Is hybrid AI suited for hybrid threats? Insights from social media analysis

Valentina Dragos, Bruce Forrester, Kellyn Rein

Is hybrid AI suited for hybrid threats? Insights from social media analysis

Valentina Dragos
ONERA - The French Aerospace Lab
Palaiseau, France
valentina.dragos@onera.fr

Bruce Forrester
Defence Research and Development Canada
Quebec City, Canada
bruce.forrester@drdc-rddc.gc.ca

Kelyn Rein
Fraunhofer FKIE
Wachtberg, Germany
kellyn.rein@fraunhofer.fkie.de

***Abstract*—Social media create the opportunity for a truly connected world and change the way people communicate, exchange ideas and organize themselves into virtual communities. Both understanding online behavior and processing online content are of strategic importance for security applications. However, high volumes, noisy data and rapid changes of topics impose challenges that hinder the efficacy of classification models and the relevance of semantic models. This paper performs a comparative analysis on supervised, unsupervised and semantic-driven approaches used to analyze social data streams. The goal of the paper is to determine whether empirical findings support the enhancement of decision support and pattern recognition applications. The paper reports on research that has used various approaches to identify hidden patterns in social data collections where text is highly unstructured, comes with a mix of modalities and has potentially incorrect spatial-temporal stamps. The conclusion reports that the disconnected use of machine learning models and semantic-driven approaches in mining social media data has several weaknesses.**

***Index Terms*—social networks, hybrid AI, defense and security**

I. Introduction

This paper addresses the exploration of the cyber-social space as defined by Shets and colleagues in [1], as a network of humans and autonomous agents and theirs links creating human, artificial or mixed communities.

In the context of this work, hybrid threats are understood as coordinated and synchronized actions that deliberately target vulnerabilities of real world systems (states, institutions), through different means supported by online platforms. For the cyber-physical space, recent examples on disinformation [2] or social manipulation [3] show that phenomena in the cyberspace have the capacity to polarize social views, making social groups form and fracture in online spaces with concrete consequences on real political and social environments.

Social media exploration for security applications presents challenges that are beyond the ability of one domain or discipline to address. From online propaganda and disinformation to influence campaigns, social networks and platforms are often the vectors of intentionally distorted and biased narratives. The lack of social accountability in many digital platforms yields plenty of incentives for unprecedented forms of misuse. Disinformation, propaganda and fake news are just a few examples of ill-uses lurking in this largely accessible technology.

It is then important to have means to mine the cyberspace and analyze its dynamics and especially the emotional contagion [4], online propaganda [5] and spread of extremist ideologies [6].

Effective solutions for cyberspace exploration require synergies from organizational sociology, human computer interaction, communication, information science, and political science to interpret and analyze the evidence. However, the vast majority of solutions are technical and divided mainly on two classes: machine learning techniques that rely upon manually tagged input, and advanced linguistic processing techniques, that in turn require extensive linguistic resources. Only a limited number of solutions take into account social indicators and developing hybrid approaches by combining learning models and semantics is currently the least explored research direction.

This paper outlines main applications of social data analysis, including specific content characterization i.e. online hate, violence and extremism. The paper also discusses several techniques developed for social stream analysis and highlights theirs limitations thanks to three use cases. Finally the paper makes the case for hybrid artificial intelligence solutions to overcome the limitations of existing approaches implemented to interconnect the physical and virtual spaces.

The paper is structured in five sections: section II discusses the main applications of social data analysis. Section III emphasises main the techniques for mining social streams. Section IV illustrates some limitations of current approaches thanks to inconclusive use cases and their causes are discussed in section V. Section VI concludes this paper.

II. Applications of social stream analysis

According to their goal, applications of social data analysis can be classified into two main classes: first, approaches are content-oriented and aim at detecting specific online content in order to gain a better understanding of concepts and notions conveyed. Second, approaches focus on the dynamics of online content, and in this case they identify influent users spreading specific ideas on social media, or virtual communities of users.

A. Analysis of specific online contents

Most of the research studies in the field of social stream analysis keep the distinction between factual and subjective aspects and focus either on topics [7], [8] and narrative detection [9], [10] or subjectivity specific tasks: detecting polarities [11], sentiments [12] or opinions [13]. Both tasks are difficult given that often content and dynamics in such platforms varies among topics as well as in the same conversation.

a) Detection of online hate, violence, and extremism: Detection of specific content aims at identifying online sources conveying specific concepts and ideas, such as hate, violence or extremism. Recent applications are in the field of security, with emphasis on extremist content [14] and users [15], hate [16], propaganda [17], right wing extremism [18] and white supremacy ideology [19].

Amongst the earliest and most popular techniques, Link Based Bootstrapping (LBB) use a semi-automatic approach to detect extremist content released on several modalities, including online discussion forums, websites or blogs [20].

Those methods start with a set of seed URLs collected from authoritative sources (intelligence or security services) and this core list is expanded by adding related or strongly connected URLs. Those approaches rely on a practical hypothesis, stating that websites or forums conveying extreme ideas will sooner or later link to each other and tend to build virtual community. Once enriched, the set of links is again manually analysed and filtered by experts in order to avoid gathering off-topic pages. The list validated is then provided to web crawlers to collect and download content.

The main difficulty of LBB approaches is to access extremist sites hidden on the Dark Web, a region of the cyberspace not indexed by regular search engines, that often comes with additional access restrictions, in the form of memberships requirements or even adversarial detection. To overcome this limitation, results of crawlers are incremented thanks to a manual procedure used to collect extremist forums on the Dark Web [21].

b) Detection of opinions, emotions, sentiments: The analysis of sentiments and opinions falls under the umbrella of a subfield of Natural Language Processing (NLP) field, with the aim of detecting valence, emotions, and other subjective states from text [22]. Solutions were developed for sentiment analysis and opinion detection, although those terms are not accurately defined and are sometimes interchangeable.

The goal of early approaches was to detect polarity in products [23] or service reviews [24], and the analysis was primarily performed at sentence [25] or document level [26]. Solutions developed range from lexicon and ontology-based methods [27] to supervised machine learning and deep learning techniques that were developed in recent years [28].

However, most work in this area focuses on overall opinion detection or sentiment analysis, regardless of the entities, targets or topics mentioned in the content. The algorithms detect sentiments or opinion by assuming a known target.

This disjoint analysis is a limitation of those approaches as meaningful clues hidden in online data are often a combination of topics and subjective aspects and their identification involves analysis of emotions conveyed towards specific topics. To overcome this limitation, several studies addressed the dynamics of emotions or opinions in time [29] as well as detecting the mapping between the emotional categories and linguistic instances [30].

Taking a step further, recent studies by Schoene and de Mel [31] investigate the correlation of topics and emotions while Vijayaraghavan and colleagues addressed the classification of topics along with the sentiment [32].

Although those approaches achieve reasonable accuracy of sentiment analysis, difficulties in processing a mix of multi-domain data and the use of manually tagged inputs have been plaguing their robustness [33].

B. Detection of communities and identification of leaders

Contributions in this category explore the activity of virtual communities, and aims at detecting users exhibiting weak signals, communities producing extremist content and their leaders, who acts as influencers within the network.

An incremental solution to detect extremist users is described in [34]. The study combines content analysis and network flow to detect users releasing extremist content; first, relevant posts are identified using a key-word driven approach and then those posts are used to identify individuals sharing content consistent with extremist views.

Following a similar direction, a behavioral model for extremist users is developed in [35]. The model takes into account attributes of the account, and the users connect with. This model is used to identify new extremist accounts and the empirical validation is done by predicting if they will be suspended for extremist activity.

The dynamics of social influence is investigated in [36] with dynamical activity-connectivity maps, based on network and temporal activity patterns. The authors draw a parallel between propaganda and epidemics spreading, highlighting that information broadcasters and influential supporters generate highly-infectious cascades of information contagion.

Ferrara and colleagues discuss in [37] a machine learning framework leveraging a mixture of metadata, network, and temporal features to detect extremist users, and predict online supporters and interaction reciprocity in social media. Aggregation of users into more complex communities is addressed in [38]. The solution detects communities based on interaction aspects such as follow or followed by, and shows that the categorization goes beyond the dual categorization, with four main types of communities being detected.

Studies based on interactions suffer from several drawbacks. First, the accuracy of virtual links is to be questioned, as by their architecture, certain media like forums might not contain explicit links between their users even though they exist. Those undisclosed relations can further bias the results of analysis undertaken at network level. Secondly, from a practical standpoint, it is difficult to clearly model behavioural patterns or models for online hate, violence or extremism and social models apply to online behaviour only to a certain extent.

III. Techniques for social stream analysis

A. Machine learning techniques

Before discussing machine learning approaches for social data mining we should emphasize that they are not specific to this type of content, but are already used in a variety of text classification tasks. Machine learning methods do not require in-depth linguistic analysis, but rather use a set of specific features such as words, frequencies of words, lexical or syntactical patterns to be associated with text categories. Machine learning methods further are divided into supervised and unsupervised approaches. Supervised algorithms are first trained to identify a set of features from annotated corpora, and then are used to detect them within distinct unseen corpora. Unsupervised solutions perform text classification based on feature analysis, and cluster text according to their similarity in terms of features. More specifically, support vector machines (SVMs), boosted decision trees (BDTs), and latent dirichlet allocation-based (LDAs) methods and strategies have been used in the past to classify political opinions.

Early machine learning approaches constructed a basic binary classifier which used n-grams and part-of-speech features, to assign positive or negative labels to text. Among them, Pak and Paroubek [39] classified tweets as objective, positive and negative by using a sentiment classifier based on the multinomial Naive Bayes, and using a combination of syntactic and linguistic features such as n-gram and POS-tags.

Barbosa and colleagues [40] implemented a two-phase classifier that first detected subjective and objective tweets, and then classified subjective tweets as positive or negative. The set of features also included platform-specific input, in the form of retweets and hashtags. Liang et al. [41] used a basic unigram Naive Bayes model to classify tweets as positive, negative, and neutral. The overall classification approach was improved by using the Mutual Information and Chi square test to eliminate useless, irrelevant features. Another solution based on Bayesian classifiers augmented with linguistic inputs is presented by Gamallo and colleagues in [42]. The authors designed two variants of Naive Bayes classifiers: Baseline was trained to classify tweets as positive, negative and neutral, while Binary classified tweets as positive and negative while ignoring neutral tweets. For bots classifiers, the set of features consisted of lemmas of nouns, verbs, adjectives and adverbs and results were also refined by using a polarity lexicon. Xia et al. [43] analysed the association of various feature sets and classification techniques. The authors used two types of feature sets (part-of-speech information and lexical relations) and three basic classifiers - Naive Bayes, Maximum Entropy and Support Vector Machines. Then, they achieved a better accuracy for sentiment classification by using different combination strategies such as weighted and meta-classifier aggregation.

Classification algorithms were widely used for sentiment analysis [44], and those techniques mostly depend on feature engineering and manually defined rules and resources, such as dependency and causality relations, n-grams or sentiment lexicons. They leverage the bag-of-words representation to convent the corpus into a term-document matrix, following the largely adopted routine of pre-processing technique, such as basic normalization and stemming. Recently, the weekly supervised approaches based on neural network techniques without feature engineering became popular for social data analysis [25]. Those solutions rely upon embedded structures, such as low dimensional word vectors which contain shallow semantic information. Similar approaches are developed by adopting supervised sequence labelling. Thus, Hidden Markov models and conditional random fields are used by Chen and colleagues to extract aspect and polarity from social data [45]. Although approaches above show promising results, opinion mining techniques making use of machine learning become problematic for social data exploration, which involves several different domains, multi languages and distinct text types, because models have to be trained for each one, and large sets of training data are required to achieve good results. Generally, most classifiers built using supervised methods perform well on polarity detection tasks, but their accuracy decreases drastically when used in new domains.

B. Lexicon-based techniques

Lexicon-based techniques for social data analysis assume the existence of cognitive categories that are independent of the language and are represented by words or more complex associations of words. These solutions rely on semantic resources modeling concepts associated to categories and their relationships that can be used to implemented procedures allowing the automatic detection of concepts and relations in texts. More specifically, resources developed highlight concepts having intrinsic and constant polarity that can be previously identified and added as an attribute of the word.

Among resources created by different teams, SenticNet [46] offers a collection of around 100,000 natural language concepts, described in terms of four affective dimensions (Pleasantness, Attention, Sensitivity, and Aptitude) and also having a polarity or orientation assignment, as a floating number between -1 and +1 , where -1 is negative polarity and +1 is positive polarity.

A Sentiment Treebank is used in [47] to provide fine grained sentiment labels for around 215 000 phrases and to allow sentiment compositionality. The Treebank is used to train a recursive neural tensor network, and the authors show that the model is able to accurately capture the effect of contrastive conjunctions and negations.

SentiWordNet [48] is a semantic resource enriching WordNet [49] by adding polarity attributes at synset level. For SentiWordNet, synsets are considered as neutral, positive or negative and the resource was used to implement incremental analysis of texts at word level, taking into account lexical categories and polarity, such as negative adverbs, positive adjectives, etc. The solution was used to analyse a set of movie reviews and the authors show that the addition of morpho-syntactic information does not significantly improve the results of the classification of texts [50].

From a different perspective, an ontology of appraisal categories was designed to capture and model concepts describing fined-grained appraisal categories expressing support, deny, rejection or endorsement [51].

Domain adaptation is still a challenge for lexicon-based approaches, and Bollgala et al. describe in [52] a solution using a distributional thesaurus to expand feature vectors during training and testing phases of a binary classifier. The lexicon provides a set of labelled data for the source domain and unlabeled data for both source and target domains, and sensitivity attributes are added for each word by measuring their distributional similarity.

Many of those resources come with limitations as they are designed to achieve broad coverage and fail to capture domain-specific concepts and standpoints [53].

IV. Illustration from social media analysis

This section discusses three studies designed to analyse social data in order to support the analysis of high-impact events. The illustrations focus on: identification of virtual communities in the aftermath of an attack, detection of unreliable information after terrorist blasts and the analysis of online propaganda.

A. Detection of virtual communities

This case study was reported by Tyshchuk and colleagues in [54] and its goal was to investigate how virtual communities arise in the cyberspace in response to major events in real environments.

Application context: For this study, Twitter data was gathered in a response to the 2013 Syrian sarin gas attack for two days following the event. Reactions of users usually divide the opinions of the population and those divisions are strong enough to give rise to virtual communities. Data collection was carried out in the light of time frame and localisation, without assigning any specific view, in the interest of one or several topics. Filters were developed distinctly for released content and users. This data set was then processed in order to investigate if more or less homogeneous communities can be identified, by taking into account the users, the content released and their interactions.

Methodology: The authors designed a processing chain in order to detect communities and to uncover their leaders and ideas propagators. Each tweet was tagged with geo-political and religious annotations; entities detected in the tweet, event, and verb, to capture actions. Moreover, directed Twitter identifiers were used to build the social network. Dynamics aspect was also considered and the nature of communications in the dataset as was carefully examined. Semantic annotations, content of data and dynamics were exploited to detect virtual communities that formed in response to the event on Twitter.

Analysis of results: The study detected around 10 virtual communities exchanged not only the most information conveyed but also the most unique, specific and important information. The study also identified several distinct leadership roles: diffuser, gatekeeper, and information broker, which were occupied by persons or organizations involved in several domains: news media, political and social. In spite of those quantitative aspects, the qualitative results show that in the two days following the event, the opinions in response to the attack were not yet fully formed. From a practical standpoint, the Twitter community's responses to the event were too sparse to establish conformity and detection of unified opinions. According to authors, using a snapshot of data may come with limitations and two days after the event is then too narrow to detect emerging polarization of opinions within and across virtual communities.

B. Detection of fake content

This case study was reported by Gupta and colleagues in [55] and the goal was to analyze the spread of fake content on Twitter during high impact events, namely the Boston marathon blasts.

Application context: The authors collected data from Twitter using the Streaming API and they also queried the Twitter Trends API after every hour to identify current trending topics, and then collect tweets corresponding to these topics as query search words for the Streaming API. In addition, twelve keywords were used to gather messages including: *BostonStrong, bostonbombing, oneboston, bostonmarathon, prayforboston, boston marathon, bostonblasts, boston blasts, boston terrorist, boston explosions, bostonhelp, boston suspect*.

The final data set included more than 7M tweets from more that 3M users. The study considered the various type of content that emerged during the event, i.e. images, texts and profiles of users.

Methodology: Data was collected for the one-hour period immediately following the first identified tweet that pertained to the bombings. The study initially looked at the types of tweets that were posted on Twitter pertaining to the Boston Marathon Bombings in the aftermath. The collected tweets were placed in groupings of 10-minute intervals and each group was broken down, separating the re-tweets from original tweets. After putting the tweets into categories, the authors also examined the tweets that were shared, or re-tweeted by other users. Tweets were annotated to the there categories: *fake or rumor*, *true* and *not applicable (NA)* and the authors applied two standard algorithms used for classification (Naive Bayes and Decision Tree) in order to detect fake content.

Analysis of results: Shortly after the event many tweet messages about the bombings were propagated, some of which created considerable confusion among the public. The results show that in the immediate aftermath of the event, only 20 percent of tweets mentioning the event convey reliable information. More than half of tweeted messages (51 percent) expressed frustration or inquiries. Informational tweets reached a peak at 63 percent of the overall amount of tweeted messages and declined rapidly to 32 percent. The results also show that 29 percent of tweets relayed rumours or fake content and that those categories are identifies as such later, thanks to the analysis of additional features.

C. Detection of online propaganda

Case study: This case study was reported by Forrester and colleagues in [56] and aimed at detecting Russian influence. Two simple filters were developed in order to detect suspected Russian-based tweets. One filter looked at content and the other examined users. The first filter consisted of a list of 200 websites taken from www.propornot.com. This site has, and continues to, identify sites that produce or propagate Russian propaganda. The second filter identified authors who were associated with the Internet Research Agency (IRA) and who have been labelled as Russian Trolls by Twitter. This data set was released by Twitter in 2018.

Application context: These filters were recently applied to a Twitter data collection that had as an aim to detect foreign influence campaigns within Canada. Past influence efforts, for instance the 2016 US presidential election, usually involve trying to divide the opinions of the population. The campaign would focus on wedge issues, where there were opposing views. Efforts concentrated on moving one or both sides to a more radical view thus creating a seemingly irreconcilable divide between people. To do this well-designed BOTs are employed. The BOTs are assigned a certain view – usually one that is in the interest of the foreign influencer. However, sometimes the aim is just to divide the people in order to cause distrust of elected officials and democratic institutions. Once the BOT focus is decided, one technique used is that the BOTs are employed to support the chosen side using relevant hashtags and linking to local websites that push extreme views for a given topic. This technique makes it seem that there is a lot of support for the views within the extreme website at a local level thus encouraging others to adopt these views.

Methodology: For this case, the authors first identified potential contentious issues that were likely to be discussed or to become wedge issues during an election. A data collection, using appropriate hashtags and keywords, was captured for each issue over several months. Care was taken to isolate the issues and ensure that mainly Canadian users were captured. For each issue the two filters were applied in order to determine if there were links with Russian propaganda or trolls. The first issue examined was a smaller data set and neither filter produced results. However, several BOTs were identified within this topic. Next a much larger topic was examined and the Russian propaganda filter produced significant results. The authors also examined the users identified by the filter and found a large proportion of BOTs and they uncovered a large BOTNET that had many BOTs involved in several issues and at least two BOTs that posted in all of the wedge issues.

Analysis of results: Applying the IRA filter did not produce any results, probably because the Russian Troll handles were compromised by their release by Twitter and thus these handles are no longer in use. A possible replacement for the user-based filter would be to use the BOT handles.

The three studies analysed in this section are sound methodologically, however their results are partially irrelevant. Cohen and colleagues already highlighted in [57] that in the case of social stream analysis reported accuracy have been systemically over-optimistic, with reporting accuracy levels nearly 30 percent higher than values expected in populations of general Twitter users. Data biases due to echo chambers [58], transferability of classifiers that cannot be used to classify outside the narrow range of data sets on which they were trained and the use of large populations of unknown users are the main reasons behind those over-estimations.

V. Discussion

Analysis of online streams comes with challenges stemming from both the specific nature of data created on digital platforms and the difficulties of social media exploration, as discussed hereafter.

Specificities of data: data collected on social media data are vast, noisy, unstructured, inherently dynamic and heterogeneous in nature. Moreover, they convey reports on real-life facts and events augmented with personal points of view, such as evaluations, attitudes, and emotions. Therefore, social data analysis is challenging for traditional data mining approaches that are often too slow and expensive, rely on sample sizes, and come with biases leading to errors.

Limitation of access data and impact of secondary sources: An important volume of online data is released on DarkWeb [59], a problematic side of Web made of encrypted portions of the Internet that are not indexed by search engines and thus cannot be listed on results pages returned by search engines to user queries. Collection, processing and sharing of such content require specific procedures to be set up or the use of secondary sources. Nevertheless, secondary sources introduce biases, and several authors emphasise that improving knowledge on the role of the Internet for hybrid threats should be mainly done by collecting primary data across multiple types of users [60].

Influence of platforms and media-induced bias: Cyberspace is an artificial, man-made environment, with data and interactions framed in a particular manner. This is the main reason any social media platform induces bias in how information is viewed by observers. Moreover, those observers can either be a part of the platform when undertaking their analysis or adopting a more direct approach to collect data via technical procedures. Those procedures are built on application programming interfaces (API) which are subroutines provided by social platforms to access their collections of data, which are also stored in proprietary formats, most of the time. This is a major drawback, as the collection requires an effort to translate data into formats easier to process and has additional limitations. As an example, although it is possible to extract data directly from Twitter archives via a request using external calls to the Twitter API, the volume accessed during one session cannot exceed 3200 tweets.

Security and ethical constraints, privacy protection: There are significant security related and ethical constraints to obtain first-hand information about sites, portals or content on social media platforms created by terrorists or extremist groups, from intelligence services, for example. Regarding privacy, data gathering and remote analysis for research purposes requires procedures and techniques that should be employed lawfully, as to make sure the overall process stay within the law.

In additional to those general challenges, social media exploration is also affected by technical bottlenecks, such as: multilingual issues, multimodality content, relevance and coherence of data sets, contextual information, aggregation and correlation of items, etc..

Although non-technical limitations can be resolved by regulatory mechanisms, technical bottlenecks can be solved thanks to hybrid approaches. In the light of limitations discussed above, three dominant hybrid approaches are identified:

- The hybrid human intelligence approach: i.e. human-centric algorithms finding and reporting information and knowledge on particular topics;
- The social and technical driven approach: i.e. indicators collected from experts who have experience with solving a rather narrow and rare empirical problem and help designing automatic procedures for those rare cases;
- The hybrid learning-semantics approach: i.e. implementing flexible learning approaches, relying on semantic inputs for a dynamic and domain-adaptive exploration of contents;

The overall implication of the analysis conducted in this paper is that social stream exploration is currently a hard problem in several key ways and addressing these aspects can significantly advance the utility and accuracy of methods and techniques.

VI. Conclusion

This paper presents on overview of techniques, approaches, methods and algorithms which are used or proposed by the research community for social data analysis. The paper also presented three use cases that show the limitations of current approaches and makes the case for hybrid artificial intelligence techniques to overcome these limitations.

Three main research directions were found for social data analysis, i.e., utilizing machine learning techniques, employing semantic-driven algorithms and the appeal to indicators from sociology, linguistics and authority-provided inputs.

The studied papers in general tend to be narrow, since they focus on solving a small task with only one type of data from one main source. The most common approach to social stream exploration is to perform text analysis using supervised machine learning.

There are also several research gaps identified, including challenges of research evaluation, with many data sets and models being not publicly available for assessment purposes, and a lack of efforts targeting hybrid solutions.

References

[1] A. Sheth, P. Anantharam, and C. Henson, “Physical-cyber-social computing: An early 21st century approach,” *IEEE Intelligent Systems*, no. 1, pp. 78–82, 2013.

[2] P. Wang, R. Angarita, and I. Renna, “Is this the era of misinformation yet: combining social bots and fake news to deceive the masses,” in *Companion Proceedings of the The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 2018, pp. 1557–1561.

[3] M. J. Mazarr, A. Casey, A. Demus, S. W. Harold, L. J. Matthews, N. Beauchamp-Mustafaga, and J. Sladden, “Hostile social manipulation present realities and emerging trends,” RAND Corporation Santa Monica United States, Tech. Rep., 2019.

[4] J. Jouhki, E. Lauk, M. Penttinen, N. Sormanen, and T. Uskali, “Facebook’s emotional contagion experiment as a challenge to research ethics,” *Media and Communication*, vol. 4, 2016.

[5] E. Ferrara, “Contagion dynamics of extremist propaganda in social networks,” *Information Sciences*, vol. 418, pp. 1–12, 2017.

[6] M. Caiani and L. Parenti, *European and American extreme right groups and the Internet*. Routledge, 2016.

[7] C. Vicient and A. Moreno, “Unsupervised topic discovery in micro-blogging networks,” *Expert Systems with Applications*, vol. 42, no. 17-18, pp. 6472–6485, 2015.

[8] B. Wang, M. Liakata, A. Zubiaga, and R. Procter, “A hierarchical topic modelling approach for tweet clustering,” in *International Conference on Social Informatics*. Springer, 2017, pp. 378–390.

[9] J. Zeng, J. Li, Y. He, C. Gao, M. R. Lyu, and I. King, “What you say and how you say it: Joint modeling of topics and discourse in microblog conversations,” *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 267–281, 2019.

[10] P. Srijith, M. Hepple, K. Bontcheva, and D. Preotiuc-Pietro, “Sub-story detection in twitter with hierarchical dirichlet processes,” *Information Processing & Management*, vol. 53, no. 4, pp. 989–1003, 2017.

[11] A. C. E. Lima, L. N. de Castro, and J. M. Corchado, “A polarity analysis framework for twitter messages,” *Applied Mathematics and Computation*, vol. 270, pp. 756–767, 2015.

[12] A. Giachanou and F. Crestani, “Like it or not: A survey of twitter sentiment analysis methods,” *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, p. 28, 2016.

[13] K. Ravi and V. Ravi, “A survey on opinion mining and sentiment analysis: tasks, approaches and applications,” *Knowledge-Based Systems*, vol. 89, pp. 14–46, 2015.

[14] H. Alvari, S. Sarkar, and P. Shakarian, “Detection of violent extremists in social media,” *arXiv preprint arXiv:1902.01577*, 2019.

[15] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. Awan, “Detection and classification of social media-based extremist affiliations using sentiment analysis techniques,” *Human-centric Computing and Information Sciences*, vol. 9, no. 1, p. 24, 2019.

[16] M. Mondal, L. A. Silva, and F. Benevenuto, “A measurement study of hate speech in social media,” in *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, 2017, pp. 85–94.

[17] S. Kannangara, “Mining twitter for fine-grained political opinion polarity classification, ideology detection and sarcasm detection,” in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018, pp. 751–752.

[18] D. O’Callaghan, D. Greene, M. Conway, J. Carthy, and P. Cunningham, “Uncovering the wider structure of extreme right communities spanning popular online networks,” in *Proceedings of the 5th Annual ACM Web Science Conference*, 2013, pp. 276–285.

[19] M. A. Wong, R. Frank, and R. Allsup, “The supremacy of online white supremacists–an analysis of online discussions by white supremacists,” *Information & Communications Technology Law*, vol. 24, no. 1, pp. 41–73, 2015.

[20] J. Qin, Y. Zhou, and H. Chen, “A multi-region empirical study on the internet presence of global extremist organizations,” *Information Systems Frontiers*, vol. 13, no. 1, pp. 75–88, 2011.

[21] J. Robertson, A. Diab, E. Marin, E. Nunes, V. Paliath, J. Shakarian, and P. Shakarian, *Darkweb cyber threat intelligence mining*. Cambridge University Press, 2017.

[22] S. M. Mohammad, “Sentiment analysis: Detecting valence, emotions, and other affectual states from text,” in *Emotion measurement*. Elsevier, 2016, pp. 201–237.

[23] T. P. Sahu and S. Ahuja, “Sentiment analysis of movie reviews: A study on feature selection and classification algorithms,” in *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*. IEEE, 2016, pp. 1–6.

[24] B. Liu, “Sentiment analysis and opinion mining,” *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[25] J. Deriu, A. Lucchi, V. De Luca, A. Severyn, S. Müller, M. Cieliebak, T. Hofmann, and M. Jaggi, "Leveraging large amounts of weakly supervised data for multi-language sentiment classification," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1045–1052.

[26] D. Tang, B. Qin, and T. Liu, "Learning semantic representations of users and products for document level sentiment classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1014–1023.

[27] M. d. P. Salas-Zárate, R. Valencia-García, A. Ruiz-Martínez, and R. Colomo-Palacios, "Feature-based opinion mining in financial news: an ontology-driven approach," *Journal of Information Science*, vol. 43, no. 4, pp. 458–479, 2017.

[28] K. Chakraborty, S. Bhattacharyya, R. Bag, and A. E. Hassanien, "Comparative sentiment analysis on a set of movie reviews using deep learning approach," in *International Conference on Advanced Machine Learning Technologies and Applications*. Springer, 2018, pp. 311–318.

[29] Q. Zhou and C. Zhang, "Emotion evolutions of sub-topics about popular events on microblogs," *The Electronic Library*, vol. 35, no. 4, pp. 770–782, 2017.

[30] Y. Hu, X. Xu, and L. Li, "Analyzing topic-sentiment and topic evolution over time from social media," in *International conference on knowledge science, engineering and management*. Springer, 2016, pp. 97–109.

[31] A. Schoene and G. de Mel, "Pooling tweets by fine-grained emotions to uncover topic trends in social media."

[32] P. Vijayaraghavan, S. Vosoughi, and D. Roy, "Automatic detection and categorization of election-related tweets," in *Tenth International AAAI Conference on Web and Social Media*, 2016.

[33] Z. Li, Y. Zhang, Y. Wei, Y. Wu, and Q. Yang, "End-to-end adversarial memory network for cross-domain sentiment classification." in *IJCAI*, 2017, pp. 2237–2243.

[34] Y. Wei and L. Singh, "Using network flows to identify users sharing extremist content on social media," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2017, pp. 330–342.

[35] J. Klausen, C. E. Marks, and T. Zaman, "Finding extremists in online social networks," *Operations Research*, vol. 66, no. 4, pp. 957–976, 2018.

[36] E. Ferrara, "Computational social science to gauge online extremism," *arXiv preprint arXiv:1701.08170*, 2017.

[37] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan, "Predicting online extremism, content adopters, and interaction reciprocity," in *International conference on social informatics*. Springer, 2016, pp. 22–39.

[38] D. O'Callaghan, N. Prucha, D. Greene, M. Conway, J. Carthy, and P. Cunningham, "Online social media in the syria conflict: Encompassing the extremes and the in-betweens," in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. IEEE, 2014, pp. 409–416.

[39] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *LREc*, vol. 10, no. 2010, 2010, pp. 1320–1326.

[40] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *Proceedings of the 23rd international conference on computational linguistics: posters*. Association for Computational Linguistics, 2010, pp. 36–44.

[41] P.-W. Liang and B.-R. Dai, "Opinion mining on social media data," in *2013 IEEE 14th International Conference on Mobile Data Management*, vol. 2. IEEE, 2013, pp. 91–96.

[42] P. Gamallo and M. Garcia, "Citius: A naivebayes strategy for sentiment analysis on english tweets," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014*. Citeseer, 2014.

[43] R. Xia and C. Zong, "A pos-based ensemble model for cross-domain sentiment classification," in *Proceedings of 5th International Joint Conference on Natural Language Processing*, 2011, pp. 614–622.

[44] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," *Artificial Intelligence Review*, pp. 1–51, 2017.

[45] L. Chen, J. Martineau, D. Cheng, and A. Sheth, "Clustering for simultaneous extraction of aspects and features from reviews," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 789–799.

[46] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, *A practical guide to sentiment analysis*. Springer, 2017.

[47] S. Lee, "Sentiment analysis system using stanford sentiment treebank," *Journal of the Korean Society of Marine Engineering*, vol. 39, no. 3, pp. 274–279, 2015.

[48] C. Hung and H.-K. Lin, "Using objective words in sentiwordnet to improve word-of-mouth sentiment classification," *IEEE Intelligent Systems*, no. 2, pp. 47–54, 2013.

[49] C. Fellbaum, "Wordnet," *The encyclopedia of applied linguistics*, 2012.

[50] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Y. Hawalah, A. Gelbukh, and Q. Zhou, "Multilingual sentiment analysis: state of the art and independent comparison of techniques," *Cognitive computation*, vol. 8, no. 4, pp. 757–771, 2016.

[51] V. Dragos, D. Battistelli, and E. Kelodjoue, "Beyond sentiments and opinions: exploring social media with appraisal categories," in *2018 21st International Conference on Information Fusion (FUSION)*. IEEE, 2018, pp. 1851–1858.

[52] D. Bollegala, T. Mu, and J. Y. Goulermas, "Cross-domain sentiment classification using sentiment sensitive embeddings," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 2, pp. 398–410, 2015.

[53] A. Esuli, "The user feedback on sentiwordnet," *arXiv preprint arXiv:1306.1343*, 2013.

[54] Y. Tyshchuk, W. A. Wallace, H. Li, H. Ji, and S. E. Kase, "The nature of communications and emerging communities on twitter following the 2013 syria sarin gas attacks," in *2014 IEEE Joint Intelligence and Security Informatics Conference*. IEEE, 2014, pp. 41–47.

[55] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Analyzing and measuring the spread of fake content on twitter during high impact events," in *Security and Privacy Symposium 2014, CSE-IIT-Kanpur*, 2014.

[56] N. Biteniece, D. V., B. Forrester, K. T., and A. Pritzgau, "Security perspectives on social media exploitation," in *Proceedings of NATO Workshop on Big Data Challenges: Situation Awareness and Decision Support*. NATO IST, 2019, pp. 1–8.

[57] R. Cohen and D. Ruths, "Classifying political orientation on twitter: It's not easy!" in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[58] P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau, "Tweeting from left to right: Is online political communication more than an echo chamber?" *Psychological science*, vol. 26, no. 10, pp. 1531–1542, 2015.

[59] G. Weimann, "Going dark: Terrorism on the dark web," *Studies in Conflict & Terrorism*, vol. 39, no. 3, pp. 195–206, 2016.

[60] M. Conway, "Determining the role of the internet in violent extremism and terrorism: Six suggestions for progressing research," *Studies in Conflict & Terrorism*, vol. 40, no. 1, pp. 77–98, 2017.