Scalable system for opinion mining on Twitter data. Dynamic visualization for data related to refugees' crisis and to terrorist attacks

Adrian Iftene *adiftene@info.uaic.ro*
"Alexandru Ioan Cuza" University, Faculty of Computer Science
Iasi, Romania

Mihai-Ștefan Dudu *mihai.dudu@info.uaic.ro*
"Alexandru Ioan Cuza" University, Faculty of Computer Science
Iasi, Romania

Andrei-Remus Miron *remus.miron@info.uaic.ro*
"Alexandru Ioan Cuza" University, Faculty of Computer Science
Iasi, Romania

Abstract

Social networks such as Twitter or Facebook grew rapidly in popularity, and users use them to share opinions about topics of interest, to be part of the community or to post messages that are available everywhere. This paper presents a system created in order to process streamed data taken from Twitter and classify it into positive, negative or neutral. The results of these processing's can be visualized in a suggestive manner on Google Maps, users can select the language of the tweets, can group tweets that present the same news and can even display a dynamic evolution of the news in terms of its appearance. With all this amount of information it is very opportune to do some data analysis to detect different types of events (and their locations) that happen worldwide, especially at the time when this data represents information related to refugee crisis or signals terrorist attacks.
Keywords: Twitter, real time processing, opinion mining, dynamic visualization.

1. Introduction

Twitter has evolved constantly from year to year (https://about.twitter.com/company), stating today that they have 313 million active users each month, 1 billion unique visitors monthly, 82 % active users on mobile devices with 3860 employees worldwide.

Tweets from Twitter began to be the input for many of the applications developed in recent years. The goals of these applications is varying from identifying the opinions of users [6], [19], [27, 28], [31], [36] (in particular for products or companies, for presidential candidates, for local elections or for an event, etc.), to identifying extreme natural phenomena [41] (earthquakes, fires, hurricanes, etc.), the condition of satisfaction of people [2] (on health services which they benefit from, or on services provided by local government, etc.), the level of depression of people or groups of people [20], etc.

A system that processes data from Twitter has three classical modules: (1) a module for search and local storage for tweets (in a database or as XML or JSON), (2) a module for data processing (usually natural language processing services which removes stop-word, identifies the POS, identifies name entities, identifies and classifies the opinions into the positive, negative or neutral, etc.), (3) a module to display the results (either statistics or information geographically visualized when this is possible).

In [7] authors define the notion of terrorism informatics and they present techniques that involve acquiring, integrating, analyzing and managing information related to terrorism. In [8] authors introduce a special issue based on nine papers about terrorism informatics containing

information related to (1) the status of the domain, existing techniques and trends [40], (2) how the web content can be exploited to identify terrorist websites [39], (3) how the collected information can be combined in order to help users in decision making [17], (4) visualization techniques used to show the obtained results [35].

2. Processing data on Twitter

2.1. Opinion mining on Twitter

In [24], the authors present the main techniques used to identify opinions in existing data on Twitter. Their study shows how the existing techniques process these tweets in order to categorize them into positives, negatives or neutrals. Existing techniques based on machine learning algorithms (Naive Bayes, Max Entropy and SVM), or lexicons based techniques are presented, analyzed and compared.

In [15] and [36] are proposed models for classifying tweets using emoticons. [3], [13], [27], [37] and [49] implemented models based on maximum entropy or Naive Bayes to classify tweets. Features spaces contained retweets, hashtags, links, punctuation marks in combination with features like words polarity and POS (Part-of-speech). In [4] authors experimented using multinomial naive Bayes, stochastic gradient and Hoeffding tree.

In [1] experimented with models based on unigrams based on features and based on trees. For the trees-based model, they represented a tweet like a tree. The feature-based model uses 100 features and unigrams-based model uses 10,000 features. [10] proposed an approach using hashtags from Twitter, punctuation marks, words, n-grams and templates for various features, which are then combined into a single vector of features used to classify tweets. They used the k-NN (k-Nearest Neighbor) algorithm to label opinions by building a features vector for each example of test data.

[46] used methods based on “bag-of-words” perform opinion mining - the relationships between words are not taken into account, and the document is represented as a collection of words. Kamps et al. 2004 used lexical database like WordNet [12] to determinate the emotional content of a word in a multi-dimensional space. They developed a metric on WordNet and determined semantic polarity of adjectives. [28] brought into focus the difficulties that can be encountered when we want to classify tweets. Spam and variety of existing languages on Twitter makes the work of identifying opinions very difficult.

2.2. Real time analysis of Twitter data

One of the problems encountered when operating with information on Twitter is related to scalability of the application you want to develop [9], [23] and [43]. Existing techniques involving the use of Hadoop machines [5], [21] and [33], HPC platforms (High Performance Computing) [20], and hardware components included EC2 (Amazon Elastic Compute Cloud). Big Data analysis techniques like Map-Reduce are used as well [32]. Machine learning techniques were implemented on these super-machines in order to perform supervised classification, in the context of having to deal with large of data (millions of tweets). Many existing approaches have focused both on processing speed and the quality of the results obtained by them, currently the accuracy of results is between 74% and 82%, depending on the used technique and input size.

2.3. Methods of visualizing information from Twitter

Visualization of identified opinions in tweets using maps like Google Maps [16] (Figure 1, left) is very similar to visualization of friendship relations between users of Facebook [29] (Figure 1 in the right).

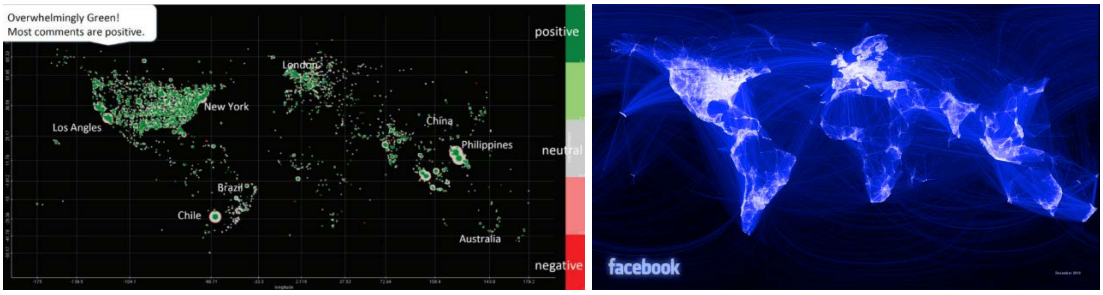

Fig. 1. (a) Displaying opinions that appear in tweets (b) in friendship relations between users from Facebook

In the paper [30], the authors present the most discussed subjects from Twitter, graphically, by using the TopicFlow application, using the time axis. Interesting are also two approaches: (1) one using hashtags from tweets to make statistics [45] (Figure 2a) and (2) another one to display opinions from posts coming from USA [11] (Figure 2b).

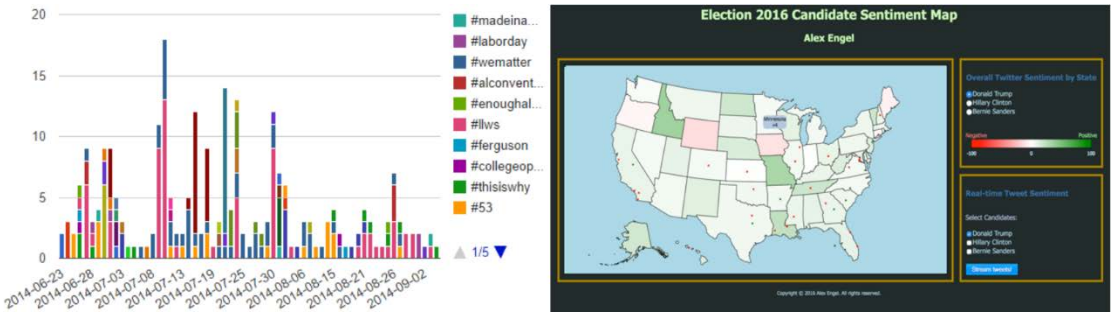

Fig. 2. (a) Statistics based on hashtags, (b) centralized displaying on US states at presidential election 2016

3. Proposed solution

For developing our tweet classification application we used data from Twitter, more specifically tweets collected between March 4, 2017 and April 3, 2017. From the collected tweets we decided to mark and analyze more carefully the tweets that are related to the refugee crisis and to the terrorist attacks. In order to filter these tweets, we prepared a list of keywords grouped into six sections: (1) NATO - contains the list of the 28 NATO member countries, (2) the European Union - the list of the six member countries of the EU, but not members of NATO, (3) Fight and attack - contains the lexicon of the words (i.e. words usually used to express the concepts of fighting and attacking), (4) Civilian - contains the lexicon of the word (5) Peace - contains the lexicon of the word and (6) European refugee crisis - contains the lexicon of the word. In addition we considered the top three origin countries of the immigrants (Syria, Afghanistan and Iraq). In total, the list contains 103 words for each of the six target languages (English, German, French, Greek, Turkish and Italian). We initially started with English, German, French languages, but, considering the refugee topic we selected for our analysis, we also included Greek, Turkish and Italian, as the languages of the most common transit countries.

3.1. Ensuring scalability

Twitter provides live data including details about the user who posts the tweet, the language of the tweet, the location etc. The connection between Twitter and our application is made using a module that constantly collects tweets. Received data is filtered in two steps (1) firstly, we keep only those in English, German, French, Greek, Turkish or Italian that contain information related to the geolocation of the user who made the posting (we collected this way 66,975 tweets), (2) secondly, we filter the tweets that contain words from the list with

information related to the refugee crisis and to the terrorist attacks (we thus marked 2,702 tweets). These 2,702 tweets were then annotated with opinions (negative/positive/neutral) (see Section 3.3). To ensure scalability, we used threads to asynchronously process the tweets that are going to be stored in the database: (1) identify opinion of the tweet and (2) grouping similar tweets. After marking a tweet as processed, it ready to be visualized on Google Map.

3.2. Grouping similar tweets

Redistribution of information on Twitter is getting more and more popular, being it a simple retweet or copy/pasting a tweet and adding hashtags, emoticons or links. The purpose of this step was to group similar tweets, in order to ensure displaying in a diversified way at the GUI interface. For this we perform experiments with four distance similarity algorithms: Jaro-Winkler, Levenshtein, Needleman-Wunsch and Smith-Waterman.

Jaro–Winkler distance is a measure of similarity between two strings [18], [48]. It is a variant of the Jaro distance metric a type of string edit distance, and was developed in the area of record linkage (duplicate detection). The lower the Jaro–Winkler distance for two strings is, the more similar the strings are. The Levenshtein distance is a string metric for measuring the difference between two sequences [26]. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other. The Needleman-Wunsch algorithm, which is based on dynamic programming guarantees finding the optimal alignment of pairs of sequences. The algorithm essentially divides a large problem (e.g. the full sequence) into a series of smaller problems and uses the solutions to the smaller problems to reconstruct a solution to the larger problem [25], [34]. The Smith-Waterman algorithm is a dynamic programming method for determining similarity between nucleotide or protein sequences. The algorithm was first proposed in 1981 by Smith and Waterman and is identifying homologous regions between sequences by searching for optimal local alignments. The Smith-Waterman algorithm is built on the idea of comparing segments of all possible lengths between two sequences to identify the best local alignment. It is based on calculation of local alignments instead of global alignments of the sequences and allowing a consideration of deletions and insertions of arbitrary length [44].

For detecting which tweets from the 2702 are similar, each tweet was compared with all that follow it. Because, the number of possible combinations is very high, in order to improve the execution time, a caching mechanism from Microsoft was used. Table 1 below presents the execution time for different algorithms.

Table 1. Execution times (in minutes) for considered distances.

	Without cache	**With cache**	**With cache after removing stop words**
Jaro-Winkler	06:05.35	03:50.11	01:46.06
Levenshtein	22:35.37	12:10.08	11:26.82
Needleman-Wunsch	51:11.47	34:21.63	21:22.49
Smith-Waterman	39:46.32	35:33.20	23:51.46

Three of the algorithms find 683 tweets with distance 0 (tweets are similar), while Smith-Waterman find 769 tweets with distance 0 (it sees a tweet and a retweet being the same). In the end, we decided to use Jaro-Winkler algorithm to group similar tweets due to its fast running time, and we consider to show on Google Maps 2.019 tweets from initial set of 2.702 tweets.

Since tweets are constantly collected, when a new tweet matching the search criteria is found, it is automatically classified, depending on the computed distance, either into one of the existing clusters, or into a new cluster, and then visualized on the map.

For further research, we will try to group tweets by their credibility [14] and based on concepts presented in them. The idea is that the same concept may be expressed using

different syntax and we want to group tweets expressing the same concepts in the same cluster using semantic similarities [38].

3.3. Opinion mining

In order to classify the collected tweets by opinions we used Sentiment140 API [42], which internally uses a (semi-supervised) trained model. Unfortunately Sentiment140 API only works with Spanish and English languages. In this situation we would have been forced to discard a big part of our collected data. To overcome this situation we decided to make use of a translation API (Yandex) [50] which would translate the tweet's content from its native language into English. Once translated in English the tweet is sent to Sentiment140 API [42] in order to be annotated accordingly, then being persistently stored in our MongoDB database for later usage (offline analysis).

Once annotated, they will be sent to a module that is responsible to distribute them to the web clients interfaces. Also, annotated tweets will be stored to database to enable offline analysis, when needed. The web interface simply takes the information from the module that distributes it and presents it to users in a graphical way.

3.4. Display tweets on Google Maps

As can be seen in Figures 3 and 7, tweets are grouped into clusters, represented by circles or hexagons distributed on the global map. The color of a cluster will be intuitive and related to its contain: white for neutral clusters (containing the same number of positive and negative tweets or only neutral tweets), a color between green and red (green representing positive tweets and red negative tweet). On the right side are displayed the latest tweets without duplicates (green being for the tweets positive annotated, and red being for the tweets negative annotated).

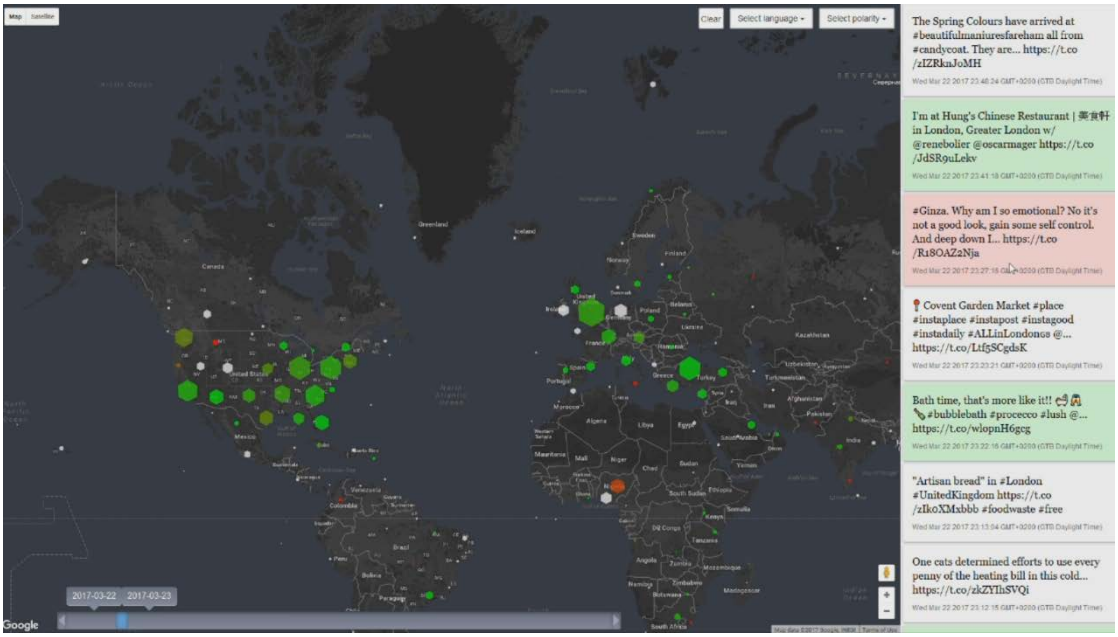

Fig. 3. Viewing clusters on Google Maps

When a cluster grows, it affects other clusters on the map, which will decrease proportionally, so the largest clusters being the most visible. To speed the display the clusters who associated a circle with a radius less than a value set by the user are displayed as dots. If the user will use the option of zooming in a given region (Figure 4 left), where he spotted a cluster higher, it will be divided into smaller clusters, until we see only clusters composed of a single tweet (Figure 4 right).

Fig. 4. (a) View details of a larger cluster, (b) zoom in for a region from the map

As shown in Figure 3 at the bottom, we can select the period for which we want to see the information on the map, the minimum being one day (by default uses the current day), and for the moment the maximum being seven days (for not expect too much the processing of data that must be shown on the map). Also, by setting the length of a period (from 1-7 days), we can activate the option to view in an animated way the progress over time of the tweets. With a specific step (being set between 1 seconds and 5 seconds) the cursor moves from the setted period to the current days, at every step, disappearing tweets that came out of current range and appearing tweets entering in the new current interval.

4. Results

4.1. Statistics

First of all it is important to note that we only had access to 1% of Twitter's Public Stream consisting of sample real-time tweets. Between 4 March and 3 April, we collected a total of 66,975 tweets that has geolocation attached. Among all languages that we monitored English and Turkish are the most popular (full languages distribution can be seen in Table 2 and in Figure 5).

Table 2. Tweets count by native language.

Language	Number of tweets	Positive / Negative tweets	Tweets with specific keywords	Positive / Negative tweets with keywords
English	28,208	7,052 / 1,399	1,138	135 / 14
German	1,115	31 / 2	45	3 / 0
French	2,305	53 / 4	93	4 / 0
Italian	2,231	40 / 5	90	2 / 0
Turkish	33,017	343 / 19	1,332	54 / 2
Greek	99	2 / 1	4	1 / 0
Total	**66,975**	**7,521 / 1,430**	**2,702**	**199 / 16**

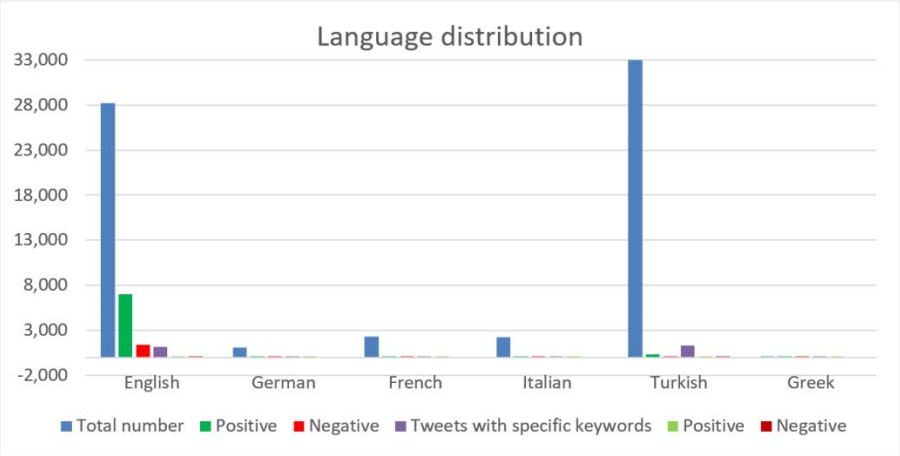

Fig. 5. Language distribution

How we can see in Figure 5, the most active languages are English and Turkish, and from this reason many from next analyzes are performed on these languages. After 4 April we still collected data from Twitter and now our database contains 160,883 tweets.

4.2. Results analysis

In Figure 6 we have a representation of tweets (that contain keywords) distribution per days (between 18 March and 4 April) and languages (English and Turkish). On the Y axis (vertical) we got total tweets count on the indicated day for the indicated language. We can instantly observe that on the second day (22nd of March 2017 - date of the Westminster Attack in London) tweets count is significantly higher than the day before. Also, we can see in Figure 7, the differences between tweets distribution for 21 and 22 March.

Higher number of tweets means higher number of keywords occurrences which means higher chances of tweets being related to the events we are interested in. Also, the content of tweets is related to this attack: 22 April: "*UK Terrorist Attack: President Buhari Sympathises With Britons*", 23 April: "*British police raid after deadly Westminster attack*", "*Another terrorist attack, this time in my hometown of London! We are mad, pissed, sad, horrified*", "*Four killed in UK parliament terrorist attack*".

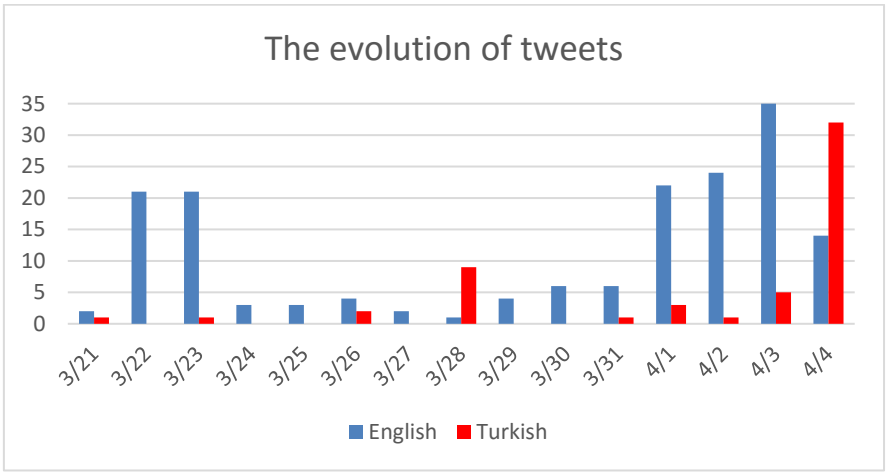

Fig. 6. The evolution of tweets for English and Turkish

The red bars corresponds to Turkish language which we found that on 28 April, correspond to relevant tweets that contain words "*bombardment, weapon, tank, arms*" and on 4 May correspond to irrelevant tweets that contain mainly the word "*Turkey*". We noticed that

French, German, Greek and Italian languages are not very popular, providing us no or very little relevant information to analyze.

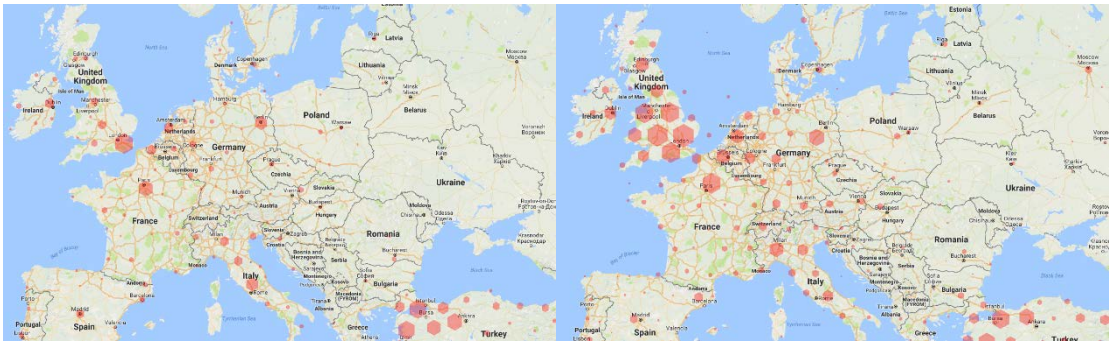

Fig. 7. Tweets distribution for 21 March (in the left) and 22 March (in the right)

To better analyze the data collected we have selected five keywords (*assault*, *attack*, *raid*, *terrorist* and *victim*) that we considered to be the most relevant for this analysis and we searched for occurrences over a period of five days (Figure 8).

As you can see there is a big increase in usage of keyword "attack" starting with 22nd of March 2017 (date of the Westminster Attack in London). What is interesting is that the day following the attack presents a bigger increase in usage of "attack" keyword than attack's day. This happens mostly because of mass media starting to cover the incident. Two days after the incident there is a clear frequency decrease of keyword "attack" and no occurrences at all for "terrorist" and "raid".

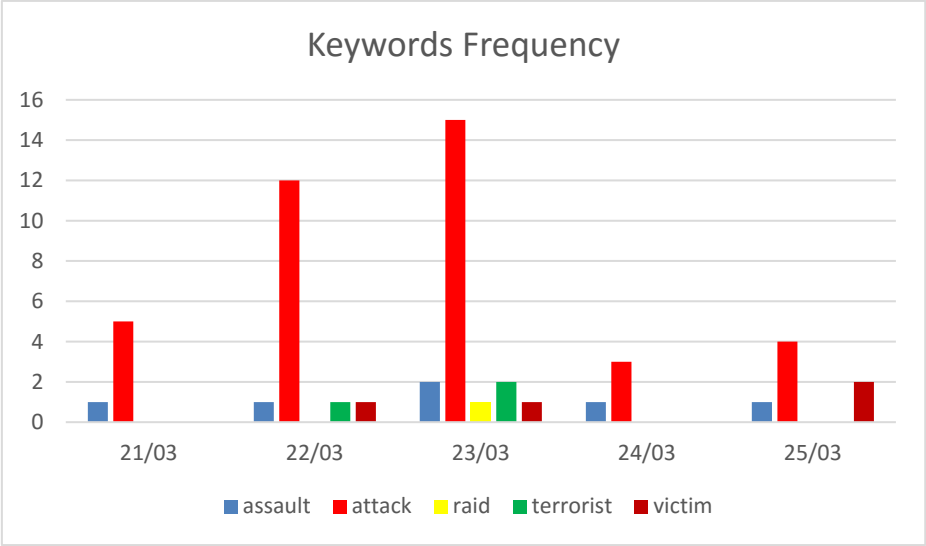

Fig. 8. Keywords Frequency (21st of March 2017 to 25th of March 2017)

We also performed opinion analysis on tweets that contains the keywords we are interested in (Figure 9). For easier comparison we kept the same date interval as in Figures 6, 7 and 8, but this time we got no relevant connection between polarity distribution and the event that took place at Westminster. As you can see in chart tweets are mostly neutral and positive. This happens due to the fact that Sentiment140 API model for polarity annotation probably was not trained on tweets related to violent acts like a terrorist attack so we end up with irrelevant distribution of polarities.

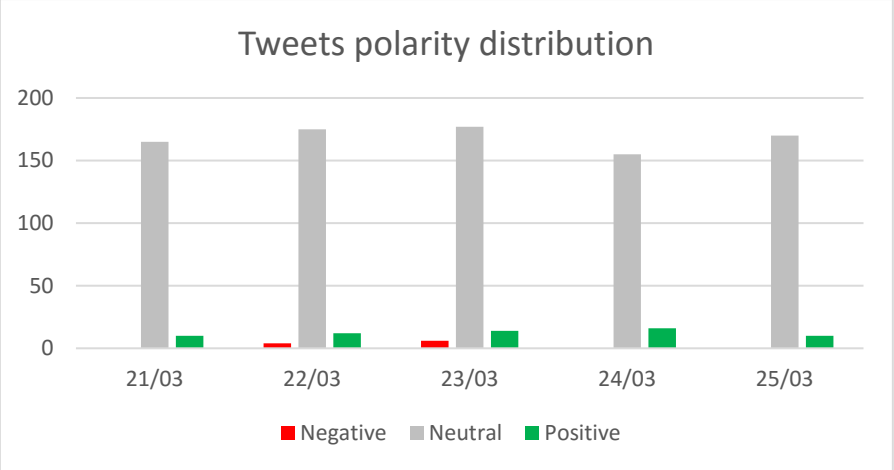

Fig. 9. Tweets polarity distribution

The gap between languages distribution is given by Twitter's active users base (mostly English speaking users) hence the regions that generates the most tweets are United Kingdom and Turkey (in Europe).

4.3. Error analysis

The small number of relevant tweets compared to the total number of tweets collected is due to massive amounts of tweets being automatically posted (weather forecasts, check-in, jobs announcements, advising, etc.). This kind of tweets are irrelevant for our analysis so there is no point to rely on their content. For the future, we need to pay more attention to how we collect information for our system to avoid such tweets coming into our database.

The biggest number of irrelevant tweets comes from Turkey - mostly check-in posts or messages of the type advertising automatically added to twitter containing no information beside user's location or advertisement. These tweets are so frequent that they make up almost 50% of tweets collected by us.

As you can see in figures 5 and 9 most tweets are classified as neutral by Sentiment140. But many of these tweets contain words or hashtags that would allow us to classify them as negative (*silly*, *death*, *fuck*, *killin*, *bad*, *lose*, *fucking*, *block*, *incident*, *protest*, etc.) or positive (*great*, *champion*, *good*, *inspiring*, *thank*, etc.). Due to this problem we created a list with triggers for negative and positive opinions and we completed the result of Sentiment140 service with an additional checking in case the identified opinion is neutral. For example, on 8th of April 2017, we collected 4,354 tweets on English and classified them with Sentiment140 service, 3,704 tweets being neutral, 569 positive and 83 negative. After additional checking we classified 363 neutral tweets into positive tweets and 72 from neutral into negative.

Furthermore our analysis was made only on tweets that contains geolocation information because we have to know the location where the data is coming from to place it on the map. Hence we could have more relevant results if more tweets would contain this data.

5. Conclusions and Future Work

Social networks can provide at a given moment a lot of information, especially in the moments when are special natural phenomena (like earthquakes, floods, fires, snow, chills or extreme heat), or when are presidential elections, or when are attacks with many casualties, that stir up panic. In such cases, the amount of data resulting from these networks are increasing exponentially, and there appears the need to apply advanced techniques to provide scalability and results in real time.

To analyze data on Twitter we used a dictionary with specific terms to monitor the refugee crisis and the terrorist attacks, with specific techniques from natural language

processing: POS, removing stop-words, identification of opinions, etc. We also saw how big data analysis techniques such as map-reduce allow us to obtain useful information from the data gathered.

The work that we have undertaken have come up with a new proposal to display information on Google Maps. Compared to previous approaches, which displays all the tweets collected over a certain period, we can manage to display tweets from a period with 1 day to 7 days. Interestingly, when we use during day and time we move from the oldest to the latest date and time we can see how it evolves from day to day posts on Twitter worldwide. In the future, we have to find a more efficient data management, to work with longer periods of seven days, while having real-time processing of data.

We believe that social media mining will become a key component used to find users' opinions about a subject, to plan marketing strategies, to decide whether a customer will buy a product or not, to help in case of emergency, etc. That is why in future any application, no matter how small it is, it should acknowledge of this kind of information and exploit it.

Acknowledgements

We would like to thank to Elena Șușnea, involved in the building of the first lists of words related to refugee crisis, and to Iuliana Minea, involved in the grouping of similar tweets.

References

1. Agarwal, B., Xie, I., Vovsha, O., Rambow, R., Passonneau, R.: Sentiment Analysis of Twitter Data. In Proceedings of the ACL 2011, Workshop on Languages in Social Media, pp. 30-38 (2011)
2. Ali, A., Magdy, W., Vogel, S.: A Tool for Monitoring and Analyzing HealthCare Tweets. HSD 2013, pp. 23-26 (2013)
3. Barbosa, L., Feng, J. (2010) Robust Sentiment Detection on Twitter from Biased and Noisy Data. COLING 2010: Poster Volume, pp. 36-44.
4. Bifet, F. E.: Sentiment Knowledge Discovery in Twitter Streaming Data. In Proceedings of the 13th International Conference on Discovery Science, pp. 1-15, Berlin, Germany: Springer (2010)
5. Bingwei, L., Blasch, E., Chen, Y., Shen, D., Chen, G.: Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier. In Big Data, 2013 IEEE International Conference on, pp. 99-104 (2013)
6. Borruto, G.: Analysis of tweets in Twitter. Webology, 12 (1), June (2015)
7. Chen, H., Reid, E., Sinai, J., Silke, A., Ganor, B.: Terrorism informatics: Knowledge management and data mining for homeland security (Integrated Series in Information Systems). New York: Springer (2008)
8. Chen, H., Zhou, Y., Reid, E. F., Larson, C. A.: Introduction to special issue on terrorism informatics. Information Systems Frontiers, 13 (1), pp. 1-3, (2011)
9. Cuesta, Á., David, F., Moreno, M.: A Framework for Massive Twitter Data Extraction and Analysis. In Malaysian Journal of Computer Science, pp. 50-67 (2014)
10. Davidov, D., Rappoport, A.: Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. Coling 2010: Poster Volume pp. 241-249, Beijing, August (2010)
11. Engle, A.: Election 2016 Twitter Sentiment Map. Stanford University. http://web.stanford.edu/class/cs448b/cgi-bin/wiki-sp16/images/5/5f/Alex_Engel_FinalPaper.pdf (2016)
12. Fellbaum, C.: WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press (1998, ed.)
13. Gamallo, G., Garcia, M.: Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets. 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, Aug 23-24, pp 171-175 (2014)

14. Gînscă, A. L., Popescu, A., Lupu, M., Iftene, A., Kanellos, I.: Evaluating User Image Tagging Credibility. In Proceedings of 6th International Conference of the CLEF Association, CLEF'15 Toulouse, France, September 8-11. Experimental IR meets Multilinguality, Multimodality, and Interaction. Lecture Notes in Computer Science, Publisher Springer International Publishing, 9283, pp. 41-52 (2015)
15. Go, R., Bhayani, L., Huang, L.: Twitter Sentiment Classification Using Distant Supervision. Stanford University, Technical Paper (2009)
16. Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D. A., Haug, L. E., Hsu, M. C.: Visual Sentiment Analysis on Twitter Data Streams. In Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST), 23-28 Oct. pp. 275-276 (2011)
17. Hayne, S., Troup, L., McComb, S.: "Where's Farah?": Knowledge silos and information fusion by distributed collaborating teams. Information Systems Frontiers, 13 (1), pp. 89-100 (2011)
18. Jaro, M. A.: Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. Journal of the American Statistical Association, 84 (406), pp. 414–420 (1989)
19. Jenders, M., Kasneci, G., Naumann, F.: Analyzing and predicting viral tweets. In Proceedings of the 22nd International Conference on World Wide Web (WWW '13 Companion). ACM, New York, NY, USA, 657-664 (2013)
20. Jiang, B., Topaloglu, U., Yu, F.: Towards large-scale twitter mining for drug-related adverse events. In Proceedings of the 2012 international workshop on Smart health and wellbeing, ACM, pp. 25-32 (2012)
21. Jimmy, L., Kolcz, A.: Large-Scale Machine Learning at Twitter. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 793-804 (2012)
22. Kamps, J., Marx, M., Mokken, R. J., De Rijke, M.: Using wordnet to measure semantic orientations of adjectives (2004)
23. Karanasou, M., Ampla, A., Doulkeridis, C., Halkidi, M.: Scalable and Real-time Sentiment Analysis of Twitter Data. http://sentic.net/sentire2016karanasou.pdf (2016)
24. Kharde, V. A., Sonawane, S. S.: Sentiment Analysis of Twitter Data: A Survey of Techniques. International Journal of Computer Applications (0975 – 8887), 139 (11), April 2016, pp. 5-15 (2016)
25. Lesk, A.: Introduction to Bioinformatics. Oxford University Press (2002)
26. Levenshtein, V. I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady. 10 (8), pp. 707–710 (1966)
27. Liang, P.W., Dai, B. R.: Opinion Mining on Social Media Data. IEEE 14th International Conference on Mobile Data Management, Milan, Italy, June 3 - 6, pp 91-96 (2013)
28. Luo, Z., Osborne, M., Wang, T.: An effective approach to tweets opinion retrieval. Springer Journal on World Wide Web, Dec 2013, DOI: 10.1007/s11280-013-0268-7 (2013)
29. Ma, D.: Visualization of social media data: mapping changing social networks. Master Thesis, University of Twente. 68 pages (2012)
30. Malik, S., Smith, A., Hawes, T., Papadatos, P., Li, J., Dunne, C., Shneiderman, B.: TopicFlow: visualizing topic alignment of Twitter data over time. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '13). ACM, New York, NY, USA, pp. 720-726 (2013)
31. Maynard, D., Funk, A.: Automatic detection of political opinions in Tweets. Proceedings of the 8th International Conference on The Semantic Web, pp. 88- 99, Heraklion, Crete, Greece (2011)
32. Michal, S., Romanowski, A.: Sentiment analysis of Twitter data within big data distributed environment for stock prediction. In Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on, IEEE, pp. 1349-1354 (2015)

33. Mohit, T., Gohokar, I., Sable, J., Paratwar, D., Wajgi, R.: Multi-Class Tweet Categorization Using Map Reduce Paradigm. In International Journal of Computer Trends and Technology, pp. 78-81 (2014)
34. Needleman, S. B., Wunsch, C. D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 48 (3), pp. 443-453 (1970)
35. Oh, O., Agrawal, M., Rao, R.: Information control and terrorism: tracking the Mumbai terrorist attack through twitter. Information Systems Frontiers, 13 (1), pp. 33-43 (2011)
36. Pak, A., Paroubek. P.: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Proceedings of the Seventh Conference on International Language Resources and Evaluation, pp.1320-1326 (2010)
37. Parikh, R., Movassate, M.: Sentiment Analysis of User- Generated Twitter Updates using Various Classification Techniques. CS224N Final Report, 2009 (2009)
38. Pedersen, T., Patwardhan, S., Michelizzi, P.: WordNet::Similarity: measuring the relatedness of concepts. In Demonstration Papers at HLT-NAACL 2004 (HLT-NAACL-Demonstrations '04). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 38-41 (2004)
39. Qin, J., Zhou, Y., Chen, H.: A multi-region empirical study on the internet presence of global extremist organizations. Information Systems Frontiers, 13 (1), pp. 75-88 (2011)
40. Roberts, N. C.: Tracking and disrupting dark networks: Challenges of data collection and analysis. Information Systems Frontiers, 13 (1), pp. 5-19 (2011)
41. Sakaki, T., Okazaki, M., Matsuo, Y.: Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development. IEEE Trans. on Knowl. and Data Eng. 25, 4 (April 2013), pp. 919-931 (2013)
42. Sentiment140 API: http://help.sentiment140.com/api. Accessed April 2017
43. Sheela, L. J.: A Review of Sentiment Analysis in Twitter Data Using Hadoop. In International Journal of Database Theory and Application, 9 (1), pp.77-86 (2016)
44. Smith, T. F., Waterman, M. S.: Identification of Common Molecular Subsequences. Journal of Molecular Biology, 147, pp. 195–197 (1981)
45. Stojanovski, D., Dimitrovski, I., Madjarov, G.: TweetViz: Twitter data visualization. Proceedings of the Data Mining and Data Warehouses (2014)
46. Turney, P. D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics (2002)
47. Twitter usage: https://about.twitter.com/company. Accessed April 2017
48. Winkler, W. E.: String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods. American Statistical Association: pp. 354–359 (1990)
49. Xia, R., Zong, C., Li, S.: Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences: an International Journal, 181 (6), pp. 1138–1152 (2011)
50. Yandex: https://www.yandex.com/. Accessed April 2017