FCTNLP: An architecture to fight cyberterrorism with natural language processing

Andrés Zapata Rozo[1], Daniel Díaz-López[1], Javier Pastor-Galindo[2], Félix Gómez Mármol[2]
[1]School of Engineering, Science and Technology, Universidad del Rosario, Bogotá, Colombia
{andresf.zapata, danielo.diaz}@urosario.edu.co
[2]Department of Information and Communications Engineering, University of Murcia, 30100, Murcia, Spain
{javierpg, felixgm}@um.es

***Abstract*—Law Enforcement Agencies (LEA) are everyday more and more concerned about illicit activities that may be found in cyberspace like cybercrimes, cyber espionage, cyberterrorism, cyber warfare, among others. In a cyberterrorism context, Hostile Social Manipulation (HSM) is a strategy that employs different manipulation methods mostly through social media to produce damage to a target state. The efforts to fight cyberterrorism could come along with new technologies that allow a faster and more effective control of offensive actions. For that reason, this paper proposes an artificial intelligence-based solution that processes posts in social networks using Natural Language Processing (NLP) techniques, applying the following three models: i) Sentiment Model to discriminate between threat and non-threat publications, ii) Similarity Model to identify suspects with similar intentions and iii) NER model that identifies entities in the text. Finally, the proposal was tested exhaustively to validate its functionality and feasibility, achieving an integrated and simple prototype.**

***Index Terms*—Cyberterrorism, Natural Language Processing, OSINT, Semantic Similarity, NER, Sentiment Analysis.**

Contribution type: *Consolidated scientific research*

I. Introduction

Cybercrimes are either committed against the integrity, availability, and confidentiality of computer systems and telecommunication networks or the use of such networks or their services to conduct traditional offenses [1]. Following the previous idea, we can consider cyberterrorism as a crime that aims to involve the generation of terror in the cyberspace with the aim of subverting the political order. Thus, it has a greater impact than conventional cybercrimes as it provokes a state of terror in the general public, a group of persons, or particular persons, for political, philosophical, ideological, racial, ethnic, religious or other nature purposes purposes[1].

One of the best-known cases of cyberterrorism began when ISIS posted a video on YouTube on August 19th 2014 titled "A Message to America", where the journalist James Foley was beheaded as a response to the authorization of offensive actions against this terrorist group by the Obama's government [2].

Another case of intimidation occurred on February 10th, 2020 when the Colombian guerrilla group ELN, through the squadron "Omar Gomez", announced an armed strike on the main roads of Colombia to be realized in the middle of that month[3]. Those announcements included publications on social networks as well. This armed strike intimidated the population forcing them to stay in their homes, and people who violated the restrictions could be victims of violence from this armed group. As a consequence, many Colombian towns and cities where ELN was present stopped most of their economic activities.

On the other hand, by applying Open Source Intelligence (OSINT), we can obtain knowledge that can be reached using publicly available data [2]. The Internet and especially social media have contributed to the growing importance of public information that can be extracted using OSINT tools. Also, these intelligence sources have been relevant for the defense enterprise due to their potential use in big data [3] [4].

The OSINT can be complemented using Natural language processing (NLP), which compounds computer science and linguistics to generate an approach to the understanding of the human language by a computer, this task is carried out through tools of artificial intelligence, statistics and grammar [5]. One of the most common examples of the application of NLP is a conversational agent that uses these techniques in order to understand the interlocutor language and emulates a functional conversation, taking into consideration variables such as the entities and the intention of the input text [6].

In this paper, we present an OSINT solution that extracts information from social networks and other resources. Then, such information is processed using NLP techniques including three models: i) a similarity model that relates text with similar semantic meaning, ii) a sentiment analysis model that estimates the polarity of a sentence, and iii) a Name Entity Recognition (NER) model that recognized relevant entities in the text and the type of these entities. All the results are integrated into a simple module that yields the output of this model generating a report that contributes actionable information to Law Enforcement Agencies.

[1]https://digitallibrary.un.org/record/631639?ln=es

[2]https://edition.cnn.com/2014/08/19/world/meast/isis-james-foley/index.html

[3]https://thecitypaperbogota.com/news/eln-announces-72-hour-armed-strike-warns-of-consequences-to-travelers/23849

II. State of the art

A compilation of remarkable works that use NLP to fight against cybercrimes is presented next. In [7] the authors use data from two datasets: one available online and another built with data from Twitter and Facebook and labeled manually to construct a cybercrime text classifier. In addition, they compare their different classifiers with sentiment analysis from the NLTK Python library.

The work proposed in [8] consists of a big data architecture that allows a real-time analysis of tweets to classify users and their respective followers as part of ISIS (a terrorist organization), according to parameters like level of activity, influence on other users, and post content. A graph was created where indicators of centrality were applied to identify the most influential users before applying analysis over the data. Finally, user profiles were obtained through Fuzzy clustering techniques.

Cyberterrorism may use hate speech as a technique to intimidate a specific target population. In [9] the detection of hate speech in the Arabic context was developed through the use of different comparative machine learning methods like Support Vector Machine (SVM), Naive Bayes (NB), Decision tree (DT) and Random Forest (RF). The data used in this work come from tweets related to racism, journalism, sports orientation, terrorism and Islam.

The detection of cyberterrorism vocabulary in web pages was proposed in [10] and in this work, the results of the following algorithms were evaluated: Random Forest, Boosting, SVM, Neural Network, K-Nearest Neighbor (KNN) and Naive Bayes, where a Random Forest approach gives the best results. In all the cases the percentage of accuracy was higher than 80% and in the case of the Random Forest approach was 95.62%. The vocabulary developed to detect the websites include information related to Al Qaeda, Supreme Truth, KKK and ETA groups.

An analysis of a historical dataset was presented in [11], where the data from Twitter messages of attacks between 2008 and 2019 by the terrorist group Boko Haram in Nigeria were analyzed using DynamicK-reference clustering algorithms. This work allowed the identification of various of strategies for Boko Haram attacks and pointed out weaknesses in security control in sectors of northern Nigeria.

The proposal presented at [12] tries to detect cyberterror and extremism in the text using Fuzzy sets-based weighting methods, Naive Bayes Multinomial (NBM) and SVM. The experimental analysis shows that the fuzzy set-based weighting method with SVM classifier gives the best classification with a 99.4% of accuracy.

As observed, the most recent works that make contributions in the fight against cyberterrorism through the use of NLP have the purpose of detecting terrorist behavior on the Internet.

III. Proposal of FCTNLP

This section describes the main aspects of the design of FCTNLP, covering the definition of requirements and the explanation of the main components. The development of this design follows the phases defined in a data science life cycle [13]: i) Business understanding, ii) Data acquisition and preprocessing, iii) Modeling, and iv) Deployment.

A. Business understanding

As seen in Section I, cyberterrorism is a real problem that affects the general population or a target state, and as a crime, it should be fought. Following this, one of the main problems in the fight against cyberterrorism is recognizing it in the middle of a large data flow, such as the one existing in social networks. In addition to the amount of data that must be analyzed, human-generated text may display different structures with different meanings [14]. Thus, one option to analyze sentences is to devote a person to extract key information from such human-generated text, however that analysis would be subjective and a very large group of people would be needed to analyze all content coming from social networks. This is why a solution like FCTNLP that automates the structuring and analysis of posts on social networks is vital in the fight against cyberterrorism. Thus, the architecture proposed for FCTNLP is expected to meet the following targets:

- Distinguish tweets: It should be capable of creating initial groups of tweets related by their semantic meaning.
- Evaluate polarity: It should contain a model that scores the polarity of each tweet.
- Recognize entities: It should be capable of extracting entities relevant to tracking cyberterrorism.
- Identify communities: It should use OSINT information related to the Twitter accounts that are generating content to identify existing relations between such actors.

B. Data acquisition and preprocessing

Amongst all social networks, Twitter has consolidated as an important source of data due to the relevance and diversity of data that can be obtained from it to be analyzed [15]. Thus, we consider Twitter as a social network with the possibility to generate a high impact beyond cyberspace and that is why such a social network was selected in this paper as the source of the raw data that feeds our proposal. In order to gather tweets, different OSINT tools and techniques may be used, which can be divided into two categories: i) Scrappers and ii) API-based tools. The first category contains scrappers that use bots, some of them emulating human behavior, to get specific data from Twitter, e.g. Octoparse[4]. These kinds of tools allow to personalize the type of data that will be extracted but, due to the policies of Twitter, most of these kinds of tools have a tweets extraction limit of 1000 tweets per day. The second category refers to tools that consume the Twitter API[5] and therefore run under the restrictions defined by such API, e.g. tweets can only be extracted from a short time window that may be up to the last 7 days previous to the date of the collection.

[4]https://www.octoparse.com/tutorial-7/scrape-tweets-from-twitter
[5]https://developer.twitter.com/en/docs/twitter-api

Once tweets are collected, many strategies of preprocessing can be used to prepare the text before feeding the NLP models. The principal objective of these strategies is to keep the meaning of the text while cleaning it from noise data that can influence in a bad way the performance of the models. The most common strategies applicable to tweets are: i) remove hashtags, mentions, URLs, strange characters and punctuation symbols, ii) normalize text that converts text into lower or upper case, iii) replace emojis for words that represent their meaning, iv) apply tokenization that divides the text in tokens that can be only words or words with punctuation symbols, and v) do lemmatization that replaces a word by their lemma, i.e the canonical form of the word. The different strategies used in the implementation will depend on the NLP model that will be fed with such preprocessed data.

C. Modeling

In this section three NLP models will be described. The first one is the sentiment analysis model that is used in the NLP tasks to determine the emotions that the author of a text expresses or the mood of the author at the moment of writing the text. Secondly, a NER model is used in NLP to extract relevant entities and their respective type from raw text. Finally, the similarity model is used to represent a text in a vector way so that the representation can contain the semantic meaning of the text. The inclusion of these models allows achieving the targets proposed in section III-A

1) Sentiment Analysis Model: A sentiment analysis model can be implemented as a classifier that discriminates text between classes according to the polarity of the text (positive, negative or neutral), classifies the subjectivity of the author (subjective, objective), or extracts the emotional state of the text (happy, angry, friendly, confident, etc.) [16].

Sentiment analysis models may also in a single function, e.x. regression function scores two or more aspects of the text like the polarity and subjectivity [17]. Finally, another way to implement a sentiment analyzer is using a rules-based algorithm using the knowledge about the language structure and the meaning of the words [18].

2) Name Entity Recognition Model: A Long Short Term Memory (LSTM) is a Recurrent Neural Network (RNN) that can take advantage of consecutive and non-consecutive terms to give the best results in the task of recognizing entities from human language. It may also offer an understanding of the relation between words and their grammatical meaning. Another architecture for this task is the Bidirectional LSTM (Bi-LSTM), where two LSTM models are concatenated in a way that the Bi-LSTM can receive information from the beginning of the text up to the end, and from the end up to the beginning [19]. Regardless of the RNN used for the implementation, a NER model is generally trained using a Begin Inside Outside (BIO) notation where a phrase is decomposed in beginning, inside and outside sections. Thus, a NER model adds actionable information to analyzed tweets such as the one that helps to describe where (location) and who (subject, organization) is involved in some specific actions being monitored.

Transformers are also used in NER tasks. They were proposed in [20] and consists of two stacks: one encoder and one decoder. Both inputs and outputs have embeddings and positional encoding. Each stack uses multi-head attention layers: a non-masked one for the encoder, and a masked one for the decoder. At the end of both stacks a fully connected feed-forward network is also placed. Finally, a linear layer and a softmax activation function are used to get the prediction.

3) Similarity Model: The similarity model uses word embeddings to represent the meaning of the words in a real space. Such representation is useful to get the relation between words that have a similar meaning as they will have a close vector representation. The metric used to calculate the similarity between words is the soft cosine distance, which is represented by Equation 1.

$$soft\_cosine(w;v) = \frac{\sum_{i;j}^{N} s_{ij} w_i v_j}{\sqrt{\sum_{i;j}^{N} s_{ij} w_i a_j} \sqrt{\sum_{i;j}^{N} s_{ij} v_i v_j}} \quad (1)$$

To generate the word embeddings two principal algorithms may be used: Word2Vec and FastText. In the case of Word2Vec, it creates a vector representation for each word in the text, keeping similar words close in the vector space [21]. This approach has the problem that words that are not included in the training set (new words included in tweets) will not be considered in the similarity calculus as they will not have a vector representation. FastText algorithm may help to solve this previous problem as it uses a vector representation that takes into account the n-grams of a word, i.e. the sequence of n characters. Thus, this last approach is capable of representing words that are composed of some n-grams contained in the training dataset, even if it losses the property of having semantic knowledge in the vector representation [21].

An illustration of how the previously described models (similarity model, sentiment analysis model and NER model) are integrated into the FCTNLP architecture is shown in Figure 1.

D. Deployment

The information generated by FCTNLP should be stored properly, so it can be recovered lately for additional processing. Amongst this information, we have certain diversity represented by: the gathered tweets, the outcomes from the polarity, the similarity and the NER models, and additional useful information that may be obtained by OSINT techniques. This scenario with heterogeneous and unstructured information suggests that graphs may be an adequate way to store the information. Additionally, graphs may be one of the most useful ways to show different types of information associated with the user accounts, the message polarity, the conformed clusters, and the identified entities, in the same space. For these previous reasons, FCTNLP incorporates in its architecture (Figure 1) a graph database that allows storing and representing nodes and relations.

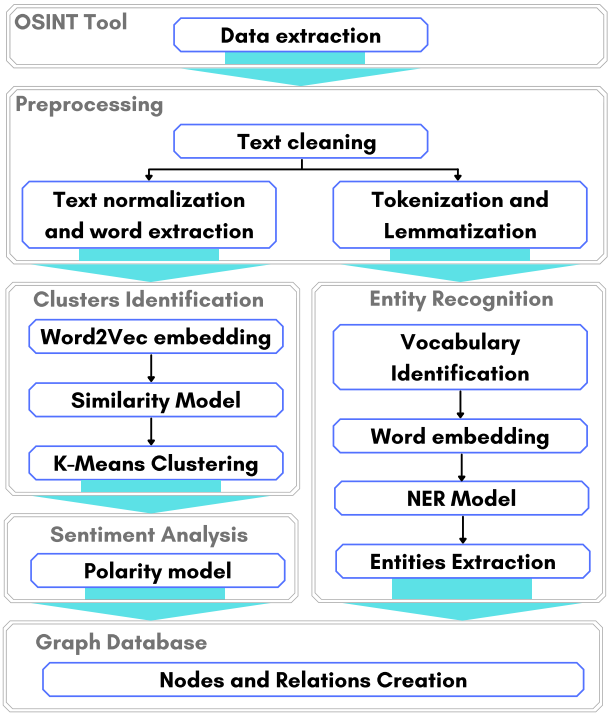

Figure 1: Components of the FCTNLP architecture

FCTNLP could be used by a Law enforcement Agency (LEA) to automatize the analysis of a human text obtained from open sources and help in the prevention of cyberterrorism. Such implementation can be a key tool for a cyber intelligence agent as it facilitates the search for spotlights of cyberterrorism. Results obtained from FCTNLP may also be enriched with the information provided by an already running commercial cyber intelligence solution.

Due to the modularity, the high cohesion, and the low coupling of the proposed architecture, each NLP model can be deployed as part of a scalable solution that allows the flow of a huge amount of data. For instance, to be deployed as a microservice. The extraction of tweets can also be set as a task to be executed in real-time or under demand. In both cases, extracted data can be saved in a data lake that will be processed by the NLP models that compose FCTNLP, and a cache solution can also be fed with the more recent and relevant results in order to have quick access to required data.

IV. Experiments

This section contains the results obtained from applying the proposal described in Section III in a scenario related to a protest that occurred on October 26, 2021, in Ecuador, being the data and code available in the project repository [6]. Twitter was the social network used to provide the raw information to be processed. The gathering was done using `TAGSv6.1`[7].

The embedding process in the English language was based on the use of Google News Embedding, which contains generic embeddings for 3,000,000 English words with a dimension of 300. On the other hand, the embedding of content in Spanish was done using the Spanish Billion Words Corpus and Embeddings[8] that contains a set of 1,000,653 words with a 300 dimensions vector representation. Both embeddings were built using the word2vec algorithm[9], as in section III. The extraction of information related to followers from Twitter accounts was made using tinfoleak[10] and the generation of the neighborhood graph for a selected cluster was made using Gephi[11].

A. Gathering tweets in the protest against economic policies in Ecuador

The straw that broke the camel's back was the announcement of the increase in fuel prices by the government of Guillermo Lasso[12], in addition to the economic crisis in which Ecuador finds itself and the fall in popularity of the Lasso government due to the investigation that he is facing due to he is appearing involved in the Pandora Papers[13].

This scenario implied the gathering of 10,608 tweets containing at least one of the following hashtags `#ParoNacional`, `#ParoNacionalEC`, `#LassoEsUnFracasso`, `#Quito`, `#LassoCorrupto`, `#LassoMentiroso` and in order to obtain only original tweets with text we use the following filters of the twitter API `-filter:images`, `-filter:videos`, `-filter:retweets`. Between October 24 and October 27 of 2021. The protest occurred between 26 October and October 27, 2021, for this reason the tweets before these dates were omitted so a total of 7,086 tweets remained.

The day of protests was marked by some acts of violence, the most serious of which was the confrontation between the demonstrators and the police in front of the presidential palace[14]. On the other hand, other disturbances occurred in various parts of Ecuador such as road blockades[15]. At the end of the protests, 37 people were arrested for acts of violence[16].

The data of this experiment is a unique collection from Twitter that uses TAGS. As we see in Section III, tools like TAGS that used the Twitter API to collect the information from Twitter have a limit to the tweets that can be collected in a window of time. Most of the time this limit is not reached, especially with specific topic queries like the use in this experiment.

[6]https://github.com/AndZapCod/NLP_Cybersecurity_Case
[7]https://tags.hawksey.info/
[8]https://crscardellino.ar/SBWCE/
[9]https://code.google.com/archive/p/word2vec/
[10]https://tinfoleak.com/
[11]https://gephi.org/
[12]https://www.argusmedia.com/en/news/2266703-ecuador-freezes-fuel-prices-update
[13]https://www.reuters.com/world/americas/ecuador-president-lasso-be-investigated-tax-fraud-after-pandora-papers-leak-2021-10-21/
[14]https://www.laprensalatina.com/quito-violence-marks-day-of-protests-against-ecuador-president/
[15]https://frontline.thehindu.com/dispatches/protesters-in-ecuador-block-roads-over-gasoline-price-hikes/article37195681.ece
[16]https://www.reuters.com/world/ecuador-demonstrators-block-some-roads-protests-over-gas-prices-2021-10-26/

Tweets were preprocessed and cleaned properly to be consumed by the models that will be used later in the pipeline. The first step in preprocessing was to construct a dictionary with the most used hashtags and mentions of the collected tweets and replace them with their meaning words. The second step was to remove URLs, mentions, hashtags, reserve words like RT and FAV, smilies and strange characters from the tweets using the python library `tweet-preprocessor`[17]. Then, emoticons were replaced by their meaning in words through the use of the Python library `emoji`[18]. Finally, empty and duplicated tweets were removed and a total of 7,077 tweets remained.

Additionally, preprocessing was required for each model. For the sentiment analysis model, punctuation symbols were removed from the text and the latter was normalized to lowercase. Regarding the similarity model, tweets were translated from Spanish to English using Google API Services[19], and punctuation symbols were also removed and the text was converted to lowercase. In the case of the NER model, cleaned tweets were tokenized and lemmatized. Using a python dictionary constructed in the training of the model, each token was converted to an integer to be used as input for the Bi-LSTM model see section III.

B. Application of the similarity model

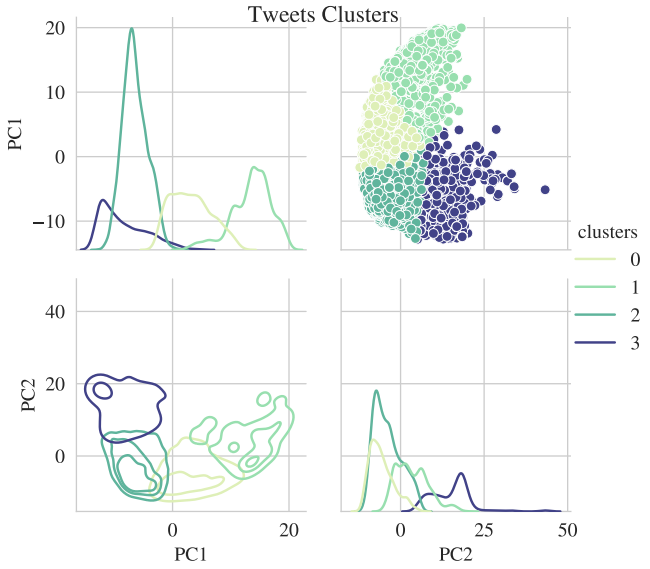

Figure 2: K-means clustering using Google News embeddings

For the analysis of the translated tweets, $(t_m = t_1; :::; t_n)$ were processed by the similarity model mentioned in Section III using the google news embedding to build a matrix of cosine distances between the different tweets. The entries of such a matrix were done by taking each tweet (t_i) and calculating their cosine distance against the remaining tweets. Afterward, The 7,077 tweets were split into four clusters according to the K-means algorithm using the similarity matrix in order to take advantage of the semantic representation of the tweets.

In Figure 2 we can see a representation of the tweets using the similarity matrix and the PCA algorithm to extract the two dimensions that represent the 86:4% of the variability of the data.

In Figure 3 we can see that the reasons for the protest are dominant in all of the clusters. In addition to this, the reason for the protest, Ecuador's president is more mentioned in clusters 0 and 1.

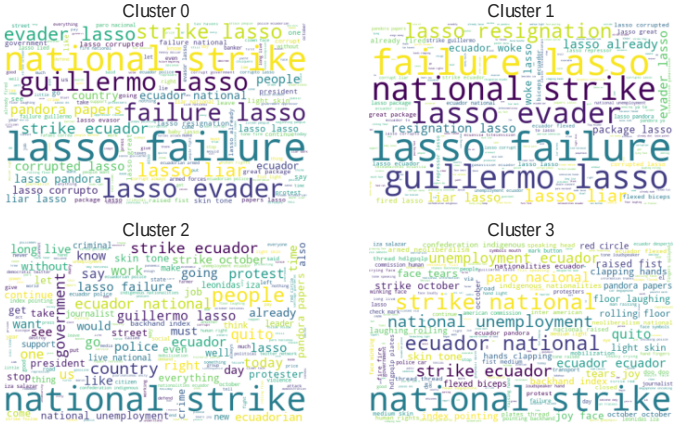

Figure 3: Word cloud for the clusters with google news embeddings

C. Application of the sentiment analysis model

For each cluster, sentiment analysis was conducted. The analysis used the `TextBlob` python library. that use a single perceptron to extract a score of the polarity between $[-1; 1]$ the values with a more negative score are also with a polarity more negative.

As we can see in the Figure 4 the cluster with a higher proportion of negative tweets is the cluster 1 with a 64% of negative tweets and as we can see in the Table I the cluster 1 has also the less polarity score, i.e we can consider a cluster of tweets that may contain cyberterrorism.

Cluster	Negative	Positive
0	-0.25	0.23
1*	**-0.33**	0.24
2	-0.21	0.22
3	-0.23	0.32

Table I: Polarity media of each cluster

D. Application of the NER model

In this case, a NER model was trained using the WikiNer dataset [22] with the Spanish corpus, this dataset contains 141; 761 sentences extracted from Wikipedia and annotated using the BIO format with three types of entities: i) LOC

[17]https://pypi.org/project/tweet-preprocessor/
[18]https://pypi.org/project/emoji/
[19]https://pypi.org/project/google-cloud-translate/

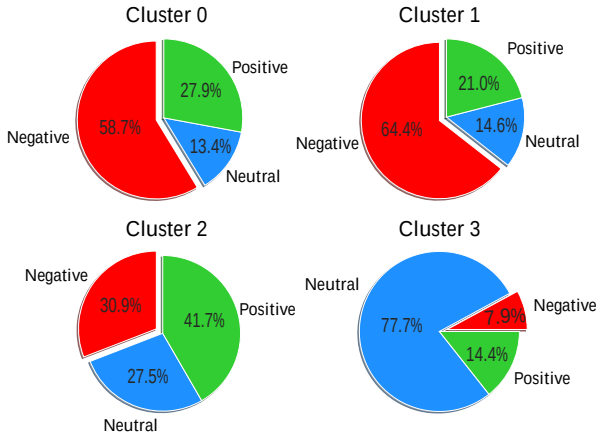

Figure 4: Polarity by cluster with TextBlob

Localization, ii) *MISC* Miscellaneous iii) *ORG* Organization and iv) *PER* Person. The model was trained using the Bi-LSTM architecture see in the section III. The metrics obtained in the train set and the test are shown in the Table II.

In Table III we can see the eight entities with the highest frequency of appearance predicted by the NER model after removing bad predictions as stops words, punctuation symbols and names of emojis. The last name of the president of Ecuador is part of the set of entities with greater frequency in all categories of entities. In the case of entities such as person and organization, this would make sense since depending on the context this word can refer to the president as a human being or as a representation of an organization, in this case, the presidency. It must also be considered that this word was not found in the training set, so the model managed to label it in any case.

	Train Set Quality		
	Precision	***Recall***	***F1-Score***
General Performance	95.37%	95.98%	95.67%
LOC	94.73%	95.76%	95.24%
MISC	92.77%	93.05%	92.91%
ORG	93.88%	93.21%	93.55%
PER	98.11%	98.56%	98.34%
	Test Set Quality		
	Precision	***Recall***	***F1-Score***
General Performance	82.45%	84.93%	83.67%
LOC	82.90%	86.77%	84.79%
MISC	67.77%	68.94%	68.35%
ORG	77.48%	77.33%	77.41%
PER	89.91%	91.50%	90.70%

Table II: Tain and Test Quality

token	**ORG**	**LOC**	**PER**	**MISC**	**Total**	**Percentage**
Paro	3426	1801	196	74	5497	62.3% (ORG)
Lasso	353	3145	2125	264	5887	36.1% (PER)
Nacional	2661	2415	204	97	5377	49.5% (ORG)
Ecuador	724	2535	14	60	3333	76.1% (LOC)
Guillermo	21	36	1299	3	1359	95.6% (PER)
Fuera	115	1067	85	17	1284	1.3% (MISC)
Renuncia	10	593	56	3	662	0.5% (MISC)
Pandora	9	205	558	8	780	0.1% (MISC)

Table III: Principal NER Predictions and percentage of accuracy

On the other hand, An example of a correctly labeled location is the word "Ecuador". In a similar way as the president's Lastname. Ecuador also can be considered as a Location entity or Organization entity, this depends on the context in which the word is used. The model labeled the leaked documents known as "pandora papers" as a person most of the time due to the name "pandora". Finally, we can see in the prediction of "Fuera" that most of the time was predicted as a Location entity. But in the context of protest, this word is used as an expression of rejection instead of a relative position.

Although the model successfully predicts several of the entities that were relevant within the context of the protests, errors also occur that are due to the difference between the training set, Wikipedia articles, and the analyzed tweets that also contain a language more informal in addition to the use of emojis.

E. Graph representation of the user network of contacts

Finally, the data of the Twitter users were collected using the `tinfoleak` tool in order to construct a network of contacts In the case of finding some type of suspicious activity. Intelligence agencies can consult a graph like the one shown in Figure 5. In this case, it is a sample of users who published tweets grouped using google news embeddings in cluster 1 (red nodes), which was the one with the most negative polarity and some of its contacts (blue nodes) from which tweets not were collected.

In this graph, we can see communities of users that follow big red nodes these communities are connected by internal nodes that are common contacts between these big nodes and other red nodes.

V. Conclusions and Future Work

Taking into consideration the influence that social networks have on society and their possible misuse of them to promote cyberterrorism, the application of NLP may be considered a way to automate the analysis of the large volume of data coming from social networks.

In this paper, we proposed FCTNLP, a solution that integrates three NLP models to process information extracted from open sources, i.e. social networks, with the purpose of identifying behaviors associated with HSM and in that way supporting LAWs agencies in the identification and prevention of cybercrimes. FCTNLP was tested through a set of experiments that process tweets related to a real scenario of

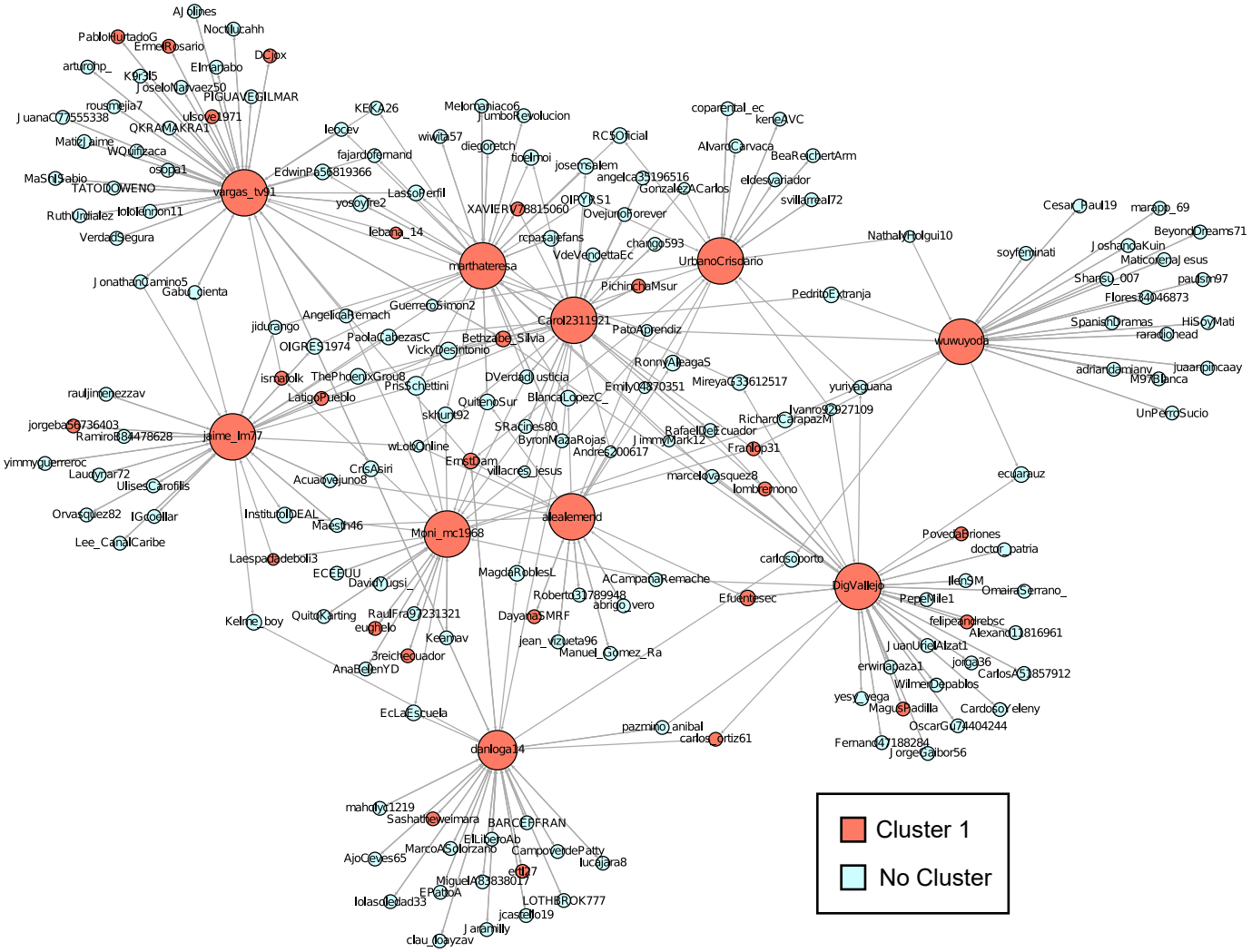

Figure 5: User network sample of user of the cluster 1

protests that occurred on October 26, 2021, in Ecuador. Such experiments demonstrated the feasibility of our proposal for a scenario of cyberterrorism and the usefulness that it may have for LAWs.

As future work, we plan to do a new set of experiments that include additional sources of information, like underground forums and other social networks, which allow enriching the information represented in the graph to discover more useful insights within cyberterrorism research. It may also be interesting to enlarge the windows of time employed for the gathering of data, including previous and subsequent moments of a protest, so it may be clearer to identify how changed the hostility in the content is generated. Finally, FCTNLP may be extended with new NLP models, e.g. one able to predict intentions so a response may also be designed with some automaticity as a way to contain a hostile campaign.

Acknowledgment

This work has been supported by Universidad del Rosario (Bogotá) through the project "IV-TFA043 - Developing Cyber Intelligence Capacities for the Prevention of Crime". This work has also been supported by an FPU contract (FPU18/00304) granted by the Spanish Ministry of Universities.

References

[1] C. of Europe, "Explanatory report to the convention on cybercrime," https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016800cce5b, 2001.

[2] J. R. G. Evangelista, R. J. Sassi, M. Romero, and D. Napolitano, "Systematic literature review to investigate the application of open source intelligence (osint) with artificial intelligence," *Journal of Applied Security Research*, pp. 1–25, 2020.

[3] H. J. Williams and I. Blum, "Defining second generation open source intelligence (osint) for the defense enterprise," RAND Corporation Santa Monica United States, Tech. Rep., 2018.

[4] J. Pastor-Galindo, P. Nespoli, F. Gómez Mármol, and G. Martínez Pérez, "The not yet exploited goldmine of osint: Opportunities, open challenges and future trends," *IEEE Access*, vol. 8, pp. 10 282–10 304, 2020.

[5] A. Thomas, *Natural Language Processing with Spark NLP: Learning to Understand Text at Scale*. O'Reilly Media, 2020. [Online]. Available: https://books.google.com.co/books?id=sJw6zQEACAAJ

[6] L. Clark, N. Pantidi, O. Cooney, P. Doyle, D. Garaialde, J. Edwards, B. Spillane, E. Gilmartin, C. Murad, C. Munteanu, V. Wade, and B. R. Cowan, *What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents*. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–12. [Online]. Available: 10.1145.3290605.3300705

[7] S. Kumari, Z. Saquib, and S. Pawar, "Machine learning approach for text classification in cybercrime," pp. 1–6, 2018.

[8] C. Sánchez-Rebollo, C. Puente, R. Palacios, C. Piriz, J. Fuentes, and J. Jarauta, "Detection of jihadism in social networks using big data techniques supported by graphs and fuzzy clustering," *Hindawi*, vol. 2019, no. 1238780, p. 13, 2019.

[9] I. Aljarah, M. Habib, N. Hijazi, H. Faris, R. Qaddoura, B. Hammo, M. Abushariah, and M. Alfawareh, "Intelligent detection of hate speech in arabic social network: A machine learning approach," *Journal of Information Science*, vol. 47, no. 4, pp. 483–501, 2021. [Online]. Available: https://doi.org/10.1177/0165551520917651

[10] I. Castillo-Zúñiga, F. Luna-Rosas, L. Rodríguez-Martínez, J. Muñoz-Arteaga, J. López-Veyna, and M. Rodríguez-Díaz, "Internet data analysis methodology for cyberterrorism vocabulary detection, combining techniques of big data analytics, nlp and semantic web," *International Journal on Semantic Web and Information Systems*, vol. 16, pp. 69–86, 01 2020.

[11] C. Oleji, N. Euphemia, G. Chukwudebe, and O. Chukwueneka Philips, "Big data analitic of boko haram insurgency attacks menace in nigeria using dynamick-reference clustering algorithm," vol. 7, pp. 1099–1107, 04 2020.

[12] V. N. Uzel, E. Saraç Eşsiz, and S. Ayşe Özel, "Using fuzzy sets for detecting cyber terrorism and extremism in the text," in *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2018, pp. 1–4.

[13] S. J. Wagh, M. S. Bhende, and A. D. Thakare, *Fundamentals of Data Science*. Chapman and Hall/CRC, 2021.

[14] S. Bazzaz Abkenar, M. Haghi Kashani, E. Mahdipour, and S. M. Jameii, "Big data analytics meets social media: A systematic review of techniques, open issues, and future directions," *Telematics and Informatics*, vol. 57, p. 101517, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736585320301763

[15] A. Karami, M. Lundy, F. Webb, and Y. K. Dwivedi, "Twitter and research: A systematic literature review through text mining," *IEEE Access*, vol. 8, pp. 67 698–67 717, 2020.

[16] Ankit and N. Saleena, "An ensemble classification system for twitter sentiment analysis," *Procedia Computer Science*, vol. 132, pp. 937–946, 2018, international Conference on Computational Intelligence and Data Science. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187705091830841X

[17] A. Poornima and K. S. Priya, "A comparative sentiment analysis of sentence embedding using machine learning techniques," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 493–496.

[18] S. Zahoor and R. Rohilla, "Twitter sentiment analysis using lexical or rule based approach: A case study," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2020, pp. 537–542.

[19] B. Jang, M. Kim, G. Harerimana, S.-u. Kang, and J. W. Kim, "Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism," *Applied Sciences*, vol. 10, no. 17, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/17/5841

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[21] J. Choi and S.-W. Lee, "Improving fasttext with inverse document frequency of subwords," *Pattern Recognition Letters*, vol. 133, pp. 165–172, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865520300817

[22] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, "Learning multilingual named entity recognition from wikipedia," Oct 2017. [Online]. Available: https://figshare.com/articles/dataset/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500