# Social Sentiment for Sustainability of Infrastructure Mega-projects

Moein Hosseini, moein.hosseini@mail.utoronto.ca
*University of Toronto, Canada*

Mazdak Nik-Bakht, mazdak.nikbakht@concordia.ca
*Concordia University, Canada*

Tamer E. El-Diraby, tamer@ecf.utoronto.ca
*Universtiy of Toronto, Canada*

**Abstract**
In this paper, it is shown how to train automated classifiers for detecting the sentiment of infrastructure-related discussions in online social media, within the context of sustainability of a project. By modeling "opinion" as a combination of "subject" and "sentiment", and by defining sentiment as a measure of being "proponent" or "opponent" to a specific aspect of a project, trained classifiers can be used in the process of stakeholder management for urban infrastructure projects. Infrastructure-related tweets collected over a course of eight months (from Aug 2012 to Mar 2013) were annotated in terms of subject and sentiment. The annotation was completed by players of a 'Game With A Purpose' (GWAP) – called 'Sustweetability'. Wisdom of the crowd resulting from the GWAP helped us overcome the subjectivity of the annotation problem. We used the results of this game as the training (and testing) dataset for the classifiers. The combination of statistical Natural Language Processing (NLP) and supervised learning methods used to develop the classifiers are addressed in this paper. We defined a new statistical measure, called Deviation from Maximum Accuracy (DMA), to control the performance of classifiers developed through different combinations of tools. Together with the work presented earlier (in CIB W78, 2015) on detection and classification of subjects, the findings can give rise to detection of public opinion on infrastructure projects.

**Keywords:** Infrastructure mega-projects, Stakeholders Analysis, Public opinion, Twitter, Sentiment classification, Natural Language Processing

## 1 Introduction

The ever-increasing trends in employing online social media for customer relationship management have recently attracted the attention of stakeholder management practitioners and researchers in domain of urban construction & management. Attributes such as wide outreach, accessibility, and openness make online social media an ideal communication channel for infrastructure owners/operators to communicate information, directions and updates about the system with its end users. Many transportation agencies, among other infrastructure sectors, actively use social media channels such as Twitter, Facebook and YouTube as part of their public relations campaign when planning, building, or operating urban infrastructure project. More importantly, these channels are two-way communication tools and employing them at a higher level of maturity can help to distill the information and knowledge (which in most cases is context-specific) from the public's inputs. This can cover a wide range of applications including level of service definition, performance evaluation, demand detection, user innovation, etc. In AEC industry, however, one major barrier to achieve these goals is the lack of automation required to classify, structure, and understand the seemingly chaotic inputs from end users/prospective end users of the system developed/under development.

Classification of user generated contents is of interest in various domains. In general such a task must be done either manually (which is expensive and error-prone), or automatically. By induction over a set of pre-classified sample instances, machine learning methods can help for automatic topic/sentiment detection. These methods generally work based on statistical pattern recognition. More sophisticated methods such as neural networks are also used in the literature for such purposes. In classification of the content in a specific domain however, the defined and limited scope of the problem helps machine learning to reach a satisfactory performance. It is shown that classifiers trained through machine learning are topic-dependent, domain-dependent, and temporally-dependent[1] (Read, 2005). Therefore, in order to 'understand' infrastructure-related discussions over online social media in terms of subject and sentiment, particular context-aware classifiers must be trained using infrastructure-related text corpora and knowledge. We have used machine learning and Natural Language Processing (NLP) for this purpose and this paper reports the outcome of our attempts to detect and classify the sentiment of the public on the sustainability of infrastructure mega-projects.

### 1.1 Contribution & impact

Billions of dollars are being wasted worldwide on ineffective public involvement campaigns and consequently project failures due to the lack of accordance to the public community's demands. This is calling for the use of more effective outreach channels in a more meaningful manner. Although online social media can offer such channels, as briefly addressed, AEC industry currently lacks formal/context-sensitive tools and systems to automate the procedure of collecting, classifying, contextualizing, and aggregating public inputs. This has resulted in wasted opportunities as well as frustrations at both sides of the communication (public communities whom are not being heard and infrastructure owners who do not find social media channels more than 'another webpage' for the project). In spite of achievements and tools developed in other domains (such as computer science and business administration), off-the-shelf solutions offered in one domain may not be directly applied to another domain. This is due to the context-sensitivity of opinion analysis among other reasons.

More specifically, when it comes to the analysis of public opinion on an infrastructure project, one must categorize and classify the characteristics of a project which create more sensitivity among public communities and therefore are discussed more frequently. There is a wide range of aspects (from technical to social, economic, political, etc.) which may create interests among the communities, and automation of stakeholder analysis must first and foremost realize their vested interests. In a former paper in CIB-W78, the authors presented automated classifiers developed to detect the topic (subject) of online discussions (Tweets) within the context of project sustainability (Nik-Bakht, et al. 2015). Upon defining the topics of interest in an infrastructure project (normally by the project owner), such classifiers can be used to automatically categorize the public's inputs based on the aspect they discuss.

The other side of the coin of opinion analysis (the more difficult side to handle) is evaluation of the sentiment of a discussion. Despite the variety of automated sentiment classifiers available off-the-shelf, none of them can be reused or recycled for infrastructure projects and that is due to the fundamental difference in definition of the sentiment within the context of stakeholder analysis. Sentiment in general applications is defined as happy (positive sentiment), sad (negative sentiment) and neutral. The aim of stakeholder analysis, however, is to identify whether a stakeholder is "proponent" or "opponent" to a project (from a certain aspect). The latter does not necessarily correspond to the former and each proponent or opponent attitude can be expressed in form of sentences with either positive or negative sentiments. For example, the tweet left on a waste water management project in the city of Montreal: "*Yayy! Montreal is going to be an even bigger trash bin!*", is expressing being against the project, through a sentence with a positive sentiment. Such issues make the problem of automated classification ultimately complicated.

On the other hand, training classifiers to detect such a sentiment cannot happen by repeating the same steps taken for general sentiment classifiers. This is first of all due to the subjectivity which is inherently involved in the problem. Also collecting training data will not be as straightforward as in

[1] Temporal dependence refers to the dependency of the trained classifier on biasedness of the training data collected over a specific period of time

general sentiment analyses (where normally sets of tweets with happy or sad emoticons are taken as training data for positive or negative sentiment)

### 1.2 Organization of the paper

The paper starts with a brief review on similar attempts inside and outside the domain of infrastructure, and then in section 3 we explain the methodology mentioning the data collection and data preprocessing, the features and algorithms used to train the classifier, and the measures used to evaluate the classifiers. 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 Related works

### 2.1 Social media and urban infrastructure projects

Detecting the stakeholders vested-interest in an infrastructure project and their position (being in-favor or against the relevant decisions) are two major objectives of stakeholder analysis (Olander, 2007). Public communities are the most challenging key players among the stakeholders of a project. Public involvement (PI) programs are designed for the purpose of involving the public in the decision making process. In recent years there has been a dramatic upsurge in the popularity of online social media among PI practitioners because of the bi-directionality nature of the communication in online social media. However, harvesting the wealth of knowledge from the corpus of user generated content on online social media requires tools and methods which do not yet exist in the field.

Recent studies have emphasized the potential of online using social media to involve the public in different stages of decision making ranging from long-term policy development and planning to daily operation (Grant-Muller, et al., 2014). For example, Collins, et al. (2013) introduced a tool to evaluate public transit rider perceptions about the quality of service. They performed a sentiment analysis on messages extracted from Twitter about the public transit services in Chicago. Therefore, the tool could be used to incorporate public opinion in decisions made for the daily operation of the transit system. As a result, an effective public involvement in the decision making process for an infrastructure project requires analyzing the social media text content both semantically and sentimentally. There are various techniques in the field of computational linguistics to perform this type of analysis which are discussed in more detail in the remainder of this section.

### 2.2 Statistical NLP

Ambiguity of the natural language for the machine has many different aspects. Word sense, word category, syntactic structure, and semantic scope are among other features which challenge automatic understanding of naturally generated texts. Computational linguistic 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 can be considered as one solution. Although working perfectly in machine interoperability, when it comes to evaluation of the naturally occurring text, such methods perform poorly (Manning & Schutze, 1999). Statistical NLP approaches on the other hand suggest a solution to this challenge by "automatically learning lexical and structural preference from 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. "Statistical models are robust, generalize well, and behave gracefully in the presence of errors and new data" (Manning & Schutze, 1999). Terms' statistics are usually related to word counts (from [simply] the most frequent terms in a text to a truly representative sample of words)

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 equivalence classes)
- Statistical estimation (Finding a good statistical estimator for each equivalence class)
- Combining multiple estimators

## 2.3 Classifiers

Training a classifier is generally involved in two main phases: pre-processing, and training. Preprocessing prepares collected tweets to train the classifier. Post processing on the other hand is the learning process and includes feature selection and training classifier.

### 2.3.1 Feature selection/extraction

Feature selection is decision making on the most distinctive set of features to be used in training a classifier. Feature selection is the cornerstone of an efficient and accurate learning process. The procedure generally involves in scoring all potential features according to particular metrics and selecting the best ones to make sure about employing adequate/not too many features. Preprocessed data must be analyzed to find variables that are indicative for each class. The goal is a correlation-based feature selection which minimizes redundancy and maximizes relevance at the same time with keeping the ability of distinction.

The first question in feature extraction is what type of features should be selected to count. Different types of features are used in literature for training classifiers including *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), semantics, etc. Given the specific nature of tweets (such as short length, brevity, use of slang, etc.), many studies have focused on the tweets corpus in particular. Go, et al. (2009) refer to the casual language used on tweeter, extensive usage of URL links, and mentioning user names as some of the challenges in training classifiers for Twitter. They suggest the unigram method to be used in order to overcome these challenges.

The second question is how many features are enough to be considered. Collection of all features' values for an instance (a tweet) is called a feature vector. Values assigned to each entree of the feature vector can have one of the three different domains: (Positive) integer (natural) numbers $\in \mathbb{N}$ (for the primary counts), Real numbers $\in \mathbb{R}$ (for retrieved scores such as tf-idf, subjectivity, or polarity), and Boolean $\in \mathbb{B}$ (for presence or absence of bag of words). The learning process will be a problem in the vector space of feature vectors; therefore, the problem dimension will be equal to the number of features considered.

Feature selection is generally a complex problem in machine learning. Blum and Langley address this problem by giving various definitions for 'relevance' of features and introducing computational methods for selecting relevant set of features for machine learning procedures (Blum & Langley, 1997). Forman compares efficiency of various metrics in feature selection through analysis of empirical data (Forman, 2003). Weaker classifiers (using less number of features) are particularly preferable when limited training data is available. This is due to the bias-variance tradeoff (Manning, Raghavan, & Shutze, 2009). Basic feature selection algorithms involve defining a utility measure $A(f,c)$ for selecting each feature ($f$), for each class ($c$); and then selecting the $k$ features with the highest utilities. Various utility measures are defined and used in the literature: mutual information (or information gain), $\chi^2$-test, and frequency can be mentioned among others.

Sentiment classifiers recognize the general feeling of the author about the subject they discuss. Vocabulary cannot be distinctive enough for such a purpose and more descriptive features are necessary to be considered. Syntactic structure, part of speech (PoS), and *n*-grams are the candidates for this purpose. Although *n*-grams work very well with the large texts, as Barbosa and Fang show, their performance is not as good when it comes to analysis of very short messages such as tweets (Barbosa & Feng, 2010).

There is a considerable amount of published work on sentiment analysis for the online/user-generated material such as blogs, reviews, and tweets. In many studies the training sets are classified based on the occurrence of happy and sad emoticons (Read, 2005), (Agarwal, et al., 2011), (Saif, et al., 2012). Table 1 summarizes some of the most important studies in this regard. Go et al.

(2009) suggest that a combination of mutual information feature selection and Naïve Bayes classifier gives the highest accuracy. They also address *Negate* and POS as additional features helping to increase the accuracy (despite increasing training time). An important result of their study is the fact that adding "Neutral" as a third class (together with positive and negative sentiment) drastically drops the accuracy (up to half!) They blame the noisy training data of their neutral class, but such a drop in accuracy is more or less confirmed by later studies.
Park and Paroubek experimented with different variants of unigram, bigrams and trigrams with POS and concluded that bigrams show the highest accuracy (more than unigram and even trigrams). They also conclude that syntactic structures and part of speech are strong features for indicating emotion in texts. Considerable points about their method include connecting the negations into the words (e.g. do not → do+not) and removing stopwords! (Park & Paroubek, 2010). Barbosa and Feng proposed frequency of POS and some tweet-specific syntax (such as retweets, hashtags, reply, links, etc.) together with punctuations, emoticons, and uppercases to be used as the features. On top of the usual POS, they also suggested two more meta-information tags for each word: prior subjectivity (weak or strong subjectivity), and polarity (positive, negative or neutral). They finally suggest positive polarity, existence of links, strong subjectivity, words starting with uppercase, and verbs as the five top indicator features (Barbosa & Feng, 2010). Agarwal et al. (2011) repeated similar experiment by calculating polarity using DAL (dictionary of affect in language) and combining prior polarity with POS. For words which are not in DAL they used the closest synonym from WordNet. They use a 5-fold cross validation and report the average of accuracy over the five folds as the overall accuracy. Recently, some studies have looked at semantic sentiment analysis. Saif et al. add semantic class as a feature for analysis of sentiment and report a 5%-6% increase in the average accuracy (Saif, He, & Alani, 2012).

**Table 1** Sentiment Analysis of Twitter; Literature Review

| Authors | Year | Number of Classes† | Features Used | Feature Selection Method | Learning Algorithm | Validation | Accuracy Range |
|---|---|---|---|---|---|---|---|
| Read | 2005 | 2-way | Unigrams, Syntax, Polarity, | Not specified | SVM, NB | Not specified | 41%-70% |
| Go, Huang & Bhayani | 2009 | 2 -way | Unigram, Bigrams, and POS | Frequency based, Mutual information and $\chi^2$-test | NB and MaxEnt. | 2-fold cross validation | 61%-83% |
| | | 3-way | | | | | 40%-45% |
| Park and Paroubek | 2010 | 3-way | Binary N-grams and POS frequency | Not specified | NB | F-measure over the whole dataset | > 60% |
| Barbosa & Feng | 2010 | 3-way | POS, polarity, subjectivity, and Tweet syntax | Mutual information | SVM | Not specified | 68%-80% |
| Agarwal, et al. | 2011 | 3-way | Unigrams, Polarity, POS, and syntax | Mutual information | SVM | 5-fold cross validation | 70% - 75% |
| Saif, He & Alani | 2012 | 2-way | Unigrams, polarity, POS, semantics | Not specified | NB | 2-fold and 5-fold cross validation | 66%-84% |

† 2-way means: Positive/Negative and 3-way means Positive/Negative/Neutral

### 2.3.2 Training classifiers

The best classification algorithm will be selected among [at least] three types of classifiers: Naïve Bayes, Support Vector Machines (SVMs), decision trees, *k* nearest neighbor (kNN) and. There are also other types of classifiers such as Maximum entropy (MaxEnt); however, given the prevalence of the first four types in the literature, we limit the scope of our experiment into those classifiers.

Naïve Bayes classifiers work based on the features independence assumption, and Bayes theorem. This method calculates the probability for belonging a document (or instance) $d$ to a class $c$ :$P(c|d)$, given the document's features ($f_1$ to $f_k$). Based on the chain rule, if features $f_1$ to $f_k$ 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. Therefore, SVM can be interpreted as maximization of margins between different classes.

Decision trees (which are 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.

K nearest neighbors is one of the most simple machine learning algorithms. First, the data points are plotted in the predefined feature space. Then, the label of each unlabeled data point is defined based on the label of its k nearest neighbors. Euclidean distance is the most common function used to find the distance between each two points in the feature space.

## 3 Methodology

Training a classifier is generally involved in two main phases: pre-processing, and training. Preprocessing prepares collected tweets to train the classifier. Post processing on the other hand is the learning process and includes feature selection and training classifier. In this part, different steps for each phase are explained.

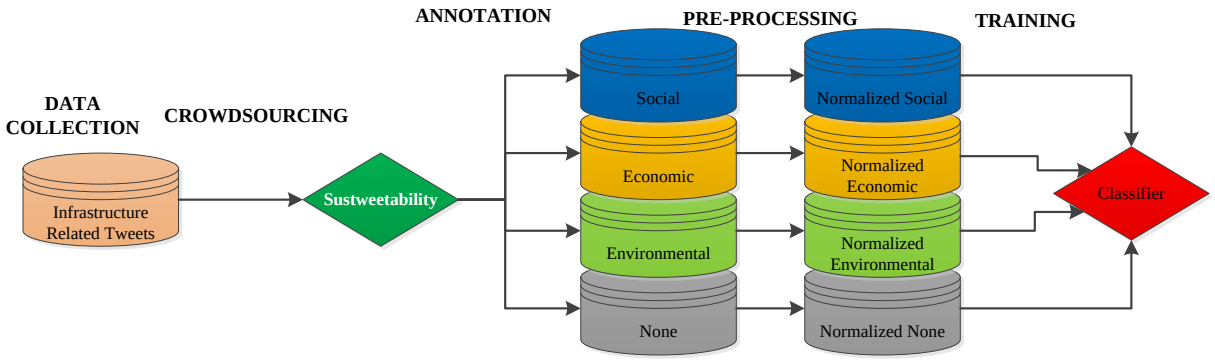


**Figure** 1 steps for training classifiers (similar preprocessing is repeated for sentiment)

## 3.1 Corpus

### 3.1.1 Collection and annotation

The dataset includes tweets about four North American light rail transit (LRT) projects which were at different phases of their lifecycle at the time period of data collection. The case studies were the Woodward Ave. LRT (Detroit, USA), Eglinton Crosstown LRT (Toronto, Canada), Atlanta Streetcar (Atlanta, USA), and Central Corridor (Minneapolis, USA) which were at the pre-construction, early construction, construction, and late construction, respectively during the data collection.

In order to collect the data relevant to these projects from Twitter an automated data collector was programmed which sent requests to the Twitter API. The requests contained a request URL, sender's electronic signature, and specific keywords related to the projects. Responses were received as .json files and five fields of the responses were used including 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). More than 40,900 tweets were collected from Aug 2012 to Mar 2013 for the four case-study projects. After preprocessing the tweets and removing irrelevant and repeated tweets 1,228 tweets were remained which were used in the annotation step.

Crowdsourcing was utilized to annotate the preprocessed tweets. We set up a Game With A Purpose (GWAP), called "Sustweetability", in which players were asked to annotate the tweets from the aspects of semantic and sentiment. Sustainability of the infrastructure system was selected as the main scope for semantic classification. Five classes were defined accordingly: *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). Players could earn points by classifying each tweet in the 'right' class. The right answer was considered as the mode(s) of the distribution. Therefore, a tweet could belong to multiple subject or sentiment groups if the answers have a multimodal distribution. 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. More details about Sustweetability GWAP is available in (Nik-Bakht & El-diraby, 2015). Table 1 summarizes the number of tweets which were annotated at each subject and sentiment classes.

Table 2 The number of tweets for each subject and sentiment classes used in the analysis

| | Sentiment | | | |
|---|---|---|---|---|
| **Subject** | Negative | Neutral | Positive | Total |
| **Environmental** | 6 | 3 | 19 | 28 |
| **Economic** | 38 | 38 | 76 | 152 |
| **Social** | 64 | 56 | 313 | 433 |
| **Engineering** | 42 | 209 | 122 | 373 |
| None | 7 | 199 | 36 | 242 |
| **Total** | 157 | 505 | 566 | 1228 |

### 3.1.2 Pre-processing

The tweets collected via Twitter API include 'noise' and are not in a form amendable to feature extraction for classification. Preprocessing is in fact the procedure of removing such noises and tokenizing the tweets. Converting the collected/annotated data into the normalized form includes the following steps:

*Clearing:*

- 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
- Replacing all the URLs with symbol (TWITTER_LINK)
- Replacing Twitter specific elements with relevant symbols (TWITTER_RETWEET, TWITTER_MENTION, TWITTER_HASHTAG)

- Replacing all the monetary values with a trackable variable $XXX
- Replacing all the percentages with a trackable variable XX%

*Tokenizing:*

- Decomposing clitics and punctuations from their hosts
- Saving ellipsis and other forms of multiple punctuations as separate tokens
- [For content classifier] Stemming each token
- [For sentiment classifier] Tagging each token by its part of speech (token/PoS)

Python codes were developed to perform these activities using RegEx (regular expression) together with parsing and tagger libraries.

## 3.2 Sentiment classifiers

### 3.2.1 Feature extraction

As mentioned, given the short length of tweets, weselected unigrams as the features. One problem with using unigram features is the issue of negations. However, as Go et al. (2009) show, using bigrams not only does not help, but also drops the accuracy (due to sparsity of negations in the data collected from Twitter.

We collected all the features and feature vectors in a file called Attribute-Relation File Format (ARFF). Such files contain a header (including the name of the relation and a list of the attributes (features), together with the type of values they accept), and the data (introducing the values of each attribute for each instance). Each row represents the feature vector for one instance. As we are counting occurrence of the features, our attributes are mostly numerical, except for the class name and the semantic class. Another python program was developed to take the preprocessed set of tweets and build the ARFF file.

Taking the works done into consideration, we experimented with unigrams and three types of features for sentiment analysis: POS, Twitter Elements, and Semantics. We particularly introduced the semantic class (social, economic, environmental, or none) as an extra feature and evaluated its contribution to the accuracy of the classifier. Table 3 summarizes the features that were tested.

**Table 3** Features to be tested for training the sentiment classifier

| Type | Candidates |
|---|---|
| **POS** | Nouns (common & Proper nouns), Verbs (person, tense, modal verbs, objective/subjective form), Adjectives (superlative and comparative), Pronouns (1st, 2nd, 3rd person), Utterances, Wh-words, Negations, Determiners, Preposition, Numbers, Modals |
| **Twitter Elements** | Hashtags (#), Mentions (@), Retweets (RT: @), URLs |
| **Semantic** | Sustainability element discussed |

### 3.2.2 Training classifiers

Data which was collected, filtered, and annotated as explained earlier was fed into the classifiers. Decision tree (DT), Naïve Bayes (NB), K-Nearest Neighbors (K-NN) and Support Vector Machine (SVM) were the main learning algorithms tested in this study. Seven combinations of features were tested as follow:

- Token: Word occurrence was considered as the feature value for a tweet.
- Token + Bigram: Occurrence of consecutive tokens was regarded as another feature.

- Token + Bigram + Twitter Elements: The occurrence of Twitter Elements was considered which added four new features to the feature set.
- Token + POS (occurrence): Occurrence of various type of POS elements was added to the occurrence of tokens.
- Token + POS (relative frequency): In contrast to the previous feature set which assigned binary values to the POS features, in this feature set POS elements possessed a continuous value ranging from 0 to 1.
- Token + POS (median threshold): A binary value was assigned to each feature defined for POS elements. The only difference with the fourth feature set is that, for each POS element, the threshold was changed from zero frequency to the median of the non-zero frequency distribution calculated for the entire corpus.
- Token + POS (occurrence and median thresholds): Both the occurrence and median thresholds were retained. Therefore, a discretized value is assigned to each POS element from the following list: "No Occurrence", "Low Frequency", "High Frequency".

Different classifiers were trained for each subject under consideration and for the entire data set. In total 252 classifiers were trained and tested by using different combinations of algorithms and feature sets. Since the classifiers were trained to detect the sentiment of a tweet, the sentiment was chosen as the decision variable which could possess three different values: "Positive", "Neutral", or "Negative".

### 3.2.3 Validation and performance measures

Various algorithmic performance measures such as learning speed, real time classification speed and classification accuracy have been introduced in the literature to compare the effectiveness of automatic learning algorithms (Dumais, Platt, et al., 1998). For our purpose however, accuracy and relevance seem to be the major performance measures. *Precision*, *recall*, and *average accuracy* are the main measures of quality in this respect. While the first two are calculated for each class separately, the last one is introduced for the whole classifier. In order to calculate these measures, first a $c \times c$ matrix (called *confusion matrix*) is formed for each classifier to summarize the classification performance in different classes. Here $c$ introduces the number of classes, and each entry $C_{ij}$ of the confusion matrix reports the number of instances originally belonging to class $i$ that have been classified as class $j$.

The accuracy ($A$), is the overall measure of classifier's quality. It is the fraction of all the hits (trace of confusion matrix: the cases which are classified in the classes they truly belong to) to all the instances:

$$A = \frac{\sum_i C_{ii}}{\sum_{i,j} C_{ij}} \qquad (2)$$

$k$-fold cross-validation will be the method of validating performance of the trained classifier. In this method, the whole dataset will be randomly split into $k$ segments of the same size and accuracy is evaluated $k$ times. Each time one of the subsamples is taken as the testing set, and the other $k - 1$ sets are used as training sets. The total accuracy is usually taken as the average of the $k$ accuracies. Value of $k$ is selected such that the mean response in all subsets is approximately equal. $k$=2, 3, 5, and 10 are typically used in the literature, resulting in 2-fold, 3-fold, 5-fold, and 10-fold cross validation respectively.

The objective is to find the best pairs of algorithm and feature set which have a relatively high performance throughout all semantic classes. In order to achieve these goals a measure is defined based on the average accuracy as follows:

$$DMA_{i,j} = \sqrt{\sum_k \left(\frac{max(A_k) - A_{i,j,k}}{max(A_k)}\right)^2} \times 100 \qquad (3)$$

Where DMA is the deviation from maximum accuracy in percentage; i is the index for feature set; j is the index for algorithm; k is the index for subject categories; and the maximum function is performed on all accuracies calculated for each subject over all analyzed feature sets and machine learning algorithms.

## 4 Results and discussion

The average accuracy calculated for the classifiers shows that the K-NN algorithm is the least suitable algorithm for training the classifier. However, because of the simplicity of the K-NN algorithm, if one decides to use this algorithm, considering only the closest three neighbors is recommended. Over the other algorithms, the average accuracy of the models were slightly above 50% when the entire data set was fed into the model. The accuracy increased by approximately 10% when the model input was restricted to only ENV, ECO, or ENG subjects. Among the pillars of sustainability the model trained by the SOC subject data reached an accuracy of 70%. Training the models with the data tagged with the NONE subject resulted in the highest average accuracy of 80%. As a result the models developed for specific subjects are more accurate than the models trained by the entire data set. The reason for such difference is that some tweets expressed mixed emotions by supporting a sustainability aspect of the project while disapproving another aspect at the same time. Including the bigram and the Twitter elements feature did not improve the performance of the classifiers.

Table 3 shows the DMA calculated for each trained classifier. The results show that the SVM is the most stable algorithm because for almost all feature sets the DMA of the classifier is less than 20%. The NB algorithm is the second most stable algorithm especially if POS is considered in the feature set either with a binary or a trinary value.

**Table 3** DMA for each classifier

| Feature Set | Machine Learning Algorithm | | | | | |
|---|---|---|---|---|---|---|
| | NB | SVM | DT | K-NN (K=2) | K-NN (K=3) | K-NN (K=4) |
| Token | 19.46 | 12.82 | 37.82 | 56.35 | 45.00 | 54.68 |
| Token + Bigram | 24.11 | 12.41 | 22.95 | 100.13 | 50.85 | 68.51 |
| Token + Bigram + Twitter Elements | 25.14 | 23.77 | 28.82 | 60.19 | 50.77 | 61.11 |
| Token + POS (occurrence) | 15.25 | 13.36 | 28.16 | 64.27 | 37.87 | 49.56 |
| Token + POS (relative frequency) | 33.73 | 12.55 | 48.51 | 68.97 | 45.62 | 50.63 |
| Token + POS (median threshold) | 20.23 | 15.04 | 33.02 | 76.87 | 33.90 | 39.17 |
| Token + POS (occurrence and median thresholds) | 15.72 | 17.08 | 38.11 | 60.94 | 39.96 | 39.11 |

To reach a reliable conclusion on the best pairs of feature set and algorithm for training a sentiment classifier, considering both the average accuracy and the DMA is essential. The best combinations could be achieved by adding POS in the feature set with discretized values for models trained by the NB algorithm, or with continuous values for models trained by the SVM algorithm.

## 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. Data about projects from transportation sector were collected, processed, and used in different combinations for generating a sentiment 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 different classifiers were examined to reach the highest level of accuracy in classification. The results show an acceptable – yet not impressive – level of accuracy by selecting SVM as the learning algorithm for the classifier along with commonly used features such as vocabulary and part-of-speech. In our data-set, SVM dominates NB, DT, and K-NN in terms of accuracy of the results and stability of the accuracy over various sustainability subjects. Further investigations using other metrics such as precision and recall could illuminate the other key differences between the classifiers developed in this study.

Pure dependency on machine coding has been criticized by former experiences due to the confusions caused by word-phrase associations and hyponyms, and because the AI systems fails to consider the 'context' in the classification task (Macnamara, 2005). However, a classification system similar to the one presented in this paper can be a good starting point to reduce the required cost and effort of human coding, and/or to control its outputs.

The work presented here can be combined with the work presented by Nik-Bakht et al. (2015) to develop an automated classifier capable of detecting both the subjects of tweets as the vested interests of the project followers, and the sentiment of the tweets as the position of project followers with respect to the project. Therefore, the combined semantic and sentiment classifier could result in a full stakeholder mapping in terms of the opinions surrounding an infrastructure project. 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. Another opportunity to extend this work is to develop a classifier which is capable of detecting mixed classification both at the semantic level and at the sentiment level.

## References


Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*, pp. 30-38.

Barbosa, L., & Feng, J. (2010, August). Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 36-44.

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence, 97*(1), pp. 245-271.

Collins, C., Hasan, S., & Ukkusuri, S. V. (2013). A novel transit rider satisfaction metric: Rider sentiments measured from online social media data. *Journal of Public Transportation*, *16*(2), pp. 21-45.

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.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research, 3(Mar), pp. 1289-1305.

Go, A., Huang, L., & Bhayani, R. (2009). Twitter sentiment analysis. Stanford NLP group.

Grant-Muller, S. M., Gal-Tzur, A., Minkov, E., Nocera, S., Kuflik, T., & Shoor, I. (2014). Enhancing transport data collection through social media sources: methods, challenges and opportunities for textual data. *IET Intelligent Transport Systems*, *9*(4), pp. 407-417.

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

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

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). Cambridge: MIT press.

Nik-Bakht, M., & El-Diraby, T. E. (2016). Sus-tweet-ability: Exposing public community' s perspective on sustainability of urban infrastructure through online social media. *International Journal of Human-Computer Studies*, *89*, pp. 54-72.

Nik-Bakht, M., Hosseini, M., & El-Diraby, T. E. (2015). Sustweetability: Infrastructure Sustainability-Related Topic Classification in Social Media. *Proceedings of the 32nd CIB W78 Conference 2015*, 27th-29th 2015, Eindhoven, The Netherlands.

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

Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the 7th conference on international language resources and evaluation LER*, 10, pp. 1320-1326.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *Proceedings of the ACL Student Research Workshop*, Ann Arbor, Michigan, pp. 43-48

Saif, H., He, Y., & Alani, H. (2012, November). Semantic sentiment analysis of twitter. In *International Semantic Web Conference*, pp. 508-524.