Can we trust this user? Predicting insider's attitude via YouTube usage profiling

Miltiadis Kandias, Vasilis Stavrou, Nick Bozovic, Lilian Mitrou, Dimitris Gritzalis
Information Security & Critical Infrastructure Protection Research Laboratory
Dept. of Informatics, Athens University of Economics & Business (AUEB)
76 Patission Ave., GR-10434, Athens, Greece
{kandiasm, stavrouv, nbozovic, l.mitrou, dgrit} @aueb.gr

***Abstract*— Addressing the insider threat is a major issue in cyber and corporate security in order to enhance trusted computing in critical infrastructures. In this paper we study the psychosocial perspective and the implications of insider threat prediction via social media, Open Source Intelligence, and user generated content classification. Inductively, we propose a prediction method by evaluating the predisposition towards law enforcement and authorities, a personal psychosocial trait closely connected to the manifestation of malevolent insiders. We propose a methodology to detect users holding negative attitude towards authorities. For doing so, we facilitate a brief analysis of the medium (YouTube), machine learning techniques and a dictionary-based approach, in order to detect comments expressing negative attitude. Thus, we can draw conclusions over a user behavior and beliefs via the content the user generated within the limits a social medium. We also use an assumption free flat data representation technique in order to decide over the user's attitude and improve the scalability of our method. Furthermore, we compare the results of each method and highlight the common behavior and characteristics manifested by the users. As privacy violations may well - rise when using such methods, their use should be restricted only on exceptional cases, e.g. when appointing security officers or decision-making staff in critical infrastructures.**

Keywords— Insider Threat; Social Media; Critical Infrastructures; Privacy; Behavior Prediction; Ethics; Legal Aspects

I. INTRODUCTION

Threats that an information system may encounter derive from either external or internal environments. In order to mitigate such threats, information security officers and researchers are often asked to identify the optimised/balanced analogy between security and functionality. One of the most demanding problems in cyber and corporate security is the insider threat [1]. In principle, the malevolent insider manifests when a trusted user of the information system behaves in a way that the security policy defines as unacceptable [2]. A malevolent insider could cause the failure of a critical information system and lead to loss of control of its supporting infrastructure, causing from minor to severe consequences. Trusted computing aims at protecting such information systems and infrastructures and making its services fully available and trustworthy to their users.

Regardless of the numerous technical countermeasures, techniques and methods proposed, insider computer abuse incidents keep occurring. As a result, research suggested [3] that technical and social solutions should be implemented so as to reduce the impact of this threat. Social learning theory [4] assumes that a person commits a crime because she has come to associate with delinquent peers. Similar approaches study the criminal computer behavior and examine the deviant computer-related behavior [5]. Furthermore, behavior analysis leads to studying employees under the prism of predisposition towards malevolent behavior, by examining personal traits that have been proved to abut to this kind of behavior. Shaw's et al. [6] research examined the trait of social and personal frustrations. The most important observation is the "revenge syndrome" they develop and the anger they feel towards authority figures. Thus, an employee negatively predisposed towards law enforcement and authorities, is considered to be more possible to manifest delinquent behavior against the organization. In the past, one should utilize stiff questionnaires and psychometric evaluations, in order to examine the above mentioned traits. On the contrary, nowadays social media offer us the opportunity to study such traits in an automated and flexible manner. Research has proved that individuals tend to transfer their offline behavior online [7], thus making it possible to perform psychometric evaluations by utilizing the content a user has made publicly available, without the user being aware of it.

In this paper we have revised and considerably extended the methodology and the results presented in a previous publication of ours [8]. We have also added further commentary over the adequacy of the YouTube social medium (via graph theoretic and content analysis), statistical analysis and common characteristics of the detected users and on legal and ethical issues. Furthermore, we have refocused the application of our method towards trusted computing. Our goal is to extract conclusions and statistics over the users, regarding the personality trait of predisposition against law enforcement and authorities. Our methodology has been found to be able to extract a user's attitude towards that trait. Along with two complementary approaches we quote proper comparison and analysis of the results. To this extend, we use data crawled by YouTube to apply our methodology on a realistic environment. We formed a Greek community of YouTube users and classified them via (a) comment classification using two approaches (machine learning and dictionary-based classification on user comments) and (b) comment classification via flat data in an assumption free basis, regarding the data mining process. In addition, users are divided in two categories, i.e., those who are predisposed negatively towards law enforcement and authorities (Category P), and those who are not (Category N). The user attitude is extracted by aggregating the individual results from the attitude expressed in comments, uploads, favorites and playlists, in the

comment classification approach and by examining each user as a flat file, in the flat data classification.

However, processing user's online data in social media, without a user's consent, interferes with the personality and privacy rights of the user and may lead to a social threat. As a result, such processing of user generated data could be ethically and legally acceptable only in cases involving critical infrastructures on user's explicit consent, where aspects like national security, economic prosperity, and national well-being are really at stake. Furthermore, trusted computing has been accused of encapsulating a core controversy; namely the system is secured for both its owner and from its owner. To this extend, our method utilizes this so-called technical "neurosis" as a feature and enables proactive protection of the system from its users in favor of its users. Thus, usage profiling could be applied as a means of protection from and in favor of key-role and decision-making personnel, security officers, and other high profile employees within security-critical infrastructures.

II. Related Work

Dealing with insider threat incidents is one of the most important challenges faced by todays organizational, industrial, and other forms of information infrastructures. To this end, researchers have proposed numerous countermeasures tackling this threat. Such countermeasures include, among others, the development of an information security common body of knowledge, in order to develop an information security curriculum [9]. Furthermore, the field of risk assessment/management and critical infrastructure protection has contributed towards an elevated understanding over the issue [10-11]. Alternative approaches have, also, been proposed to this extend [12-13].

The area of insider threat prediction involves various methods and techniques [14]. Magklaras et al. [14] introduced a threat evaluation system based on certain profiles of user behavior. Kandias et al. [15] proposed a combination of technical and psychological approaches to deal with insider threats, while Greitzer et al. and Brdiczka et al. also take into consideration the psychosocial perspective of an insider. Greitzer et al. [16] developed a psychosocial model to assess employees' behavior, while Brdiczka et al. [17] proposed an approach that combines Structural Anomaly Detection from social and information networks and Psychological Profiling of individuals so as to identify threats. Personal factors that may increase the likelihood someone to develop malevolent behavior are presented by the FBI, too [18]. Personal traits such as anger, revenge or greed along with certain circumstances present in an organization could lead to the manifestation of an insider. An approach of studying personality traits, described by Shaw, has been introduced by studying the trait of narcissism using Graph Theoretic Analysis [19].

III. Methodology and Testing Enviroment

In our research, we focus on a Greek community of YouTube and on information gathered from our previous work [20]. In comparison with our previous work, this paper builds upon our previous research and poses a significant extension mainly in terms of further analysis and comparison of our methods. We utilized the dataset we crawled during our previous work. Dataset collection was performed using YouTube's REST-based API (https://developers.google.com/ youtube/v3/), which simplifies and accelerates the procedure and does not cause any harm to YouTube's infrastructure. With respect to user privacy, we added an anonymisation layer to the collected data, thus usernames have been replaced with MD5 hashes, so as to eliminate possible connections between collected data and real life users. However, it is in principle possible, to reverse this process by using indirect means. The dataset includes: (a) 12.964 users, (b) 207.377 videos, and (c) 2.043.362 comments. The time span of the collected data covers a period of 7 years (Nov. 2005 - Oct. 2012). In addition, data was divided into three categories: (a) user-related information (profile, uploaded videos, subscriptions and favorites), (b) video-related information (license, number of likes and dislikes, category and tags) and (c) comment-related information (comment content and number of likes/dislikes).

A. Graph-Theoretic and Content Analysis

The main conclusions from the graph-theoretic and content analysis, as initially presented in [20]**,** based on the forms of interactions in YouTube**,** are:

(a). The *small world* phenomenon does apply to the collected Greek community, as every user of the community is 6 hops away from everyone else**.** (b). A *small group of users have the most subscribers*, while the rest of the users have considerably fewer subscribers. Also, user content is generated mostly by a small fraction of users, while the rest is significantly lower than the one from the above mentioned small fraction. (c). *Users join YouTube to participate*. We detected one large and strongly connected component consisting of 175.000 users, 2 small connected components with approximately 20 users, and 3.795 consisting of 1 user. As a result, most nodes in the graph have an outgoing tie to another node. Only a considerably small number of nodes had no outgoing ties (were inactive). (d). *Users of interest may be easily spotted within the limits of the medium.* Tag cloud and manual content analysis of our dataset concluded to a core axis of the medium's users' frames of interest. Most videos are related to *music* and *entertainment*. Several of them refer to *comedy*, *sports*, and *political issues*.

B. YouTube

We carried out the experimentation on a real environment i.e. YouTube, in order to prove that such an approach is feasible and a social medium could be used to tackle the insider threat issue. Also, we offer a real-life proof-of-concept application of the proposed methodology. Observations indicate that users tend to participate in the medium and generate personalized content. YouTube videos and comments contain political characteristics and the appropriate phraseology of interest (anti-authority and negatively predisposed towards law enforcement). Thus, we formed the hypothesis that attitude towards law enforcement may be extracted via content analysis. Based on these observations, we consider that: (a) YouTube often contains content expressing negative attitude towards law enforcement and the authorities, (b) users often feel free to express their opinions, especially when it comes to law enforcement and even work place superiors (because of the anonymity they assume and the emotional content of the medium), (c) most users join YouTube to participate, so one can reasonably expect that they will reveal, among others, their personal data.

IV. PROPOSED APPROACH

Below we propose a method that enables the identification of a user's attitude towards law enforcement and authority. Social media offer the ability to monitor users' online behaviour, record their online life imprint, and identify the attitude, which is expressed via their videos, comments and likes. Video is YouTube basic module. As we cannot process the video itself, we draw conclusions via video comments. This method has been thoroughly described in a previous, short version of our work [8] and, thus, it shall be illustrated in a simpler manner.

We detect the attitude towards law enforcement expressed in a comment by performing text classification into two categories: (a) category P (negative attitude towards law enforcement), and (b) category N (neutral or absent attitude towards law enforcement and authority). Category N may also contain expressions holding positive attitude towards law enforcement. However, we are interested only in the prevalence or absence of negative attitude towards law enforcement.

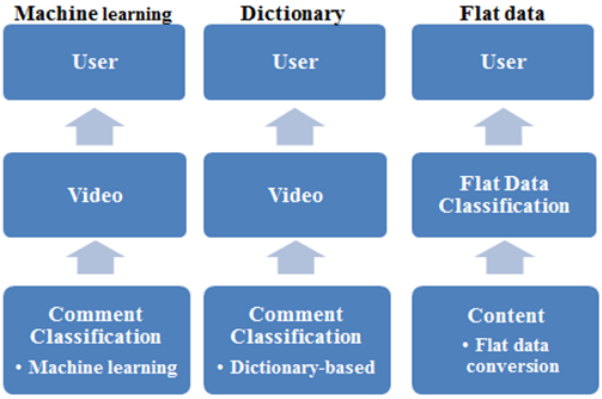

Fig. 1. Classification approaches

Comment classification enables the extraction of conclusions over the attitude expressed in a video's comments. This approach facilitates classification of videos and further categorization into one of the predefined categories. So, assignment of a comment into a corresponding category implies the attitude of the user who uploaded it, commented it, or liked it. The same applies to a list of videos (favorite videos and playlists). By acquiring the category a video falls into, a conclusion can be drawn for the attitude expressed in the list. Being able to classify a user's content, we may extract conclusions for user's comments, uploaded videos, favorite videos or playlists. By aggregating these results we can elicit adequate information to draw a conclusion over the user's overall attitude.

A. Comment Classification

In order to further analyze the collected dataset, we have chosen to store it in a relational database. We classify comments into the predefined categories using two techniques, i.e. (a) machine learning, and (b) dictionary-based detection.

A.1. *Machine Learning Approach*

In order to classify comments into one of the predefined categories of attitude towards law enforcement, training of an appropriate classifier is required. The machine is trained using input text examples, along with the corresponding categories to which they belong. The machine's output is a prediction of the label that is assigned to a text seen for the first time. Label assignment requires the assistance of an expert, who can distinguish and justify the categories each text belongs to.

An interesting characteristic of the Greek YouTube community is that users write Greek words using Latin alphabet in their communication (i.e., "greeklish"). This is the dominant way of writing in the collected community (the 51% of our dataset comments are written in greeklish), thus most users prefer to use greeklish instead of Greek. In order to address the issue of the two types of writing we decided to analyze them as two different languages. Furthermore, we observed a series of comments that were comprised of both Latin and Greek characters, a characteristic which seemed to be problematic for our method. In order to overcome this limitation, we merged the training sets and trained solely one classifier. Forming Greek and greeklish training sets requires two processes: The first is comment selection from the database and the second is proper label assignment to comments according to the category they belong. The later categorization was supported by a domain expert (i.e., Sociologist), who could assign and justify the chosen labels on the training sets. Thus, we developed a reliable classification mechanism. We chose 430 comments from category P and 470 from category N of the training set for each language. The expert contributed by assigning a category label to each comment. Apart from the training set, we also created a test set, which is required to evaluate the efficiency of the resulting classifier. The test set consists of pre-labeled data that are fed to the machine to check if the initial assigned label of each comment is equal to the one predicted by the algorithms. The test set labels were also assigned by the domain expert.

We performed comment classification using: (a) Naïve Bayes Multinomial (NBM), (b) Support Vector Machines (SVM), and (c) Logistic Regression (LR), so as to compare the results and pick the most efficient classifier. We compared each classifier's efficiency based on the metrics of precision, recall, f-score, and accuracy [8]. TABLE I. presents each classifier's efficiency, based on accuracy, precision, recall, and f-score metrics. LR and SVM achieve the highest accuracy.

TABLE I. METRICS COMPARISON OF CLASSIFICATION ALGORITHMS

	Metrics					
Classifier	***NBM***		***SVM***		***LR***	
Classes	P	N	P	N	P	N
Precision	71	70	83	77	86	76
Recall	72	68	75	82	74	88
F-Score	71	69	79	79.5	80	81
Accuracy	70		80		81	

SVM and LR achieve similar results regarding all metrics, so we chose LR because of the better f-score value achieved for each category. By choosing an algorithm based solely on the recall or precision metrics, we would favor one metric at the expense of another. A better f-score assessment indicates a balanced combination of false positive and false negative classifications. LR was, also, chosen because of its ability to provide us the probability of each category's membership instead of the purely dichotomous result of SVM. We utilized the probability of LR to determine a threshold above which a comment is classified into Category P. We chose to apply a threshold

solely on the comments characterized by negative attitude, in order to reduce false positive classifications and avoid classifying a comment into category P without strong confidence. The threshold we determined is 72%, i.e., if a comment belongs to category P and its probability of category membership is lower than 72%, it is classified as non-negative. We tested several threshold values, and after utilizing all information provided by the Social Scientist involved in the above mentioned procedure, we verified that the set of comments with probability of category membership lesser than 72% included enough false positive classifications.

A.2. *Dictionary-Based Approach*

An alternative approach we implemented was to classify comments using a specific dictionary. Comments classification was conducted with the help of a field expert (Sociologist), so as to form a dictionary containing approximately 65 terms and phrases, part of which belong to various Greek jargons, expressing negative attitude towards law enforcement and authorities. In correspondence to the machine learning approach we faced the "greeklish" usage problem, which was solved by treating them as a different language. Thus, if a comment contains a term of the dictionary, it is assigned to the P-category. We only check if the comment belongs to the category P, otherwise it is categorized as non-negative (category N). A disadvantage found in dictionary approaches is the low resulting recall value. On the other hand, they manage to achieve a high precision value. This phenomenon happens due to simple term occurrence checking, instead of taking into account other parameters that machine learning algorithms do.

B. Video Classification

For determining the attitude expressed in a video, we analyze its comments, thus we are able to extract its viewer's attitude instead of the content of the video per se. If the video contains comments with negative attitude towards law enforcement the video falls into category P. If the video does not contain any comment, which does not express negative attitude towards law enforcement, it is classified as non-negative. Users' understanding of the video is more important in this case than the content of the video, itself. This is so because we aim at classifying users instead of videos. During video classification we eliminated comments that had no likes and more than one dislikes, in order to avoid classifying a video into category P due to comments with no acceptance from other users. If a comment receives only dislikes, then it is possible that the content of the video has nothing to do with negative attitude towards law enforcement and is not calculated in the video classification process.

C. List Classification

The procedure followed to extract a conclusion about a list of videos is similar to the formerly mentioned video conclusion extraction method. The only difference is that we utilize videos instead of comments. So, if the list contains videos that belong to the category P, the list is also classified as possessing negative attitude towards law enforcement. The list classification method applies to the user's uploaded videos, favourite videos and her playlists. Also, we eliminate these videos from the list that had only dislikes and no likes.

D. User Classification

For drawing a conclusion over the user's overall attitude towards law enforcement, we examined the user according to the following procedure. The procedure takes into account a user's comments, uploaded videos, favourite videos, and playlists. A user is able to: (a) write a comment to express feelings or opinion, (b) upload a video, (c) add a video to favourites list, and (d) create a playlist and add videos to it.

Based on these observations, one may look for indications of the attitude towards law enforcement within the user generated content. A user is classified to category P, if there are comments or videos detected to her content that contain negative attitude towards law enforcement. A user may not express any attitude towards law enforcement via her comments or uploaded videos; however, the user may have added videos with similar content to playlists. The result for the user will be that she belongs to category P. It is likely to detect users who have added only a limited amount of content expressing negative attitude towards law enforcement. Similar content may also be detected in the favourite videos or playlists. In an uncommon case the user may have added to her favourites a video that contains comments unrelated to its content that express negative attitude towards law enforcement and the user is unfairly penalized. In these cases, it is possible that the user is not predisposed negatively towards law enforcement and the authorities. However it is considered as an indication that further examination is needed in order to decide whether the user shares this psychosocial characteristic or not. Inductively, the proposed method could be useful as a median in the hands of an experienced field expert who is able to weight the results and extrapolate the conclusions of our method within the limits of a critical infrastructure.

Regarding the execution time of the aforementioned approaches, it grows explosively if the user's material includes more comments and videos, as they are processed sequentially in order to draw the overall result. In order to scale well the system needs to be redesigned, as it was implemented and run on limited computing sources, namely a PC with a Core2Duo processor (3GHz) and 4GB of RAM.

E. Methods comparison

Here we present the results obtained for each classification approach utilized (i.e. Logistic Regression as LR, Dictionary approach as Terms). TABLE II. depicts the number of users per attribute who have been classified as expressing negative attitude towards law enforcement and authorities. As noted before, depending on the scenario applied, the security officer may decide which of the above mentioned attributes should be processed, so as to reveal a user attitude. TABLE III. depicts the number of comments, videos, and users that have been classified as predisposed negatively towards law enforcement.

The deviation that appears between the results of LR and the dictionary-based approach indicate that LR has detected correlations from the training set that simple word-checking could not. The detected deviation could also be a result of false positive or false negative classifications. However, the false positive/negative classifications are rather low for both categories according to metrics appearing in Table 1.

TABLE II. MACHINE LEARNING VS. DICTIONARY APPROACH

	Category P Users/Classification Attribute			
Method	***Comments***	***Uploads***	***Favorite***	***Playlists***
LR	1400	838	2300	2200
Terms	1070	516	1900	1500

TABLE III. MACHINE LEARNING VS. DICTIONARY CLASSIFICATIONS

	Category P Classifications		
Method	***Comments***	***Videos***	***Users***
LR	11986	7195	4850
Terms	11389	4784	3845

V. FLAT DATA CLASSIFICATION

The method was designed to address the same problem from a slightly different perspective. The aim was to design a method that was assumption free and could scale well. For that we chose a flat representation of the data, something that can ease their partitioning and distributed processing. We also chose a method in which we make no assumption for the users, the videos or their comments. The only input of the system is a list of negatively and non-negatively predisposed users that are used to train our system. This enables us to use the system irrespective of language or other barriers as long as we provide the system with a good initial pool of identified users, as they were indicated by the field expert. To evaluate our system's performance we conducted extensive experiments using the same dataset as in the evaluation of the above mentioned method (Section IV.A.1), and compared the results.

A. Data Transformation

The data collected and used for the experiments were stored in a relational database, so testing the alternative method required decoupling the data in a flat structure. The data where represented in one relation consisting of user information bundled with each individual comment. This flat approach was chosen because it is significantly different from our previously proposed methods thus enabling us to prove or disprove the correctness of our method. This method can also facilitate the partitioning and distributed processing of the data. Also, since this method is capable of evaluating a single post/user tuple at a time, it is a better starting point for a real-time warning system. The focus of the method is on the comment content, as that is augmented with user data as collected from the crawler. The flat relation consists of the following attributes: username, comment's text content, unique video identifier that the comment refers to user's country, age, genre, and number of subscribers and video views. Lastly, there is an additional field filled with text information that the user has chosen to publicly share on her profile. During our experiments with this model, we tested the possibility of the content and video id attributes containing the complete writings and list of commented videos respectively, of each user. This representation resulted in a huge search space where the differences between users where minimized. This led to this approach producing poor results.

B. Methodology

After formatting the data, we separated the comments into two classes based on the writer's class as evaluated by a human expert. This is a method that uses no assumptions on our part, except of the knowledge and skill of a human expert. So, by using this 44.000 strong, two class, training set we trained a simple Bayesian classifier. The resulting predictions for each user/comment tuple were combined trough a simple mechanism, to provide insight into the behavior of the individual user. More specifically, since the classification produces results for user/comment tuples, it was important to find a way to use that information for marking the interesting users as such.

In contrast to the method used before and in order to avoid marking a user over a single post (that may, or may not, be a true positive), we used a simple metric to overcome this case. We opted to use the division of the number of *P* comments a user has made by his total number of comments. This method can help us differentiate the users in an efficient and resource cheap way. This quotient acts as a threshold for identifying users that warrant closer inspection and, as such, may be chosen for appointing a security officer. In our case, the number of users flagged (blue line) in relation to the chosen threshold is summarized in Fig. 2, which also depicts the number of predisposed users (red line) detected by the above mentioned machine learning-based approach (Section IV.A.1). For each threshold depicted in Fig. 2, we present the number of users detected by both methods. As the confidence increases, the fractions of users detected by both methods begin to equalize. After manual survey of the detected users, we found out that both sets (flat data and comment classification) contain the same users. Especially in the case of 100% negatively predisposed comments percentage, the sets were almost identical.

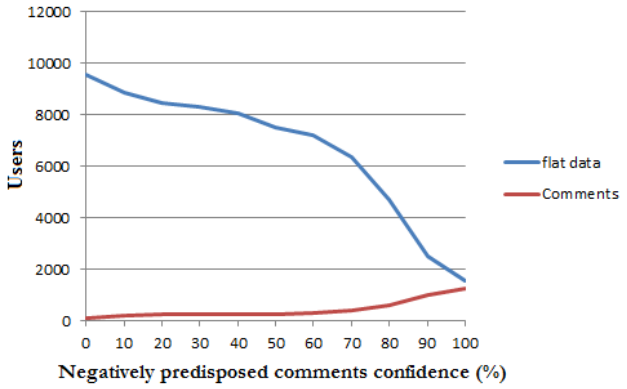

Fig. 2. Relation between detected users and comments percentage

The 1400 users, classified into category P by the machine learning-based approach, are verified by the flat data approach. The threshold of 72%, that we set in the machine learning-based methodology, above which a comment is classified as predisposed negatively, can be enhanced by the Fig. 2. As one may notice, the threshold above which both lines start to converge is above 70%, with the blue line dropping increasingly. At the threshold of 100%, the fraction of users detected by the machine learning-based approach is 10% lower than the flat data one. This variation could be explained by the fact that the two approaches are based on different classification algorithms (the core method is based on LR and flat data classification is based on Naïve Bayes). The results of the evaluation of the alternate system, that used the specified training set, are encouraging. A ten-fold cross validation showed a precision of 73% and a recall of 93% on the user/comment relati-

on. Detailed results and an overview of the system's performance with different classifiers are presented in TABLE IV.

TABLE IV. FLAT DATA ALGORITHM METRICS COMPARISON

	Naïve Bayes Metrics			
Classes	***Precision***	***Recall***	***F-score***	***Accuracy***
P	72	92	81	81
N	93	73	82	

The results are comparable and, in some cases, better than the previous two methods. Thus, the choice lies again with the security officer on which one to pick, as they have different strengths and are best used in different scenarios. The execution time of this approach is faster than the other two approaches, as it takes advantage of the flat data representation. In this case each user is represented and processed as a single tuple, which is faster than sequential processing of comments and videos. Flat data classification is likely to scale better than other approaches, as it makes use of technologies of distributed processing that accelerate the whole procedure [21].

VI. USERS' COMMON CHARACTERISTICS PRESENTATION

According to the above mentioned methods, we are able to extract a series of results regarding the crawled YouTube users. The core decision of these processes is, however, the predisposition of the user towards law enforcement and authorities. Based on the analysis we performed using the LR algorithm classifier, the 0,6% of all collected comments contain negative attitude. Regarding videos, 3,7% are classified as videos containing negative attitude. The percentage of users that indicate negative predisposing towards law enforcement is 37%. The users classified in category P have been automatically located and manually verified that express negative attitude towards law enforcement and authorities.

Regarding users classified as negatively predisposed towards law enforcement, we found that they tend to prefer the greeklish way of writing (i.e., 28% of the comments expressing negative attitude are written in Greek, 52% in greeklish, and 20% use both Greek and Latin alphabet). In TABLE V. we present the average number of characters that a comment consists of. Manual analysis of the classified comments indicates that comments written in greeklish tend to be shorter and more aggressive. On the contrary, comments written in Greek tend to be larger, more explanatory, and polite. Another finding is that the aggressive comments often tend to be misspelled.

TABLE V. AVERAGE NUMBER OF CHARACTERS IN A COMMENT

Alphabet	***Average no. of characters (P)***	***Average no. of characters (N)***
Greek	186	108
Greeklish	173	105
Both	196	159

Users predisposed negatively towards law enforcement tend to comment on the same videos and add them to their list of favorites. We detected that these videos contain unequivocal political issues, such as political events, political music clips, sports, incidents of police brutality, and content regarding the recent Greek financial crisis. Fig. 3 depicts the tag cloud generated by the videos the users predisposed negatively towards law enforcement have uploaded. One may notice that the above mentioned tags refer to politics, sports teams and music bands known to be against any form of authority. Regarding the users' demographic characteristics, 75% of the ones predisposed negatively towards law enforcement are males, 23% females, and a 2% has not declared its genre. 55% of users are between 16 and 35 years old, and 13% has not declared age.

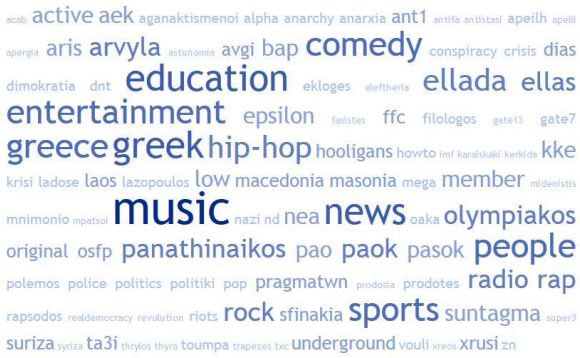

Fig. 3. Tag cloud negatively predisposed user's videos

VII. ETHICAL AND LEGAL ISSUES

Such methods raise, inevitably, ethical and legal issues, which relate at first to the lawfulness of aggregating and assessing information and content produced in different context and for other purposes, as well as the ethical/legal acceptability of mining in social media content, which reflects the life and personality of the user and her interaction with other users, in order to anticipate future threats. Moreover we have to deal with issues that are inherent in every kind of profiling, the ethics' and democracy boundary of classification and predictability of human behaviour. These issues are interrelated and they do have a common denominator: Insider threat prediction and prevention raises ethical and legal issues concerning the protection of employee privacy, personality, and dignity.

It is true that if a person generates digital/online content about herself or comment on other users' content - especially if using her real identity - could hardly claim that others refrain from judging her based on information that is made publicly available by her [22]. If theory and jurisprudence in the US recognizes no "reasonable expectation of privacy if data is voluntarily revealed to others" [23-24], also the European data protection law allows data processing even of sensible categories of personal data if this data "are manifestly made public by the data subject" (Art. 8, §2e of the European Data Protection Directive). In this context, employers that have a legitimate interest to protect their organisation from threats and ensure a secure environment and organisation are, in principle, not prohibited to consider information about a person, which is documented in YouTube, videos, "likes", and public comments.

However, major concerns have to be expressed in case a person is not revealing her identity but can be identified, detected, correlated, and profiled, even "after her key attributes (name, affiliation, address) have been removed, based on her web history" [25]. Especially in such a case we have to take into consideration that persons tailor their (digital) presence and behaviour towards particular audiences and within specific contexts. Informational privacy lies in the capacity to be a multiple personality, to maintain a variety of social identities and roles and to share different information depending on context

[26], on condition that they cause no harm to other legitimate rights and interests. Being detected and judged out of context and for entirely different and unintended purposes bears the risk of "oversimplification" [27] and in any case constitutes an interference with a person's life choices and the possibility to act and express freely in society.

Profiling, i.e. gaining probabilistic knowledge from data and inferring, on the basis of data relating to an identified or identifiable person, new data which are in fact those of the category to which she belongs, may be a useful tool to identify risks and propose predictions but, on the other side, it carries far-reaching consequences in terms of privacy infringement and unjustified and often invisible discrimination. Studies conveyed how profiling and the widespread collection and aggregation of personal information may increase social injustice and generate even further discrimination [28]. Furthermore, we should reflect on the effects of profiling methods on other democratic rights. Profiling of persons on the base of their comments and views expressed in social media and networks could lead to self-censorship and impede the exercise of fundamental rights such as the freedom of speech [29]. If individuals fear that information pertaining to them might lead to false incrimination or reprisals they would probably hesitate to engage in communication and to abstain from participation [28].

These rights and freedoms are not only essential components of persons' dignity but also, and moreover, they correspond to fundamental constitutional values and principles in democratic societies. The use of such methods should be regarded as exceptional and be applied on the ground of the so called proportionality test. A proportionality test may include the possibility of deciding that a method could be deemed as unacceptable because it may harm the essence of a fundamental right or of the constitutional order (democracy test: "necessary in a democratic state" test), even if it can be shown that this method can effectively realize another legitimate interest. Another aspect of a proportionality test lies in the obligation to explore if there are alternative measures that allow for the realization of the legitimate interest in a way that does not affect the fundamental rights in the same way as the proposed method. Such a method as the proposed one can be justified on the basis of concrete identified risks and or/in specific endangered environments, such as the case of critical infrastructures or where common goods of major importance are at stake. In the final analysis, the challenge that have to be faced is if - and to what extent - security of an organisation should be preserved at the expense of dignity and freedom.

VIII. CONCLUSION

In this paper we deal with the insider threat prediction issue so as to enhance trusted computing in critical infrastructures. Malevolent insiders and predisposition towards computer crime has been closely linked to the psychosocial trait of negative attitude towards law enforcement and authorities. In specific, we proposed a detection method for users holding this trait towards law enforcement, based on comment classification via machine learning in YouTube social medium. The desirable results are achieved by classifying each video's comments into the predefined categories. Thus, we were able to extract a conclusion over the user's attitude towards law enforcement and authorities as expressed in the content by aggregating partial classification attributes (comments, uploads, favorites, etc.).

We also presented two additional methods for user detection, along with the one described above; i.e. a dictionary-based comment classification and an assumption free flat data comment/user classification. The former approach utilizes a dictionary containing terms and expressions holding negative attitude towards law enforcement, while the rest of the procedure remains the same. The latter utilizes a sum of nine user parameters aggregated in a single tuple for each user. Then, we trained a simple Bayesian classifier who proved to have similar results to the core machine learning technique, described in our previous work [8]. Thus, using the flat data classification results, we compared the users our machine learning-based approach indicated as negatively predisposed via their comments. The comparison indicated that both methods have almost identical results, which demonstrates the robustness and adequate performance of the proposed method. In order to demonstrate the efficiency of the methodology, we used a data set that we got from YouTube, to which we applied the above mentioned approaches.

According to the above mentioned observations, one could argue over the applicability of the proposed method. The insight offered by these classification techniques over the psychological aspects of the users could be proved useful but violate several human rights. Such violations may not only set user's privacy and dignity at stake, but generate and reproduce social and labor discriminations. The asymmetry of power that characterizes the nature of employment relationship favors the battle against user privacy and leads to an increasing will for user monitoring. Thus, the dialectic over the issue should follow the road of selecting the appropriate field of application for these techniques, similarly to other proactive detection techniques, even of differentiated scope [30-34], while in some cases in order to defend, it is necessary to examine the attacker's techniques [35]. To this end, we interpolate the principle of proportionality regarding the implementation of the proposed method. Under a legal and an ethical point of view, it should be clear that such checks could be implemented under strict terms and conditions. Thus, they could be used (a) in a pre-final recruitment stage (for shortlisted candidates, with a compromised possibility of subject identifications), (b) on users' explicit consent, or (c) with the use of a trusted third party as an arbiter who orders the background check of selected candidates. Thus, these checks could be used to assist recruitment of critical infrastructures experts.

For future work we plan on improving our methods by classifying content based only on Greek or greeklish language. This is feasible by transforming greeklish words into Greek ones. In addition, we plan on applying meta-training techniques in our methodology, so as to detect common behavioral patterns among the users and determine a threshold correlation between negatively predisposed users and their comments.

REFERENCES

[1] Randazzo, R., Keeney, M., Kowalski, E., Cappelli, D., Moore, A., "Insider threat study: Illicit cyber activity in the banking and finance sector", CMU/SEI-2004-TR-021, 2005.

[2] Theoharidou, M., Kokolakis, S., Karyda, M., and Kiountouzis, E., "The insider threat to information systems and the effectiveness of ISO17799", Computers & Security, vol. 24, no. 6, pp. 472-484, 2005.

[3] Lee, J. and Lee, Y., "A holistic model of computer abuse within organizations", Information Management & Computer Security, vol. 10, no. 2, pp. 57-63, 2002.

[4] Akers, R., Krohn, M. D., Lanza-Kaduce, L., and Radosevich, M., "Social learning and deviant behavior: A specific test of a general theory", American Sociological Review, pp. 636-655.

[5] Rogers, M., Smoak, N., and Liu, J., "Self-reported deviant computer behavior: A big-5, moral choice, and manipulative exploitative behavior analysis", Deviant Behavior, vol. 27, no. 3, pp. 245-268, 2006.

[6] Shaw, E., Ruby, K., and Post, J., "The insider threat to information systems: The psychology of the dangerous insider", Security Awareness Bulletin, vol. 2, no. 98, pp. 1-10, 1998.

[7] Amichai-Hamburger, Y., and Vinitzky, G., "Social network use and personality", Computers in Human Behavior, vol. 26, no. 6, pp. 1289-1295, 2010.

[8] Kandias, M., Stavrou, V., Bosovic, N., and Gritzalis, D., "Predicting the insider threat via social media: The YouTube case", 12th Workshop on Privacy in the Electronic Society, 2013.

[9] Gritzalis, D., Theoharidou, M., and Kalimeri, E., "Towards an interdisciplinary InfoSec education model", 4th IFIP World Conference on Information Security Education, pp. 22-35, 2005.

[10] Theoharidou, M., Kotzanikolaou, P., and Gritzalis, D., "Risk assessment methodology for interdependent critical infrastructures", International Journal of Risk Assessment and Management, vol. 15, no. 2, pp. 128-148, 2011.

[11] Coles-Kemp, L., and Theoharidou, M., "Insider threat and information security management", Insider Threats in Cyber Security, Springer, pp. 45-71, 2010.

[12] Kandias, M., Virvilis, N., and Gritzalis, D., "The insider threat in Cloud computing", 6th International Conference on Critical Infrastructure Security, Springer, pp. 95-106, 2011.

[13] Kandias, M., Mylonas, A., Theoharidou, M., and Gritzalis, D., "Exploitation of auctions for outsourcing security-critical projects", 16th IEEE Symposium on Computers and Communications, IEEE, pp. 646-651, 2011.

[14] Magklaras, G., and Furnell, S., "Insider threat prediction tool: Evaluating the probability of IT misuse", Computers and Security, vol. 21, no. 1, pp. 62-73, 2011.

[15] Kandias, M., Mylonas, A., Virvilis, N., Theoharidou, M., and Gritzalis, D., "An insider threat prediction model", 3rd International Conference on Trust, Privacy and Security in Digital Business, Springer, pp. 26-37, 2010.

[16] Greitzer, F., Kangas, L., Noonan, C., Dalton, A., and Hohimer, R., "Identifying at-risk employees: Modeling psychosocial precursors of potential insider threats", 45th Hawaii International Conference on System Science, IEEE, pp. 2392-2401, 2012.

[17] Brdiczka, O., Liu, J., Price, B., Shen, J., Patil, A., Chow, R., and Ducheneaut, N., "Proactive insider threat detection through graph learning and psychological context", 33rd IEEE Symposium on Security and Privacy, IEEE, pp. 142-149, 2012.

[18] FBI, "The Insider Threat: An introduction to detecting and deterring an insider spy", 2012. http://www.fbi.gov/about-us/investigate/counterintelligence/ the-insider-threat.

[19] Kandias, M.., Galbogini, K., Mitrou, L., and Gritzalis, D., "Insiders trapped in the mirror reveal themselves in social media", 7th International Conference on Network and System Security, Springer, pp. 220-235, 2013.

[20] Kandias, M., Mitrou, L., Stavrou, V., and Gritzalis, D., "Which side are you on? A new Panopticon vs. privacy", 10th International Conference on Security and Cryptography, SciTekPress, pp. 98-110, 2013.

[21] Borthakur, D., The hadoop distributed file system: Architecture and design, O'Reilly, USA, 2007.

[22] Abril-Sánchez, P., Levin, A., and Del Riego, A. "Blurred Boundaries: Social Media Privacy and the 21st Century Employee", American Business Law Journal, vol. 49, no. 1, pp. 63-124.

[23] Solove, D., A taxonomy of privacy, University of Pennsylvania Law Review, vol. 154, no. 3, pp. 477, 2006.

[24] Henderson, S., "Expectations of Privacy in Social Media", Mississippi College L. Rev., vol. 31, pp. 227, 2012.

[25] Castelluccia, C., Druschel, P., Hübner, S., Pasic, A., Preneel, B., and Tschofenig, H., "Privacy, accountability and trust: Challenges and opportunities", Technical Report, ENISA, 2011.

[26] Nissenbaum, H., "Privacy as Contextual Integrity", Washington Law Review, vol. 79, pp. 119-157, 2004.

[27] Dumortier, F., "Facebook and Risks of "de-contextualization" of Information", Data Protection in a Profiled World, Springer, pp. 119-137, 2010.

[28] Mylonas, A., Tsoumas, B., Dritsas, S., and Gritzalis, D., "A secure smartphone applications roll-out scheme", 8th International Conference on Trust, Privacy and Security in Digital Business, Springer, pp. 49-61, 2011.

[29] Cas, I., "Ubiquitous Computing, Privacy and Data Protection: Options and limitations to reconcile the unprecedented contradictions", Computers, Privacy and Data Protection: An Element of Choice, Springer, pp. 139-170, 2011.

[30] Mylonas, A., Meletiadis, V., Tsoumas, B., Mitrou, L., Gritzalis, D., "Smartphone forensics: A proactive investigation scheme for evidence acquisition", 27th International Information Security and Privacy Conference, Springer, pp. 249-260, 2012.

[31] Soupionis, Y., and Gritzalis, D., "Audio CAPTCHA: Existing solutions assessment and a new implementation for VoIP telephony", Computers & Security, vol. 29, no. 5, pp. 603-618, 2009.

[32] Theoharidou, M., Mylonas, A., and Gritzalis, D., "A risk assessment method for smartphones", 27th IFIP International Information Security and Privacy Conference, pp. 428-440, Springer, 2012.

[33] Gritzalis, D. (Ed.), Secure Electronic Voting, Kluwer Academic Publishers, USA, 2003.

[34] Gritzalis, D., "Embedding privacy in IT applications development", Information Management & Computer Security Journal, vol. 12, no. 1, pp. 8-26, 2004.

[35] Lekkas, D., and Gritzalis, D., "Long-term verifiability of the electronic healthcare records' authenticity", International Journal of Medical Informatics, nol. 76, nos. 5-6, pp. 442-448, 2006.