Sustweetability: Infrastructure Sustainability-Related Topic Classification in Social Media

Mazdak Nik-Bakht, Post-Doctoral Research Associate, mazdak.nikbakht@mail.utoronto.ca
Moein Hosseini, MASc Student, moein.hosseini@mail.utoronto.ca
Tamer E. El-Diraby, Associate Professor, tamer@ecf.utoronto.ca

Department of Civil Engineering, University of Toronto, Ontario, Canada

Abstract
In this paper we try to automate detection and classification of the topics discussed over online social media, in the context of sustainability of infrastructure projects. By focusing on the micro-blogging website Twitter, supervised learning was used to train classifiers for detecting the subject of infrastructure-related tweets. Crowdsourcing was used in form of a Game With A Purpose (GWAP) to annotate the training set through collective intelligence of participants. Different forms of classifiers including Naïve Bayes, K-Nearest Neighbors algorithm, Decision Tree, and Support Vector Machine were then tested and combined through different architectures. The outcomes of the research can help the PI practitioners to detect latent patterns of opinion in public inputs over Twitter; and the lessons learned can pave the way for researchers towards automation of understanding sustainability-related contents from the public input.

Keywords: Infrastructure discussion network, online social media, Twitter, computational linguistics, supervised learning, classifier

1 Introduction

Stakeholder management in infrastructure planning and construction is principally interested in understanding stakeholders of the infrastructure project. More specifically, detecting the stakeholders vested-interests in the project, and their positions (in terms of being in-favor or against the decisions made) are two of the core aims of stakeholder analysis in urban infrastructure projects (Olander, 2007). Public communities, as the prospective users of the system are big players of this game, and somewhat the most challenging ones to be handled in the process of stakeholder mapping. Public involvement (PI) and social engagement practices are primarily designed as a response to this challenge. Many PI agencies have recently found interest in using online social media as a bi-directional communication channel with the public. Taking advantage of listening to what citizens say about their community can provide planners with an opportunity to engage the public in a completely different way (Evans-Cowley & Griffin, 2011). However, harvesting relevant items from the corpus of user generated content on social media, and analyzing them to achieve meaningful results require tools and methods which do not yet officially exist in the field.

As a part of modern social trends, knowledge-enabled communities discuss different aspects of an infrastructure projects in online social media and seek a more active participation in the process of related decision making. Along with the evolution of the knowledge economy, Infrastructure Discussion Networks (IDN) on social media are poised to be a source of creative ideas regarding project scope, funding, design and operation plans. Indeed, this could also be the starting point for a new realm in innovation democratization, and more importantly, a bottom-up public decision making procedure. Although this creates great opportunities for social engagement, the lack of means to analyze these seemingly chaotic discussions wastes these opportunities and is becoming frustrating to both communities and public decision makers. The chaos is resulted from the free participation of hundreds of thousands (even up to millions) of nodes and the unstructured nature of their inputs (which are normally in natural language) in social media. The present study is an

attempt to detect the patterns of order beneath the chaos of communication with the public over the micro-blogging website Twitter.

With more than 645 million active users, and a growth rate of 135,000 new users per day, Twitter records an average of 58 million tweets every day (Twitter Statistics, 2015). People express their opinions on many different issues (including the built environment) in less than 140 character statements, and this can be a significant opportunity for decision makers to communicate with citizens and detect their demands or feedback. This tremendous pool of online users together with its open API (Application Programming Interface) has made Twitter the subject of many research works in different domains. Based on the transit cooperative research program-synthesis 99 of the Transportation Research Board (TRB), major transportation service providers who use online channels to involve the public in the USA and Canada find Twitter in many aspects the most (and in some aspects the second most) convenient communication tool. Apart from connecting with the customers/community, where online social media plays the role of a tool for real-time communication, advocacy, feedback collection, etc., it puts the customers in power and as mentioned, creates opportunities for user innovation (Bregman & Watkins, 2013). On the other hand, among social media tools used in construction industry, statistics show that Twitter has jumped from the fourth most effective (back in 2012) to the second most effective one in 2013, and stands bellow LinkedIn that has been mainly used for recruitment purposes (Bennett, 2013). Later studies claim that, Twitter is found to be "the most preferred social media tool" among the construction companies who use online social media (Azhar & Abeln, 2014).

The present study aims to develop an automated tool to detect the public communities' vested interests in infrastructure construction projects through analysis of the main subject in their tweets. Sustainability was selected as the main context of the analysis, and Artificial Intelligence (AI) through supervised learning was utilized as the method. We have examined the performance of various combinations of methods and algorithms using empirical data collected on a set of case-study projects, and we offer a subject classifier at the end. The paper start with a brief review on similar attempts inside and outside the domain of infrastructure, and then in section 3 we explain the methodology followed for detection and analysis of data. Section 4 reports the major findings, and finally the concluding remarks along with the future steps of the study are presented in section 5.

2 Background literature and related works

Evans-Cowley & Griffin (2011) identified the values of reviewing discussions over micro-blogging website Twitter for infrastructure project-related ideas in the sector of transportation. They analyzed public discussions from content (type and theme) and sentiment points of view in form of a program called SNAPP (Social Networking and Planning Project). SNAPP was seeking an answer to the question of whether micro-blogs can be analyzed to help understand the public's views on transportation issues. They used linguistic analysis and word count to assess emotional cognitive and structural components of more than 8,300 relevant tweets they had collected around transportation related issues in Austin, Texas. The results of this study indicated that aggregation of micro-blogs can provide meaningful and helpful data in order to understand the communities' perspective on infrastructure. while this research used commercial software for linguistic analysis, Evans-Cowley and Griffin admit that context-specific tools are required for this purpose. SNAPP report is closed with an emphasis on the demand for further empirical investigations to find ways in which information extracted through micro-blogging can be weighted and analyzed in the specific context of infrastructure planning.

Analysis of natural language by the machine is involved in ambiguity from different aspects. Word sense, word category, syntactic structure, and semantic scope are among other features which challenge automatic understanding of naturally generated texts. Computational linguistics is the matter of selecting disambiguation strategies to detect the correct content out of the user created context. Creating an ontology which is based on rule creation and hand-tuning may be considered as one solution. Although working perfectly in machine interoperability, when it comes to evaluation of the naturally occurring text, such a method performs poorly (Manning & Schutze, 1999). Statistical NLP approaches on the other hand solve this challenge through learning the lexical structure from the corpora. They try to create a shortcut to semantic relationships by

counting co-occurrence of words (lexical co-occurrence) or syntactic structures in the corpora; they are normally robust and generalize well in the encounter of new data.

Models in statistical NLP work essentially based on the stationary model assumption which states that the future can be predicted by looking at the past behavior. They aim to infer about the structure of data which is generated by the natural language with originally no particular probability distribution. It is usually aimed to predict a target feature based on a set of classificatory features. For this purpose, a training set is required to be classified into partitions which have the same value for the classificatory features. Then pattern recognition algorithms are used to estimate the common patterns existing in each class. The whole procedure can be followed under three main steps (Manning & Schutze, 1999):

- Forming equivalence classes (Dividing the training data into classes with the same value for the target feature);
- statistical estimation (Finding a good statistical estimator for each equivalence class); and
- combining multiple estimators

Decision making on the most distinctive set of features to be used in training a classifier is called feature selection. Feature selection is normally involved in scoring all potential features (data attributes) according to particular metrics and selecting the best ones to make sure about employing adequate-yet-not too many features. The goal is a correlation-based feature selection which minimizes redundancy and maximizes relevance at the same time with keeping the ability of distinction. There are two core questions to be answered in this regard: "what type of features should be selected to count?" and "how many features are enough to be considered?" The typical options to answer the former question are attributes such as Terms (unigrams) in a document/tweet; n-grams (sequence of n tokens), lexical properties (such as polarity: being positive, negative, or neutral and subjectivity), Part of speech (POS), syntactical properties (such as punctuations and clitics), and Semantic class of terms. It has been shown in the literature that using unigram 'terms' as classification features in topic classification performs as good as more sophisticated syntactic and morphological analyses (Dumais, et al. 1998). This was particularly approved for the case of classifying tweets (Horn, 2010). As Manning et al. (2009) indicated, selecting a reduced vocabulary (a subset of more relevant terms in different classes) may increase the efficiency not only by reducing dimensions of the problem space but also by eliminating the noise features and preventing the over-fitting problem. Forman (2003) suggested using the high-frequency terms only for this purpose. As the Zipf's law normally governs the probability distribution of words in a document, this will significantly reduce dimensions of the learning problem. On the other hand, stemming (removing the inflectional endings and pre-fixes from words and grouping them into their lexical roots) has theoretically ben considered as another way to not only reduce the number of features, but also increase the domain of queries which can be classified by a trained classifier.

Regarding values of feature vectors, although term-count metrics such as term frequency (or more sophisticated measures such as TF-IDF: term frequency-inverse document frequency) can be thought of as advanced attribute values, Dumais et al. (1998) have successfully shown that even a simple Boolean measure of term occurrence can guarantee the required efficiency and efficacy. In particular, Forman (2003) suggested when the instances are short documents, terms are not likely to repeat and therefore Boolean word indicators are nearly as informative as the counts.

Regardless of the selected features and their values, there are various supervised learning techniques used in text classification. The most popular ones include: Decision Tree (DT), Naïve Bayes (NB), K-Nearest Neighbors (K-NN), and Support Vector Machines (SVMs). They are generally developed for problems with wider data domains such as quantitative or categorical data. By selecting term frequency as the feature value, most of the quantitative classification algorithms will be also applicable to text classification. There are also other classifiers which are not as prevalent as the ones mentioned above; those include Maximum entropy, the Pattern-based classifier, the Neural Network classifier, and the Genetic Algorithm-based classifiers (Aggarwal & Zhai, 2012).

Naïve Bayes classifiers work based on the features independence assumption and Bayes theorem. This method calculates the probability of a document (or instance) d to belong to a class c :P(c | d),

given the document's features (f1 to fk). Based on the chain rule, if features f1 to fk are independent from each another, then this probability can be calculated as:

$$P(c|d) \propto P(c) \times \prod_{1 \le i \le k} P(f_i|c) \quad (1)$$

The conditional probabilities are calculated from the training data, and update the prior into the posterior for each class. Finally, the class with the highest probability will be introduced as the class of the test instance.

Support Vector Machines (SVM) work based on separating the data belonging to different classes by passing hyper-planes (or hyper surfaces in general) between them in the feature space of the problem. These hyper-surfaces are loci of points with maximum distance from the data points of different classes. The position of hyper surfaces is apparently specified based on a subset of the frontier data-points. These points (which are the closest points to the separator surface) are referred to as support vectors, and the closest distance between the support vectors of two different classes determines margin of the classifier.

Decision Trees (sometimes referred to as random forests) break the classification problem into a hierarchy of decisions. Each quarry is passed through all the decision nodes and at each decision node the instance is classified based on one feature. In other words, at each iteration the feature space is decomposed into more discriminating subspaces, through splits parallel to one of the features' axis. The class with the highest number of votes at the end is specified as the class the quarry instance belongs to. A stopping criterion (usually in form of a minimum number of steps passed or a minimum number of features covered) is introduced, and the instance is classified at the stop point. For a comprehensive review on classifiers and using different algorithms for topic classification, one can see Aggarwal & Zhai (2012).

3 Methods

In order to achieve subject classifiers which can specifically detect the main infrastructure-related topics discussed on Twitter, the training procedure must be performed on the empirical/domain-specific data. By focusing on a set of case-study project, we collected data and trained various types of classifiers to reach the one with an acceptable performance. This section explains different components of our methodology in detail.

3.1 Case study projects & data collection

Our data-set includes tweets about a set of Light Rail Transit (LRT) projects in different North American cities. The projects were at different phases of their lifecycle while the data collection was in progress. In this part, the projects are briefly introduced and the data-collection procedure is explained.

The Eglinton Crosstown LRT in Toronto, Ontario, is part of one of the largest transit projects currently underway in North America. The project was announced in 2007 and has been under long debate since then. It is an east-west line in a total length of 19.5 Km passing through a congested corridor of Toronto's midtown and running underground in major parts. The total cost of the project is estimated around CAD8.4 Billion. Central Corridor (Metro Green Line) LRT is built on over 18 Km of exclusive right of way and links five major centers of activity in the Twin Cities St. Paul and downtown Minneapolis, Minnesota. Construction began in late 2009 and the operation started in June 2014. Construction was funded by federal, state, and local governments. The Atlanta Streetcar is an East-West light rail route in Atlanta, Georgia, shared with other traffic on-street lanes in a total length of 4.3 Km and having 12 stops. The project is the result of a public-private partnership. The construction started in early 2012 and was performed in three major phases. Its operation began in December 2014. Finally, the Woodward Ave. LRT (or M1-Rail) is a 5.3 Km long light rail in the public right-of-way within the city of Detroit, Michigan, which is planned to connect the downtown and the new center of the city. The project is composed of 5.3Km long railway and is estimated to cost $140milion which will be granted through a public-private partnership. A Michigan non-profit corporation called M-1 Rail was formed by local business leaders in 2007 to develop and potentially operate the system over a term of 10 years. As mentioned in the M1-Rail business plan (April 2012), the project does not require any business or residential dislocations, and the streetcar service will be co-mingled with vehicular traffic (M1-Rail Streetcar

project business plan, 2012). Construction of the project was bid in the form of a design-build contract in May 2013, and two more contracts will be awarded for construction of a vehicle storage and maintenance facility, and for the streetcar vehicles themselves.

Table 1 summarizes the four projects used for data collection, along with the number of tweets collected in each project, which were involved in training and testing our subject classifier. It is worth to mention that the total number of tweets collected for each project was a lot more than what is shown in this table. The collected tweets underwent a pre-processing and an annotation process, through which, a considerable number of repeated or irrelevant tweets were detected and filtered from the data-set.

Table 1 Case study projects and the number of tweets from each project used in the analysis

			Number of tweets						
Project	**Location**	**Phase♣**	**Environmental**	**Economic**	**Social**	**Engineering**	**None**	**Total**	**Dataset Name**
Central Corridor (Metro Green Line)	Minneapolis, Minnesota, USA	Late Construction	6	10	57	46	23	142	CC
Atlanta Streetcar	Atlanta, Georgia, USA	Construction	5	16	125	58	18	222	ATL
Eglinton Crosstown	Toronto, Ontario, Canada	Early Construction	13	91	168	233	176	681	CT
Woodward Ave. LRT (M1 Rail)	Detroit, Michigan, USA	Pre-Construction	4	35	83	36	25	183	M1
♣ Project Lifecycle Phase during data-collection					**Total:**			1228	

In general, communicating with Twitter API to collect data requires encoding requests (including the request URL, specific keywords under quarry, and the electronic signature of the sender for authentication), and submitting them as a request to the website's database. One important advantage of Twitter for research purposes is the open nature of its API. Despite some limitations, data on many different aspects of the contents generated and shared on Twitter is openly accessible. Although being open, working with Twitter API has its own limitations and quarry-rate-limit is one of the most restrictive ones; there is a cap on the number of requests that Twitter API responds over the time.

An automated data collector was set up to search for tweets anchored by hashtags related to the projects, or mentioning their IDs. The collector was created in server-side scripting language *PHP*, using a class called *Twitter-Search-api.php*, written by Faerman (2009), to send queries through Twitter API. The collector was fed by the set of keywords, to embed in form of queries, sent requests and received the results back as a *.json* (JavaScript Object Notation) file. Five main components were extracted from .json files and were copied to our database: user_id (numerical ID of the person/organization who is tweeting), date, text_id (numerical ID of the tweet), user (username on Twitter), and text (content of the tweet). Data collection was performed as an ongoing process over almost ten months and as a result, a total of more than 40,900 tweets was collected for these four case-study projects.

3.2 Annotating training set through crowdsourcing

Crowdsourcing was utilized in the present study for this purpose. Crowdsourcing can add value by ensuring the pluralism of perspectives in the classification (which in some cases may be a subjective task). We took advantage of 'wisdom of the crowd' in refining and fine-tuning the results. We set up a Game With A Purpose (GWAP), called "Sustweetability", in which players were provided with a set of tweets and were asked to annotate them from the aspects of subject and sentiment. Sustainability of the infrastructure system was selected as the main scope for subject classification, and the players could score points by classifying each tweet in the 'right' class. In order to decide

which answer is right and which one is not, we looked at the distribution of answers by all players and selected the mode as the correct answer. In the rare cases that there were multi-modal distributions, two classes were considered as the right answer. The game was running from June through August 2013, and there were prizes for the winners to motivate the players to get involved and provide quality answers. Figure 1 shows a snapshot of the game homepage. More details about Sustweetability GWAP can be found in (Nik-Bakht & El-diraby, 2015).

The GWAP resulted in detection of irrelevant tweets (tweets which were not related to the infrastructure system at all) and tagging the relevant tweets under five main classes: *Engineering* (discussing technical issues related to the projects), *Environmental*, *Economic*, *Social* (the three pillars of sustainability), and *None* (tweets discussing the infrastructure project-related issues from other aspects). As Table 1 highlights, The Environmental class had the lowest number of tweets. On the other hand, Social and Engineering classes were the two largest classes. Also, as it is seen, a majority of collected tweets have been either irrelevant or repeating and at the end, a total of 1,228 tweets were used for training our classifiers.

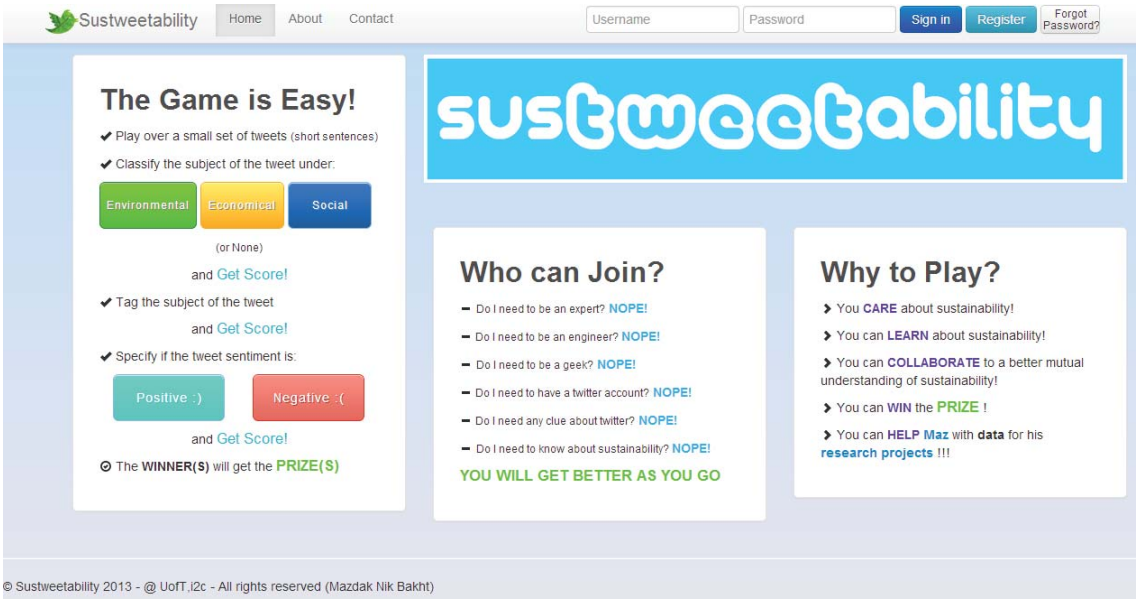

Figure 1 Homepage of the GWAP run for annotating the training set through crowdsourcing

3.3 Training classifiers

Data which was collected, filtered, and annotated as explained above, underwent a pre-processing in the next step. Pre-processing helps with cleaning the data-set and removing noise (objects which do not add specific value or may result in confusion for the classifiers). The noise removal in this study included the following steps and was handled through RegularExpression (RegEx):

- Removing html tags and attributes which are not visible on a browser
- Replacing all html character codes (such as &amp, &quot, etc.) with their ASCII equivalents
- Removing all the URLs (as they are most of the time associated with advertisements and commercials)
- Removing Twitter-specific characters such as hash-tags (#) and mentioning (@)
- Replacing all the monetary values with a trackable variable ($XXX)
- Replacing all the percentages with a trackable variable (XX%)

The next phase of pre-processing was tokenizing tweets and preparing the data-attribution table. The data-attribution table is the input of classification through supervised learning. We took the following steps for tokenizing:

- Transforming all characters to lower case
- Tokenizing the text at "non-letter" characters (such as white spaces, hyphens, etc.) as splitting points

- Removing Stop-Words (the words with no specific sense such as a, the, and, etc.)
- Filtering unneeded tokens (tokens with a length of lower than 3 or higher than 25 characters)
- Forming unigrams as well as bi-grams (although as mentioned before, bi-grams do not add much value in the current analysis)

Moreover, we considered the effect of stemming; however, as classification of stemmed data did not eventually show a higher level of accuracy, we decided not to include stemming as a part of the pre-processing. Decision tree (DT), Naïve Bayes (NB), K-Nearest Neighbors (K-NN) and Support Vector Machine (SVM) were the main classifiers tested in this study. Term and bi-grams were selected as the features and term frequency was considered as the feature value. We split our data-set into four sub-sets (one per each project, as introduced by Table 1) and examined the performance of different combinations of algorithms in classifying tweets within each data-set, as well as on the full data-set (including all 1,228 annotated tweets).

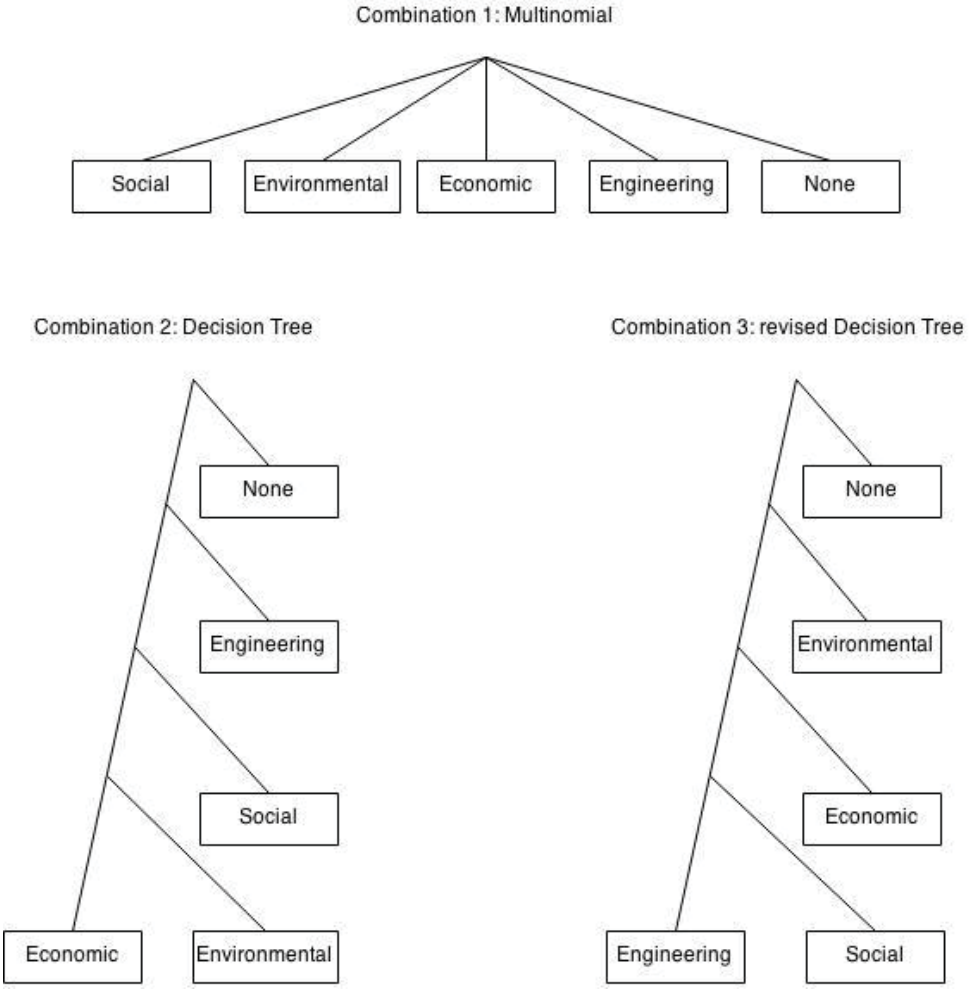

Figure 2 Classification combinations tested in this study

Figure 2 depicts the main combinations tested in this study. The first combination was a multinomial model assigning a label to each instance at one level. This method does not use the capabilities of a decision-tree model; therefore the model developed based on the Decision Tree algorithm had a low performance. In the K-NN model (taking k=2 through a trial and error procedure), the two nearest neighbors were determined based on Mixed Euclidian Distance. The second combination constitutes a predefined Decision Tree and the other models (K-NN, SVM, or NB). This type of model benefits from the hierarchical feature of the Decision Tree model. At each

level of the tree, a decision is made on whether the instance belongs to a specific class or not (binary decisions). The first level decides on the relevancy of the Tweet to the pillars of sustainability and/or engineering context. Since the tweets at each level are classified in two categories, unlike the first combination, the K-NN model uses Cosine Similarity as a numerical measure to find the two nearest neighbors of the instance under consideration. The third combination was a revised version of the second combination. Since the algorithms are categorized as supervised model, the available data plays an important role in the performance of the models. Based on this fact, the hierarchy of the decision tree was revised in order to attain the highest possible performance in terms of accuracy, recall, and precision. Therefore, the third combination can be considered an optimized structure, for the data-set in hand.

In all combinations, cross validation was performed to assess the performance of the models for each data-set as well as for the combined data-set. The model was trained by 90 percent of the data and the remainder was used to test the trained model's performance. Stratified sampling was utilized to ensure that the class distribution in the training and testing subsets are the same. The models were compared by forming confusion matrices and through their accuracy, recall, and precision. More information about the comparison and the results are provided in the following section.

4 Results

As explained above, the optimum classifier is a decision tree with four binary decision points; at the first layer, the classifier detects the tweets which do not belong to any of the semantic classes of interest. At the second layer, the decision tree decides if the tweet is discussing environmental sustainability; given the low number of these tweets in our data-sets, this can be interpreted as filtering out such tweets. The third decision point separates tweets discussing the economy of the project form the rest of the tweets, and finally the classifier decides whether the tweet is focusing the project from the aspect of social sustainability or is discussing an engineering/technical aspect of the project.

Table 2 Performance of different classifiers (in terms of accuracy)

			Decision Tree (Comb. 2)				Optimized DT (Comb. 3)			
Data-set	Method	Multinomial (comb. 1)	Layer 1 (NONE)	Layer 2 (ENG)	Layer 3 (SOC)	Layer 4 (ECO /ENV)	Layer 1 (NONE)	Layer 2 (ENV)	Layer 3 (ECO)	level 4 (ENG/SOC)
All	K-NN	38.69	68	54.76	70	82.78	68	96.96	83.19	56.58
All	NB	35.27	80.38	53.95	64.93	77.22	80.38	91.38	72.55	55.59
All	SVM	N.A.	79.64	59.44	65.74	81.11	79.64	95.95	80.06	58.68
CC	K-NN	43.05	68.29	73.11	69.64	60	68.29	94.17	88.48	62.36
CC	NB	31.76	74.48	51.14	61.43	55	74.48	83.26	78.48	52.55
CC	SVM	N.A.	83.86	61.44	71.07	70	83.86	91.67	87.73	60
ATL	K-NN	52.71	86.03	74.55	83.52	20	86.03	97.55	91.97	70.56
ATL	NB	40.12	87.37	53.36	73.19	30	87.37	87.19	81.95	53.1
ATL	SVM	N.A.	91.92	65.71	84.95	71.67	91.92	97.07	91.47	63.36
CT	K-NN	39.35	62.7	52.44	63.62	87.55	62.7	97.43	80.1	55.59
CT	NB	37.16	78.41	56.58	65.11	78.82	78.41	91.72	72.14	58.83
CT	SVM	N.A.	74.89	56	59.23	82.91	74.89	96.45	76.63	53.12
M1	K-NN	39.44	85.29	74.04	55.77	87.5	85.29	97.46	62.38	70.53
M1	NB	35.03	80.23	69.04	61.28	86.67	80.23	93.04	68.17	65.53
M1	SVM	N.A.	85.29	75.25	64.49	90	85.29	96.83	72.13	71.36

(All numbers are percentage)

The relatively large number of tweets which are annotated by the crowd at both engineering and social-sustainability classes can justify why this decision is the last level of the tree (between 9% and

14% of tweets in different data-sets belong to both of these classes). The observation which gives more weight to this speculation is that data-sets with lower levels of overlap between these classes have higher levels of accuracy at the last layer. Table 2 compares the three combinations depicted in Figure 2 for different data-sets, based on the accuracy (introducing the overall effectiveness of the classification system as the percentage of the results that have been classified correctly).

Decision at any of the decision points can be made through any of the classification methods (K-NN, NB, or SVM). In this table, the results based on each of these algorithms are presented and compared. As seen, the SVM at each decision point has a better performance compared to K-NN and NB (with only a few exceptions); however, a clear preference cannot be highlighted between K-NN and NB classifiers. Nonetheless, the optimal order of classification is independent of the classifier type used at decision points. Studying the distribution of subjects in data-sets suggests that higher standard deviation for class sizes in a data-set is generally resulting in a higher overall accuracy for the classifier. Investigation in the correlation between the data-set size and accuracy suggests a lower accuracy for small and large sized data-sets. The highest level of accuracy on the other hand, is seen in classifying tweets with the subject of environmental sustainability. This may be a result of the low size of this class rather than a more accurate distinction for this subject. In general, the results show that after filtering out irrelevant data, classes with lower size can be detected at a higher accuracy.

Last but not the least, Figure 3 depicts the performance of the selected classifier (the optimal decision tree in which each decision node works based on an SVM) for different data-sets of this study. As seen in this figure, while there is no drastic differences between the overall performance of the classifier in various data-sets, ATL streetcar has a higher level of accuracy at all four layers. The figure also shows accuracy above 50% at all levels and the accuracy of above 70% at the first three layers for all datasets.

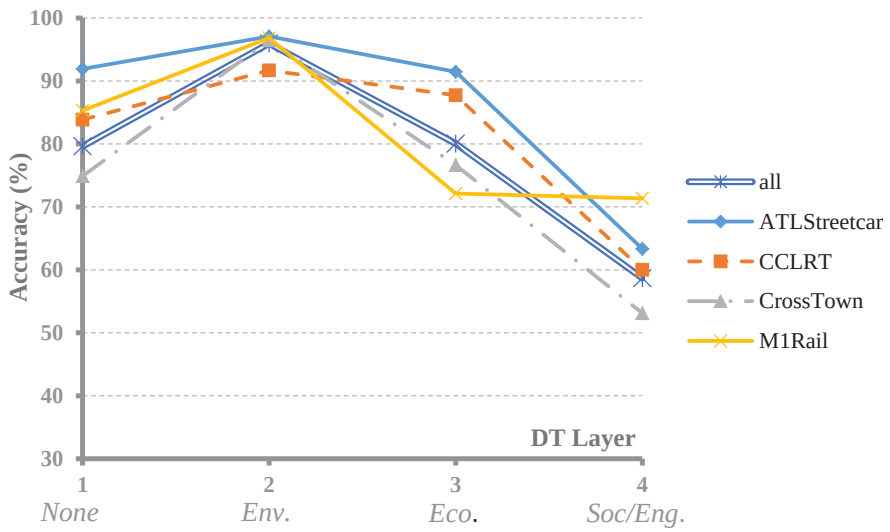

Figure 1 Accuracy of the optimal classifier (optimal DT, with SVM) in different data-sets

5 Conclusion and future work

This paper reported the initial steps towards developing an automated detector and classifier for infrastructure-related opinion over social media. Projects form transportation sector (and LRT sub-sector) were collected, processed, and used in different combinations for generating a subject classifier in the context of sustainability of the infrastructure system. Crowdsourcing along with the wisdom of the crowd was used for annotating the training and test sets, and Decision Tree combined with different classifiers was examined to reach the highest level of accuracy in classification. The results show an acceptable – yet not impressive – level of accuracy by selecting a simple feature such as vocabulary, and term frequency as the attribute value. In our data-set, and for the selected analysis context, the order of retrievable accuracy was: Environmental, Economic, and Social sustainability, and then Engineering/Technical issues; however, this may be dependent on the

data-set. Although the order of classification at the decision tree can be optimized independently from the classifier type in use, SVM dominates K-NN and NB in terms of accuracy of the results.

Some former experiences have criticized the performance of pure machine coding due to the confusions happening for AI systems with respect to issues such as word-phrase associations and hyponyms, as well as being disconnected from the 'context', which challenge the reliability of interpretations (Macnamara, 2005). Such studies suggest human coding as an alternative, or a complement to AI-based classification. However, a classification system such as what was suggested by this paper can be a good starting point to reduce the required cost and effort of human coding, and/or to control its outputs.

Subjects of tweets by followers of an infrastructure project can be matched with their vested interests with respect to the project. Sentiment of their tweets on the other hand can be correlated with their position with respect to the project. Therefore, in combination with an effective sentiment classifier, what was presented in this paper can result in a full stakeholder mapping in terms of their opinion. A full analysis over time can then lead to detection of their opinion dynamics. These are the future steps of the research which is currently underway.

References

M1-Rail Streetcar project business plan. (2012). Retrieved July 15, 2013, from http://www.m-1rail.com/wp-content/uploads/2013/02/Business-Plan-for-FTA.pdf

Twitter Statistics. (2015, December 27). Retrieved from Statisticbrain: http://www.statisticbrain.com/twitter-statistics/

Aggarwal, C. C., & Zhai, C. (2012). *Mining Text Data.* Springer-Verlag New York.

Azhar, S., & Abeln, J. M. (2014). Investigating social media applications for the construction industry. *Procedia Engineering*, 42-51.

Bennett, S. (2013, Nov 12). *Social media and the construction industry.* Retrieved Feb 09, 2015, from Social Times: http://www.adweek.com/socialtimes/social-media-construction/493101?red=at

Bregman, S., & Watkins, K. (2013). *Best practices for transportation agency use of social media.* CRC Press.

Dumais, S., Platt, J., Hecherman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. *Proceedings of teh 7th international conference on information and knowledge management*, (pp. 148-155). New York.

Evans-Cowley, J., & Griffin, G. (2011, Feb 12). *Micro-Participation: The Role of Microblogging in Planning.* Retrieved Feb 27, 2014, from SSRN: http://ssrn.com/abstract=1760522 or http://dx.doi.org/10.2139/ssrn.1760522

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 1289-1305.

Horn, C. (2010). *Analysis and classification of Twitter messages.* Graz: Graz University of Technology.

Macnamara, J. (2005). Media content analysis: Its uses; benefits and best practice methodology. *Asia Pacific Public Relations Journal*, 1-34.

Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing.* MIT Press.

Manning, C. D., Raghavan, P., & Shutze, H. (2009). *An introduction to information retrieval.* Cambridge, England: Cambridge University Press.

Nik-Bakht, M., & El-diraby, T. E. (2015). SUSTWEETABILITY: Exposing Public Community's Perspective on Sustainability of Urban Infrastructure through Online Social Media. *International Journal of Human-Computer Studie* (in press)

Olander, S. (2007). Stakeholder impact analysis in construction project management. *Construction management and exonomicsa, 25*, 277-287.