Detecting Citizen Problems and Their Locations Using Twitter Data

Gizem Abalı
Department of Computer Engineering
Muğla Sıtkı Koçman University
Mugla, Turkey
gizemabali93@gmail.com

Enis Karaarslan
Department of Computer Engineering
Muğla Sıtkı Koçman University
Mugla, Turkey
enis.karaarslan@mu.edu.tr

Ali Hürriyetoğlu
Centre for Language Studies
Radboud University
Nijmegen, Netherlands
a.hurriyetoglu@cbs.nl

Feriştah Dalkılıç
Department of Computer Engineering
Dokuz Eylul University
Izmir, Turkey
feristah@cs.deu.edu.tr

***Abstract*—Twitter is a social network, which contains information of the city events (concerts, festival, etc.), city problems (traffic, collision, and road incident), the news, feelings of people, etc. For these reasons, there are many studies, which use tweet data to detect useful information to support the smart city management. In this paper, the ways of finding citizen problems with their locations by using tweet data is discussed. Tweets in Turkish language from the Aegean Region of Turkey were used for the study. It is aimed to form a smart system, which detects problems of citizens and extracts the problems' exact locations from tweet texts. Firstly, the collected data was analyzed to get information of any city event, citizen's complaint or requests about a problem. After the possibility of detecting tweets, which have any city problem, was ensured, two datasets were created. The first one consists of the tweets that have an event information or a problem and the second one has the tweets, which have other information not related to our study. Then Naive Bayes classifier was trained on the annotated tweets and was tested on a separate set of tweets. Accuracy, precision, recall, and F-measure of the classifier is given. A location recognizer, which finds the Turkish place names in a text, is created and applied on the tweets that are marked as information-containing by the classifier to detect the location of the problem precisely. The first findings of the project is promising. The high accuracy, which is obtained by the classifier, shows that it is proper to use this classifier for our study. The location recognizer is planned to be improved and place names on the real-time tweet data is to be detected.**

***Index Terms*— Data Analysis, Machine Learning, Smart Cities, Social Media Analysis, Text Mining.**

I. INTRODUCTION

In the past years, the technology has developed and brought a big change to the lives of people. Many people started to migrate to the cities. The increment in the number of citizens brought along many city problems like pollution, accidents, traffics, etc. While changes were occurring, many social networking services (like Facebook, Snap chat, etc.) which people use to communicate were created and rapidly grew. Users have started using them to not only communicate with the others, but to report the problems of the cities where they live in. Because of this usage of the social platforms, the social media analysis became a popular research area in the urban projects and some research teams were formed to work to create better urban areas. In order to get information about an area, people tend to use many social network platforms. One of these platforms is Twitter. Twitter has 330 million monthly active users as of the third quarter of 2017 according to the statistics of the Statista Statistics Portal. Twitter is different from other social networks; as the main purpose is not to interact with friends; it is formed for sharing and seeking information [1]. Owing to the fact that Twitter is used by people to proclaim their wishes or to announce problems, it becomes a research tool that can be used to analyze the situation of an area and find the problems.

II. RELATED WORKS

When the past studies using Twitter data were studied, it was seen that many of them are about analyzing traffic tweets [2-4], some of them were done to understand the happiness in a city [5], finding the locations of where a user tweets[6], analyzing disaster tweets [7, 8], and detecting place names that are passed on a tweet text [9].

Some studies [10, 11] show that, Twitter is the most important social network which people use to communicate with other users and give information to them during a disaster or a critical event. What's more, it is observed that tweet texts

978-1-5386-4478-2/18/$31.00 ©2018 IEEE

consist of information about traffics, transportation in a city, security, environmental factors in these studies [10, 11]. When we consider these researches and our aim to form a smart system in this study together, they show us that Twitter data can also be used in Turkish language to perform our work.

In the next section, basic concepts; smart city concept, tweet data and the usage of machine learning in the tweet data analysis will be discussed. Then, the methodology and the implementation will be given with the results. In the last section, conclusion will be given and possible future works will be discussed.

III. BASIC CONCEPTS

A. Smart City

In the past years, the numbers of people who live in cities and rural areas changed and a big increment was observed in the cities' population. Cities' population count has increased from 746 million to 3.9 billion from 1950 to 2014 [12] and it is predicted to reach 69% of the world population by 2050 [13].

A smart city is still a fuzzy concept. When we explain the features of smart cities, we can list it as the following [14]:

- A city that uses smart computing technologies to create its all infrastructures and services (including health-care, education, transportation).
- A city that connects its all infrastructures together to enhance the intelligence.
- A city that watches its all critical transportation, communication, energy and water infrastructures and also major buildings.
- A city which is more effective, liveable, fair and sustainable.

A smart city can be explained as the area where uses intelligence functions to collect the data and synthesize it to improve the efficiency of services, equity, sustainability and quality of life [15].

B. Tweet Data

Tweet data provides details of tweets (including tweet text, images, videos, retweet counts, favorite counts, location information, date, time, user name, user screen name, user picture, font color, ids, reply id if the tweet is a reply to another tweet and etc.). The data can be obtained by creating a Twitter application and using a token and access ids.

C. Machine Learning

Machine learning is a field of computer science, which allows computers to learn without be programmed. It has been used in many research areas (including spam detection, search engines, optical character recognition). It is seen that machine learning is also a popular technique that is used in tweet analysis studies [16 – 18] including Twitter sentiment classification, opinion finding, etc.

IV. IMPLEMENTATION

A. Using The Tweet Data For Smart Cities

For this project, we aim to detect the tweets that include a problem about a part of a city and find exact location of the problem. In Figure 1, some tweets that are related to our work were given. In the first tweet, the user says, “Küçükbakkalköy Beyaz Street, Nergis, Çiğdem and Yenidoğan Avenues are closed. We have a hard time. Please immediately help us”. In the second tweet the user says, “Geothermals in Aydın Germencik and İmamköy harm the agriculture and the ecology. We wait for you to come to Aydın”. In the last tweet consists the text, “It is said that the road is close around Aydın, Muğla, Yatağan tunnel. Please be careful. We have neither snow tire nor chain in our cars. Don't take a risk”. On these texts, Muğla, Aydın, Germencik, İmamköy, Yatağan, Küçükbakkalköy, Nergis, Çiğdem and Yenidoğan are the place names. As you can see in the tweets, users talk about their needs or requests. Moreover, the common feature of these tweets is that all of them have location information. After detecting tweets, which can have critical information, a process for finding location information can be applied and the most problematical parts of a city and sometimes suggestions of the citizens might be found.

Figure 1. Example tweets which have critical information.

B. Dataset

In this study, we used the tweet data that were collected from a selected area (Aegean region) in Turkey. In order to collect data, Twitter API is used. The Twitter API provides nearly 1% of tweet data that streams from the selected area. This is the amount, which Twitter gives as free. MongoDB is used to save the data in JSON format. The continuity of the data is provided by a Linux script.

C. Methods

The aim of this project is to detect the tweets that have critical information and the citizens' requests about the city. First, we accumulated tweets that have information about the intended area and also the tweets that do not have any information to help us. Then, we created two separate datasets as related ones and not related ones to our work. We trained the system by using Naive Bayes Classifier to detect information-containing tweets about the area. We put the tweets that the classifier found on another collection apart from related dataset and inserted really information-containing tweets into related dataset. In the Figure 2, the methodology for the classification of tweets is explained. After obtaining tweets truly related to our study, we applied location recognizer that we created using Python programming language on them to find the related location.

D. Training Data

In order to train new tweet data, we used two datasets that were formed as related and not related. Related dataset has tweets that consist of a problem about an area. The dataset named as "Not Related", consists of tweets that are not related to our study. We used Naive Bayes Classifier on the 100 example tweets that we marked as related (77 tweets) and not related (23 tweets). The results of the text (TP, FP, FN, TN) are found according to the confusion matrix in Table 1.

TABLE I. CONFUSION MATRIX

		Actual Class	
		Related	Not Related
Predicted Class	Related	**T**rue **P**ositive (**TP**)	**F**alse **P**ositive (**FP**)
	Not Related	**F**alse **N**egative (**FN**)	**T**rue **N**egative (**TN**)

In the confusion matrix, "True Positive" stands for correctly predicted related values, "False Positive" stands for incorrectly predicted related values, "True Negative" stands for correctly predicted not related values and "False Negative" stands for incorrectly predicted not related values. We found that true positive count is 71, false positive count is 8, the number of true negatives is 15 and the number of false negatives is 6. We calculated the accuracy, precision, recall and F-measure values by using the equations (1)-(4). Accuracy, precision, recall and F-measure values obtained as 0.86, 0.90, 0.92 and 0.91, respectively.

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

$$\text{F-measure} = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

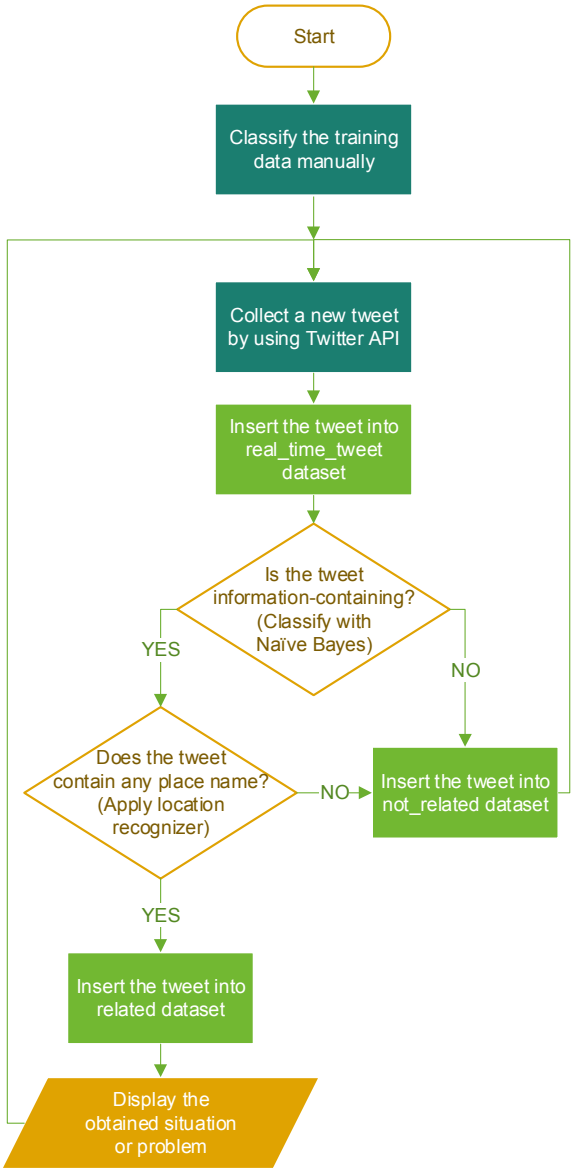

Figure 2. Flow diagram of the tweet classification methodology.

E. Location Analysis

A location recognizer that finds Turkish place names in given texts was created and applied on tweet texts. If we look at an example tweet "Efemçukuru Altın Madeni bütün bir İzmir'in suyunu kirletiyor" (it means "Efemçukuru Gold Mine pollutes the water of İzmir"), "İzmir" and "Efemçukuru" that are included in the tweet text are place names (Izmir is a city in Turkey and Efemçukuru is a street that is connected to a town of Izmir). The location recognizer uses a geonames file[1] which contains all place names from Turkey (including city names, town names, street names, etc.). In order to detect place names, language structures were created by using pyparsing [19] and regular expression libraries of Python programming language. The source code for our location recognizer can be found in the Bitbucket repository[2] [9].

[1]http://www.geonames.org/
[2]https://bitbucket.org/hurrial/placenames/branch/turkish_location_recognizer

During the process of creating the location recognizer, we encountered a few problems because of the Turkish language word structure. The problems that we have faced are the following:

- Some place names are homophones with common words in Turkish (example; tahta (wood), yağmur (rain), siyah (black), sandık (box), etc.)
- Some place names are also used as surnames in Turkish (example; Akalın, etc.).

In order to solve first problem, we removed the place names that stand for common homophone words from the place name list. For the second problem, we formed a language grammar with regular expressions to separate correct location names from people's surnames. (Example; if we say that "Metin Akçapınar değerli bir oyuncuydu." ("Metin Akçapınar was a precious actor."), "Akçapınar" is the surname of the person whose name is Metin but if we say that "Akçapınar, Muğla ilinin Ula ilçesine bağlı bir mahalledir." ("Akçapınar is a neighborhood that is connected to Ula District of Muğla Province."), "Akçapınar" is a neighborhood of a city called as Muğla.)

After separating tweets that were marked as information-containing by the classifier, we applied the location recognizer on these tweets to find the precise related locations.

V. CONCLUSION

In this study, it is discussed how to detect any city problem or citizen requests by analyzing the tweet messages which are sent in the city coordinates. The possibility of detecting city problems by using the tweet data is found as convenient as the results of the classifier is promising. In addition, the tweet messages related to our work were analyzed and it was observed that all of them has location information. This common feature shows us that the tweet data can be used to find the locations of a city where problems occur and it can be useful to support the city management. The importance of detecting the location names is seen and a location recognizer is formed. In the future work, it is planned to use the tools through the real-time tweet data. We also aim to find the topics of what citizens talk about with respect to hashtags and we believe that these will help in city management. In addition, a web application will be created to show the results.

ACKNOWLEDGMENT

This work is supported by the Scientific Research Project Fund of Muğla Sıtkı Koçman University under the project number 16/160.

REFERENCES

[1] I. L. B. Liu, C. M. K. Cheung, and M. K. O. Lee, "Understanding Twitter usage: What drives people to continue to tweet," in *Proc. 2010 Pacific Asia Conference on Information Systems (PACIS)*, pp. 928-939.

[2] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-atikom, and P. Chaovalit, "Social-based traffic information extraction and classification," in *Proc. 2011 International Conference on ITS Telecommunication (ITST)*, pp. 107-112.

[3] M. Hasby and M. L. Kodra, "Optimal path finding based on traffic information extraction from Twitter social-based traffic information," in *Proc. 2013 International Conference on ICT for Smart Society (ICISS)*, pp. 1-5.

[4] S. B. Marupudi, "Framework for semantic integration and scalable processing of city traffic events," M.Sc. Thesis, *Wright State University*, 2016.

[5] W. Guo, N. Gupta, G. Pogrebna, and S. Jarvis, "Understanding happiness in cities using Twitter: Jobs, children and transport," in *Proc. 2016 IEEE International Smart Cities Conference*, pp. 1-7.

[6] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating Twitter users," in *Proc. 2010 ACM International Conference on Information and Knowledge Management*, pp. 759-768.

[7] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: Real-time event detection by social sensors." in *Proc. 2010 International Conference on World Wide Web*, pp. 851-860.

[8] A. Acar and Y. Muraki, "Twitter for crisis communication: Lessons learned from Japan's tsunami disaster," *International Journal of Web Based Communities*, vol. 7.3, pp. 392-402, 2011.

[9] G. Abalı, A. Hürriyetoğlu, and E. Karaarslan, "Event information based location name analysis in the Twitter data: A preliminary study," in *Proc. 2016 International Conference on Computer Science and Engineering (UBMK)*.

[10] B. D. M. Peary, R. Shaw, and Y. Takeuchi, "Utilization of social media in the east Japan earthquake and tsunami and its effectiveness," *Journal of Natural Disaster Science*, vol. 34(1), pp. 3-18, 2012.

[11] P. Anantharam, P. Barnaghi, K. Thirunarayan, and A. Sheth, "Extracting city traffic events from social streams," *ACM Transactions on Intelligent Systems and Technology*, vol. 6(4), pp. 43:1–27, July 2015.

[12] United Nations, "World Urbanization Prospects: The 2014 Revision, Highlights," Department of Economic and Social Affairs, Population Division, 2014.

[13] A. Barresi and G. Pultrone, "European strategies for smarter cities," *Tema, Journal of Land Use, Mobility and Environment*, vol. 6(1), pp. 61-72, 2013.

[14] H. Chourabi, and et al., "Understanding smart cities: An integrative framework," in *Proc. 2012 Hawaii International Conference on System Science (HICSS)*, pp. 2289-2297.

[15] M. Batty, and et al., "Smart cities of the future," *The European Physical Journal Special Topics,* vol. 214(1), pp. 481-518, 2012

[16] M. Pennacchiotti and A. M. Popescu, "A machine learning approach to Twitter user classification," in *Proc. 2011 International AAAI Conference on Weblogs and Social Media*, pp. 281-288.

[17] A. Z. H. Khan, M. Atique, and V. M. Thakare, "Combining lexicon-based and learning-based methods for Twitter sentiment analysis," *International Journal of Electronics, Communication and Soft Computing Science & Engineering*, vol. 4(4), pp. 89-91, 2015.

[18] G. Sidorov, and et al., "Empirical study of machine learning based approach for opinion mining in tweets," in *Proc. 2012 Mexican International Conference on Artificial Intelligence*, Springer Berlin Heidelberg, pp. 1-14.

[19] P. McGuire, *Getting started with pyparsing*, California: O'Reilly Media Inc., 2007, p. 65.