Title

OCICATS (Online community input classification to advance transportation services) – a GIS-based decision-support tool

Authors

D., Brownstone
Michael, McBride

Publication Date

2016-04-01

Sidebar Info

Program Steering Committee (PSC): None

June 2016

Title: OCICATS (Online community input classification to advance transportation services) – a GIS-based decision-support tool

Task Number: TO 028
Start Date: May 1, 2015
Completion Date: April 1, 2016

Product Category: --

Task Manager: Christine Azevedo
Christine Azevedo, Associate Transportation Planner
christine.azevedo@dot.ca.gov

TITLE
OCICATS (Online community input classification to advance transportation services) – a GIS-based decision-support tool

AUTHORS
Lourdes V. Abellera, California State Polytechnic University, Pomona
Anand Panangadan, California State University, Fullerton

ACKNOWLEDGMENTS
The California Department of Transportation (Caltrans) provided financial support for this project through the UCCONNECT (University of California Center on Economic Competitiveness in Transportation) Faculty Research Grant for FY 2014-15.

The following are the main student contributors to this work:
Pritesh Pimpale, California State University, Fullerton
Daniel Zhao, California State Polytechnic University, Pomona

regulation. This report does not constitute an endorsement by the Department of any product described herein.

For individuals with sensory disabilities, this document is available in alternate formats. For information, call (916) 654-8899, TTY 711, or write to California Department of Transportation, Division of Research, Innovation and System Information, MS-83, P.O. Box 942873, Sacramento, CA 94273-0001.

ABSTRACT

This project developed software based on machine learning techniques to summarize discussions regarding transportation in California on social media. This tool is intended to reveal factors important to transportation users that may not be evident to transit agencies. The software uses topic modeling to cluster public messages related to transportation on the Twitter social media platform.
Sentiment analysis techniques were then utilized to assign a polarity (positive, negative, or neutral) to each message and then aggregated by topic. The software is thus able to summarize the sentiment towards each automatically identified topic. It was found that the quality of topic identification depends on the size of the dataset, the number of topics that has to be specified to the topic modeling algorithm, and the positive/negative thresholds for the sentiment analysis algorithm.

1. INTRODUCTION

Public transportation agencies can obtain large amounts of information regarding timeliness, efficiency, cleanliness, ridership, and other performance measures. However, these metrics are based on the interests of these agencies and do not necessarily represent the concerns of the customers. Recently, social media have become a platform for people to show their satisfaction or discontent about particular services and products (e.g., Twitter feeds, Yelp reviews, Change.org petitions).

1.1 Problem Statement

Transit agencies rely on regular rider surveys to determine their performance. These traditional metrics include safety, timeliness, efficiency, and cleanliness, among others. As emerging technologies advance and values of societies change, there are many other factors that affect the quality of transportation experience of individual riders. Social media in the form of blogs, Twitter feeds, reviews etc. have become a platform to express these diverse customer needs. For example, Yelp reviewers would prefer that cell phone service is available in underground train stations; they say they do not feel secure if the doors through the stations are not locked because this will encourage anyone, particularly homeless people, to enter the station; they say they would appreciate a station that provides visual interest through art. These are clearly new factors that must be studied by transit

authorities to increase ridership in shuttles, buses and trains. The challenge is therefore to develop a method that can **completely automatically** identify the changing factors that community members consider relevant to a desirable transportation system from the online publicly available social media communications of the members.

The goal is therefore to develop a software tool that will harvest unstructured text data from social media and analyze this data to identify the factors important for riders. Since the size of this data will be large, the ridership factors must be discovered in an automated manner (by using machine learning and text processing techniques). The aggregate subjective opinion (sentiment) expressed towards each of these factors will then be quantified by the software. Ideally, these results will be output in geographic information system (GIS) layers so that the discovered information can be easily used by transportation planners using their existing GIS software.

1.2 Relevance

The project addresses topic category 2, Data collection and use. Social media data are "big data" that contain tons of information about rider sentiments. Although these data are available, it is a challenge to use them because of their sheer volume. These data are largely unstructured which are very different from the historical structured surveys used by transit agencies for many years [1]. The software tool that we developed is the first step toward crowdsourcing these data and present them in a functional way that can be intuitively understood by transportation planners. This tool is meant to identify the factors important for users. Transit authorities can then focus their resources on these factors. For example, if it has been determined that the locks must definitely work for a particular train station so that riders can feel safe, the transit agency must send mechanics immediately to fix the locks, and that this task should be a priority for this particular location. Another example is the use of the new app where a rider can determine the time of the next coming bus. Because this is a new transit technology, and only a certain demographic group uses this service, it will be difficult to conduct a paper survey to determine the performance of this app. Hence, crowdsourcing rider experiences through social media will be a better alternative to know the effectiveness of this new technology. Because the software can help transit agencies prioritize tasks, it is also a decision-support tool, which is in topic category 1.

We also assert that topic category 5, institutional culture and relationships, is also addressed by our project. At present, stakeholder meetings are being held the traditional way, where people gather at a particular time and place to discuss ways to solve a specific problem. OCICATS can be extended to gather sentiments of people for particular topics, for example, how they feel about the creation of a station near their neighborhood. Instead of individual people airing their individual views in these stakeholder meetings, OCICATS can collect and count these views, and select the strength of these sentiments.

Effective use of software such as OCICATS can improve understanding of factors that are important to a transit agencies users and potentially increase the number of riders in shuttles, buses and trains. The increase in ridership reduces the costs of maintaining public transportation services and infrastructure, and hence contributes to the economic competitiveness of a region. This work advances research and practice in the selected categories because currently, there are few studies harvesting social media for enhancement of transportation services [2, 3]. Chen and Krishnan [2] and Collins et al. [3] use sentiment analysis methods on Twitter messages to identify specific problems on public transportation in real-time. In contrast, we intend that OCICATS be used as a planning tool that can identify sentiment relating to the multiple factors that affect utilization of public transit. OCICATS can extract information on factors that may be expressed in different ways and will require both short-term and long-term solutions.

Sentiment analysis coupled with topic modeling holds a lot of promise in automating the collection of reasonable amount of feedbacks from riders. It has advantages over the traditional survey method and stakeholder meetings. The cost of data collection is insignificant. Myriad of data can be obtained in real-time. Concerns of riders can be evaluated based on their words and word patterns.

2. METHODOLOGY

Figure 1 shows the components of the OCICATS software system. Text data from relevant online social media is first harvested by the software. The dataset is defined using keywords so as to contain only those messages which are related to transportation in California. Location information regarding these online messages is also extracted where available. This dataset is then input to *topic modeling* algorithms which can identify statistically co-occurring groups of words (the *topics*), which in the ideal case will each be related to a single aspect of transportation usage as perceived by the social media users. The social media messages relating to each of these aspects is then input to a *sentiment analysis* algorithm which will quantify the strength and polarity expressed by the online community towards these aspects. These components are next described in greater detail.

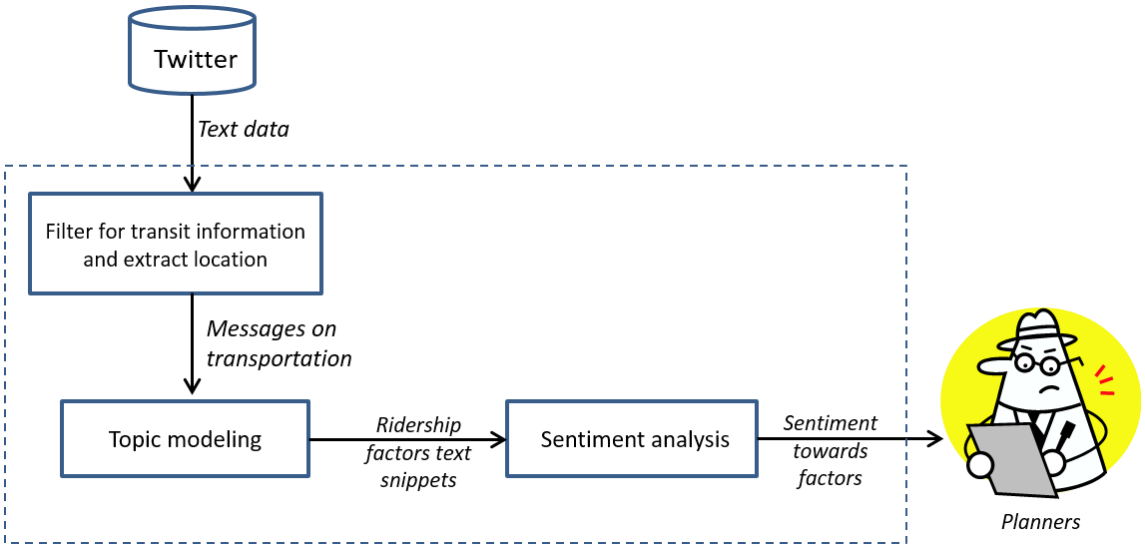

Figure 1: Components of OCICATS system for extracting sentiment expressed in online forums regarding public transit

Dataset: The OCICATS software will operate on publicly available online social media where community members express their opinion of their local transportation issues. Currently, the software processes messages from the Twitter social media platform. Twitter messages ("tweets") will be collected using the API provided by Twitter [4]. Relevant tweets are collected by keyword search and location-based searches, both of which are supported by the Twitter API. Table 1 is a list of the keywords used to capture the live tweets from Twitter in the current version of the code. Tweets have been continuously collected in real-time for a period of one month.

Table 1: Keywords used to capture the live tweets from Twitter

#caltrans	California trains	#expresslanes California
caltrans	California freight	fastrak California
California transportation	high speed rail California	transponder California
California traffic	#hsr California	#bicyclelanes California
California cars	bullet train California	bicycle lanes California
California rail	expresslanes California	

The software for capturing the real time tweets from the Twitter API [4] was developed in Python and uses the Python library called Tweepy [5]. Tweepy is efficient in maintaining a continuous http connection with the server while consuming minimum resources. MongoDB [6] is used for storing the tweets offline as the data obtained from Twitter is in the JSON format which can be easily stored in the NoSQL database like MongoDB. It provides Document-oriented data structure which helps in directly storing the Twitter data without any change in it.

Topic modeling: Topic modeling is a method of representing semantic concepts as a statistical model of word occurrences for each concept. Topic modeling algorithms learn a set of such statistical models from a set of text documents and assign a relative score for the occurrence of each of these topics in a new document. Topic models provide one way to analyze collections of text documents. A "topic" is a cluster of words that tend to occur together. For instance, the ridership topic of safety could be represented with a relatively high probability of the occurrence of words such as “safe”, “danger”, “secure” in a sentence. We used the Latent Dirichlet Allocation (LDA) algorithm [7] to learn the topic models. The LDA algorithm models each document as a mixture of various topics and the probability distribution is assumed to follow a Dirichlet distribution prior. Topic modeling has been proposed as a tool for analyzing text data for social science applications [8] and topic modeling algorithms are now available as software toolboxes that are ready to be applied to a new domain [9]. In OCICATS, we used the Java-based MALLET [10] framework for topic modeling.

Sentiment analysis: After a set of topics have been identified from the text documents/Twitter tweets, the next step is to quantify the subjective opinion expressed by the author of the tweet to each of these performance measures. Several algorithms have been developed by computer scientists over the last decade to automatically identify subjective opinions from text [11,12]. Typically, sentiment is reported as polarity – a one-dimensional value indicating the amount of positive or negative sentiment expressed in the text data. Automatic extraction of sentiment from messages on the Twitter social network has already been applied to measure rider satisfaction [3] and identify transit safety issues [2]. In those approaches, general sentiment was extracted from the Twitter messages.

For OCICATS, we used the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis software. VADER is a lexicon and rule-based tool that is specifically designed to identify sentiment in social media messages [13]. VADER is available as part of the Natural Language Toolkit (NLTK), a Python-based platform for analyzing text data [14]. The sentiment analysis component of OCICATS is written in the Python programming language.

The entire code is submitted to Caltrans in a zip file.

3. RESULTS AND DISCUSSION

3.1 Results

We applied the OCICATS software to a set of 10,400 tweets downloaded over a period of three weeks using the method described earlier. The tweets were pre-processed in the following manner.

1. All characters were converted to lowercase

2. URLs were replaced with a single keyword (i.e., we do not distinguish between URLs)
3. Stop words (very common words in English that appear in many contexts and hence are not useful to discriminate between topics)
4. The tweet is broken into a sequence of words/tokens

For sentiment analysis, if the neutral score returned by the VADER sentiment analysis algorithm for a tweet is greater than 0.8, then that tweet is labeled as neutral. Otherwise, if the positive score is greater than the negative score, then the tweet is labeled positive (and vice-versa).

The number of sampling iterations in the LDA algorithm was set to 2000. The alpha and beta parameter for the LDA algorithm was set to 0.01. The main parameter that needs to be specified in a topic modeling algorithm is the number of topics. We experimented with specifying 10, 20, and 50 topics. The results are given below.

Table 2 shows the five most correlated words defining each of the 10 topics. Here, URL_HTTPS denotes a URL that was replaced in the pre-processing step.

Table 2: Five most correlated keywords for 10 topics

Topic number	Five most correlated words				
0	california	traffic	worst	tacos	insane
1	URL_HTTPS	california	#job	#transportation	#jobs
2	URL_HTTPS	caltrans	california	cars	police
3	URL_HTTPS	cars	faraday	california	self-driving
4	URL_HTTPS	cars	california	light	green
5	URL_HTTPS	faraday	future	cars	test
6	URL_HTTPS	california	converter	catalytic	magnaflow
7	URL_HTTPS	california	rail	high-speed	caltrans
8	california	traffic	URL_HTTPS	cars	bad
9	URL_HTTPS	california	cars	#cars	caltrans

Table 3 shows the average sentiment and a representative tweet for each of the 10 topics. The number of tweets for each topic varies between 705 and 1557 (note the total number of tweets is 10,400) which shows a reasonable distribution of tweets to topics. However, the number of polarizing tweets (i.e., non-neutral tweets as identified by the sentiment analysis algorithm) varies from 47 to 568. The most strongly positive topic is Topic 3 (with words: cars, Faraday, self-driving) while the most strongly negative topic is Topic 0 (with words: California, traffic, worst, tacos, insane).

Table 3: Average sentiment, number of tweets, and representative tweet for each of the 10 topics

Topic Number	Average sentiment	Number of tweets	Number of polar tweets	Representative tweet
topic=0	-0.537	811	568	"RT @conradjwilson: Before Columbia River Gorge Derailment Union Pacific Lobbied Oregon Against Tougher Oil Train Rules, And Won https://t"
topic=1	0.105	760	144	"Clifford the Big Red Dog Kaboom Trix Crunch Berries Matchbox cars Starsky and Hutch Mash Lavern Shirley Fonzy Warhol Dali California"
topic=2	0.009	1125	150	RT @AKHLASS1: Los Angeles California #heatwave don't leave kids & pets in cars Be Alert @ parking lots #Guciffer2 #OurRevolution https://t
topic=3	0.226	1540	350	Tesla rival Faraday approved to test self-driving cars on California roads: DETROIT (Reuters) - Faraday Future plans to begin testing...
topic=4	-0.059	999	65	"#SocialMedia #Seo Faraday Future gets green light to test driverless cars in California repo... URL_HTTPS #tech #BUSINESS"
topic=5	0.0263	1557	47	Faraday cleared to test self-driving cars on California roads: Faraday Future plans to begin URL_HTTPS #Autos #Car #Trucks
topic=6	0.017	705	62	California High-Speed Rail Authority agrees to explore 'green' technology - Progressive Rail Roa... URL_HTTPS #green #tech

topic=7	-0.0023	848	284	"RT @abc7alysha: #TRAFFICALERT tow truck on scene to clear 3 car crash on 405 NORTH past Mulholland slows from Wilshire #caltrans https:/"
topic=8	0.0334	1316	466	Fuck California. Fuck your traffic. Fuck this tobacco law. Fuck your expensive ass gas. Fuck you for letting Hilary win. Fuck this state.
topic=9	0.1082	739	180	#car Ford: Mustang GTA 1967 ford mustang gta fastback s code white california car rust free: $18... URL_HTTPS #ebay #cars

Table 4 shows the five most correlated words defining **20** topics.

Table 4: Five most correlated keywords for 20 topics

Topic number	Five most correlated words				
0	URL_HTTPS	california	cars	flying	circuit
1	URL_HTTPS	faraday	cars	self-driving	roads
2	california	URL_HTTPS	traffic	cars	#california
3	california	traffic	cars	jam	real
4	URL_HTTPS	cars	california	bmw	angeles
5	URL_HTTPS	california	rail	high-speed	#california
6	URL_HTTPS	california	people	traffic	oregon
7	URL_HTTPS	faraday	future	cars	test
8	URL_HTTPS	cars	test	self-driving	future
9	URL_HTTPS	caltrans	california	traffic	plan
10	URL_HTTPS	light	green	cars	california
11	URL_HTTPS	faraday	cars	future	california
12		rail		high-speed	

	URL_HTTPS		california		california's
13	URL_HTTPS	california	cars	#cars	police
14	traffic	URL_HTTPS	california	bad	made
15	URL_HTTPS	caltrans	california	update	cars
16	URL_HTTPS	california	king	transportation	caltrans
17	URL_HTTPS	california	#job	#transportation	#jobs
18	URL_HTTPS	california	converter	catalytic	magnaflow
19	california	traffic	worst	insane	tacos

Table 5 shows the average sentiment and a representative tweet for each of the 20 topics. The number of tweets for each topic varies between 261 and 1196 (note the total number of tweets is 10,400) which shows a reasonable distribution of tweets to topics but more skewed than with 10 topics. However, the number of polarizing tweets (i.e., non-neutral tweets as identified by the sentiment analysis algorithm) varies drastically from 2 to 470. The most strongly positive topic is Topic 4 (with words: cars, California, bmw, angeles) while the most strongly negative topic is Topic 19 (with words: California, traffic, worst, tacos, insane). Note that the most negative topic is consistent with topic modeling with 10 topics. While, the most positive topic is not the same, an examination of the underlying tweets indicate many are related to electric cars.

Table 5: Average sentiment, number of tweets, and representative tweet for each of 20 topics

Topic Number	Average sentiment	Number of tweets	Number of polar tweets	Representative tweet
topic=0	-0.03304	454	31	"Caltrans staging at key on & off ramps of Hwy 101, ready to shut down lanes at moment's notice due to #ScherpaFire (#SherpaFire). #KSBYNews"
topic=1	0.229933	1196	335	Tesla rival Faraday approved to test self-driving cars on California roads - Yahoo Tech URL_HTTPS #awesome #tech #news #tech
topic=2	0.023613	847	380	Fuck California. Fuck your traffic. Fuck this tobacco law. Fuck your expensive ass gas. Fuck you for letting Hilary win. Fuck this

				state.
topic=3	0.007663	261	24	Faraday Future seeks approval to build electric cars in California URL_HTTPS ibrido elettrico idrogeno. Dott ssa Roberta Cell
topic=4	0.304217	332	135	RT @IronMillTech: lease don't buy electric cars - Tesla Model X Lemon Law Suit Filed In California URL_HTTPS #EV $TSLA https:
topic=5	0.171004	269	82	Carson Palmer : 2001... URL_HTTPS #art #ball #california #cars #cat #dow #fin #foot #football #game #hr #league #om #school
topic=6	-0.05461	293	92	"RT @conradjwilson: Before Columbia River Gorge Derailment, Union Pacific Lobbied Oregon Against Tougher Oil Train Rules, And Won https://t"
topic=7	0.008159	858	9	Faraday Future cleared to test self-driving cars in California: Faraday Future isn't waiting ... URL_HTTPS #UAE #KSA #EURO
topic=8	0.034409	465	24	Sign the petition to stop the deadly bomb trains hauling crude oil through California neighborhoods. URL_HTTPS via @APEN4EJ
topic=9	-0.01766	453	156	Detour alert! Caltrans closing Airport Blvd exit Friday & Monday evening till early morning. URL_HTTPS URL_HTTPS
topic=10	0	838	2	"#SocialMedia #Seo Faraday Future gets green light to test driverless cars in California, repo... URL_HTTPS #tech #BUSINESS"
topic=11	0.136201	558	76	Faraday cleared to test self-driving cars on California roads: Faraday Future plans to begin URL_HTTPS #Autos #Car #Trucks
topic=12	-0.02494	401	122	RT @AKHLASS1: Los Angeles California #heatwave don't leave kids & pets in cars Be Alert @ parking lots #Guciffer2 #OurRevolution https://t
topic=13	0.064386	497	116	#car Ford: Mustang GTA 1967 ford mustang gta fastback s code white california car rust free: $18... URL_HTTPS #ebay #cars
topic=14	-0.08314	421	57	RT @CaltransDist7: #Caltrans removing tree that fell onto hwy. Expect delays & consider alt rte. @CityMalibu @TopangaChamber @LHSLASD http
topic=15	0.045375	573	118	Sallys Cozy Cone Motel Food Options in Disneylands California Adventure Cars Land:

				Youve spent the day exp... URL_HTTPS
topic=16	0.057637	347	44	"Setaluna Premier Soft & Silky Satina Bed Sheet Set, California King, Beige Cream URL_HTTPS URL_HTTPS"
topic=17	0.042135	356	29	"Border rail line to connect U.S., Mexico: Officials expect deal to finally connect Baja California factories to US URL_HTTPS"
topic=18	-0.02954	474	14	"RT @abc7alysha: #TRAFFICALERT tow truck on scene to clear 3 car crash on 405 NORTH past Mulholland, slows from Wilshire #caltrans https:/"
topic=19	-0.8284	507	470	RT @smasher_blakes: In case anyone was curious as to what 5 oclock traffic looks like in California:-) pls stop moving here thx & gig em ht

Table 6 shows the average sentiment and a representative tweet for 19 of the 50 topics when the number of topics was preset to 50. Only the most polar topics are shown – those with average polarity greater than 0.1 (positive) or lesser than -0.1 (negative). The number of tweets for each topic varies between 89 and 470 (note the total number of tweets is 10,400) which shows a reasonable distribution of tweets to topics but more skewed than with 10 topics and less skewed than with 20 topics. However, the number of polarizing tweets (i.e., non-neutral tweets as identified by the sentiment analysis algorithm) varies from 18 to 428. The most strongly positive topic is Topic 31 (with words (not shown in the tables): cars, #california, bmw, free) while the most strongly negative topic is Topic 14 (with words (not shown in the tables): California, traffic, worst, tacos, wildfires). Note that the most negative topic is again consistent with topic modeling with 10 and 20 topics. However, what the most positive topic represents is less clear as compared to topic modeling with 10 and 20 topics.

Table 6: Average sentiment, number of tweets, and representative tweet for 19 of 50 topics

Topic Number	Average sentiment	Number of tweets	Number of polar tweets	Representative tweet
topic=31	0.542	131	87	10 Ft. Curved Double Rail Span Bridge Curved Double Rail Sealed Constructed of Hand-Selected California Redwood URL_HTTPS
topic=27	0.368	144	55	"Cars Land (Disney California Adventure) Avec les nons en soire c'est magnifique ... source Disneyland

				Resort Images"
topic=37	0.352	352	124	RT @DominionHarbor: Tesla $TSLA rival Faraday approved to test self-driving cars on California roads-bring on competition & #innovation htt
topic=43	0.309	470	145	Tesla rival Faraday approved to test self-driving cars on California roads - Yahoo Tech URL_HTTPS #awesome #tech #news #tech
topic=12	0.200	110	24	TRAFFIC ALERT: The Grand Detour is in effect. Avoid Grand Ave. Use California or Santa Rosa for campus access. URL_HTTPS
topic=49	0.182	154	38	"RT @conradjwilson: Before Columbia River Gorge Derailment Union Pacific Lobbied Oregon Against Tougher Oil Train Rules And Won https://t"
topic=22	0.167	162	49	"1st Rick Scott fail. Florida should've got Fed funds not California. #Rail #HighSpeedRail #Florida #RickScott #Fail URL_HTTPS"
topic=40	0.161	273	44	Tesla rival Faraday approved to test self-driving cars on California roads: DETROIT (Reu... URL_HTTPS #expo #expo2015 #biz
topic=10	0.136	132	40	"FULL AUTOMATIC Blog for BLOGGER BLOGSPOT URL_HTTPS #serp #traffic California Fire Crews Fight Sherpa Wildfire Brace for"
topic=44	0.125	208	26	Faraday cleared to test self-driving cars on California roads: Faraday Future plans to begin URL_HTTPS #Autos #Car #Trucks
topic=26	0.117	325	38	Faraday Future Approved to Test Self-Driving Cars in California: Faraday Future Approved to Test Self-Drivin... URL_HTTPS
topic=38	0.116	121	34	RT @IronMillTech: lease don't buy electric cars - Tesla Model X Lemon Law Suit Filed In California URL_HTTPS #EV $TSLA https:

topic=8	0.112	89	18	LAPD Adds 100 BMW Electric Cars to Its Fleet - German automaker BMW has beaten California-based Tesla Motors In... URL_HTTPS
topic=13	0.107	122	31	"Clifford the Big Red Dog Kaboom Trix Crunch Berries Matchbox cars Starsky and Hutch Mash Lavern Shirley Fonzy Warhol Dali California"
topic=48	-0.130	254	41	RT @OWCalifornia: I made this map of California where traffic will be bad at 5 o'lock. Traffic is highlighted in red. URL_HTTPS
topic=45	-0.186	188	55	#Auctions #Ferrari Ferrari: California Ferrari California V8 loaded leather nav Crave Luxury Auto URL_HTTPS #Cars #Automotive
topic=32	-0.252	123	35	pmarca: RT Lebeaucarnews: BREAKING: Chinese electric car company Faraday Futures gets approval by California DMV to test self-driving cars
topic=4	-0.336	125	58	RT @AKHLASS1: Los Angeles California #heatwave don't leave kids & pets in cars Be Alert @ parking lots #Guciffer2 #OurRevolution https://t
topic=14	-0.988	431	428	RT @OWCalifornia: California Facts: - California wildfires are the worst - LA traffic is insane - 68 is cold - No one says no to tacos

3.2 Strengths and Limitations

The combination of text pre-processing, topic modeling, and sentiment analysis resulted in the identification of topics agree with currently debated topics on social media and their expected sentiment (i.e., traffic in California is considered very negatively; electric cars and self-driving cars have a positive image). The main strength of the tool is that this processing was done **completely automatically**. The software was not provided any pre-defined topics of interest, only the raw tweets from Twitter collected using a set of generic transportation-related keywords (listed in Table 1). This analysis can easily be repeated and interpretable results obtained with a different set of keywords or for a longer time period (the current analysis was done with about three weeks of data).

Such a tool will become increasingly necessary to analyze the large number of social media messages and which is only expected to grow. For instance, we were able to download over 10,000 messages related to transportation in California from the Twitter social network in just three weeks.

We have identified three limitations of the current version of our software:

1. The results are highly dependent on the topics that are currently being discussed; hence, topics that are highly polarizing may not be identified if the time period of social media data collection did not include a time when these topics were being discussed. We are therefore continuously collecting tweets since we expect that we will get a wider coverage of topics from messages collected over a longer period of time.
2. While the software does not require the end-user to provide any input requiring expert information, the results are somewhat sensitive to the number of topics that are preset in the topic modeling step. There are methods that can be used to objectively set the number of topics (for example, [15]); however, these have not been implemented.
3. We were not able to get a significant number of useable location information. A majority of users on Twitter do not disclose their locations while many who do reveal location information do so, specify a large area (such as Los Angeles) which makes it difficult to draw location-specific conclusions from this type of analysis.

3.3 Recommendations and Policy Implications

This project demonstrates that completely automatic analysis and summarization of the overall public sentiment on transportation-related topics as expressed in social media platforms is possible. The software developed by the project can with some effort (primarily by adding an appealing graphical user interface) be extended into a ready-to-use tool that can be used by the planners and engineers of Caltrans and the transit agencies to gauge public opinion on specific topics and act accordingly. This tool can also be used to provide input when prioritizing projects.

3.4 Other Work

Many of the efforts leading to the goals of this project have been presented and the details are shown in Appendices 1, 2, 3, and 4. The supporting materials are available upon request.

4. CONCLUSIONS AND FUTURE WORK

We developed a software tool that can automatically analyze social media communications and provide an interpretable summary of the transportation-related topics that are being discussed and the overall sentiment of the online users towards these topics.

Some limitations of the tool have been identified and the tool can be improved with some additional work. The tool can also be relatively easily extended into a directly useable tool by the personnel of Caltrans and the transit agencies who are interested in public relations.

5. REFERENCES

[1] S. Kaufman, "Co-Monitoring for Transit Management " NYU Wagner2014.

[2] F. Chen and R. Krishnan, "Transportation Sentiment Analysis for Safety Enhancement," 2013.

[3] C. Collins, S. Hasan, and S. V. Ukkusuri, "A Novel Transit Rider Satisfaction Metric: Rider Sentiments Measured from Online Social Media Data," *Journal of Public Transportation,* vol. 16, 2013.

[4] Twitter API. https://dev.twitter.com/overview/api

[5] Tweepy. http://www.tweepy.org/

[6] MongoDB. https://www.mongodb.com/

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research,* vol. 3, pp. 993-1022, 2003.

[8] D. Ramage, E. Rosen, J. Chuang, C. D. Manning, and D. A. McFarland, "Topic modeling for the social sciences," in *NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond*, 2009.

[9] D. Ramage and E. Rosen. *The Stanford Topic Modeling Toolbox*. Available: http://nlp.stanford.edu/downloads/tmt/tmt-0.4/

[10] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

[11] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, pp. 115-124.

[12] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval,* vol. 2, pp. 1-135, 2008.

[13] C.J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[14] S. Bird, E. Loper and E. Klein, Natural Language Processing with Python. O'Reilly Media Inc., 2009.

[15] T. L. Griffiths and M. Steyvers, "Finding scientific topics." Proceedings of the National academy of Sciences 101, no. suppl 1 (2004): 5228-5235.

APPENDIX 1- Oral Presentation of Related Work

Hao, L., Panangadan, A., and Abellera, L.V., "Understanding the Public Sentiment toward I-710 Corridor Project from Social Media Based on Natural Language Processing," Cal Poly Pomona 1st Annual Creative Activities and Research Symposium, August 19, 2015 (Pomona, California, U.S.A.).

Abstract:

With social media like Twitter gaining popularity in our daily life, bulk of data are emerging online. These data implies the information closely relevant to the preferences, tastes, attitudes and beliefs of the general public. Such information is useful in making public strategies, budgets or plans. In order to provide Caltrans with the information on the 710 Corridor Project and the five alternatives from Twitter users, we investigate the tweets related to the 710 Project during the time period from 2009 to August 3rd, 2015, analyze the sentiment evolution over time, and examine the sentiment for the five alternatives of the 710 study in this paper. The results can be used as a reference for the stakeholders to make decisions on the 710 Project in the future.

APPENDIX 2- Poster Presentations of Related Work

Zhao, D., Bravo, J., Vigus, C., Hao, L., Panangadan, A., and Abellera, L.V., "Discovery of Non-traditional Ridership Factors with Sentiment Analysis Tools," UCCONNECT Student Conference, Feb. 11-12, 2016 (Riverside, California, U.S.A.).

Abstract:

Public transportation agencies can obtain large amounts of information regarding timeliness, efficiency, cleanliness, ridership and other performance measures through paper and online surveys. However, these metrics are based on the interests of these agencies and do not necessarily represent the concerns of the customers. Recently, social media have become a platform for people to show their satisfaction or discontent about particular services and products (e.g., Twitter feeds, Yelp reviews, Change.org petitions).

The goal of this project is to develop a tool using machine learning techniques to discover factors affecting public transit rider satisfaction using social media data. This tool that will interface with Geographic Information Systems (GIS) is intended to reveal features of ridership that are not evident to transit agencies. For instance, a sense of community and pride are positive aspects of ridership that are not measured by traditional surveys. Specifically, sentiment analysis techniques will be utilized to classify numerous sets of rider sentiment data over a period of time and for particular locations (e.g., a Metrolink station). This presentation shows our initial efforts to create this tool.

Giancristofaro, G.T. and Panangadan, A., "Predicting Sentiment in Social Media Posts Using Visual and Textual Features," Cal State Fullerton Engineering and Computer Science Student Projects Showcase and Awards, May 10, 2016 (Fullerton, California, U.S.A.).

Abstract:

We applied "machine learning" (analysis that automatically classifies data based on previously seen examples) to identify if public posts made to Instagram expressed positive or negative sentiment toward the California Department of Transportation and its projects. The study applied machine learning to photographs, captions and comments from the social network, and showed that including both image and text data increased the accuracy of identifying sentiment.

APPENDIX 3- Publications of Related Work

Hao, L., Panangadan, A., and Abellera, L.V., "Understanding Public Sentiment toward I-710 Corridor Project from Social Media Based on Natural Language Processing," IEEE 19th International Conference on Intelligent Transportation Systems (ITSC 2016), November 1-4, 2016 (Rio de Janeiro, Brazil), under review.

Abstract:

With many social media platforms like Twitter gaining popularity, extremely large amounts of user-generated content are becoming available online. These contents are closely related to the preferences, attitudes and beliefs of the general public on a particular topic. To provide Caltrans with opinions from Twitter users on the five alternatives of the 710 Corridor Project, we investigate relevant tweets from 2009 to August 3, 2015, analyze how public sentiment evolves over time, and examine the general sentiment for the five alternatives. The results can be used as a reference for the decision-makers to select the most acceptable alternative for the 710 Project in the future.

Giancristofaro, G.T. and Panangadan, A., "Predicting Sentiment toward Transportation in Social Media using Visual and Textual Features," IEEE 19th International Conference on Intelligent Transportation Systems (ITSC 2016), November 1-4, 2016 (Rio de Janeiro, Brazil), under review.

Abstract:

Social media platforms can be used by transportation agencies to receive feedback from their customers, thus creating two-way communication between the service provider and its consumers. Sentiment analysis is one method of aggregating overall polarity (positive or negative) towards a topic. However, most sentiment analysis methods rely on text processing, thus ignoring the large amount of image data present in popular social networks.

The primary aim of this study is to exploit image data in conjunction with text and to evaluate this integrated approach for sentiment analysis for transportation. This study used image, captions, and comments data from the Instagram social network that were marked as being relevant to California Department of Transportation (Caltrans) and attempted to predict the expressed sentiment towards this agency. Approximately 1,000 posts were collected from Instagram which included the hashtag #caltrans and were classified manually as to whether they expressed a neutral, positive, or negative sentiment towards Caltrans. A Web-based application was developed to enable efficient human annotation to create labeled data. Supervised machine learning algorithms are used to train a classifier on this dataset. A set of high-level features were extracted from images using the web-based Miscrosoft Cognitive Services APIs. These features included the detection of faces and emotion recognition. Text features included the set of individual words and structural features.

The experiment results of different machine learning techniques show a gain in precision when images and texts are combined compared to text-only approaches, thus confirming the relevance of visual content usage. This method is also applicable

to posts that do not contain any text data. The precision reaches a performance close to human classification agreement (typically approximately 80%). However, the results do not indicate that visual features are more informative than text features.

APPENDIX 4- News Report of Related Work

Under the supervision of Dr. Anand Panangadan, Gabriel T. Giancristofaro presented his project, "Predicting Sentiment in Social Media Posts Using Visual and Textual Features" at the Cal State Fullerton Engineering and Computer Science Student Projects Showcase and Awards on May 10, 2016. He won the award for Best in College – Computer Science for which he received $1,000 in prize money. Gabriel T. Giancristofaro is an exchange student from the University of Sao Paulo in Brazil where he is a computer science major. Dr. Lourdes Abellera and her student Galya Clein from Cal Poly Pomona worked on the human annotation phase of this project.

The work is described briefly in a news item which can be found in the following link.

http://news.fullerton.edu/2016sp/engineering-showcase.aspx