A LEXICON BASED MECHANISM FOR IDENTIFYING AND MONITORING SECURITY THREATS ON ROADS

Munazza Zaib* Shahzad Nizamani** Muhammad Ali Nizamani*** Intesab Sadhayo****

*Department of Software Engineering, MUET, SZAB Campus, Khairpur Mir's, Sindh, Pakistan ** Department of Software Engineering, MUET, Jamshoro, Sindh, Pakistan *** Department of Electrical Engineering, Isra University. Hyderabad ****Department of Computer Systems Engineering, QUEST, Nawabshah, Sindh, Pakistan

ABSTRACT

Given the booming expansion of social media, it is not surprising that the field of sentiment analysis has seen advancements rapidly in recent years. Nevertheless, the use of sentiment analysis is quite limited in the field of transportation to assess the safety of an area. This research paper propose the sentiment analysis of traffic or crime information as a new way to handle this problem. To achieve this, we have used one of the user generated contents i.e. Twitter as our source of information. Twitter has emerged as an essential new tool to make social measurements. Millions of tweeps express their thoughts and sentiments about any topic imaginable on daily basis voluntarily. This heap of data is quite significant from both research and business perspectives. Thus, we intend to design an application through our research with which the categorization of data publically available at Twitter can be done, so that the users can have access to the customized and useful information related to the areas they are planning to visit. To carry out this research practically, data from Twitter was collected for a particular source and destination and sentiment analysis was performed using SentiWordNet. The result yielded in overall polarity of the tweets informing users about the safety of all the available routes. This study will help greatly in the development of intelligent transportation systems and our experimental results demonstrate the effectiveness of the system.

Keywords—sentiment; polarity; safety; lexicon

1. INTRODUCTION

Many researches in Europe and America have been conducted in the field of route planning and transportation that focuses entirely on providing traffic information, which unfortunately, is not critical and of utmost importance in our society. In Middle East (Asia specifically), the main concern is the safety of a road user. In improving well-being in human mobility, one can think of other criteria that could be meaningful, such as the friendliest, most scenic, safest, most enjoyable, or matching one's interest the best [2]. There has been an alarming rise in street crimes during last few years and countries are once again in the grip of suchcrimes, thus creating a sense of insecurity among the road dwellers. The crux of our idea is whether real-time, geo-tagged social media streams can be used to enhance everyday experience in the way we interact with places, for instance by avoiding crime-prone areas.

There exists a lot of useful information regarding incidents that might possess risk while travelling from one place to another such as mobile snatching, firing or abduction etc. However, a mechanism is needed to retrieve and categorize the information before presenting it to the user in a useful manner. Thus, the objective of this study is to use public opinions and sentiments to help people to decide about the route they are planning to take when travelling from one place to another.

The remaining paper is organized as: the related work is discussed in Section-II. The implementation methodology is discussed in Section-III. Results has been discussed in Section-IV and the paper is concluded in Section-V along with the recommendations about future work in Section-VI.

2. LITERATURE

Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text mining [1]. In other words, it is a measure of opinion and subjectivity in textual data [15].

Sentiment analysis systems are being applied in almost every business and social domain because opinions are central to almost all human activities and are key influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are largely conditioned on how others see and evaluate the world. For this reason, when we need to make a decision we often seek out the opinions of others [1]. Different researches have conducted in the last few years that attempt to employ the concept of

Sentiment Analysis in different aspects. As we know that with the rapid rise and popularity of social media, sentiment analysis has developed quite speedily in recent years [3] because of its ability to find emerging trends and topics. In the field of sentiment analysis, many people have worked in this aspect using different techniques and algorithms. Every project has its own importance and significance in its own way and perspective.One can find number of projects and research work done on this topic.

3. IMPLEMENTATION METHODOLOGY

Typically, the methodologies for lexicon based sentiment analysis are based on the fact that he overall polarity of the textual data can be calculated by the polarity of the individual words which compose it [9].

Fig. 1. Illustrates the architecture of our Android application; the architecture is based on a number of stages which includes: a) collection of data from Twitter, b)segmentation of sentences(tweets), c)extraction of sentiment polarities, d) sentiment calculation, and e) evaluation of results.

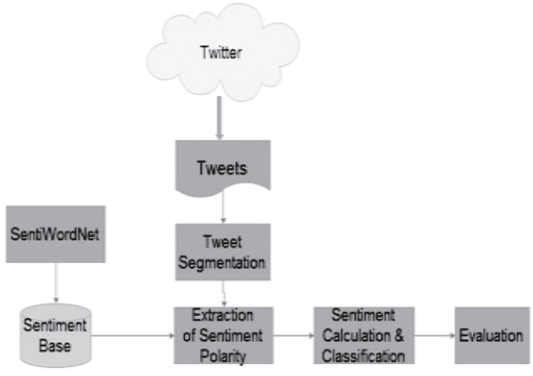

Fig. 1. Architecture of the system.

a) Data Collection:

To ensure that conclusions are purely public opinion based, we have gathered all the data from Twitter (represented by upper block in Fig. 1. We have used Twitter4j library to get data from Twitter.

b) Segmentation of Sentences:

Sentence segmentation is further sub divided into two phases:

i. Tokenization that decomposes the sentences into small tokens such as words, numbers of symbols of varying types.

ii. Process of speech tagging that assigns a part of speech tag on every symbol or word.

c) Extraction of Sentiment Polarities

The polarity of the words can be extracted from sentiment base. However, the sentiment base needs to be updated from time to time since words and topics discussed on internet changes quickly.

d) Sentiment Calculation:

The process of sentiment analysis is then performed on the extracted data. Individual score of each sentence is calculated first which serves as an input to calculate total score of the data.

e) Evaluation:

The final evaluation of the collected score is carried out in this step and one of the ratings shown in Table 1 is then assigned based on the total score to the area that the user is planning to travel.

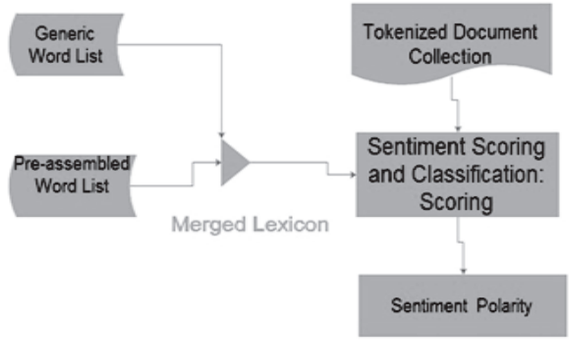

Fig. 2. Sentiment Classification Process

f) Setiment Base:

The process of sentiment analysis consists of two linked segments, i.e. , sentiment lexicon and the sentiment polarity of those words. Since our major task is to focus on information about crime andtraffic conditions, we have developedasentiment base that consists of keywords related to our desired domain and are extracted from SentiWordNet. SentiWordNet is an extension of WordNet and is available freely for research purposes. SentiWordNet[7] is a lexical resource used in the field of opinion mining. SentiWordNet assigns to each Sysnet of WordNet[8] three sentiment scores namely objectivity, positivity and negativity, showing how Objective, Positive or Negative the words are. The score of these sentiments ranges from 0.0 to 0.1

and their overall sum is 1.0 for each sysnet. These words are arranged as a set of synonyms[13]. As we are only interested in knowing how safe or dangerous the routes are, therefore, we have considered only positive and negative scores of the words as shown in Fig. 3. The words taken from SentiWordNet are stored in "Preassembled word list" whereas general words such as very, totally, completely etc are stored "Generic Word List" as shown in Fig. . These two lists are merged to form ou crime and traffic lexicon to which tokenized document set is passed as an input and the process of classification is carried out as shown in Fig. 2.

```
# POSID PosScore        NegScore        SynsetTerms
Gloss
a       00064787        0.625   0       good#5
beneficial#1    promoting or enhancing well-being; "an
arms limitation agreement beneficial to all countries";
"the beneficial effects of a temperate climate"; "the
experience was good for her"
a       00106020        0       0       good#2 full#6
having the normally expected amount; "gives full measure";
"gives good measure"; "a good mile from here"
a       00231252        0       0.75    worse#1
(comparative of `bad') inferior to another in quality or
condition or desirability; "this road is worse than the
first one we took"; "the road is in worse shape than it
was"; "she was accused of worse things than cheating and
lying"
a       00478311        0       0.75    tough#6 bad#3
feeling physical discomfort or pain (`tough' is
occasionally used colloquially for `bad'); "my throat
feels bad"; "she felt bad all over"; "he was feeling tough
```

Fig. 3. Fragment of SentiWordNet

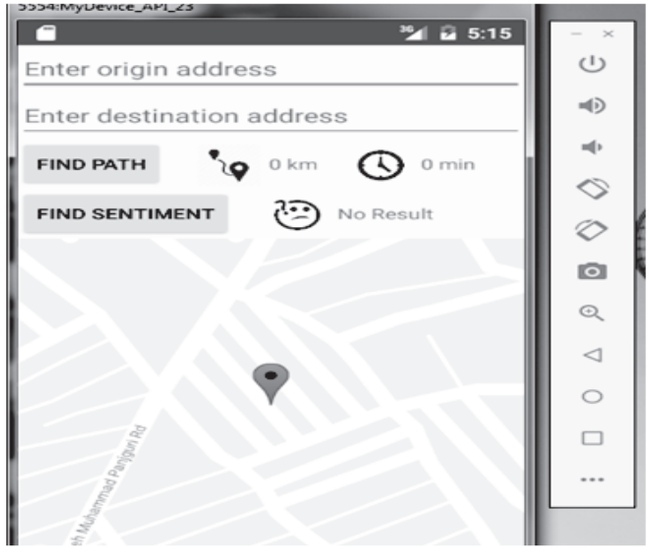

Fig. 4. Final interface of application.

- Table 1. Classification of Sentiment

	Range	Sentiment
Total Score	>=0.75	Strong Positive
Total Score	> 0.25 AND <= 0.5	Positive
Total Score	> 0 AND <=0.25	Weak Positive
Total Score	< 0 AND >= -0.25	Weak Negative
Total Score	< -0.25 AND >= -0.5	Negative
Total Score	<= -0.75	Strong Negative
Total Score	0	Neutral

4. RESULTS:

Fig. 4 shows the final interface of our application. We have tested the application and it works well in predicting the sentiment of routes. When the find path button is clicked, all the possible routes will be drawn between source and destination provided by user on Google Map using Google Map Direction API as shown in Fig. 5. The brighter shade of green shows the path as being safer rather than the darker one. The screen also shows the distance and time needed to travel between the two points. On clicking the 'Find Sentiment' button, the rating of each individual path (from source to destination) is calculated as shown in the Fig. 5 that helps user in making decision about choosing he route. The term 'Neutral' is only for the main route. The sentiment for alternate routes is only shown through lines Google map. Fig 6 shows two paths, one safe and the other unsafe, from Lyari to Sohrab Goth. The path that passes through Nazimabad had negative tweets about target killing and other incidents that is why it is unsafe and is shown by pink line. Fig. 7 shows the tweets extracted by the application having information about source and destination entered by user.

```
.googlemapdemo I/System.out: Move back to Chicago for school or stay in Florida ..
.googlemapdemo I/System.out: Nothing like booking a spur of the moment Chicago trip to fit nicely in between the Florida & Colorado
.googlemapdemo I/System.out: Chicago feels like winter time in Florida
.googlemapdemo I/System.out: I have cousins on snap from all over the U.S so i know what's going down in Cali,Chicago,New York, Flori
.googlemapdemo I/System.out: Massive Shootings: Sandy Hook, Charleston S.C, Boston, Aurora, Dallas, Florida, Oregon, San Bernardino,
.googlemapdemo I/System.out: #SlingTV move to Florida from Chicago. If you actually let me watch Chicago programming I would actuall
.googlemapdemo I/System.out: Another friend ex friend was in airport in Florida getting ready to board plane for Chicago. Lived that
.googlemapdemo I/System.out: 6:30 in Florida which means it's 5:30 in Chicago...debating getting payback and waking up @CoreyHollis
.googlemapdemo I/System.out: Folks from, Boston, New Jersey, Ohio, Florida,Chicago Mexico, Missouri, Illinois, india, Tenneessee. Wo
.googlemapdemo I/System.out: Chicago then Pensacola Florida gon be a highlight
.googlemapdemo I/System.out: Only time I've been out of Texas was when I went to chicago and that was like 3 years ago so you know I'
.googlemapdemo I/System.out: KJRI = General Consulate of Republic of Indonesia Office..in the US such as in LA,SF,Houston (handle als
.googlemapdemo I/System.out: The Chicago Bears were defending Super Bowl champs, the last time #Kentucky beat #Florida. □ □ #UK
.googlemapdemo I/System.out: Professional Interstate Movers from Chicago to Florida
.googlemapdemo I/System.out: Florida / Kentucky game on CBS in Chicago □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
.googlemapdemo I/System.out: Florida is not close to being as cool as chicago, and will never ever be.
.googlemapdemo I/System.out: I wanna go to chicago and cali and florida and vegas again but w friends this time
.googlemapdemo I/System.out: Like to Chicago , California , st.louis , Colorado , Arizona , Florida plus more  all just for soccer □
.googlemapdemo I/System.out: One more full day in Florida tomorrow, fly to providence on Monday, fly to Chicago on Tuesday, home on
.googlemapdemo I/System.out: In chicago you have people crossing the roads and protesting. In Florida you have iguanas crossing the
.googlemapdemo I/System.out: Hey Chicago Wildcats! Join your fellow #BBN'ers at @PonyInnChicago for UK vs Florida. Free pizza and st
.googlemapdemo I/System.out: My Florida followers; look into the Chicago scene, you'll definitely find someone you would want to bri
.googlemapdemo I/System.out: Chicago; do the same
```

Fig. 7. Tweets extracted using Twitter4j

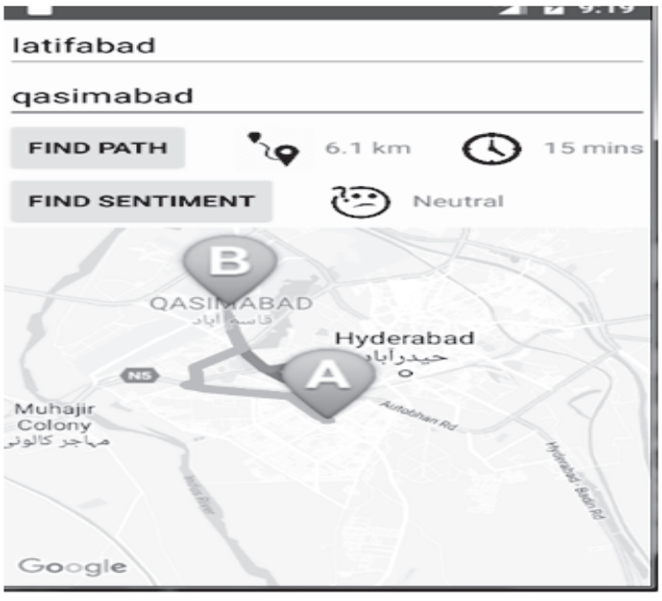

Fig.5. All possible routes from A to B

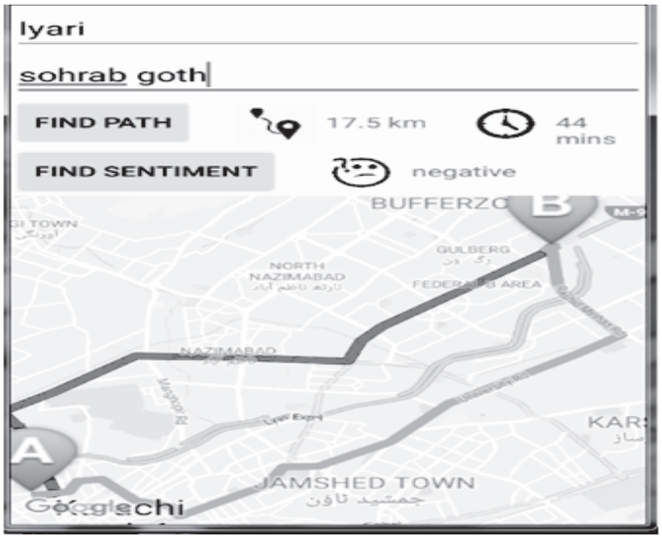

Fig. 6. Safe and unsafe route between source and destination.

5. CONCLUSION

To our best knowledge this is the first attempt to use sentiment analysis to gather the data about criminal activities as well as traffic unlike old experiments that completely focused on collecting information related to transportation. Twitter serves as the fast and real time source of information as compared to news articles thus providing users with the access to timely updates about the situation of crime and transportation in different areas. However, the reliability of data collected from Twitter is low. This project has potential applications in developing intelligent transportation system. It can be used to accommodate traveler's best interest along with ensuring his safety.

6. FUTURE WORK

Though our application seem to work properly, still there are some aspects and areas that needs to be worked on for further enhancement.

Firstly, our project focus on gathering information about routes using manually saved waypoints.. This can be improvised by taking points of interest or landmarks in to account rather than waypoints. The application will look for all the possible routes automatically that exists between source and destination, perform sentiment analysis on each of them individually and suggest the route that is safest of them all .The results would be more accurate and beneficiary then.

Secondly, we have used Twitter4j library for the extraction of tweets from Twitter that returns only newest data available on the website. In order to get all the previous information available on Twitter about a particular area, we can use Streaming API along with Twitter4j.

REFERENCES

[1] Liu, Bing. Sentiment Analysis and Opinion Mining. N.p.: Morgan & Claypool, 2012. Print.

[2] Kim, Jaewoo, Meeyoung Cha, and Thomas Sandholm. "SocRoutes: safe routes based on tweet sentiments."Proceedings of the companion publication of the 23rd international conference on World wide web companion. International World Wide Web Conferences Steering Committee, 2014.

[3] Cao, Jianping, Ke Zeng, Hui Wang, Jiajun Cheng, Fengcai Qiao, Ding Wen, and Yanqing Gao. "Web-Based Traffic Sentiment Analysis: Methods and Applications."Intelligent Transportation Systems, IEEE Transactions on 15, no. 2 (2014): 844-853.

[4] Megally, Mirna. "Information extraction from social media for route planning." (2012)

[5] Wanichayapong, Napong, Wasawat Pruthipunyaskul, Wasan Pattara-Atikom, and Pimwadee Chaovalit. "Social-based traffic information extraction and classification." In ITS Telecommunications (ITST), 2011 11th International Conference on, pp. 107-112. IEEE, 2011.

[6] Pant, Kireet, Dibyendu Talukder, and Pravesh Biyani. "TrafficKarma: Estimating Effective Traffic Indicators using Public Data."Proceedings of the 2nd IKDD Conference on Data Sciences. ACM, 2015.

[7] A. Esuli and F. Sebastiani, "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining." Proceedings from International Conference on Language Resources and Evaluation (LREC), Genoa, 2006.

[8] A. Esuli and F. Sebastiani. "Determining term subjectivity and term orientation for opinion mining." In Proceedings of EACL-06,11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, IT. Forthcoming , 2006.

[9] Musto, Cataldo, Giovanni Semeraro, and Marco Polignano. "A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts."Information Filtering and Retrieval 59 (2014).

[10] Kharde, Vishal, and Sheetal Sonawane. "Sentiment Analysis of Twitter Data: A Survey of Techniques."arXiv preprint arXiv:1601.06971 (2016).

[11] Chen, Feng, Ramayya Krishnan, and CONTRACT No DTRT12GUTG11. "Transportation Sentiment Analysis for Safety Enhancement." (2013).

[12] Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews."Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

[13] Miller, George A. "WordNet: a lexical database for English."Communications of the ACM 38.11 (1995): 39-41.

[14] Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. 2013.

[15] Taboada, Maite, et al. "Lexicon-based methods for sentiment analysis."Computational linguistics 37.2 (2011): 267-307.