# Building an Arabic Social Corpus from Suspicious Content: Collection and Annotation Guidelines

Amal Rekik[1,2], Hanen Ameur[1,2], Amal Abid[1,2], Atika Mbarek[1,2], Wafa Kardamine[2], Salma Jamoussi[1,2] and Abdelmajid Ben Hamadou[1,2]

[1] Multimedia InfoRmation systems and Advanced Computing Laboratory, MIRACL
[2] Digital Research Center of Sfax DRCS, Sfax, Tunisia
{rekik.amal91, ameurhanen, abidamal90, mbarek.atika91, jamoussi}@gmail.com
wafabensaid2010@hotmail.com
abdelmajid.benhamadou@isimsf.rnu.tn

**Abstract.** Social networks are considered today as revolutionary tools of communication that have a tremendous impact on our lives. However, these tools can be manipulated by vicious users namely terrorists. The process of collecting and analyzing such profiles is a considerably challenging task which has not yet been well established. For this purpose, we propose, in this paper, a new method for data extraction and annotation of suspicious users from social networks threatening the national security. Our method allows constructing a rich Arabic corpus designed for detecting terrorist users spreading on social networks. The amendment of our corpora is ensured following a set of rules defined by a domain expert. All these steps are described in details, and some typical examples are given. Also, some statistics are reported from the data collection and annotation stages as well as the evaluation of the annotated features based on the intra-agreement measurement between different experts.

**Keywords:** Data collection, Annotation guidelines, Social networks, Suspicious content, Terrorist users, Arabic social corpus.

## 1 Introduction

Social media have invaded our daily life providing easy tools for users to express their personal opinions and exchange information from all over the world. These networks seem to be tremendous means of communication with their ability to reach a large number of internet users. However, this heavy power of communication can easily turn destructive with the presence of malicious profiles; in other words, its use can go beyond the simple exchange of information to become a means of propaganda and recruitment of jihadists around the world. Actually, malicious users on social networks can use short messages mentioning suspicious words to target specific events. Thus, several attacks can be planned through suspicious profiles that aim to

disseminate a particular agenda via creating groups adhering to their networks. For these reasons and more, nowadays, this field attracts many researchers who try to tackle this challenging issue by mining social data [1, 2].

In the literature, very few studies dealt with such terrorist data [3, 4], mainly due to the lack of resources like abnormal profile information and labeled corpus. This field is still in his early stages, and it has not been yet well established. Otherwise, despite the shared terroristic content on social networks is more likely in Arabic than in other languages, these sources endure of a big vacuity in the literature. Indeed, collecting this kind of data seems very difficult since terrorist users often try to trick others and conceal their malicious intents. To do that, researchers generally exploit intelligent tools. These tools require well annotated data. Hence, it seems very important to collect and annotate data following a set of rules defined by a domain expert.

In this context, we propose a new methodology for collecting and annotating suspicious textual data from several social media sites. Therefore, it is necessary to target suspicious content and understand the behavior of extremist users to protect national security by analyzing Arabic terrorist content especially which is adopted by ISIS. In fact, the strength of our data collection methodology resides in the proposition of a unification model which combines different structures of social media. Our data annotation guideline is the first of its kind in this field. At this stage, we refer to a sociologist who is a domain expert that plays a lead role to analyze and extract knowledge from user profiles. Evidently, our method ensures the construction of the first Arabic corpora containing terrorist users' information spreading on social networks.

The reminder of this paper is planned as follows: the next section is dedicated to describe the methodology used during the data collection to create a corpus for suspicious content. In section 3, we explore the annotation guidelines in further details. Section 4 reports the findings of some statistics obtained during these major steps and the evaluation of the data annotation. Finally, section 5 concludes the paper.

## 2 Data collection

Over the last decade, social media sites have become popular and diversified. Twitter has recently become among preferred sites for terrorist organizations to disseminate their propaganda [5]. Unfortunately, these organizations also proved their online wicked presence in other social media like Facebook and Youtube. For this purpose, we address these three social media to collect suspicious data. Collecting data task consists in extracting and structuring malicious profile information from different social media sites. To do so, we follow mainly three steps:

### 2.1 Suspicious data collection and abstraction

This step consists in collecting suspicious content from three social media sites (Twitter, Facebook and YouTube) using APIs. We adopt a keyword-based method which focuses on searching data (posts, comments, videos) related to the predefined keywords. We prepare a set incremented of keywords, judged dangerous by an expert,

which are for example: الدولة الإسلامية (Islamic state), أسود الدولة(Tigers of state )… Then, we adopt another strategy based on occurred terrorist events. For this purpose, we selected a set of dangerous attacks which have already occurred in different locations like. Then, for each event, we selected all related posts within the interval of seven days before to one month after. Next, we eliminate media accounts and extract profile information of only active users towards the suspicious content. Then, we transform the extracted textual data from different social media into structured XML file depending on the social network it is extracted from. This file contains two blocs of information: (1) <Head> concerns the personal information of the user, (2) <Body> contains information of user's interactions (shares, comments and likes of the user).

### 2.2 Data unification

Following the data abstraction step, we obtain different XML structures form three social media sites where each one has its own specificities. Thus, to unify the mining treatments using these XML files, we propose to produce a common structure and to modelize a unified XML-tags by concatenating and combining all tags presented in the three generated XML files of each social media at the head level or body level.

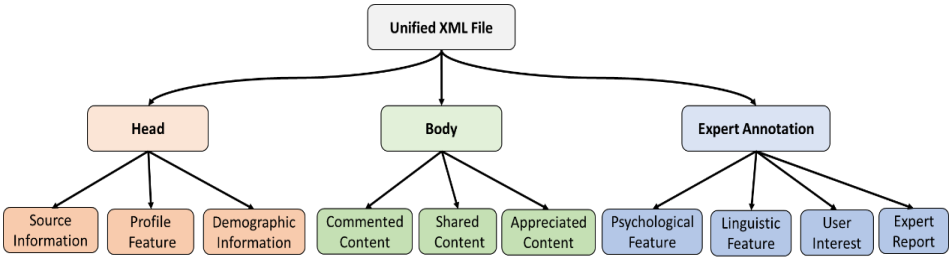


**Fig. 1.** The generic tree of the unified data structure

The unified data structure contains three major parts, as shown in Figure1:

- The head part: presents source information (user id, user name...), demographic information (age, gender...), and profile features (profile description, skills…)
- The body part: contains shared content (post information e.g. title, description...), commented content (comment text and information of commented post), and appreciated content (post information and user's reaction such as like, dislike...).
- Expert annotation part: contains the mined information by the expert namely the user interests, some psychological features (violence and terrorism class), his linguistic features (native land, user native Arabic and user dialect language) and some expert notes and remarks.

## 3 Data Annotation

During the data annotation step, expert annotators use generally different annotation schemes to fulfill their goals [6, 7]. In fact, during this stage, we refer to a sociologist as an expert who plays a key role in analyzing and extracting knowledge from users'

profiles. The sociologist has mainly two objectives. The first consists in verifying the available data and completing the missing ones (Demographic information e.g. location, age and gender). The second objective aims at interpreting and analyzing the available and obtained data in order to deduce for each user his psychological features, linguistic features, user interests, and give some observations and reports. For this purpose, the sociologist follows his intuition, also, he adopts a strategy that is based on the content analysis and which consists on several steps. First, for each user, he focuses on their commented, shared and appreciated contents. Then, he verifies if he is tricking with false information to hide his genuine personality. In fact, terrorists in social networks spread within communities. Therefore, when the expert deduces from the user profile content that he is a dangerous one, he proceeds to focus on his relationships which can lead to reveal other dangerous users within the same terrorist community. These profiles will be considered as potential malicious users that also have to be analyzed. In addition, to emphasis his required role, the expert analyzes the list of the extracted profiles based also on the concepts used by each user. More specifically, the strategy adopted by the sociologist is based on two types of analysis: qualitative and quantitative [8]. The qualitative analysis aims at understanding the user's beliefs from the meaning of the used concepts. Meanwhile, the quantitative analysis consists in studying the user's behavior relying on the frequencies of the used keywords. Indeed, regarding the unified data structure mentioned in figure 1, we have in overall 3 types of information:

- Information to be extracted and analyzed without any treatment: source information and profile feature from the head part as well as the entire body part.
- Information to be verified or completed: demographic information of the head part.
- Information to be mined and inserted by the expert: the expert annotation part.

So, during the annotation step, the expert will focus on the two last categories of information which are: (1) demographic information for verifying or completing personal data content and (2) annotated information for completing implicit data.

### 1.1 Demographic information

**Location.**

Users on social networks may not always declare their location or may aver incorrect geographic location data. For this purpose, the sociologist aims to verify or identify the user's location through different ways. To do that, he can rely on his current shared contents. For example when the user broadcasts a video commented by "Live from Syria", this can demonstrates implicitly that he is located in Syria. Moreover, the expert can even deduce the user's location explicitly based on his shared status content. He focuses also on the areas that the user is interested in his shared content including topics that are specific to certain locations (see table3).

**Table 1.** Location annotation based on some examples

| Examples | Translation | Location |
|---|---|---|
| Share a photo with description:<br>احييكم من هنا العاصمة #الجزائر | I greet you from the capital #Algeria | Algeria |

| | | |
|---|---|---|
| Explicit information from status:<br>الطقس عندنا بارد في فرنسا | The weather is cold here in France. | France |
| User interests:<br>مهرجان سوق واقف، الخطوط الجوية القطرية | Souq waqif festival, Qatar airways, | Qatar |

**Age.**

Generally, user's information about age is not available due to the social networks APIs restrictions or due to the user's choice, as it can also be tricky. Therefore, it is almost impossible to extract the exact age of the user. For this purpose, the sociologist aims to extract this information implicitly. However, the expert cannot assign an exact age value for each user. Therefore, he proposes three ranges of age to verify or to complete this information namely: teenager, adult and old. This range is determined by analyzing the user civil status (e.g. married, having children, having grandchildren...) as well as his professional status (e.g. student, employee, retired…) declared throw his shared content. Furthermore, the expert relies on the recent pictures of users and their declared age (if it exists). Table 2 illustrates some status examples that the expert relies on to identify the age range of the user:

**Table 2.** Age annotation based on status examples

| Examples | Translation | Age |
|---|---|---|
| أبحث عن تربص لمشروع ختم الدروس | Searching for an internship for the ESP | Teenager |
| دعواتكم لابنتي بالشفاء | Prayers for my daughter to heal | Adult |
| حفيدتي الغالية | My precious granddaughter | Old |

**Gender.**

In order to verify or complete the user's gender, the sociologist is based mainly on verifying the personal pronouns, the adjectives and the conjugation of Arabic verbs. In fact, this strategy is very adaptable since it is a basic rule in the case of the Arabic language. Indeed, when the user is a female, the pursuant present verb will be inevitably appendix by T. of femininity (ة). This rule is applicable for each adjective of the Arabic language. Furthermore, the expert can also identify the user's gender based on his photo. He can even verify this field based on the user's interests when this later is concerned only by an association of interests which are specific generally to solely feminine or masculine user. For example, topics concerning football, sports channels or mechanics interest mostly men. Similarly, women are interested, for example, in makeup, dresses and nails designs more than men. Table 6 contains some examples and their suitable gender annotation assumed by our expert:

**Table 3.** Gender annotation based on status examples

| Examples | Translation | Gender |
|---|---|---|
| موش فاهمة | I do not understand | Feminine |
| إني مهتم لمستقبل بلادي | I am interested in the future of my country | Masculine |

### 3.1 Expert annotation

During this step, the sociologist is originally familiar with the available data in order to make the appropriate decisions for the creation of the corpus. Based on the body information which include the commented content, shared content and appreciated content, the expert aims at mining four types of labels as follow:

**Psychological features.**

During the psychological features mining step, the sociologist aims mainly at extracting the violence and terrorism class of each user. In fact, these fields are considered as the most important features for identifying malicious users for cyber-security requirements.

*User violence*

Users on social networks can be terribly malicious, violent and vindictive in their shared content that can express for example the harassment, or encouragement to fight and Jihad. In fact, this violent content, which is transmitted either through pictures, videos or sentences, is an inference to the degree of violence embedded in the user's thoughts and intentions. For this purpose, mining such feature is very important to deeply understand the beliefs and behaviors of each user especially in the case of the cyber security requirements. So, for each user, the sociologist aims at identifying either the user is violent or not. This can be determined by analyzing the harmful scenes contained in the transmitted photos and videos. Also, the expert can mine the user's violence by focusing on the used lexicon and basing on the quantitative analyze. This later aims chiefly at focusing on the frequencies of the used vulgar concepts. These violent concepts can be for example: ذبح (slaughter), قتل (killing), الدعس (tread), القصف (bombing)، قطع رأس (Cut off the head),

*User class of terrorism*

Based on the content analysis, the sociologist can also determine the degree of terrorism for each user. This feature is considered as the most important factor since our main objective is to determine the terrorist users on social networks and to analyze their behaviors. The degree of terrorism can be categorized into three main classes as follow:

Not viable: A user can be identified as not viable to be a terrorist in two cases:

- If he doesn't have any attitude concerning the terrorism.
- If he reviles the terrorists and blackguard all the terrorism acts all over the world.

Viable: A social network user can be identified as viable to be a terrorist if:

- He is interested in terrorists and shares all their novelties with a delicate tone.
- He is sympathetic with the terrorism acts and admires their attacks but secretly. Here the expert's intuition interferes so as to reveal the user's viability.
- He has a disordered and fragile personality which can explain his vulnerability to terrorists. In fact, the sociologist deduces that a user is disordered when he is contradictor in his point of views and opinions about a given subject over the time as well as he can also describe his disorder state in his shares.

Terrorist: A user is defined as terrorist if one of the following criteria is mentioned:

- If the user belongs to the terrorist community but he is not present with them on the spot. He helps them remotely. He favors and facilitates the terrorists' tasks either logistically or by the diffusion of information.
- If he is the leader of the terrorist community. In fact, a user is defined as a leader if he is the supervisor of the group of terrorists. He assumes the responsibility to make the decisions and fix the goals. He guides the terrorists, makes connection between them and provides tips to organize and motivate them. In addition, he is frequently using his subject personal pronoun (أنا) since he is tyrant.
- If he frequently uses terrorism concepts (e.g. الاستشهاد martyrdom, الجهاد Jihad…) and he performs any terrorist acts within the community other than the massacres.
- If he is the member who is charged of carrying out the horrific massacres. He often shares very violent content including slaughter acts, torture scenes as well as horrible videos and pictures. He is able to move to the act of non-humanitarian tasks. They share different signs and codes that are understood only by them. The decipherment of these signs reveals dangerous plans like attacks and slaughters.
- If the user once belonged to the terrorists and then left them. Most of them use generally a hidden name due to their fear or mental disorder. They express their regrets and shame on their shares.

The following table contains some examples of the shared content of each user depending on his degree of terrorism:

**Table 4.** Terrorism class annotation based on status examples

| Examples | Examples translation | Terrorism'class |
|---|---|---|
| داعش أينما حلت حل الخراب والدمار قلبي يعتصر ألما وحزنا على ضحايا الإرهاب | Wherever there is ISIS, there is ruin and destruction. My heart is in great pain and sorrow for the victims of terrorism. | Not viable |
| مقتل مرتزق بالمعارك ضد الدولة الإسلامية | Murder of a mercenary against the Islamic state | Viable |
| #غزوة_تونس_المباركة# غزوة_تونس #تونس_الخلافة# غزوة_متحف_باردو# الدولة_الإسلامية 75 72 96 66 | #Blessed_Tunis_Battle #Tunis_Battle #Tunis_Caliphate #Bardo_Meseum_Battle #State_Islamic 75 72 96 66 | terrorist |

**Linguistic features.**

These features refer to:

- Native land: Refers to the place where the user was born.
- Arabic language: requires verifying if the user's native language is Arabic or not.
- Dialect language: Refers to the user's spoken language.

These features are closely related to each other. For this reason, the sociologist, first, performs an overview on the user's profile to detect the used language based on its shared status. Then, he verifies the user's declared land, his native land (if it is available) and his used dialect (from his shared contents). At this point, the sociologist is based mainly on his expertise on the different Arabic dialects to distinguish the specific country-dialect. To do so, he focuses on the used accent and

lexicon and more specifically the words which are specific to each dialect and exclusively used by the people of a certain country. The following table represents examples of status on which the expert is based in order to extract the native country of the user and its dialect:

**Table 5.** Location annotation based on status examples

| Examples | Examples translation | Native land | Dialect |
|---|---|---|---|
| مشهد عنيف وايد | A very violent scene | Kuwait | Kuwaiti |
| كيف نشفى من حب بلادي #تونس ـ المزيانة | How to heal from the love of my country #beautiful_Tunisia | Tunisia | Tunisian |

Moreover, to extract the native land, the expert focuses even on the user's interests that can concern, for example, Morocco, Tunisia… Thus, he can extract in most cases his dialect and his native land. From these observations, the expert can identify if Arabic is the user's mother tongue or his second language which he may miss spell. Thus, he can conclude if the native language of the user is the Arabic or not.

**The user interests.**
During the annotation step, the sociologist intends also to extract the user's interests. For this purpose, the expert defines from his observations a generic and unlimited list of interests which can be for instance: Religion, health, politic, sport, artistic… This list can evolve every time when the user is interested in a new topic. In fact, each topic of interest has a set of sub-topics defined by the expert (e.g. Sport: Handball, Basketball, Football…)

**The expert report.**
The last task affected by the sociologist consists in redacting an expert report containing the terrorist and violent keywords which reflects the terrorism and violence of each user. Moreover, this expert report contains the dangerous content which the sociologist relies on to detect the terrorism and violence degree of the user. After content analysis step, the expert can also detect fake profiles and irony data.

## 4 Statistics and evaluation

In this section, we discuss the statistic results of our proposed method during the data collection and annotation steps. So, in order to perform our data collection methodology, we used different APIs where each one is specific to a social media network and has its own restrictions. In fact, the implementation of these APIs requires a set of development tools to automate their main tasks. For this purpose, we have used the R and Java languages to efficiently collect our data. Furthermore, we have evaluated the performance of our data annotation by referring to other experts in order to estimate their inter-annotation agreement with our sociologist.

### 4.1 Data collection statistics

Using a set of APIs, we came to collect several data from the different social media networks. Table 6 contains some statistics about the collected data concerning the head part information which are declared by users. Some fields namely the Female and Male number in Twitter and friends' number on Facebook are not filled due to the restrictions of the social networks API. These statistics are as follow:

**Table 6.** Statistics about the collected data

| | Twitter | Facebook | Youtube |
|---|---|---|---|
| Accounts number | 992 | 550 | 854 |
| Posts or comments /account | 1087 | 6 | 112 |
| Friends/account | 36623 | NA | 3271 |
| Female number | NA | 99 | 17 |
| Male number | NA | 451 | 241 |

After the elimination of the Medias' accounts, we have preserved 992 users from Twitter, 550 channels from Facebook and 854 accounts from Youtube. Otherwise, as we mentioned before, the Facebook API prevents the extraction of the posts shared by the users. For this purpose, during the Facebook accounts analysis, we considered the comments of each user on the posts of potential suspicious pages and groups as his shared content. In fact, these comments can reflect effectively the user intention with the same manner as its shared posts. Furthermore, the Facebook API also precludes the extraction of all the personal information of each collected user namely his gender. For this purpose, we have used the GenderizeR package [9] on the RStudio framework to identify the gender of each user. This package uses genderize.io Application Programming Interface to predict gender from first names extracted from text corpus. These information will be verified by the expert during the annotation step in order to preserve only the reliable information. Moreover, we can also conclude from the table 3 that the female accounts' number of the collected data is very low comparing to the number of the male accounts' number. This gap is justified by the nature of the topic that we are targeting during the data collection which concerns mostly men (terrorism, violence...). Otherwise, the percentage of the missing information on all the social networks is high. Hence, the utility of the annotation step is strongly raised in order to enrich our corpus.

### 4.2 Data annotation

The potential terrorist profiles collected from Twitter, Facebook and Youtube have several missing information due to the APIs restrictions as well as users' concealing. That is the reason why the data annotation step is very crucial to construct our data corpora. For the fulfillment of this step, the expert completed these missing information, that he has no doubt about them, as well as verified the existing information of a set of profiles collected from Twitter. In fact, the data annotation attainment is a very hard task that consumes much effort and requires lots of time, due to which, the expert would rather annotate in the beginning just 490 profiles. Those

annotated data will be considered as the core of our learning step. Analyzing this core, we will have the opportunity to make the decision to continue on a semi-supervised learning or to perform an active annotation. This active annotation consists mainly to select and give a set of samples to the expert for annotation. Thus, in the following sub-section, we will report some statistics about the annotated data constructing our learning core:

**Annotated data statistics.**

Having the annotated data of the extracted profiles, we will report a set of statistics that describe our data corpus collected from Twitter. The following sectors contain some of these statistics on a set of features.

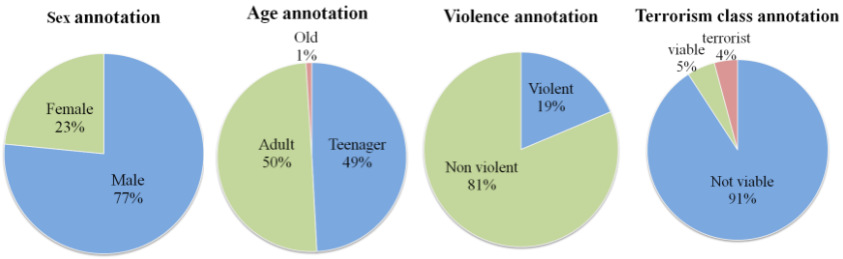


**Fig. 2.** Statistics about the sex, age, violence and terrorism class annotation

**Table 7.** Matching between violence and terrorism

| | Not viable | Viable | Terrorist |
|---|---|---|---|
| Violent | 59 | 12 | 17 |
| Non Violent | 376 | 12 | 3 |

The analyzed profiles extracted from Twitter are retrieved mainly based on the active users in terms of sharing content concerning terrorist attacks, as well as containing ISIS-specific words. In fact, females are generally not very concerned with this topic as much as it is a topic of interest for several men. So, it is very reasonable that the number of collected accounts which are of females is slight ahead comparing to the male' accounts number since this target topic of research is explicit mostly for men. Furthermore, we can notice from the second sector of figure 2 that the number of old users is very small comparing to the number of other users. This statistic is justified by the fact that active users on social networks are generally either teenagers or adults. In addition, most of the collected users are not violent and not viable to be terrorists, 5% of users are viable to be terrorists and 4% are terrorists. These percentages can be explained by the fact that most of users who are active in sharing content about terrorist attacks are reviling the terrorists and blackguarding the terrorist attacks. On the other hand, the low number of terrorist users retrieved by our method motivates us to use the semi supervised learning in order to avoid the loss of time. In addition, we reported the statistics of the matching between violence and terrorism. Therefore, we can note from table 7 that most users who are not viable to be terrorists

are as well not violent. However, users who are viable to be terrorists can be violent as well as not violent. In addition, we notice that terrorist users are mostly violent. Thus, there is a high dependence between these two features.

**Evaluation:**

In order to evaluate the performance of our annotation step, we conducted an experiment which is based on the resort to an annotator other than our expert for the aim of estimating the inter-annotation agreement. This agreement rate is dependent on the number of information having the same annotation by the expert and our sociologist on a test corpus composing of 30 profiles from Twitter. To do so, we resorted to the use of the Cohen's kappa coefficient (κ) as a statistic measurement of inter-rater agreement for different annotators. The Cohen's kappa coefficient is measured using the following formula [10]:

$$k = \frac{Po - Pc}{1 - Pc} \quad (1)$$

Where Po is the proportion of observed agreements and Pc is the proportion of agreements expected by chance.

In order to evaluate the annotation of our sociologist for each user, we should assess its annotation on different features namely: Range of age, sex, degree of violence and degree of terrorism. These features are also annotated by 3 experts other than our sociologist. The total intra-annotation agreement between our sociologist and each expert is measured using the tool of the intra-annotation agreement in each feature using the Cohen's kappa coefficient as follow:

$$k_{Total} = \frac{\sum_{i=1}^{nbrExpert} k_{expert_i}}{nbrExpert} \quad (2)$$

Where nbrExpert is the number of the experts and $k_{expert_i}$ is the Cohen's kappa coefficient between our sociologist and the $expert_i$. The obtained results are described in the following table:

**Table 8.** The Cohen's kappa coefficient for each feature

| Feature | Kappa expert 1 | Kappa expert 2 | Kappa expert 3 |
|---|---|---|---|
| Age | 0.8 | 0.89 | 0.79 |
| Sex | 0.63 | 0.86 | 0.63 |
| Violence | 0.8 | 0.86 | 0.86 |
| Terrorism class | 0. 76 | 0.8 | 0.8 |
| Average kappa | 0.75 | 0.85 | 0,77 |

We obtain as result the Cohen's Kappa coefficient between the 4 experts is 0.79. Thus, based on the table proposed by Landis and Koch [11], we can conclude that this intra-agreement between the annotators is rather strong since it is between 0.6 and 0.8. This result demonstrates the efficiency of our sociologists which proves the credibility of our annotated corpus.

## 5 Conclusion

In this paper, we presented our data collection and annotation guideline steps. We have mainly addressed the domain of cyber-security as a very sensitive domain. In this matter, social media sites have become the suitable space for terrorist groups to spread their radical ideas. For this purpose, we collected suspicious content from several online social media. Our corpus is annotated by an expert who defines annotation guidelines. Furthermore, we provided some statistics about the collected data as well as the annotated data adopted from user's profile. In addition, we resorted to 3 annotators other than our expert for the aim of estimating their inter-annotation agreement. The results highlight the strong agreement between the experts which demonstrate the efficiency of our sociologist.

**ACKNOWLEDGMENT**

This publication was made possible by NPRP grant #9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

**References**

1. Jaffali, S., Ameur, H., Jamoussi, S., & Hamadou, A. B.: Glio: A new method for grouping like-minded users. In Transactions on Computational Collective Intelligence XVIII, 44-66, Berlin (2015).
2. Rekik, A., and Jamoussi, S: Deep Learning for Hot Topic Extraction from Social Streams. In: International Conference on Hybrid Intelligent Systems, 186-197, Morocco (2016).
3. Stefanidis, A., Crooks, A., and Radzikowski, J.: Harvesting ambient geospatial information from social media feeds. GeoJournal 78(2), 319-338 (2013).
4. Abid, A., Ameur, H., Mbarek, A., Rekik, A., Jamoussi, S., and Hamadou, A. B.: An extraction and unification methodology for social networks data: an application to public security. In: International Conference on IIWAS, 176-180, Austria (2017).
5. Klausen, J.: Tweeting the Jihad: Social media networks of Western foreign fighters in Syria and Iraq. Studies in Conflict & Terrorism 38(1), 1-22 (2015).
6. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, A., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.: Part-of-speech tagging for twitter: Annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 42-47 (2011).
7. Nakov, P., Rosenthal, S., Kiritchenko, S., Mohammad, S. M., Kozareva, Z., Ritter, Stoyanov,V., Zhu, X.: Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. Language Resources and Evaluation 50(1), 35-65 (2016).
8. Shwartz, H., & Jacobs, J.: Qualitative and quantitative methods: two approaches to Sociology. Qualitative sociology. A Method to the madness", (1979).
9. Wais, K., VanHoudnos, N. and Ramey, J. 2016. Package 'genderizeR'. https://cran.r-project.org/web/packages/genderizeR/index.html, last checked on 11/05/ 2016.
10. Sim, J. and Wright, C.: The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical therapy 85(3), 257-268 (2005).
11. Landis, J. R. and Koch, G. In: The measurement of observer agreement for categorical data. biometrics, 159-174 (1977).