Identifying tweets from Syria refugees using a Random Forest classifier

Smarti Reel
Health Informatics Centre
University of Dundee
Dundee, UK
s.reel@dundee.ac.uk

Patrick Wong
School of Computing and Communications
Open University
Milton Keynes, UK
patrick.wong@open.ac.uk

Belinda Wu
School of Politics, Philosophy, Economics, Development, Geography
Open University, UK
belinda.wu@open.ac.uk

Soraya Kouadri Mostefaoui
School of Computing and Communications
Open University, UK
soraya.kouadri@open.ac.uk

Haiming Liu
Department of Computer Science and Technology
University of Bedfordshire,
Luton, UK
Haiming.Liu@beds.ac.uk

***Abstract*— A social unrest and violent atmosphere can force a vast number of people to flee their country. While governments and international aid organizations need migration data to inform their decisions, the availability of this data is often delayed due to the tediousness to collect and publish this data. Recent studies recognized the increasing usage of social networking platforms amongst refugees to seek help and express their hardship during their journeys. This paper investigates the feasibility of accurately extracting and identifying tweets from Syria refugees. A robust framework has been developed to find, retrieve, clean and classify tweets from Syria. This includes the development of a Random Forest classifier, which automatically determines which tweets are from Syria refugees. Testing the classifier with samples of historical Twitter data produced promising result of 81% correct classification rate. This preliminary study demonstrates the potential that refugees' messages can be accurately identified and extracted from social media data mixed with many unwanted messages, and this enables further works for studying refugee issues and predicting their migration patterns.**

Keywords— Social media, Syria, refugees, classification

Short Papers, *CSCI-ISNA*

I. INTRODUCTION

When a large-scale disaster or conflict breaks out, vast numbers of people in the affected areas often migrate and seek refuges in safe havens elsewhere. Such huge sudden movement of people can cause complex humanitarian, social and economic issues to the refugees, the nearby and hosting countries as well as the various governmental and aid agencies. If the patterns of these migrations can be correctly understood and accurately predicted, the governments and aid agencies of the countries concerned can be better prepared to help the refugees. Recent studies have recognized the increasing use of social media such as Twitter and Facebook amongst refugees/migrants [1]. They often use social networking platforms to express the hardship and difficulties they face and report their experience, as well as using such platforms as a tool in migration planning and finding out information during their actual journeys to destined host countries, including the most up-to-date situations and regulations on the way. This creates an opportunity for studying large-scale international migrations using the social media and data analytic methods, which have been successfully applied in modelling various human behavior such as traveling tracking [2] and Crowd dynamics [3]. However, the migration pattern predication relies heavily on accurate identification of messages from refugees. This paper therefore focuses on identifying social media messages from refugees, which was confirmed as a challenging problem due to refugees reluctant to post on open social media platforms in the fear of getting caught by their government authority or human smugglers [4]. To prevent refugees' identities from being accidentally revealed by this study, their social media messages had been anonymized by replacing their usernames with unique index numbers.

Recent studies have analyzed the trends in mobility and migration flows using geolocated twitter data. For example, Zagheni [5] investigated the use of Twitter data to analyze the international and internal migration patterns. However, such analysis is entirely dependent on the availability of geolocated data, which is often unavailable in refugees' messages due to limitations on their computing devices or other reasons.

Gillespie [4] explored the refugee media journeys through smartphones and social media networks. It obtained information on trusted sources and groups on Facebook. While some of the Facebook groups studied contained news about refugees in Syria, other groups concerned general migration issues. These groups only revealed partial information about refugees' experiences. The study also found that most of the trusted sources communicates in Arabic.

Ali [6] discussed the significance of big data for various applications and development purposes. It provided a brief background of relevant techniques to understand the applications in humanitarian development. It proposed to use predictive analytics to avoid or mitigate humanitarian emergencies before they happened.

Brouckman and Wang [7] investigated the use of supervised machine learning along with Natural Language processing methods to classify downloaded tweets (with keyword 'refugee' in four languages) using the Twitter API. They used unigram features to model the tweets and then trained five classifiers namely, - Support Vector Machines, Logistic Regression, Random Forest, Naive Bayes, and Ensemble to predict the sentiments towards refugees as either positive, negative, or neutral.

Despite the above research studies, accurately identifying social media messages from refugees remains a challenge, while this is the important step in predicting migrating patterns.

II. Methodology

This pilot study aims to identify refugees' social media messages from data openly available on social networking platforms using data analytic techniques. Due to the public nature of Twitter data, it was chosen as the social media messages for this study. Given the large scale of human movement happened during the recent Syria crisis, it was chosen as an example for this pilot study. The main challenge of identifying refugees' messages (tweets) is that there are vast amount of messages discussing the Syria crisis from non-refugees such as journalists and the general public who are concerned with the situation in Syria. These messages compose of similar keywords to those from refugees. Therefore the main objective of this pilot study is to develop a robust framework for identifying tweets from Syria refugees and this includes processes such as finding, extracting, cleaning and labelling tweets, extracting key features from the tweets and classifying them. For this study, Syria refugees are defined as Syrians who are considering to leave Syria and seek refuge elsewhere, those are on-route to their destinations or those who have recently reached their destinations.

A. Extraction and labelling of tweets

A number of tools were available for extracting tweets including Twitter API [8] streaming API [9], the R Project for Statistical Computing [10] and the Rselenium [11]. The Rselenuim was chosen as it could be modified to extract older tweets and easy to use. A customized code for Rselenuim is developed to extract tweets. The code utilizes cascading style sheets (CSS) selector function to locate and extract the given fields of tweets such as time of tweets, username and tweet content. The tweets extracted for this pilot study were from year 2011 to 2017 as the Syrian migration was more intense in this period of time. The usernames of the extracted tweets were anonymized after extraction.

The tweets were extracted based on different combinations of keywords which were expected to be used by refugees. The initial keywords used were based on general knowledge about the Syria war such as 'Syria', 'escape', 'leave', 'drowning', 'fear', 'asylum' and 'help me'. The extracted tweets were manually checked and studied during the development phase. When a refugee or potential refugees had identified, their other tweets were extracted as well for further analysis.

The extracted tweets contain anonymized username, time of tweet, retweets, replies and likes. To identify the main key words in refugees' tweets, the extracted tweets were used to form a word cloud [12]. Figure 1 shows the main keywords and their significance in the word cloud. A simple analysis indicated that Syria refugees inclined to seek asylum in Canada, Sweden and Germany. Word clouds could provide an insight about the most used terminology on twitter by refugees. Another interesting observation is that there are common spelling errors, e.g. Germani, in the extracted tweets. These unexpected spelling errors can make the identification of refugees' tweets more difficult.

Figure 1: Word cloud highlighting important keyword associations.

Around 40 different keywords were used in various combinations to extract relevant tweets from twitter. This resulted in around 5000 tweets. The extracted tweets were cleaned by removing all non-English words such as URLs, usernames, punctuation, extra white space and symbols. For the purpose of creating a training set for identifying tweets from refugees, the cleaned tweets were manually classified into two classes and labelled as: 'refugee' and 'commentator'. The tweets labelled as "refugee" are those whose authors explicitly express their own or family's desire to escape from Syria, on- route to their destination country or reached their destination recently. The commentators are non-refugees who have an opinion on the Syria crisis, e.g. tweets from journalists and general public from outside of Syria. The training set contains 212 tweets, which were randomly chosen with an aim of having a roughly equally representatives from each of the two classes. These tweets were mostly from different authors. Out of the 212 tweets, 93 were from commentators, 119 from refugees. The tweets were labelled based on common sense. For example, if a person is showing an intention to leave a country, they will be marked as "refugee".

B. Feature Extraction and classification

The key features of the cleaned tweets were extracted using the Bag of Words (BoW) method [13]. BoW is a natural language processing technique and can be used to categorize textual information and represent it as an unordered set of words with their respective frequencies. Each frequency of word is used as a feature. A total 107 features were extracted from the 212 tweets. The 107 BoW features along with 8 different metadata, such as the number of followers, retweets and likes, form a total of 115 features which were used for training a classifier.

A number of classification tools were considered and tested but the Random forest classifier was chosen for this study mainly due to the fact that it has the ability to compensate for

overfitting to their training set [14]. The performance of the classifier is validated using the 5-fold cross validation (CV) due to its robustness. In 5-fold CV, the original dataset is randomly partitioned into 5 subsets, where a single subset is retained as the validation data for testing and the remaining 4 subsets are used as training data. The CV process is then repeated 5 times with each of the 5 subsets used exactly once as the validation data. The results from the folds are averaged to produce a single estimation [15].

III. Experiment Results and Discussions

The classification results of the Random Forest classifier using the 5-fold CV are that 140 of the 212 were correctly classified which resulted in 66% classification accuracy.

To further analyze the classification performance, the confusion matrix was used to cross-check the correctly and incorrectly classifications, as shown in Table 1. The diagonal values in these matrices signify the correctly classified tweets while the other adjacent values highlight the incorrectly classified tweets.

Table 1 Confusion Matrix for 5-fold Cross Validation.

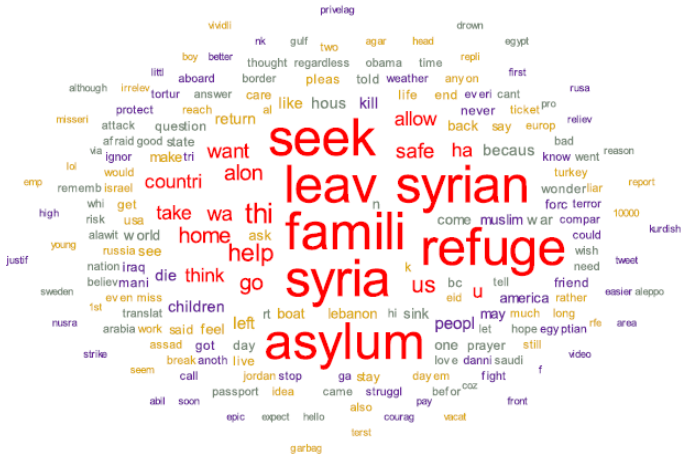
		Predicted by Classifier		
		Commentator	Refugee	Total
Manual Labelled	Commentator	**40 (43%)**	53 (57%)	93
	Refugee	19 (16%)	**100 (84%)**	119
	Total	59	153	**212**

It is observed that the classifier performed well on classifying tweets from refugees, with an accuracy of 84%. As for tweets from commentators, the performance of the classifier was moderate, with 43% accuracy, whilst 53% of them were misclassified as 'refugees'. It was believed the high misclassification rate was caused by the imbalance number of representative samples between two classes, i.e., more tweets from refugees than commentators. To verify this hypothesis, the Random-forest classifier was trained again but adopting a resampling method, which ensure the number of representative samples from each class is exactly the same by randomly duplicating some samples from the class with fewer samples and randomly removing some samples from the class with higher number of samples. The results of the classification with resampling data is shown in the confusion matrix in Table 2. Overall, 171 out 212 tweets were correctly classified, giving an average classification rate of 81%. For tweets from refugees, 77% of them were correctly classified, while 23% of them were misclassified as tweets from commentators. For tweets from commentators, 89% of them were correctly classified, but 17% of them were misclassified as tweets from refugees.

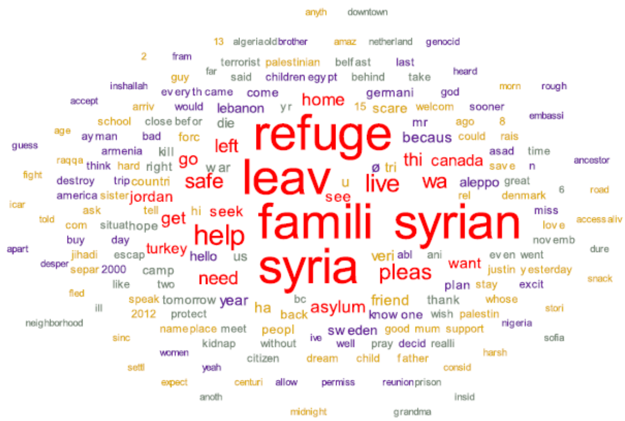
Table 2 Confusion Matrix for 5-fold Cross Validation.

		Predicted by Classifier		
		Commentator	Refugee	Total
Manual Labelled	Commentator	**89 (84%)**	17 (16%)	106
	Refugee	24 (23%)	**82 (77%)**	106
	Total	113	99	**212**

From these preliminary results, it appears the random-forest classifier trained with data resampling gives significant better overall classification results. Comparing with the classifier that trained with the original data (i.e., more refugee samples), the classifier trained with data resampling significantly reduced misclassifications of commentators' tweets as refugees' (from 57% to 17%), but also mildly increase misclassifications of refugees' tweets as commentators' (from 19% to 24%). The reason for this behavior could be due to the fact that insufficient unique features in the training samples of two classes and this makes them difficult to be well distinguished from each other. This could be caused by the overlapping of key words in the two classes of tweets. Figure 2(a) and (b) shows the word clouds of the tweets from refugees and commentators respectively. It is observed that many keywords appear in both word clouds, e.g. Syria, Syrian, refuge, leav and famili. Further study is needed on the search terms and feature extraction.



(a)



(b)

Figure 2: Word clouds showing the keywords and their significances in the tweets from (a) refugees and (b) commentators.

IV. Conclusion and Future Work

A robust framework was developed to find, extract, clean and label tweets related to Syrian refugees based on key words. A Random Forest classifier was developed and tested using the labelled tweets. Experiment result shows the developed classifier achieved an overall average of 81% classification rate. The results suggest that the developed classifier can be used to reliably extract refugees' tweets from data mixed with many unwanted commentators' tweets. This study concludes that applying suitable data analytic techniques on information extracted from social media platforms such as twitter has a great potential for identifying refugees' messages, which has previously been identified as a difficult task. With further research on classifying refugees into those sub-classes such as those intends to leave their home country, on-route to their destination and reached their destinations recently and working out their current location from their messages, these techniques together can potentially predict the migration patterns of the refugees effectively.

As for future work, it was found that some refugees sought helps from famous figures, who sometimes lent their popular social media accounts to refugees such that they can publicize their stories, so it is suggested to broaden the search term to include the messages from pro-refugee famous figures. Another observation was that many refugees restrained from openly posting or disclosing their location information on public platforms such as Twitter because of fears from getting caught by human smugglers or their government agencies. However, they were willing to share their whereabouts on trusted close groups on other social media networks such as Facebook. It is therefore suggested to seek permission to access and analyse messages in closed groups for refugees. However, safeguard must be in place to protect refugees' identities, e.g. anonymize their messages. Furthermore, many messages from Syria refugees were in Arabic. It is therefore suggested to translate the Arabic message into English and include them for analysis.

References

[1] Cassar, C.M., Gauci, J.P. and Bacchi, A. 2016. Migrants' Use of Social Media in Malta. The People for Change Foundation.

[2] Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp. Res. Part C Emerg. Technol. 68, 285–299. https://doi.org/10.1016/j.trc.2016.04.005

[3] Bellomo, N., Clarke, D., Gibelli, L., Townsend, P., Vreugdenhil, B.J., 2016. Human behaviours in evacuation crowd dynamics: From modelling to "big data" toward crisis management. Phys. Life Rev. 18, 1–21. https://doi.org/10.1016/j.plrev.2016.05.014Ali, A., Qadir, J., Rasool, R. ur, Sathiaseelan, A., Zwitter, A., Crowcroft, J., 2016. Big data for development: applications and techniques. Big Data Anal. 1, 2. https://doi.org/10.1186/s41044-016-0002-4

[4] Gillespie, M., Ampofo, L., Cheesman, M., Faith, B., Iliadou, E., Issa, A., Osseiran, S., Skleparis, D., 2016. Mapping Refugee Media Journeys: Smartphones and Social Media Networks. The Open University / France Médias Monde.

[5] Zagheni, E., Garimella, V.R.K., Weber, I., State, B., 2014. Inferring International and Internal Migration Patterns from Twitter Data, in: Proceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion. ACM, New York, NY, USA, pp. 439–444. https://doi.org/10.1145/2567948.2576930

[6] Ali, A., Qadir, J., Rasool, R. ur, Sathiaseelan, A., Zwitter, A., Crowcroft, J., 2016. Big data for development: applications and techniques. Big Data Anal. 1, 2. https://doi.org/10.1186/s41044-016-0002-4

[7] Brouckman, L., Wang, A. 2017. Analyzing Twitter Sentiment On the Refugee Crisis in 2016.

[8] Jäderberg, D. 2016. Sentiment and topic classification of messages on Twitter and using the results to interact with Twitter users. Uppsala Universitet.

[9] Wickham, H., 2016. rvest: Easily Harvest (Scrape) Web Pages.

[10] R: The R Project for Statistical Computing [WWW Document], n.d. URL https://www.r-project.org/ (accessed 11.13.17).

[11] Harrison, J. 2017. RSelenium: R Bindings for "Selenium WebDriver".

[12] Lohmann, S., Ziegler, J., Tetzlaff, L., 2009. Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration, in: Human-Computer Interaction – INTERACT 2009, Lecture Notes in Computer Science. Presented at the IFIP Conference on Human-Computer Interaction, Springer, Berlin, Heidelberg, pp. 392–404. https://doi.org/10.1007/978-3-642-03655-2_43

[13] Zhang, Y., Jin, R., Zhou, Z.-H., 2010. Understanding bag-of-words model: a statistical framework. Int. J. Mach. Learn. Cybern. 1, 43–52. https://doi.org/10.1007/s13042-010-0001-0Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79. https://doi.org/10.1214/09-SS054

[14] James, G., Witten, D., Hastie, T., Tibshirani, R., 2017. An Introduction to Statistical Learning: with Applications in R, 1st ed. 2013, Corr. 7th printing 2017 edition. ed. Springer, New York.

[15] Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79. https://doi.org/10.1214/09-SS054

[16] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical Machine Learning Tools and Techniques, 4 edition. ed. Morgan Kaufmann, Amsterdam.