Chapter 11
Analysis of Suspended Terrorism-Related Content on Social Media

George Kalpakis, Theodora Tsikrika, Ilias Gialampoukidis, Symeon Papadopoulos, Stefanos Vrochidis, and Ioannis Kompatsiaris

Introduction

Several popular social media platforms that emerged during the past decade have revolutionized modern communications among people worldwide cutting across different nationalities, cultures, and residences, and have resulted in the development of online communities providing the means for the open sharing of information. However, due to their broad reach, social media are also being used with subversive intentions. For instance, in recent years they have been employed by terrorist and extremist groups for further supporting their goals of spreading their propaganda, radicalizing new members, and disseminating material targeting potential perpetrators of future attacks. Therefore, online social networks present a digital space of particular interest to governments, law enforcement agencies and social media companies in their effort to suppress terrorism content.

Identifying terrorism-related content in social media is a challenging task. Social media platforms host overwhelming amounts of discussions, posted daily by millions of people, making it practically impossible to detect extremist or terrorism-related content by solely relying on manual inspection performed by content moderators. Therefore, an interesting research question is whether the analysis of social media content can result in identifying multiple complementary weak

G. Kalpakis · T. Tsikrika · I. Gialampoukidis · S. Papadopoulos · S. Vrochidis (✉)
I. Kompatsiaris
Information Technologies Institute, Centre for Research and Technology Hellas
Thermi-Thessaloniki, Thermi, Greece
e-mail: kalpakis@iti.gr; theodora.tsikrika@iti.gr; heliasgj@iti.gr; papadop@iti.gr; stefanos@iti.gr; ikom@iti.gr

G. Leventakis, M. R. Haberfeld (eds.), *Community-Oriented Policing and Technological Innovations*, SpringerBriefs in Criminology,
https://doi.org/10.1007/978-3-319-89294-8_11

signals revealing the distinctive nature of terrorism-related content, and thus be an additional means towards the automatic or semi-automatic detection and subsequent removal of such content.

In this context, this work aims at analyzing the particular traits of terrorism-related content published on Twitter, a popular channel among terrorist groups, with the goal to distinguish terrorism-related accounts from others. To this end, a dataset of terrorism-related content was collected from Twitter through searches based on terrorism-related keywords provided by domain experts. In our study, we analyzed several textual, spatial, temporal and social network features of the gathered posts and their metadata and compared them against "neutral" Twitter content.

Our study unveiled a number of distinct characteristics of extremism and terrorism-related Twitter accounts and paves the path towards the development of automated tools that aim at leveraging the distinct traits of terrorism-related accounts for the early detection of terrorist and extremist content in Twitter, with the goal of alerting social media companies about the presence of such posts and speeding up the process of removing them from their platforms. This work is particularly timely given the recent pledges both by major social media platforms and governments around the world to step up their efforts towards countering online abusive content (Twitter Public Policy 2017).

Section "Related Work" summarizes related work in the area. Section "Data Collection and Analysis" describes the methodology and the dataset collected for our analysis. Section "Experimental Results" presents the findings of our analysis. Finally, Section "Conclusions" concludes with an outline of future research directions.

Related Work

Related research conducted in the past years has focused on examining the nature of terrorism-related content published by participants in extremist Web forums. Specifically, several research efforts have proposed methods for analyzing extremist Web forums for detecting users representing potential lone wolf terrorists and perpetrators of radical violence (Johansson et al. 2013; Scanlon and Gerber 2014). Additionally, the use of social media by terrorist and extremist groups and the resulting social network perspectives have also been studied. Recent works have examined the use of social media platforms by terrorist groups and organizations (Chatfield et al. 2015; Klausen 2015). Moreover, key player and key community identification in terrorism-related Twitter networks has been addressed through the use of different centrality measures and community detection algorithms (Gialampoukidis et al. 2016, 2017). Complementary to the aforementioned research

efforts, our paper analyzes several textual, spatial, temporal and social network features which, when combined, are capable of characterizing the terrorism-related nature of Twitter accounts.

Data Collection and Analysis

Methodology

Our investigation focused on the Twitter platform given its popularity among terrorist groups as a means for spreading their propaganda and recruiting new members (Klausen 2015). Under the pressure put during the recent years by governments around the world to combat online extremism, Twitter has made significant efforts towards blocking accounts that promote terrorism and violence (Twitter Inc. 2016). In particular, Twitter has been suspending user accounts based on whether they are exhibiting abusive behavior that violates its rules,[1] including posting content related to violent threats, hate speech, and terrorism. To this end, Twitter has suspended 636,000 accounts between August 2015 and December 2016, with more than half of them occurring in the last 6 months of 2016 (Larson 2017).

In this context, we consider that Twitter accounts that have been posting content related to terrorism and are suspended at some point in time represent in principle users who have been exploiting the social media platform for serving their subversive intentions and promoting terrorism in general, e.g., disseminating propaganda, etc. Twitter accounts that have been posting content related to terrorism and have not been suspended are generally considered as users interested in the domain (e.g., posting news related to terrorism attacks), but without subversive intentions. There may of course be cases where non-suspended users also have darker motives and are actively engaging in propaganda and radicalization efforts, but have not thus far been detected so that they can be suspended by Twitter. In this work, we consider that this phenomenon might occur indeed but that it is less likely given Twitter's efforts in this direction.

Our analysis is based on the comparison of various characteristics of suspended Twitter accounts against those of non-suspended accounts. Both types of account post content relevant to the terrorism domain. The goal of our study is to determine the key factors that are capable of providing weak signals for distinguishing among ordinary Twitter users and those with subversive behavior based on the analysis of a variety of textual, spatial, temporal and social network features. The comparison is performed by examining the lifetime of suspended accounts, analyzing user accounts from the social network perspective (i.e. based on their connectivity with other user accounts), and exploiting geolocation information extracted from the textual content of user posts.

[1] https://support.twitter.com/articles/1831

Data Collection

The data for our study were collected using a social media crawling tool (Schinas et al. 2017) capable of running queries on the Twitter API[2] based on a set of five Arabic keywords related to terrorism propaganda. These keywords were provided by law enforcement agents and domain experts in the context of the activities of the EC-funded H2020 TENSOR[3] project and are related to the Caliphate, its news, publications and photos from the Caliphate area.

The crawling tool ran for a 7-month period, and specifically from February 9 to September 8, 2017, collecting tweets relevant to the provided keywords, along with information about the user accounts that published this content. Our dataset consists of 60,519 tweets posted by 33,827 Twitter users, with 4,967 accounts (14.70%) having been suspended by Twitter within this period.

For each tweet in our dataset, we stored its textual content together with relevant metadata, such as its URL address, the language used, its creation date, and the number of likes, shares, comments, and views. Similarly, for each user account having posted at least one tweet within our collection, we have captured its name, username, and creation date along with the number of its friends, followers, items (i.e. the total number of posts), favorites, and public lists that they are a member of.

Additionally, each user account was monitored on a daily basis to determine whether it has been suspended by Twitter. Given that Twitter does not provide information regarding the exact suspension date and time, this was determined based on the latest post published by a suspended account. Finally, after processing the data gathered, we built a social network graph representing the connectivity among Twitter accounts based on user mentions.

Experimental Results

This section presents the findings of our comparison between the suspended and non-suspended Twitter accounts on our dataset.

User Account Lifetime

First, we discuss the suspended Twitter account lifetime (see Fig. 11.1). Their lifetime is determined by computing the difference between the suspension date and the creation date of an account. The majority of suspended accounts (61.26%)

[2] https://dev.twitter.com/

[3] http://tensor-project.eu/

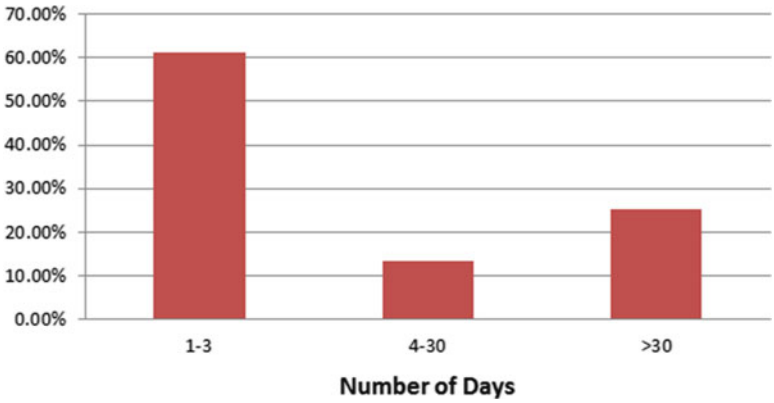

Fig. 11.1 Lifetime of suspended accounts

have very short lifetime, fluctuating between 1 and 3 days, which is explained by the efforts put by Twitter towards removing extremist content the moment it is posted. However, an interesting finding is that a significant portion of the suspended accounts (25.35%) have a lifetime longer than 30 days, which indicates that some accounts manage to evade the monitoring processes of Twitter for longer periods.

Analysis of Mention Networks

The analysis of mention networks formed by users in our dataset provides insights to the comparison between suspended and non-suspended accounts. The connectivity for the two account types differs with respect to the interconnection of accounts of the same type. In particular, 42.22% of the suspended accounts mention other suspended users, whereas 52.66% mention non-suspended accounts (the remaining 5.12% of the suspended accounts mention users are not included in our dataset and hence their suspension status is unknown). On the contrary, only 2.67% of non-suspended accounts mention suspended users, whereas 89.91% are connected with non-suspended users; again, the remaining 7.42% mention users not included in the dataset. This behavior reveals a community-like behavior, where accounts of the same type work together to fulfill their goals.

The connectivity pattern observed on the mention network is illustrated in the suspended to non-suspended mention ratio plot (see Fig. 11.2). The peak observed for mention ratio values fluctuating between 1 and 1.5 in the graph referring to suspended accounts indicates that a significant part of the them is connected to a larger number of suspended than non-suspended accounts, despite the fact that the vast majority of accounts gathered in our dataset are non-suspended users.

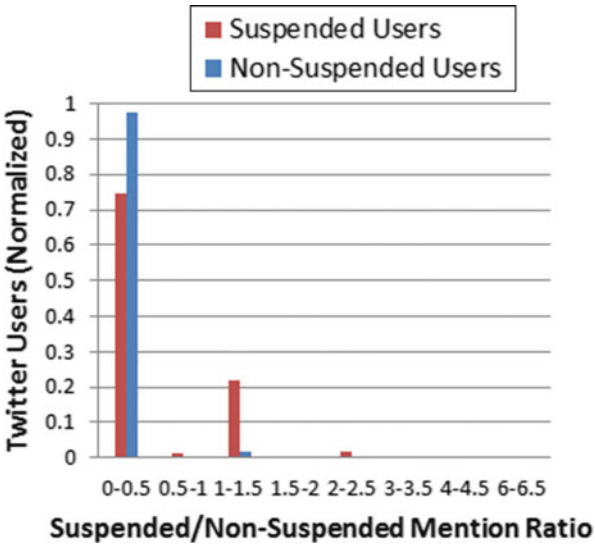

Fig. 11.2 Suspended to non-suspended mention ratio

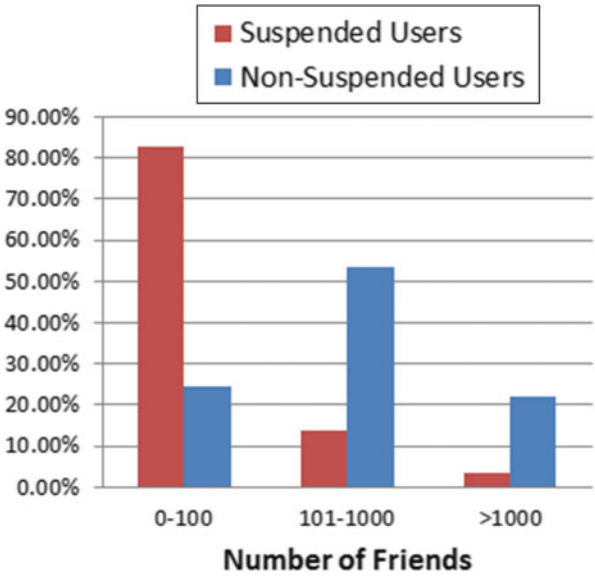

Fig. 11.3 Twitter account friends

Friends and Followers

Figures 11.3 and 11.4 illustrate the distribution of numbers of friends and followers per account type, respectively. In both cases, the vast majority of suspended users have less than 100 friends or followers, which comes in contrast with the connectivity of non-suspended accounts. The short lifetime of terrorism-related accounts (due to their suspension by Twitter) could be a determining factor regarding their number of connections.

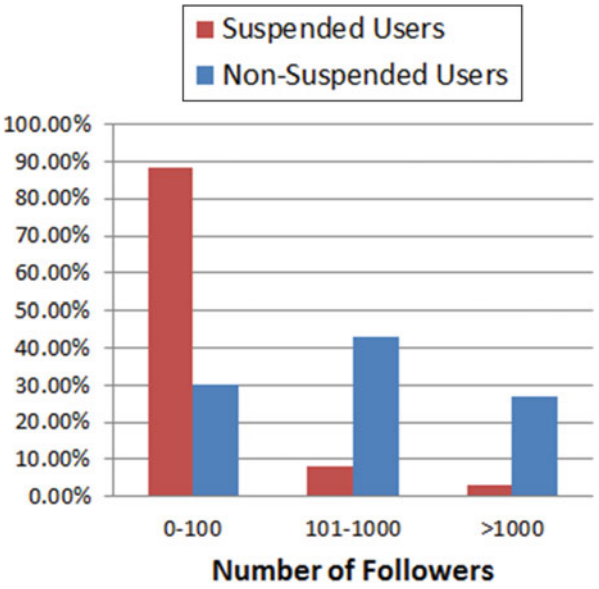

Fig. 11.4 Twitter account followers

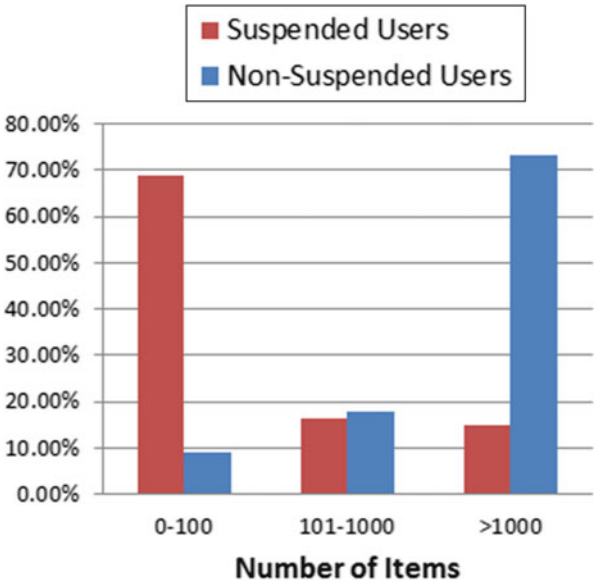

Fig. 11.5 Twitter account posts

Posts, Favorites, and Lists

A similar trend is observed regarding the number of posted items and favorites per account type (see Figs. 11.5 and 11.6, respectively). The number of posts and favorites for the majority of suspended users is less than 100, whereas non-suspended accounts exhibit the inverse behavior.

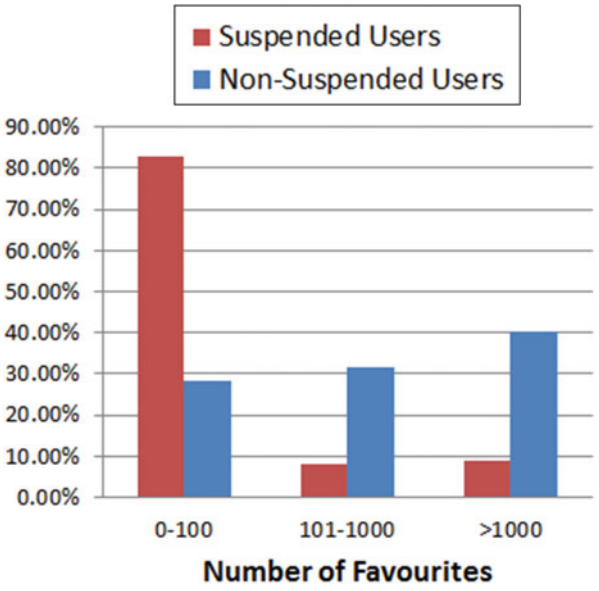

Fig. 11.6 Twitter account favorites

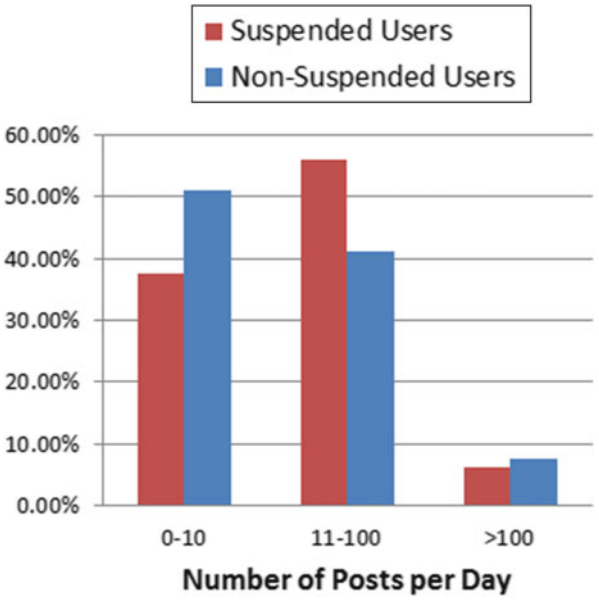

Fig. 11.7 Twitter account post rate

On the other hand, a different behavior is observed regarding the post rate (i.e. the number of posts per day) per account type (see Fig. 11.7). The majority of suspended accounts exhibit a post rate between 11 and 100 posts per day, whereas more than half of ordinary Twitter accounts post less than 10 tweets per day. This indicates that

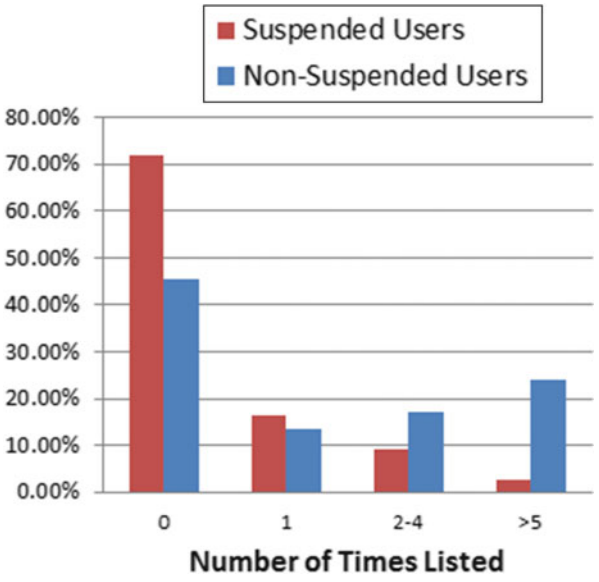

Fig. 11.8 Number of times listed

during their short lifetime, suspended accounts tend to post a relatively large number of tweets, possibly in an effort to disseminate many different pieces of information for spreading their propaganda.

Significant differences are also observed with respect to the number of public lists an account is a member of (see Fig. 11.8). The vast majority of suspended accounts (71.89%) are not a member in any list, whereas more than half of the non-suspended users are part of at least one list (54.48%), with almost one quarter of ordinary Twitter accounts being members in more than five lists.

Spatial Distribution of Accounts

To delve into the spatial distribution of accounts, we performed text-based analysis of the textual content of Twitter posts. We inferred the location of posts, even in cases when it was not explicitly available through the geotagging metadata accompanying a tweet. Geolocation inference from text was based on the approach by Kordopatis-Zilos et al. (2017), which employs refined language models learned from massive corpora of social media annotations. The results of the geolocation extraction for the posts of suspended and non-suspended users are presented in Figs. 11.9 and 11.10 respectively. Given that our dataset is retrieved based on a set of Arabic keywords, the geolocation information extracted for posts produced by both account types refers to countries from the Middle East and Northern Africa, whereas posts coming either from the United Arab Emirates or Syria are mostly associated with suspended accounts.

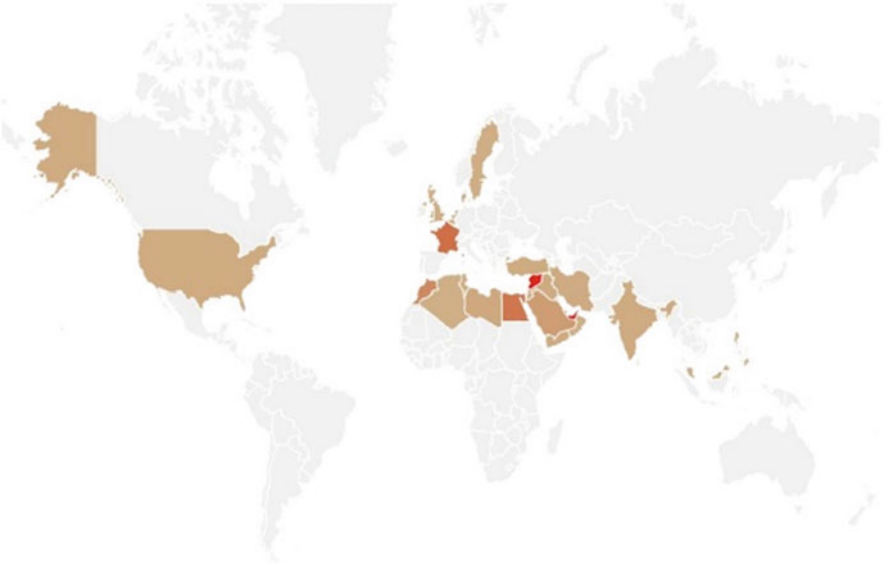

Fig. 11.9 Inferred locations from posts by suspended Twitter accounts

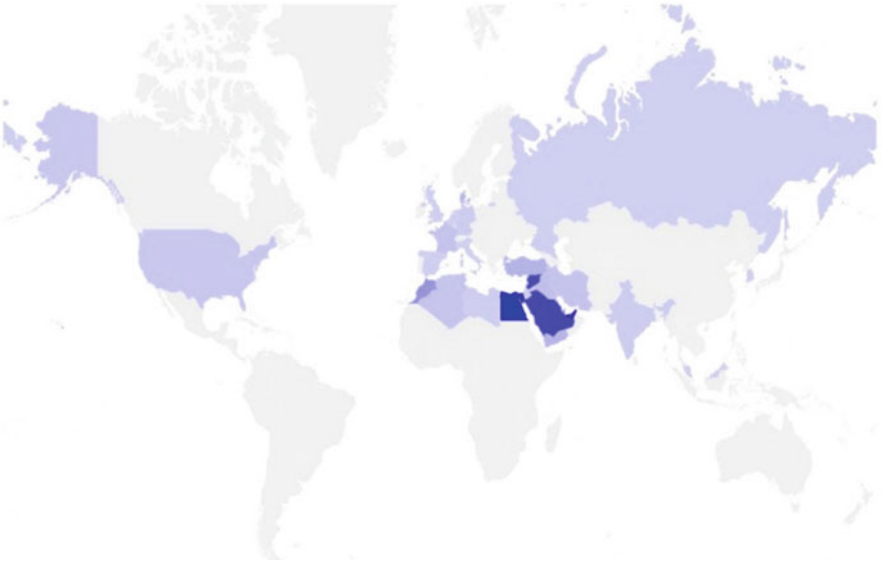

Fig. 11.10 Inferred locations from posts by non-suspended Twitter accounts

Conclusions

This paper aimed at understanding terrorism-related content on social media given their increasing employment by terrorist organizations for spreading their propaganda. We conducted an analysis on terrorism-related content posted on Twitter focusing on the differences between suspended and non-suspended accounts. Our analysis suggests that the traits observed in suspended users are different from non-suspended ones from several different perspectives, namely textual, spatial, temporal and social network features. These findings have the potential to set the basis for automated methods that detect accounts that are likely associated with abusive terrorism-related behavior. To this end, future work includes a more in-depth and large-scale analysis of features presented here, as well as taking into account additional features, including multimedia content such as images and videos.

Acknowledgements This work was supported by the TENSOR project (H2020-700024), funded by the European Commission.

References

Chatfield, A. T., Reddick, C. G., & Brajawidagda, U. (2015). Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks. In *Proceedings of the 16th Annual International Conference on Digital Government Research* (pp. 239–249).

Gialampoukidis, I., Kalpakis, G., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2016). Key player identification in terrorism-related social media networks using centrality measures. In *Intelligence and Security Informatics Conference (EISIC), 2016 European* (pp. 112–115).

Gialampoukidis, I., Kalpakis, G., Tsikrika, T., Papadopoulos, S., Vrochidis, S., & Kompatsiaris, I. (2017). Detection of terrorism-related twitter communities using centrality scores. In *Proceedings of the 2nd International Workshop on Multimedia Forensics and Security* (pp. 21–25).

Johansson, F., Kaati, L., & Shrestha, A. (2013). Detecting multiple aliases in social media. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 1004–1011).

Klausen, J. (2015). Tweeting the Jihad: Social media networks of western foreign fighters in Syria and Iraq. *Studies in Conflict & Terrorism, 38*(1), 1–22.

Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, I. (2017). Geotagging text content with language models and feature mining. In *Proceedings of the IEEE*.

Larson S. (2017). *Twitter suspends 377,000 accounts for pro-terrorism content.*http://money.cnn.com/2017/03/21/technology/twitter-bans-terrorism-accounts/index.html. Accessed 15 Sept 2017.

Scanlon, J. R., & Gerber, M. S. (2014). Automatic detection of cyber-recruitment by violent extremists. *Security Informatics, 3*(1), 5.

Schinas, M., Papadopoulos, S., Apostolidis, L., Kompatsiaris, Y., & Pericles, M. (2017). Open-source monitoring, search and analytics over social media. In *Proceedings of Internet Science Conference.* Springer.

Twitter Inc. (2016). *Combating violent extremism.*https://blog.twitter.com/official/en_us/a/2016/combating-violent-extremism.html. Accessed 15 Sept 2017.

Twitter Public Policy. (2017). *Global internet forum to counter terrorism.*https://blog.twitter.com/official/en_us/topics/company/2017/Global-Internet-Forum-to-Counter-Terrorism.html. Accessed 15 Sept 2017.