeous											
queer	queen	nance	fairy	king	pouf	queen	regnant	pansy	tabby	poof	queen
cute	cunning	precious									
cheerfulness	sun	temperatenes	sunshine	cheer	sunniness	sunlight	fair	weather			
prince											
married	woman	wife									
brainsick	disturbed	smitten	loony	soft	on	softheaded	enamored	half-baked	in	love	mad
loved	one	earnest	good	affectionately	dear	beloved	honey	love	dearly	dearest	pricey
jolly	pretty										
mollycoddle	pamper	infant	babe	cocker	child	sister	spoil	baby	cosset	indulge	featherbed
princess											
sugarcoat	glaze	candy	confect								
dancing	trip	the	light	fantastic	toe	terpsichore	saltation	trip	the	light	fantastic
beautiful											
life	story	life	sentence	life-time	life	aliveness	lifespan	lifetime	animation	life	history
splice	tie	espouse	conjoin	hook	up	with	marry	wed	get	hitched	with
fancy	visualize	fantasy	figure	phantasy	visualise	partiality	image	illusion	picture	project	take
husband	conserve	married	man	economize	hubby	economise					
sweetly	odorous	mellisonant	angelic	Sweet	gratifying	sweetness	honeyed	sweet-scentec	odoriferous	sugariness	fresh
puzzle	vex	beat	stick	get	bewilder	perplex	astonish	dumbfound	astonishing	gravel	flummox
lovely	endearing	pin-up	cover	girl	adorable						
moving-pictur	show	moving	picture	film	flick	movie	pic	picture	show	picture	motion
rattling	marvellous	wondrous	grand	terrific	wonderful	fantastic	tremendous	marvelous	howling		

Fig: 1 Database Design

Keyword matching is performed on the above database and then proceeds to semantic analysis phase.



run

The noun run has 16 senses (first 7 from tagged texts)

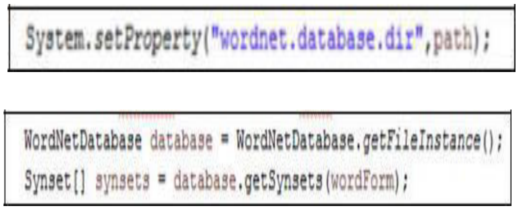
1. (18) run, tally -- (a score in baseball made by a runner touching all four bases safely; "the Yankees scored 3 runs in the bottom of the 9th"; "their first tally came in the 3rd inning")
2. (5) test, trial, run -- (the act of testing something; "in the experimental trials the amount of carbon was measured separately"; "he called each flip of the coin a new trial")
3. (2) footrace, foot race, run -- (a race run on foot; "she broke the record for the half-mile run")
4. (1) streak, run -- (an unbroken series of events; "had a streak of bad luck"; "Nicklaus had a run of birdies")
5. (1) run, running, running play, running game -- ((American football) a play in which a player attempts to carry the ball through or past the opposing team; "the defensive line braced to stop the run"; "the coach put great emphasis on running")
6. (1) run -- (a regular trip; "the ship made its run in record time")
7. (1) run, running -- (the act of running; traveling on foot at a fast pace; "he broke into a run"; "his daily run keeps him fit")
8. run -- (the continuous period of time during which something (a machine or a factory) operates or continues in operation; "the assembly line was on a 12-hour run")
9. run -- (unrestricted freedom to use; "he has the run of the house")
10. run -- (the production achieved during a continuous period of operation (of a machine or factory etc.); "a daily run of 100,000 gallons of paint")
11. rivulet, rill, run, runnel, streamlet -- (a small stream)
12. political campaign, campaign, run -- (a race between candidates for elective office; "I managed his campaign for governor"; "he is raising money for a Senate run")
13. run, ladder, ravel -- (a row of unravelled stitches; "she got a run in her stocking")
14. discharge, outpouring, run -- (the pouring forth of a fluid)
15. run -- (an unbroken chronological sequence; "the play had a long run on Broadway"; "the team enjoyed a brief run of victories")
16. run -- (a short trip; "take a run into town")

The verb run has 41 senses (first 29 from tagged texts)

1. (106) run -- (move fast by using one's feet, with one foot off the ground at any given time; "Don't run--you'll be out of breath"; "The children ran to the store")
2. (38) scat, run, scarper, turn tail, lam, run away, hightail it, bunk, head for the hills, take to the woods, escape, fly the coop, break away -- (flee; take to one's heels; cut and run; "If you see this man, run!"; "The burglars escaped before the police showed up")
3. (21) run, go, pass, lead, extend -- (stretch out over a distance, space, time, or scope; run or extend between two points or beyond a certain point; "Service runs all the way to Cranbury"; "His knowledge doesn't go very far"; "My memory extends back to my fourth year of life"; "The facts extend beyond a consideration of her personal assets")
4. (20) operate, run -- (direct or control; projects, businesses, etc.; "She is running a relief operation in the Sudan")
5. (10) run, go -- (have a particular form; "the story or argument runs as follows"; "as the saying goes...")
6. (9) run, flow, feed, course -- (move along, of liquids; "Water flowed into the cave"; "the Missouri feeds into the Mississippi")
7. (9) function, work, operate, go, run -- (perform as expected when applied; "The washing machine won't go unless it's plugged in"; "Does this old car still run well?"; "This old radio doesn't work anymore")
8. (7) range, run -- (change or be different within limits; "Estimates for the losses in the earthquake range as high as $2 billion"; "Interest rates run from 5 to 10 percent"; "The instruments ranged from tuba to cymbals"; "My students range from very bright to dull")
9. (7) campaign, run -- (run, stand, or compete for an office or a position; "Who's running for treasurer this year?")
10. (5) play, run -- (cause to emit recorded sounds; "They ran the tapes over and over again"; "Can you play my favorite record?")

Overview of run

Fig: 2 Graphical User Interface of WordNet

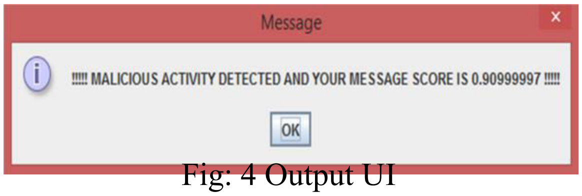


```
System.setProperty("wordnet.database.dir",path);
```

```
WordNetDatabase database = WordNetDatabase.getFileInstance();
Synset[] synsets = database.getSynsets(wordForm);
```

Fig: 3 Code Snippets to retrieve the Synsets of a particular word

In the frequency counting phase, the individual scores are computed for each single message. After a count of ten messages, an average score is computed. The average score is compare now with a predefined threshold value in order to specify whether the above messages signify malicious activity or not. If the average score crosses the threshold value it implicitly generates an alert to the user.



!!!!! MALICIOUS ACTIVITY DETECTED AND YOUR MESSAGE SCORE IS 0.90999997 !!!!!

Fig: 4 Output UI

Future scope:

Behavioural analysis, emotion generation & analysis have a number of practical applications.

Tools need to be developed to mine particular emotions from variable sources of data. Also specialized tools to process multilingual data sets can also be developed.

8. References

[1] Computational Approaches for Emotion Detection in Text, 4th IEEE International Conference on Digital Ecosystems and Technologies 2014, Haji Binali, Chen Wu,VidyasagarPotdar Digital Ecosystems Business Intelligence Institute Curtin University of Technology

[2] Automatic Generation of Emotions for Social Networking Websites using Text Mining, TejasviniPatil, Sachin Patil, IEEE Conference 2013, Tiruchengode

[3] Emotion Analysis of Internet Chat, Shashank and Pushpak Bhattacharyya, CSE, IIT Bombay.

[4] Aman, S. and Szpakowicz, S. "Identifying expressions of emotion in text", In V. Matou sek and P. Mautner, editors, Text, Speech and Dialogue, volume 4629 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, pp. 196205, 2007.

[5] Alm, C. O., Roth, D., and Sproat, R. "Emotions from text: Machine learning for text-based emotion prediction", In Proceedings of the Joint Conference on Human Language Technology / Empirical Methods in Natural Language Processing (HLT/EMNLP-2005), Vancouver, Canada, pp. 579-586, 2005.

[6] Aggarwal C., "Text mining in social network", In Social Network Data Analytics, 2nd edition. Springer, pp. 353-374, 2011.

[7] Baumer E. P. S., Sinclair J. & Tomlinson B., "America is like Metamucil: Fostering critical and creative thinking about metaphor in political blogs", In Proceedings of 28th International Conference on Human Factor in Computing Systems ACM, Atlanta, GA, USA, pp. 34-45, 2010

[8] Baatarjav E., Phithakkitnukoon S. &Dantu R, "Group Recommendation System for Facebook", 2nd edition Springer, 2008.

[9] Bellegarda, J. "Emotion analysis using latent affective folding and embedding", In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Los Angeles, California, 2010

[10] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction", In Proceedings of HLT/EMNLP"05, pp. 579–586, 2005.