Modeling of online social networks for automated monitoring system

Yu.B. Savva[1], Yu.V. Davydova[1]

[1]Orel State University, 95, Komsomol'skaya, 302026, Orel, Russia

Abstract

Monitoring using keywords is necessary step in solving the problem of detection of users' illegal behavior such as drug use, extremist propaganda in online social networks. Analysis of text posts is difficult because of using jargon and making mistakes in communications. In paper model of online social networks for automated monitoring system is presented. This model focuses not on communications between users but on text posts. Features of Russian text posts are given. Problem of text posts obfuscation by users involved in illicit fields of activities is discussed.

Keywords: online social networks; monitoring; text analysis; information retrieval; fuzzy search

1. Introduction

Advantages of online social networks (OSNs) such as high speed of information dissemination which can be compared with a virus and ease of use make them a convenient tool for information influence and propaganda of deviant and illegal actions. Threats of OSNs, such as extremist and terrorist groups, were discussed in [1]. For providing information and psychological security of users, automated monitoring system of OSNs is required. Most existing monitoring systems [2] are used for business goals and find out users' attitude to brands. Sharing their opinions about products, service or social events users try to use hashtags – correctly written mentions with special label or metadata. It makes it easier for users to find messages with a specific theme or content. As for illegal activity, it isn't advertised or advertised for closed groups of users though propaganda can be an exception. It is more difficult to find posts connected with searching topic without hashtags especially if they are written with mistakes. Spelling errors and typos are common for informal writing on the whole and for text posts in OSNs in particular. Also informal writing is often characterized by using slang and different abbreviations. As for illicit fields of activity, communications often contain specialized jargon. We consider jargon as a highly specialized slang, which is often used in closed communities and is hard to understand. Taking these features into consideration it can be said without doubt that monitoring process is difficult and requires special methods for analysis of text posts.

Process of monitoring involves a kind of information retrieval, text posts from OSNs are gathered and fuzzy search by keywords is organized. Keywords represent lexics, which is used in communications in illegal fields of activity. This paper describes model of online social networks used in automated monitoring system. According to system's goal, the emphasis of the model is made on text posts.

2. The object of the study

Usually online social network is defined as a graph $G(N,E)$, where $N=\{1,2,..n\}$ is a set of vertices (agents - users, communities) and E is a set of edges which represents interaction of agents [3]. Tasks of users' behavior modeling, users' interaction modeling, analyzing features of subgraphs of friendship are popular. The main goal of current automated monitoring system is decision support in detection of illegal behavior in OSNs, wherein information retrieval and analysis of text posts play a big role. Thus, text posts should be included in model.

Let us denote $I=\{i_1,i_2,...,i_{ic}\}$ is a set of identifiers of OSNs users or communities, where ic is the number of identifiers. $M=\{m_1,m_2,...,m_{mc}\}$ is a set of posts, mc is the number of posts. Posts are gathered into groups: $M=\bigcup_{k=1}^{ic} M_k$. Every post can be represented as follows:

$m_j=\langle i_k, text_j, t_j, type_h, parent_j\rangle$, $i_k\in I$, $j=\overline{1..mc}$, $h=\overline{1..3}$,

$$parent_j=\begin{cases}\varnothing, type_j=type_1\\ i_n\in I, n=\overline{1..ic}\end{cases},$$

where: – i_k is identifier of user who posted the message m_j;

– $text_j$ is text of the post m_j, $text_j=\langle w_{j1,}w_{j2},...,w_{jg}\rangle$, w_{ji} is the *i*-th word in text;

– t_j – date and time of the posted message m_j;

– $type_h\in Type$, $Type=\{type_1,type_2,type_3\}$ is a set of post types, where $type_1$ is original post (which means that user who posted message is its author), $type_2$ is reposted message (which means that user posted somebody's message), $type_3$ is a comment to original or reposted message;

– $parent_j$ is a user's or community's identifier. If type of current post is a repost or a comment then $parent$ contains identifier of author who posted original message.

$P = \{p_1, p_2, ..., p_{ic}\}$ is a set of pages of OSNs, number of pages is equal to number of users' and communities' identifiers as every page belongs to user or community. Page is defined as follows:

$$p_k = \left\langle i_k, tt_q, c_k, M_k = \left\{ m_{kz} \middle| z = \overline{1..x} \right\} \right\rangle, i_k \in I,\ x < mc,\ q = \overline{1..2},$$

$$c_k = \begin{cases} \varnothing, tt_k = tt_1 \\ \{i_n\} \subset I, n = \overline{1..ic} \end{cases},$$

where: – i_k is the identifier of user or community of current page p_k;

– $tt_k \in TT$, $TT = \{tt_1, tt_2\}$ is a set of pages type. tt_1 is a personal page and tt_2 is a community page;

– c_k is a set of user's identifiers. If current page p_k is a personal page then c_k is an empty set as page p_k belongs to one user. If p_k is a community page then c_k keeps user's identifiers who are owners or managers of community (it can be one user, so c_k keeps one element);

– M_k is a group of posts which are posted on the page p_k. It can be empty $M_k = \varnothing$, that means OSNs page doesn't contain any posts at the moment.

Set of keywords is given $L = \{l_1, l_2, ..., l_{lc}\}$, where lc – the number keywords. Every keyword is represented by its grammatical, semantic information and word forms (according to inflection rules in Russian language) $l_s = \{GR_s, SM_s, WF_s\}$. This keywords storage model was described in [4]. In this work we are focused on word forms of keywords. They was defined as a language WF over the alphabet A. $WF \in A^+$.

The goal of automated monitoring system of OSNs is to find set of pages $PF \subset P$ which contains required amount of keywords, therefore these pages are indicators of potential illegal actions of their owners. Conceptually it can be presented as follows: $PF = \left\{ \left\langle i_k, tt_q, c_k, M_k = \left\{ m_{kz} \middle| z = \overline{1..x} \right\} \right\rangle \middle| \left(\sum_{z=1}^{x} \sum_{q=1}^{lc} f\left(m_{kz}, l_q\right) \geq \delta \right) \wedge \left(k = \overline{1..ic} \right) \right\}$,

where: – $f\left(m_{kz}, l_q\right)$ is a function which is defined as $f\left(m_{kz}, l_q\right) = y\left(text_{kz}, WF_q\right)$;

– δ is a threshold of presence of keywords in text posts of current user. It can be defined by decision maker.

$y\left(text_{kz}, WF_q\right)$ is a function of fuzzy search matching, conceptually it can be presented as follows: $y\left(text_{kz}, WF_q\right) = \sum_{i=1}^{g} \sum_{j=1}^{r} d(w_{zi}^k, wf_{qj}),\ d(w_{zi}^k, wf_{qj}) \leq \varphi$,

where: – $d(w_{zi}^k, wf_{qj})$ is a distance measure, which shows similarity between two words w_{zi}^k and wf_{qj}. Initial states are: $d(0, wf_{qj}) = wf_{qj}$ and $d(w_{zi}^k, 0) = w_{zi}^k$,

– φ is a threshold of distance measure, it can be defined by decision maker. The less is the value of distance measure, the higher is similarity between words. That means that current word in text post is a keyword written with mistakes with great probability. By choosing the value of threshold of distance measure, decision maker can manage the levels of precision and recall of information retrieval. The less is the value the more precise search is, but in this case, some relevant text posts will be lost and recall will be lower.

As the result of Russian text posts analysis from OSNs it was revealed that users use informal style of writing and often neglect the language rules.

3. Features of Russian text posts of OSNs users

Text posts in OSNs have the following features:
- use of conversational style in writing, slang and jargon use, abbreviations use;
- short length of average text post with weak formal syntactic relations;
- use of smileys, different special symbols;
- intentional and unintentional garble of words, including spelling errors and typos;
- borrowings from English language, like "4u" (For You).

These features characterize modern informal communications, where there is a high speed of information exchange and additional expression. Thus, text posts of OSNs users can be considered as unstructured sequences of letters symbols and images combining with each other. This fact should be taken into account in text analysis. As for communications in illegal fields of activity, additional features should be noted. To avoid detecting by law-enforcement agencies people use jargon. The main difficulty of text analysis in case of jargon use consists in high degree of homonony. Words of common used lexics may be organized in collocations and thus get new semantics as a result. There is a constant appearance of new jargon. The most glaring example is jargon in the field of illicit traffic of narcotic drugs as new substances appear rather quickly. Also it should be considered that OSNs users involved into illegal activities obfuscate posts with the same aim to prevent their detection.

4. Problem of text posts obfuscation and methods dealing with it

First mention of obfuscation method appeared in [5]. Authors suggested confusing of program code by adding extra variables and constructions with aim to prevent algorithm analysis and deter reverse engineering. Also obfuscation can be used to optimize code. Analysis of obfuscation methods of computer program is given in [6], deobfuscation methods are presented in [7]. Later obfuscation was applied to creating spam emails, spam messages on different web sites. In this case obfuscation allows to pass through content filtering. Obfuscated words can't be found during exact matching between words from message and words from dictionary. Dictionary contains words, which are indicators of spam messages.

Text posts in OSNs can be obfuscated by users involved in illegal fields of activity, for instance terrorist and extremist propaganda, illicit drug sales. In this case as it was discussed in [8] users obfuscate their posts to prevent effective linguistic analysis of texts and avoid detection of their destructive actions and influence on other OSNs users. For text obfuscation generally the following methods are used:

- intentional garble of words, including spelling errors, typos, wrong word boundaries (space insertions and deletions);
- letter substitution by digits, symbols which look like substituted letters;
- insertion extra not meaningful symbols;
- transliteration use.

Thus, text posts deobfuscation is the actual and difficult issue. Solution by computer means is not a trivial task as there are many ways of obfuscation of even one word. Thereby such methods as spell checking, deleting non-alphabetic symbols and constructing variants by possible substitutions are not so effective. Applying Hidden Markov Model to the task of spam emails deobfuscation showed good results [9]. Also, statistical models can be useful, for instance, model based on Bayesian rule, n-gram model [10, 11].

5. Using model of OSNs in automated monitoring system

Automated monitoring system includes the following main subsystems:

- data collection;
- fuzzy text search which includes linguistic knowledge base, keywords database, algorithmic search and deobfuscation modules;
- results processing and report generation modules;
- database of text posts and database of search index.

According to model, data collection subsystem gathers identifiers and text posts with additional attributes like type of messages, time and date of posting. This information is stored in database of text messages. Decision maker can specify settings of OSNs crawl strategy.

Subsystem of fuzzy text search takes information from database of text posts and implements the goal of automated monitoring system, trying to detect illegal behavior by using linguistic knowledge base and keywords database. At first stage tokenization is held, text is deobfuscated if it requires. The second stage is fuzzy text search using keywords. The use of linguistic knowledge base helps to make information retrieval not so sensitive to mistakes. Linguistic knowledge base contains information about inflection paradigms, models of mistakes, typos. Keywords database stores grammatical, semantic information and word forms of keywords lexemes. In case some text post contains threshold amount of keywords, it is indexed and is sent to database of search index. Processes of gathering information by data collection subsystem and searching by subsystem of fuzzy text search are parallel. Report generation modules show different slices of results to user of monitoring system such as topic distribution, age and location distribution of OSNs users and some others. Results are grouped according to threshold of similarity distance measure.

6. Results and discussions

At the present time automated monitoring system is to be used in detection of drugs use propaganda and illicit drug sales in OSNs [12], though system can be used in different fields, it depends on keywords database. Linguistic database of keywords used in the field of illicit traffic of narcotic drugs and psychotropic substances was developed [13]. It allows to store not only word forms but semantics of jargon. Deobfuscation method using Hidden Markov Model was developed [14], example of algorithm is presented at 0

Corpus of text posts is gathered from OSN Vkontakte. Currently algorithms of fuzzy search using keywords from developed linguistic database and models for linguistic knowledge base are developed. Features of algorithms and default values for distance measure should be tested on text corpus and corrected in case of need as they are a kind of empirical data because natural language is not a good formalized object [10, 11].

7. Conclusion

For providing information and psychological security of users, it is necessary to organize online social networks monitoring. Monitoring process has many difficulties like short messages in OSNs, informal communications using jargon, text posts obfuscation. To detect users' illicit activities and destructive influence effective text analysis and search by keywords should be organized. Thus, in OSNs modeling emphasis should be on text posts, corresponding model was presented in this paper. Main

subsystems of automated monitoring system using aspects of the model were described. The results of the work done were discussed and features of future work were given.

```
====================== RESTART: C:/Python34/deobf.py ======================
>>> D=Deobf(connection)
>>> D.deobfuscate(tree,"""schanc
так кт о ж мы, на конец? уа - chасть той силы, чтоо вeĥno hочет зла и вечно сов
ершает бллаг^о
Товарищ, друг, не скупись, купи немножко конопли
Ой m@ma mь|ла rra3y лала voda водафон
я-пришёл-к-тебе-с-приветом
это мыло давно и неправда н е п р а в д а
ч
у
ш
ь
весь коотрый соабка рпишёл дмой анольд коова кроова
жи3нь - б0/\ь и прочие радости страдание исчо""")
шанс так что мы наконец я часть той силы что вечно хочется и вечно совершает бл
аго товарищ друг не скупись купе немножко коноплиой м мыла разу вода вода он я
пришел тебе с приветом это мыло давно и неправда не правда чушь весь который со
бака решил домой анольдкоова крова жизнь больше и прочие радости страдание исче
>>>
```

Fig. 1. Example of text deobfuscation.

References

[1] Davydova YuV. To the issue of need for automation of threats search process in virtual social networks and communities. Actual problems in modern science in XXI century: proceedings of the 6th international scientific-practical conference. Makhachkala: “Aprobaciya” Publisher, 2014; 25–26. (in Russian)

[2] The top 25 social media monitoring tools. URL: http://keyhole.co/blog/the-top-25-social-media-monitoring-tools/ (19.01.2017).

[3] Gubanov DA, Novikov DA, Chhartishvili AG. Online social networks: models of information influence, control and confrontation. Moscow: “Fizmatlit” Publisher, 2010; 228 p. (in Russian)

[4] Savva YuB, DavydovaYuV. Linguistic database for monitoring system of online social networks in providing information and psychological security. European integration: justice, freedom and security: proceedings of VII scientific and professional conference with international participation: in 3 volumes. Belgrade: “Criminalistic-Police Academy” Publisher, 2016; 1: 145–154.

[5] Diffie W, Hellman M. New directions in cryptography. IEEE Transactions on Information Theory 1976; IT-22(6): 644–654.

[6] Korobejnikov AG, Kutuzov IM, Kolesnikov PYu. Analysis of obfuscation methods. Cybernetics and programming 2012; 1: 31–37. (in Russian)

[7] Kasperski K, Rokko E. The art of disassembling. SPb: BHV-Peterburg, 2008; 892 p. (in Russian)

[8] Savva YuB, Eryomenko VT, Davydova YuV. About the problem of the linguistic analysis of the slang in the problem of the automated search of threats of spread of drug addiction on virtual social networks. Information systems and Technologies 2015; 6(92): 68–75. (in Russian)

[9] Honglak L, Andrew YNg. Spam Deobfuscation using Hidden Markov Model. Proceedings of the Second Conference on Email and Anti-Spam, 2005. URL: http://ai.stanford.edu/~ang/papers/ceas05-spamdeobfuscation.pdf (11.07.2016).

[10] Ingersoll GS, Morton TS, Farris AL. Taming text. How to find, organize and manipulate it. NY: Manning Publications Co., 2013; 320 p.

[11] Manning CD, Raghavan P, Schutze H. Introduction to information retrieval. Cambridge: Cambridge University Press, 2008; 496 p.

[12] Savva YuB, Eryomenko VT, Davydova YuV. Design of information system identification of persons which participate illicit in field of narcotic drugs and psychotropic substances in the virtual social networks using the database jargon. Information systems and Technologies 2016; 1(93): 68–75. (in Russian)

[13] Savva YuB, Davydova YuV. Certificate of state registration database no. 2016620197. Jargon in the field of illicit traffic of narcotic drugs and psychotropic substances. Registered 10 February 2016.

[14] Nikol’skaya AN, Savva YuB. About the problem of opening of obfuscated Russian-language texts of participants of online social networks. Information systems and Technologies 2016; 6(98): 44–55. (in Russian)