Data Mining Usage for Social Networks

Hanna Martyniuk[1][0000-0003-4234-025X], Serhii Lazarenko[1] [0000-0003-3529-4806], Valeriy Kozlovskiy[1] [0000-0002-8301-5501], Yuriy Balanyuk[1] [0000-0003-3036-5804] and Ivan Yakoviv[1] [0000-0003-1455-5747], Pavlo Skladannyi[2] [0000-0002-7775-6039]

[1] National Aviation University, Kyiv, Ukraine
ganna.martyniuk@gmail.com
[2] Borys Grinchenko Kyiv University, Kyiv, Ukraine
p.skladannyi@kubg.edu.ua

Abstract. The authors present in this work information about social media and data mining usage for that. It is represented information about social networking sites, where Facebook dominates the industry by boasting an account of 85% of the internet users worldwide. Applying data mining techniques to large social media data sets has the potential to continue to improve search results for everyday search engines, realize specialized target marketing for businesses, help psychologist study behavior, provide new insights into social structure for sociologists, personalize web services for consumers, and even help detect and prevent spam for all of us. The most common data mining applications related to social networking sites is represented. Authors have also given information about different data mining techniques and list of these techniques. It is important to protect personal privacy when working with social network data. Recent publications highlight the need to protect privacy as it has been shown that even anonymizing this type of data can still reveal personal information when advanced data analysis techniques are used. A whole range of different threat of social networks is represented. Authors explain cyber hygiene behaviors in social networks, such as backing up data, identity theft and online behavior.

Keywords: social media, data mining, cyber hygiene, media platforms, threats of social networks, data mining techniques.

1 Introduction

Social media plays a vital role in our daily life. Websites like Facebook, Twitter, and Instagram are the most common social channels used to connect with our loved ones. With over 2.77 billion social media users today, such social media websites make a perfect platform for identity thefts. With huge user database of private information, it is the responsibility of social media platforms to keep personal information safe [2].

Many companies are eager to analyze huge amounts of social network data to take advantage of this social phenomenon. Social network data mining is one of the hottest research topics. The application of efficient data mining techniques has made it possible for users to discover valuable, accurate and useful knowledge from social network data [1, 5, 9]. But today there is a whole range of different threats in social networks. At this paper will be described data mining usage as threat in social networks.

2 Publications analysis. Problem statement

The use of the Internet social networks makes it possible to communicate with old friends, make new acquaintances, express their thoughts on a very wide audience, join groups of interest. By coverage of audiences some groups in social networks and popular bloggers can compete with many media. According to efficiency information transmission social networks are often superior to most of media, they are able to disseminate information around the world in seconds, thereby expediting the progress of operation, but this does not mean that television and radio have lost their popularity [1, 3, 5, 8, 10].

The structure of the social media data is unorganized and is displayed in different forms such as: text, voice, images, and videos [7]. Moreover, the social media provides an enormous amount of continuous real time data that makes traditional statistical methods unsuitable to analyze this massive data [15]. Therefore, the data mining techniques can play an important role in overcoming this problem.

But in [14, 13] describes that data mining can be used in conjunction of social media to deliver malware for cybercrime. So authors present in this paper data mining usage and describe threats for social networks.

3 Social networking sites

A social networking site is an online platform which people use to build social networks or social relationship with other people who share similar personal or career interests, activities, backgrounds or real-life connections [6].

The social network is distributed across various computer networks. The social networks are inherently computer networks, linking people, organization, and knowledge. Social networking services vary in format and the number of features. They can incorporate a range of new information and communication tools, operating on desktops and on laptops, on mobile devices such as tablet computers and smartphones. They may feature digital photo/video/sharing and "web logging" diary entries online (blogging). Social networking sites provide a space for interaction to continue beyond in person interactions. These computers mediated interactions link members of various networks and may help to both maintain and develop new social ties.

Social networking sites provide excellent sources of data for studying collaboration relationships, group structure, and who-talks-to-whom. The most common graph structure based on a social networking site is intuitive; users are represented as nodes and their relationships are represented as links. Users can link to group nodes as well.

Social networking sites allow users to share ideas, digital photos and videos, posts, and to inform others about online or real-world activities and events with people in

their network. Depending on the social media platform, members may be able to contact any other member. In other cases, members can contact anyone they have a connection to, and subsequently anyone that contact has a connection to, and so on.
In today's social networking era, Facebook dominates the industry by boasting an account of 85% of the internet users worldwide (Fig. 1) [10].

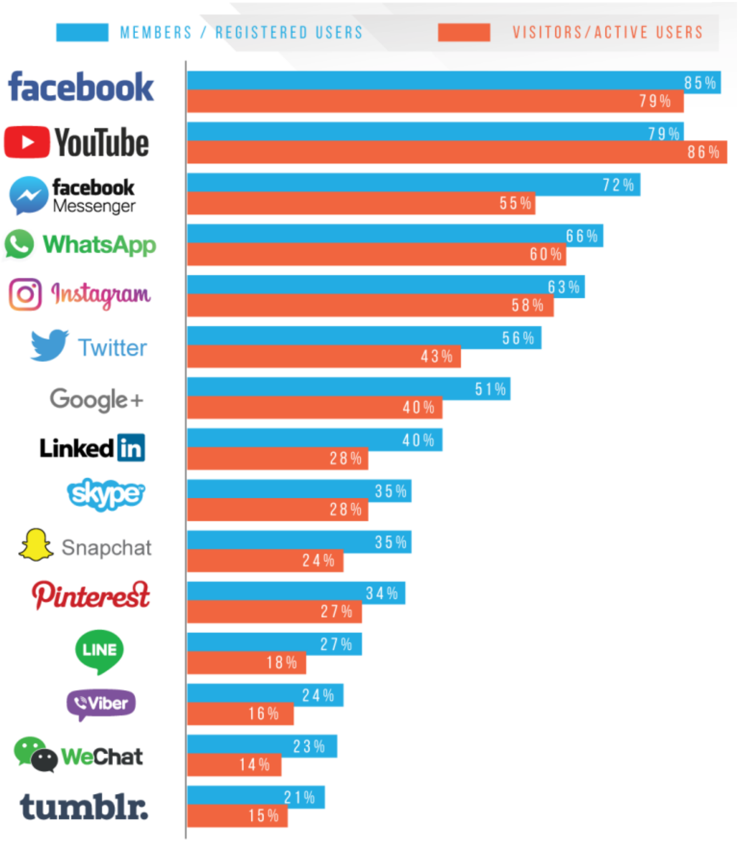

Fig. 1. The most popular social media platforms 2019

Social networking apps are going to grow even bigger as people adopt them into their everyday lives. Here we have listed the mobile-first social media platforms. But the Facebook mobile app would dominate this list with 1.37 billion monthly active users. As smartphones' adoption continues, the share of the desktop use of social media platforms will fall [6].

4 Data mining in social media

Applying data mining techniques to large social media data sets has the potential to continue to improve search results for everyday search engines, realize specialized target marketing for businesses, help psychologist study behavior, provide new insights into social structure for sociologists, personalize web services for consumers, and even help detect and prevent spam for all of us [4]. Additionally, the open access to data provides researches with unprecedented amounts of information to improve performance and optimize data mining techniques. The advancement of the data mining field itself relies on large data sets and social media is an ideal data source in the frontier of data mining for developing and testing new data mining techniques for academic and corporate data mining researchers.

The driving factors for data mining social networking sites is the "unique opportunity to understand the impact of a person's position in the network on everything from their tastes to their moods to their health." [3]. The most common data mining applications related to social networking sites include:

1. Group detection – One of the most popular applications of data mining to social networking sites is finding and identifying a group. In general, group detection applied to social networking sites is based on analyzing the structure of the network and finding individuals that associate more with each other than with other users. Understanding what groups an individual belongs to can help lead to insights about the individual such as what activities, goods, and services, an individual might be interested in.

2. Group profiling – Once a group is found, the next logical question to ask is 'What is this group about' (i.e., the group profile)? The ability to automatically profile a group is useful for a variety purposes ranging from purely scientific interests to specific marketing of goods, services, and ideas [3]. With millions of groups present in online social media, it is not practical to attempt to answer the question for each group manually.

3. Recommendation systems – A recommendation system analyzes social networking data and recommends new friends or new groups to a user. The ability to recommend group membership to an individual is advantageous for a group that would like to have additional members and can be helpful to an individual who is looking to find other individuals or a group of people with similar interests or goals. Again, large numbers of individuals and groups make this an almost impossible task without an automated system. Additionally, group characteristics change over time. For those reasons, data mining algorithms drive the inherent recommendations made to users. From the moment a user profile is entered into a social networking site, the site provides suggestions to expand the user's social network. Much of the appeal of social networking sites is a direct result of the automated recommendations which allow a user to rapidly create and expand an online social network with relatively little effort on the user's part.

Data mining is a powerful tool which will facilitate to seek out hidden patterns and various relationship between the data. Data processing discovers hidden facts from massive databases. The overall objective of the data mining technique is to extract information from a huge data set and transform it into a comprehensible structure for more use. The different data Mining techniques are [5]:

I. Characterization – used to generalize, summarize and possibly different data characteristics.

II. Classification – is a process in which the given data is classified into different classes.

III. Regression – is process similar to classification, the major difference is that the object to be predicted is continuous rather than discrete.

IV. Association – discovers the association between various data bases and the association between the attributes of single database.

V. Clustering – involves grouping of data into several new classes such that it describes the data. It breaks large data set into smaller groups to make the designing and implementation process to be simple.

VI. Change Detection – this method identifies the significant changes in the data from the previously measured values.

VII. Deviation Detection – focuses on the major deviations between the actual values of the objects and its expected values. This method finds out the deviation according to the time as well the deviation among different subsets of data.

VIII. Link Analysis – traces the connections between the objects to develop models based on the patterns in the relationships by applying graph theory techniques.

IX. Sequential Pattern Mining – involves the discovery of the frequently occurring patterns in the data.

Social network finds its application in several business activities like Co-innovation, Customer service, General promoting, increasing spoken promoting, marketing research, plan generation and new development, publicity, worker communication and reputation management.

In [9] identified 19 data mining techniques that had been applied by researchers in the area of social media. The list of these techniques is presented below:

- AdaBoost,
- Artificial Neural Network (ANN),
- Apriori,
- Bayesian Networks (BN),
- Decision Trees (DT),
- Density Based Algorithm (DBA),
- Fuzzy,
- Genetic Algorithm (GA),
- Hierarchical Clustering (HC),
- K-Means,
- k-nearest Neighbors (k-NN),
- Linear Discriminant Analysis (LDA),
- Linear-Regression (Lin-R),
- Logistic Regression (LR),
- Markov,
- Maximum Entropy (ME),
- Novel,

- Support Vector Machine (SVM),
- Wrapper.

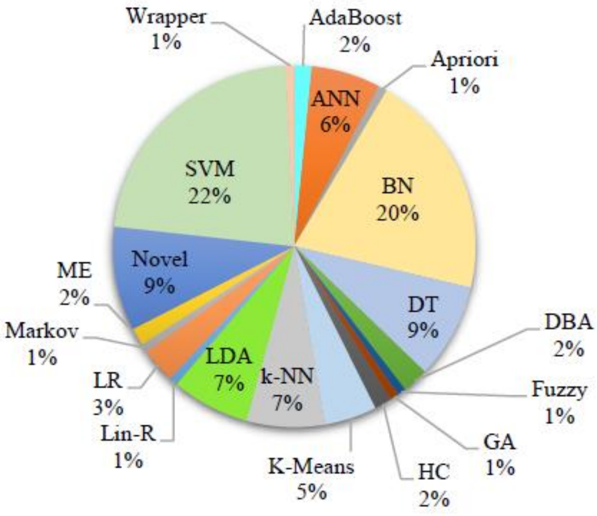

Fig. 2. Data mining techniques in social media

Fig. 2 shows that SVM, BN, and DT are the most applied techniques in the area of social media with a percentage of 51% of the selected articles. Novel techniques with the percentage of 9% were not considered as the one of the highest; because each article has its dedicated novel technique.

It is important to protect personal privacy when working with social network data. Recent publications highlight the need to protect privacy as it has been shown that even anonymizing this type of data can still reveal personal information when advanced data analysis techniques are used [3]. Privacy settings also can limit the ability of data mining applications to consider every piece of information in a social network. However, some nefarious techniques can be employed to usurp privacy settings.

5 Cyber hygiene behaviors in social media

It is important to protect personal privacy when working with social network data. Recent publications highlight the need to protect privacy as it has been shown that even anonymizing this type of data can still reveal personal information when advanced data analysis techniques are used [3]. Privacy settings also can limit the ability of data mining applications to consider every piece of information in a social network. However, some nefarious techniques can be employed to usurp privacy settings.

These days is a whole range of different threat in social networks [8].

1. Social engineering is the most popular tactic for cyber criminals. Social networks allow attackers to find confidential information that can be used for property and moral damages.

2. Friends. The trust to those who entered in the “friends” list is always higher than to random people. On the one hand, this is good, since forming a loyal audience around the company, brand or person. But on the other hand, it is an opportunity for attackers.

3. Possibility of substitution of person or masquerade: for sure it is not clear exactly who hide their actions behind the name of friends or hiding behind photos friends in social network profile.

4. Stealing passwords and phishing. As the identification of social networks uses passwords, it is sufficient to know the sequence of characters and can be possible to send advertising, some information on behalf of others, or to motivate recipients to any negative action, in particular to pass on the link and run the malicious code, and do other (often illegal) cases.

5. URL shortening services usage. In recent years, URL shortening services allow to mask unwanted website address under the short link are especially popular.

6. Using the same user names and passwords on the corporate network and external social resources. As a result, hacking profiles of social network users significantly increases the risk of penetration to corporate resources on behalf of one of the company’s employees.

7.Web-attack. As social networks are web-based applications, they can be used by hackers to organize attacks on vulnerabilities in browsers. The tools for such attacks can be Trojan applications, fake antiviruses, social worms, which are used to spread own friends lists and other. Their main goal is to get into the information system of social network visitor and gain a foothold in it.

8. Information leakage and compromising company employee’s behavior. Social networks can be used to organize leaks of important information for the company, as well as to undermine its reputation.

According these threats in [14] was prepared cyber hygiene behaviors (fig.3).

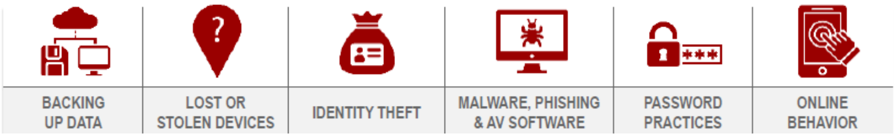

Fig. 3. Cyber hygiene behaviors

More information of these behaviors will be presented below.

5.1 Backing up data

Data backup is a process of duplicating data to allow retrieval of the duplicate set after a data loss event. Today, there are many kinds of data backup services that help enterprises and organizations ensure that data is secure and that critical information is not lost in a natural disaster, theft situation or other kind of emergency.

In the early days of personal computers (PC), the common data backup method was to download data from a computer's hard drive onto a set of small floppy disks, which were stored in physical containers. Since then, the emergence of solid-state technologies, wireless systems and other innovations have led to situations where IT managers have the option of backing up data remotely or downloading huge amounts of data into small portable devices. Cloud services and related options facilitate easy remote data storage, so that data is secure if an entire facility or location is compromised, while RAID, or mirror, technologies provide automated backup options.

In addition to remote data backup, there are new methods, such as failback and failover systems that automatically switch the destination of data when a primary destination is negatively affected in any way. All of these new options help make data security stronger as many business and government operations become increasingly reliant on various types of stored data.

Today, more than 3 in 4 (78%) persons are backing up their data using one of the methods below. However, most of them (57%) are still leaving themselves susceptible to risk by only backing up using one method, rather than backing up online (cloud) and offline (external hard drive, USB memory, etc.). Among those who are backing up their information by uploading it to the cloud, only 2 in 5 (43%) are taking the extra step in ensuring that it's stored in an encrypted format.

5.2 Identity theft

Identity theft is the unauthorized collection of personal information and its subsequent use for criminal reasons such as to open credit cards and bank accounts, redirect mail, set up cellphone service, rent vehicles and even get a job. These actions can mean severe consequences for the victim, who will be left with bills, charges and a damaged credit score.

There are many ways in which an individual's identity can be stolen, but people may be particularly vulnerable to this crime online, where savvy criminals can gain access to personal information through a number of avenues. According to the U.S. Federal Trade Commission, approximately nine million Americans have their identities stolen each year. This theft is increasingly being perpetrated electronically.

Identity thieves have a number of avenues for stealing personal information via electronic means. These include:

Retrieving stored data from discarded electronic equipment such as PCs, cellphones and USB memory sticks.

Stealing personal information using malware such as keystroke logging or spyware.

Hacking computer systems and databases to gain unauthorized access to large amounts of personal data. Phishing, or impersonating trusted organizations (such as the IRS, a bank or large retailer) via email or SMS messages and prompting users to enter personal financial information.

Compromising weak login passwords (often through calculated guesswork) to gain access to a user's online accounts. Using social networking sites to attain enough personal details to guess email passwords or impersonate the victim in other ways online.

Diverting victims' emails to attain personal information such as bank and credit card statements, or to prevent the victim from discovering that new accounts have been opened in his or her name.

There are some steps consumers can take to protect their identities, including ensuring that any transactions they make online use secure data encryption, limiting the amount of personal information they share online, remaining alert to phishing scams and keeping a close eye on their banking and credit card statements.

5.3 Online behavior

According to [14] only 36% of Americans who use social media are making sure that all of their accounts are private. The most common social media platform that Americans have kept public is Facebook (fig. 4).

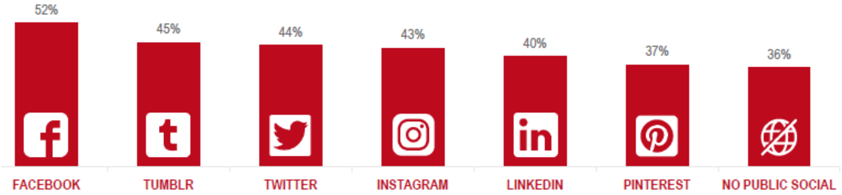

Fig. 4. Percentage who have kept social media public among those who use each social media platform

Most Americans are not yet adopting some of the key online habits that help ensure proper cyber hygiene. Less than half (49%) are regularly practicing at least 5 of the online habits shown below (Fig. 5).

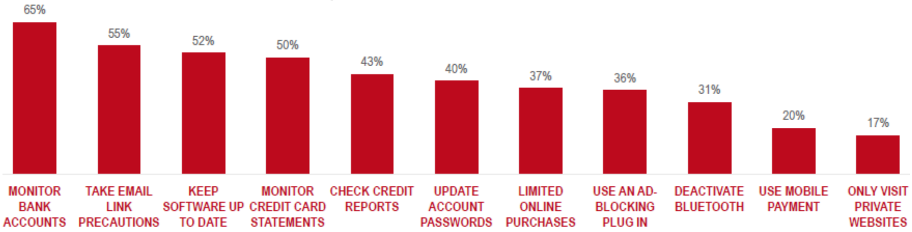

Fig. 5. Percentage regularly practicing the following cyber-hygiene habits

Meanwhile, threats go beyond technology and external hackers. Human fallibility is often the root cause of breaches. Cyber hygiene must become engrained in an organization’s daily routine to be effective.

6 Conclusions

The rise of social networks gives very strong effects to set of techniques developed for mining social networks. Social media and its analysis is an important field for research emerged from sociology, psychology, statistics and graph theory. But, on the other hand, Social media-enabled cyber crimes are generating at least $3.25bn a year in global revenue and one in five organizations has been infected with malware distributed via social media. The authors present in this work information about social networking sites, the most common data mining applications and a whole range of different threat of social networks. After that authors explain cyber hygiene behaviors in social networks, such as backing up data, identity theft and online behavior.

References

1. Aggarwal C (2011). Introduction to social network data analytics. Springer US. Retrieved from DOI:10.1007/978-1-4419-8462-3.
2. Alexandra S (2019). 10 Tips for Keeping Your Personal Data Safe on Social Media. In: Cybersecurity Today. Retrieved from https://securitytoday.com/Articles/2019/04/04/10-Tips-for-Keeping-Your-Personal-Data-Safe-on-Social-Media.aspx?Page=2.
3. Barbier G, Liu H (2011) Data Mining in Social Media. In: Aggarwal CC (ed) Social Network Data Analytics, Watson Research Center, New York, p 327-352.
4. Cortizo JC, Carrero FM, Gomez JM, Monsale B, Puertas P (2009) Introduction to mining social media. In: Cortizo JC, Carrero FM, Gomez JM, Monsale B, Puertas P (eds) Proceedings of the 1. 1st International Workshop on Mining Social Media, p 1-3.
5. Jan S, Ruby R, Najeeb PT, Muttoo MA (2017) Social Network Analysis and Data Mining. International Journal of Computer Science and Mobile Computing 6(6): 401-407.
6. Kallas P (2018) Top 15 Most Popular Social Networking Sites and Apps. Retrieved from https://www.dreamgrow.com/top-15-most-popular-social-networking-sites.
7. Kavanaugh AL, Fox EA, Sheetz SD, Yang S, Li LT, Shoemaker DJ, et al. (2012) Social media use by government: From the routine to the critical, Gov. Inf. Q. 29: 480–491. DOI: 10.1016/j.giq.2012.06.002.
8. Kirichenko L, Radiviliva T, Carlsson A (2017) Detecting Cyber Threats Through Social Network Analysis: Short Survey. Socio Economic Challenges 1(1): 20-34. DOI: 10.21272/sec.2017.1-03.
9. Injadat MN, Salo F, Nassif AB (2016) Data Mining Techniques in Social Media: A Survey. Neurocomputing. DOI: 10.1016/j.neucom.2016.06.045.
10. Irfan A (2019) The Most Popular Social Media Platforms of 2019. In: Digital Information World. Retrieved from https://www.digitalinformationworld.com/2019/01/most-popular-global-social-networks-apps-infographic.html.
11. Murthy D, Gross A, Takata A (2013) Emergent Data Mining Tools for Social Network Analysis. In: Bhatnagar V (ed) Data Mining in Dynamic Social Networks and Fuzzy Systems, Information Science Reference, Hershey, pp. 41-56.
12. Russell MA (2011) Mining the Social Web: Analyzing Data from Facebook,Twitter, LinkedIn, and Other Social Media Sites. O'Reilly, 332 p.
13. Torsten G (2019) Cyber Hygiene 101: Implementing Basics Can Go a Long Way. https://www.securityweek.com/cyber-hygiene-101-implementing-basics-can-go-long-way.
14. Wakefield Research: Cyber Hygiene Risk Index. Assessment of Americans' Cybersecurity Practices (2019). Retrieved from https://www-cdn.webroot.com
15. Zuber M A Survey of Data Mining Techniques for Social Network Analysis, Int. J. Res. Comput. Eng. Electron. 3 (2014), pp. 1-8.
16. Anisimova O., Vasylenko V., Fedushko S. Social Networks as a Tool for a Higher Education Institution Image Creation. CEUR Workshop Proceedings. Vol 2392: COAPSN-2019. P. 54–65 (2019). http://ceur-ws.org/Vol-2392/paper5.pdf

17. Yavorska T., Prihunov O., Syerov Y. Libraries in Social Networks: Opportunities and Presentations. CEUR Workshop. Vol 2392: COAPSN-2019. P. 242–251 (2019).
18. S. Gnatyuk, Critical Aviation Information Systems Cybersecurity, Meeting Security Challenges Through Data Analytics and Decision Support, NATO SPS Series, D: Information and Communication Security. IOS Press Ebooks, Vol.47, №3, pp. 308-316, 2016.
19. Gnatyuk S., Kinzeryavyy V., Kyrychenko K., Aleksander M. et al, Secure Hash Function Constructing for Future Communication Systems and Networks, Advances in Intelligent Systems and Computing, Vol. 902, pp. 561-569, 2020.