Detection of Traffic Events from Finnish Social Media Data

Hang Do Minh

University of Tampere
School of Information Sciences
Computer Science / Soft. Development
M.Sc. thesis
Supervisor: Jyrki Nummenmaa
November 2016

University of Tampere
School of Information Sciences
Computer Science / Software Development
Hang Do Minh: Detection of Traffic Events from Finnish Social Media Data
M.Sc. thesis, 48 pages, 5 index and appendix pages
November 2016

Social media has gained significant popularity and importance during the past few years and has become an essential part of many people's everyday lives. As social media users write about a broad range of topics, popular social networking sites can serve as a perfect base for various data mining and information extraction applications. One possibility among these could be the real-time detection of unexpected traffic events or anomalies, which could be used to help traffic managers to discover and mitigate problematic spots in a timely manner or to assist passengers with making informed decisions about their travel route.

The purpose of this study is to develop a Finnish traffic information system that relies on social media data. The potential of using social network streams in traffic information extraction has been demonstrated in several big cities, but no study has so far investigated the possible use in smaller communities such as towns in Finland. The complexity of Finnish language also poses further challenges. The aim of the research is to investigate what methods would be the most suitable to analyse and extract information from Finnish social media messages and to incorporate these into the implementation of a practical application.

In order to determine the most effective methods for the purposes of this study, an extensive literature research was performed in the fields of social media mining and textual and linguistic analysis with a special focus on frameworks and methods designed for Finnish language. In addition, a website and a mobile application were developed for data collection, analysis and demonstration.

The implemented traffic event detection system is able to detect and classify incidents from the public Twitter stream. Tests of the analysis methods have determined high accuracy both in terms of textual and cluster analysis. Although certain limitations and possible improvements should be considered in the future, the ready traffic information system has already demonstrated satisfactory performance and lay the foundation for further studies and research.

Key words and terms: traffic data, social media, Twitter, Finnish grammatical framework, text classification

Contents

List of figures

List of tables

1. Introduction

In the past years, social media applications have gained immense popularity and have become an essential part of many people's everyday life. These applications have not only revolutionized human interactions, but have also collected a vast amount of information provided by users from all over the world, which makes them suitable for all types of data research. Numerous studies have examined the potential of social media to measure public opinion ([*1*], [*2*], [*3*]) or to detect events and phenomena ([*4*], [*5*], [*6*]).

Awareness of the current state of traffic has always been an important issue in transportation. By gaining immediate knowledge about events and problems, traffic operators can optimize traffic management and minimize the risk of congestions, thus improving the overall traffic flow and transportation services provided to passengers. Sensors and cameras could provide a possible solution for continuous traffic monitoring, but their instalment all over the city would be too costly and sensors might also produce imprecise and unreliable data at times. Therefore, other alternative sources of real-time information need to be considered and social media can be one option.

The potential of social media as a source of real-time traffic information has already been examined in several studies. Mai and Hranac [*7*], for example, studied the correlation between the occurrence of traffic incidents and the number of social media messages related to roadway events in the same area and determined a strong connection, which implies that social media should be considered as a new alternative source of traffic related information. Kosala et al. [*8*] went further and developed a system that extracts, interprets and visualizes publicly available Twitter posts containing traffic data from the city of Jakarta. Similar systems have been implemented in other big cities too: Wanichayapong et al. [*9*] developed a method to extract and classify traffic information from tweets submitted from the Bangkok area and Steiger et al. [*10*] attempted to model the flow of public transportation in London based on data extracted from various social media channels such as Twitter, Instagram, Foursquare and Flickr.

However, all of the research concentrates on cities with a significantly large population and also more intensive social media activity. No research has so far attempted to implement similar systems and analyses in smaller localities such as Finnish towns. Compared with the cities mentioned in the previous studies, cities in Finland have a noticeably smaller population of usually no more than a few hundred thousand inhabitants. As smaller population often also results in less intensive social media activity, attempting to develop a social media based real-time data extraction system can introduce a whole range of new challenges. Moreover, the Finnish language is also significantly more complex than many of the languages used in previous projects, which makes the processing of Finnish texts more difficult as well. However, although these difficulties might seem discouraging, implementing a social media based real-time traffic event detection system is not necessarily impossible. The aim of this research is to create such a system for the city of Tampere.

1.1. Research goals

As mentioned above, the main goal of this research is to develop a system which is able to extract traffic related information in real-time from social media messages written in Finnish. This requires a thorough study of data extraction and text analysis methods in order to determine which approaches would be the most appropriate for this project. As the system should be able to group data referring to the same event correctly, some research of clustering methods is needed as well.

Therefore, the research questions this thesis attempts to answer are the following:

- How can important events be detected in social media?
- What is the most effective method to process Finnish texts?
- How can detected events be correctly classified?
- What is the most effective method to cluster data referring to the same event?

1.2. Structure of the thesis

This thesis consists of four more chapters. In the next chapter, an extensive literature review will be presented in order to describe previous research and existing solutions to the reader. The third chapter will introduce the system developed as part of this research as well as the approaches and methodologies used to solve the research problem. The fourth chapter will present and evaluate the results of the study and finally, the fifth chapter, being the last, will summarize the main conclusions and suggest ideas for future development.

2. Background

The problems of gaining real-time traffic information in order to assist urban transportation management and data mining of social media streams for various purposes have inspired numerous studies and projects. This chapter discusses some of the most interesting research in these fields as well as presents some already existing traffic information systems used in other countries. As knowledge of textual analysis and clustering methods is also crucial for this research, they are going to be described here as well.

2.1. Extracting information from social media

The use of social networking and micro-blogging platforms has become a natural part of many people’s lives. Every day a vast number of posts and other forms of user interaction are created on these sites. This huge mass of data available through social media streams offers the possibility to perform various analyses and data-mining procedures.

2.1.1. Opinion mining and sentiment analysis

Social media data can provide information of all sorts. As users share various details about themselves over social networking and micro-blogging sites, including opinions about certain products, topics, events or political parties, social media platforms are excellent corpora for opinion mining. The results of such analyses can be highly useful and valuable for marketing or political purposes, for example.

There are numerous studies discussing opinion mining and sentiment analysis in social media. Pak and Paroubek [*11*], for example, built a sentiment classifier which can automatically collect a corpus of textual posts used for training from Twitter. Texts containing positive and negative emoticons and neutral posts published by popular newspapers and magazines were collected for the corpus and linguistic analysis was performed on them in order to create a training set. This method yields a satisfactorily high accuracy if a sufficiently large dataset is provided. However, the rather simplified method used for labelling training data (determining the sentiment of the text based on the presence of one emoticon) often fails to recognize sarcasm and it can result in a significantly lower accuracy in classification.

Khan et al. [*12*] attempted to solve this problem. They used a hybrid classification scheme combining three different classifiers (Enhanced Emoticon Classifier, Improved Polarity Classifier and SentiWordNet Classifier) in order to increase accuracy. With this classification method and a detailed pre-processing of acquired data (including stop words removal, spell checking and correction, lemmatisation and expansion of abbreviations and slang words) a high average accuracy (85.7%) can be achieved.

Another approach towards enhancing sentiment classification accuracy was proposed by Kontopoulos et al. [*13*]. Their research studied the deployment of original ontology-based methods and the effect of grading each distinct notion of a post instead of simply characterizing texts with a single sentiment score. The study found that this architecture could achieve a higher accuracy compared to other existing methods when analysing opinions regarding a specific topic.

As social media platforms host all sorts of content, the possibilities of social mining are nearly endless. When using social media sites such as Twitter, for example, for opinion mining, the most prevalent topics seem to be political sentiment and brand perception. Tumasjan et al. [*1*] examined political debates and sentiments towards German politicians in Twitter and attempted to predict the results of the German elections by analysing political tweets. Even with simplified collection and textual analysis methods, their research was able to produce more accurate results than those of traditional election polls.

Younus et al. [*2*] also studied the potential of political opinion mining in Twitter: they analysed tweets about major political events in the developing world and proposed a new model for public opinion analysis based on social features instead of traditional text analysis methods. Experimental evaluations demonstrated that with this new approach it is possible to achieve a high level of accuracy.

While the previously mentioned studies concentrated on the political application of opinion mining, Mostafa [*3*] focused on consumer brand sentiment analysis. He analysed Twitter posts mentioning major global brands such as Nokia, T-Mobile, IBM, KLM and DHL in order to determine consumer sentiments towards these brands. The analysis used lexicon-based classification combined with quantitative and qualitative methodology. The research highlighted the potential of social media mining for marketing purposes and the importance of social media presence of companies. However, it also had a few serious limitations. For example, the analyser was able to determine simple sentiments only and not the underlying reason for the sentiment. In addition, it was also unable to cope with sarcasm or other linguistic disambiguities.

Ghiassi et al. [***14***] also analysed brand sentiments through Twitter in their research. They developed a Twitter specific sentiment lexicon that reduced problem complexity (compared to traditional lexicons, it contained fewer entries, but it was still able to cover over 90% of the brand specific corpus) and increased classification accuracy. The lexicon was used to perform DAN2 sentiment analysis, which yielded an exceptionally high level of accuracy.

2.1.2. Detecting events and phenomena

Apart from public opinion analysis, social media platforms also offer the possibility to observe unfolding events and phenomena in real-time. As millions of people use social media actively to share various events in their lives with others, major events and phenomena are likely to receive wider social media coverage. Therefore, social networking and micro-blogging sites offer good resources for detecting and analysing certain events.

2.1.2.1 Social event detection systems

Several studies have examined the possibility of event detection in social media and proposed systems and methods for extracting event information from social streams. Weng et al. [*4*], for example, constructed an event detection system to identify events from the Twitter stream by applying wavelet analysis. The system computes signals for individual words by using wavelet analysis to represent the frequency of the word's appearance. Afterwards the cross correlation between signals is measured and events are detected by clustering signals using modularity-based graph partitioning. The main concept behind this approach is that certain related words are commonly used to describe certain types of events. Therefore, an important event with wide social media coverage would significantly increase the usage of these words, generating a noticeable burst in their appearance. These bursts can be easily represented by wavelet-based signals; thus, traditional wavelet analysis can be used for event detection. This system has demonstrated a fairly good performance, which shows the potential of this approach; however, it also has its shortcomings. For example, it treats every word individually and therefore in some cases might falsely group words associated with different events together. It also fails to consider social features and factors, which could improve the reliability and accuracy of the system.

Watanabe et al. [***15***] attempted to build a real-time local event detection system by developing an automatic geo-tagger to locate non geo-tagged tweets. The geo-tagger relies on a place name database that is built partly by using geo-tagged posts in order to assign a location to tweets mentioning bigger events or well-known places. This method enables identifying more messages referring to local events even if they are not associated with GPS coordinates. Therefore, it increases detection accuracy. However, this approach is only able to locate posts with explicit mentions of easily identifiable locations or events, which is only a relatively small portion of tweets, so other possible ways of extracting location (e.g. analysing history of the user's movements or textual context) need to be further researched.

Li[1] et al. [***16***] proposed a Twitter-based event detection and analysis system which collects tweets referring to crime or disaster related events, affixes them with geographic location and ranks them using several social features such as the author's credibility, number of retweets and certain characteristics of the content (e.g. presence of important words and the format of the post). This approach allowed for the analysis of spatiotemporal patterns of events and the detection of various incidents with a satisfactory efficiency. The novelty of the proposed method compared with earlier works is the consideration of social features, which helps to improve reliability.

Thom et al. [***17***] developed a workbench for spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages. The workbench uses the Streaming API service of Twitter to collect tweets, which are passed on to the spatiotemporal anomaly detector after tokenization and removal of web links and stop words. The detector continuously monitors term usage and searches for anomalies in order to identify potentially relevant events. The analyser then performs cluster analysis on the extracted term artefacts in order to aggregate data referring to the same event and the results of the clustering are used to generate a graphical representation of anomalies. This proposed method showed a good efficiency in detecting bigger events and natural disasters. However, the limited accessibility of the Streaming API (only 1% of the total public stream is available through this API) and the overall scarcity of geolocated messages often pose the risk of completely ignoring smaller scale events. As this thesis deals mainly with relatively small-scale events, this method is not perfectly suitable for the research objectives. However, certain elements, especially the event clustering solutions, should still be considered.

Li^2 et al. [***18***] proposed a segment-based event detection system for Twitter. The novelty of this approach lies in analysing tweet segments instead of unigrams, which results in improved performance compared to other Twitter-based event detection systems [***18***]. To tackle the noisiness of the data, realistic events are distinguished from non-relevant babblings by scanning Wikipedia for the same events [***18***]. This method enhances the reliability and informativeness of the system; however, in case of an actually important event not being covered by Wikipedia or other online information sources (which is often the case with traffic events in Finland, for example) the efficiency is questionable.

McCreadie et al. focused on the problem of real-time detection and scalability in their research [***19***]. As popular social media streams receive millions of messages every second, processing the complete or a larger part of the stream is computationally costly. Therefore, social media messages cannot be processed on a single machine rapidly enough, which makes real-time event detection almost impossible. In big data analysis, partitioning is a common solution to alleviate computational costs. However, using traditional partitioning methods can result in spreading events across multiple partitions, which might generate problems such as identifying the same event multiple times or on the contrary, completely ignoring certain events. McCreadie et al. [***19***] attempted to tackle this problem by proposing a new approach for automatic distributed real-time detection by using a novel lexical key partitioning strategy, which can distribute the computational costs across multiple machines without partitioning the document itself. This approach can help in reducing response latency, but it preserves efficiency. Experimental results demonstrated that the proposed solution was able to scale up to large streams of data (e.g. the complete Twitter Firehose stream) without degrading event detection efficiency.

Walther and Kaisser [***6***] concentrated on the detection of local small-scale events. Such events can be, for example, house fires, robberies, accidents etc. The study proposed a novel algorithm which preselects tweets based on their geographical and temporal proximity (the original data set being already limited a pre-specified geographic area – in the case study it was New York City), then applies machine-learning mechanisms to further evaluate candidate event clusters. In order to separate real-life events from irrelevant “babbles”, the machine-learning scheme analyses 41 different features [*6*]. The identified events are shown to the user on a map GUI. The proposed system proved to be able to detect whether a cluster describes a real event or not with high reliability. However, the recall of the system in terms of actual real-life events (in other words: how many of all the events happening in the specified area can be detected with the system?) is yet to be determined.

Although there are numerous studies discussing event detection in social media (especially Twitter), only a few have examined the impact of analysing social features besides the actual text contents. Guille and Favre [*5*] aimed to target this problem and highlighted the utility and importance of social factors. Their research concentrated on the significance of mentions and their analysis in order to measure the impact of tweets over the crowd. Another novel element of their proposed approach is that it relies solely on statistical data extracted from tweets instead of including external resources, which means that the method can operate solely on Twitter data without the need of any form of external knowledge [*5*]. Unlike most other existing methods, it also estimates the period of time during which events occurred dynamically rather than assuming a predefined fixed duration, thus it provides more accurate information. During experimental evaluation the proposed method proved to be highly efficient, which demonstrates the relevance of this approach.

2.1.2.2 Harvesting traffic information from social media

Social media streams can provide all sorts of information and therefore the list of possible application areas of social network based event detection systems is practically endless. One of these possibilities is the area of transportation and the identification of traffic events and other irregularities.

Numerous studies exist which discuss the possible solutions for harvesting real-time traffic information from social media streams. Wanichayapong et al. [*9*], for example, developed a method to extract and classify traffic information appearing on Twitter. This method first extracts tweets possibly describing traffic conditions by searching for traffic related words, then performs syntactical analysis on the selected posts and categorizes them with the aid of specific wordlists [*9*]. This approach demonstrated high accuracy in the experimental setting in Bangkok. However, the rather simplistic approach of the syntactical analysis (classifying tweets based on the presence of certain prepositions and predefined words) would prove to be significantly less effective with more complex languages than Thai – such as Finnish, for example. Also, the non-exhaustive nature of the predefined word lists poses the risk of ignoring actually relevant messages and thus degrading efficiency.

Instead of turning to the general crowd for detecting traffic events, Endarnoto et al. [*20*] concentrated on extracting traffic information from a well-respected authority's

Twitter stream. The primary aim of the research was to automatically identify various traffic related events and provide an improved interface for presenting the acquired data. Traffic information was extracted from the collected tweets by parsing Indonesian sentences using context-free grammar and a predefined set of rules and vocabulary [***20***]. The extracted information was then visualized in a mobile application by using a map view. The extraction method demonstrated a fair efficiency, however, out of rule and out of vocabulary problems posed a significant risk of ignoring important messages and these problems remained mainly unsolved by the authors. Also, concentrating on one official source only deprives us from the main benefit of social media based information systems: access to a vast number of "social sensors", which facilitates real-time and exhaustive event detection. When relying on one sole source, the question always arises whether all the important events are reported and in a timely manner.

In another study, researchers Kosala et al. [***8***] also used preselected Twitter accounts for extracting traffic information. Their research differs from the previous one in the sense that it did not concentrate on only one account, but collected traffic information scattered over several channels. It also abandoned the use of natural language processing methods, which decision is supported by the fact that most Twitter posts use ungrammatical language [***8***]. Therefore the information extraction is mainly based on keyword and lexicon-based traffic information analysis [***8***]. The proposed method was able to detect traffic events described in the collected tweets with a high accuracy, however, relying solely on a few pre-specified accounts can result in a significant degradation of efficiency when certain traffic events are not reported by any of these users.

In a research performed by IBM [***21***], social media was used to confirm information forwarded by sensors. As sensor data usually contains a high amount of noise, it is important to distinguish relevant information from irrelevant in order to facilitate the work of traffic operators and to enable them to mitigate crisis situations in a timely manner [***21***]. The method proposed by this study collects traffic related tweets from authorative sources on Twitter and analyses them in order to detect anomalies. The anomaly detection method relies on statistical change detection (using an algorithm similar to CUSUM adapted to Markov chains) as well as information provided by experts on Twitter [***21***]. A spatial scope of the incident is also defined by examining the GPS coordinates assigned to tweets. When an anomaly is detected, the information is compared to the sensor data received from sensors in the geographic vicinity of the extracted location of the identified event [***21***]. Experimental evaluations demonstrated that combining sensor and social media increased event detection accuracy. However, the same problem as in the previous research remains: because of relying on a low number of sources, some events might remain unreported and therefore undetected.

A rather interesting research is that of Mai and Hranac [*7*] who studied the correlation between Twitter usage and officially reported transportation incidents in order to determine the efficiency of Twitter as a data source for traffic event detection. They compared official incident data obtained from the California Highway Patrol to tweets collected through the Streaming API [*7*]. After applying volume and semantic analysis, the research found significant correlations between real incidents and Twitter usage and content, which highlight the potential of Twitter as a corpus for event detection [*7*]. However, the analysis used in the research is quite unsophisticated, it uses a rather broad spatial and temporal scope and it fails to examine mentions of specific freeways in the text or the most direct driving route between the extracted location and a potential incident match in order to verify the realisticity of the match.

Daly et al. [***22***] attempted to harvest traffic data from Twitter in order to find relevant information and possible explanations for congestions. They developed a system called Dub-STAR which combines official data from city authorities and crowd-sourced information extracted from social media [***22***]. The system uses historical observations, semantic analysis and natural language processing in order to match events with congestion alerts [***22***]. Experimental evaluation demonstrated a fair accuracy, which shows the relevance of this approach. However, one shortcoming of this solution is that it allows the matching of only one location with a single message, whereas it was noted that many of the received messages mentioned several locations [***22***].

Steiger et al. [***10***] attempted to model the public transport flow of London based on information extracted from several social media channels such as Twitter, Foursquare, Instagram and Twitter. The novelty of this research is that it uses heterogeneous data: in addition to textual messages, images were also collected and analysed. The study attempted to infer human mobility and public transport flow from social media streams by applying semantic pre-processing, LDA topic modelling, spatial clustering and station matching on the extracted dataset [***10***]. In the experimental use case the approach demonstrated a fairly high accuracy, however, it has its limitations. For example, it automatically assumes that social media post are direct indicators of public transport usage, whereas this question needs to be further investigated and researched. The analysed dataset is also seriously limited: only georeferenced Twitter posts are examined, although most messages do not have coordinates associated with them, and only those Foursquare check-ins are collected which are posted through Twitter. In the case of images, only the caption is analysed, not the image itself, whereas the picture might carry more information than the accompanying text (if exists). These limitations might not cause a significant degradation in efficiency in case of areas with large population and a great amount of social media content generated (such as London), but they probably would prove to be problematic in smaller settings.

2.2. Analysis of textual messages

Knowledge and implementation of text analysis methods is crucial in order to extract information from textual social media posts such as tweets. For example, in order to determine what event or phenomena these messages describe, the application of certain text classification methods is needed. The briefness of social media messages, especially in the case of Twitter (which has a character limit of 140), and the complexity of Finnish language also necessitate further grammatical and morphological analyses. As Finnish language has a linguistic structure that greatly differs from many other widely spoken languages, in order to achieve maximum efficiency, it is important to examine grammatical and morphological analysis frameworks specifically designed for Finnish.

2.2.1. Classification of tweets

The classification of tweets into different categories can be formalized as a text classification problem. However, the shortness (maximum 140 characters) and the unrestricted linguistic structure of Twitter messages pose certain challenges when working with traditional text analysis methods. For example, the bag-of-words model, which is commonly used in document classification problems, has been shown to be significantly less efficient in the case of tweets [*23*]. Therefore it is important to examine classification methods that consider the special characteristics of short social media messages.

The problem of classifying tweets has already arisen in several studies. Sriram et al. [*23*], for example, attempted to provide a more personalized feed by assigning categories such as news, events, opinions, deals and private messages to incoming tweets. Instead of using traditional methods, the authors developed their own intuitive approach, which relies on a small set of discriminative features [*23*]. This method significantly outperformed the bag-of-words model; however, this approach relies on some rather simplistic presumptions (e.g. that the presence of "@username" always implies private message) and the non-incremental nature of features and categories might also decrease accuracy.

In another interesting study Nishida et al. [*24*] investigated the possible use of data compression in relation with identifying new tweets that are relevant to a certain topic. Their method used a simple learning algorithm relying on a set of positive and negative examples in order to evaluate the compressibility and relevance of a tweet: if the compression score achieved with the positive examples is higher than that with negative examples, the tweet can be considered interesting in terms of the given topic [*24*]. In experimental evaluation this method has outperformed confidence-weighted linear classification and online passive-aggressive algorithms (two popular machine learning approaches), which demonstrates the potential of this approach.

Although several studies implied that certain traditional text classification methods perform poorly with short social media messages due to their special characteristics that cannot be found in conventional text documents, it does not mean that all traditional methods are ineffective when analysing tweets. A good example for this is the study performed by Batool et al. [*25*], where the researchers applied keyword-based knowledge extraction methods combined with a knowledge enhancer and synonym binder in order to classify tweets containing information of different categories such as e.g. diabetes, dengue, food, movies, education etc. and also determine the sentiments present in them. Keywords were extracted from tweets using machine learning methods provided by the publicly available Alchemy API, a deep-learning cloud platform developed by the company IBM [*26*]. Possible omissions of important details were also compensated by a using a knowledge enhancer, which was mainly responsible for entity extraction and part-of-speech tagging, and a synonym binder, which increased detection rate and accuracy [*25*]. Experimental evaluation showed that the use of knowledge enhancer resulted in a 55% increase in accuracy, achieving a rate of 89 % overall, with the missing 11% of tweets not being detected only due to spelling mistakes, which most classifiers fail to deal with as it is a difficult problem to solve. These results demonstrate the significant benefits of utilizing knowledge enhancer methods in short text classification.

A really interesting study performed by Schulz and Janssen [*27*] suggested the use of semantic abstraction to solve certain tweet classification and generalisation problems. The authors pointed out that regional differences were often ignored in previous research attempting to generalise certain classification and information extraction systems, even though they can have a significant impact on detection accuracy. Their research investigated the problem of traffic incident detection in different cities – a problem very relevant to the topic of this thesis – and proposed a generalised model consisting of a set of abstract features created by combining different feature groups such as Linked Open Data (LOD), location and temporal mention groups [*27*]. Evaluations showed that combining different methods in semantic abstraction improved classification accuracy even when used on only one set of data collected from the same city [*27*]. This approach performed well also when analysing datasets from different locations, which proves its utility in generalisation.

2.2.2. Grammatical analysis frameworks for Finnish language

As Finnish is a complex and not actually widely spoken language, linguistic studies and analysers developed for Finnish are rather scarce. However, it is still possible to find

Finnish grammatical or morphological analysis frameworks that provide high efficiency and accuracy even with such a highly complex language.

One of the Finnish morphological analysers worth mentioning is Omorfi [*28*], which is an open source project based on the Helsinki Finite-State Transducer Technology (HFST) developed at the University of Helsinki [*29*]. The analyser supports various natural language processing applications and it has a modular structure in order to facilitate extension and modification [*28*]. It also combines lexical data acquired from different sources, which ensures wide lexical and morphological coverage.

Another interesting project is the Grammatical Framework Resource Library for Finnish [*30*]. Grammatical Framework (GF) is a functional language used to define grammars of natural languages [*31*]. Its particularity lies in the fact that it separates abstract grammatical formalisms from their concrete language-dependent implementation and thus facilitates development of multilingual applications. Although GF grammars are purely declarative, they can be easily used in different natural language processing applications such as morphological analysers, spell checkers or translators. The GF Resource Library is a collection of grammars packaged as software libraries and at the moment covers 45 languages including Finnish. The Finnish Resource Library is one of the most complete grammars in the set and its extensive grammatical and morphological coverage makes it perfectly suitable to use in analysers. Compared to simple morphological frameworks, the Grammatical Framework allows for more complex and thorough analyses as it can also recognize the grammatical structure of texts and the relationships between the different parts and elements. It is also easy to integrate into different types of applications: it provides APIs for mainstream programming languages such as Java, Python and Haskell and it can be also parsed to use in JavaScript based web applications.

2.3. Clustering data entries

Familiarity with clustering methods is important in order to be able to correctly group tweets referring to the same event together. Clustering is a rather wide concept and it involves a broad range of different unsupervised methods that partition data into groups based on similarity in order to uncover its inherent structure [*32*]. This section presents some of the most common clustering techniques as well as several examples of using clustering in traffic data analysis.

2.3.1. Cluster analysis methods

As clustering algorithms are very diverse, they can be classified in several ways. One of the most common distinctions, however, is whether they perform hard or fuzzy clustering. Hard clustering means that each object belongs to only one group whereas in fuzzy clustering objects have a degree of membership associated with each group [***32***]. Other distinctions based on the types of partitions created are also possible, for example, hierarchical clustering techniques create a nested series of partitions (e.g. general categories can be divided into more concrete subcategories) whereas partitional techniques produce only one level of partitions [***32***]. There are also other classifications of clustering approaches based on different characteristics such as algorithmic structure, use of features, incrementality etc. For example, clustering algorithms can be agglomerative or divisive: agglomerative approaches begin with each object belonging to a distinct cluster and successively merge similar clusters together whereas divisive approaches start with all objects being grouped into one single cluster and perform splitting based on dissimilarity [***32***].

There is a huge variation in use of clustering methods among applications, as different needs often require different approaches and the vast family of clustering techniques offers a large and very diverse pool to choose from. However, it seems that some of the most popular algorithms are K-means, mean shift and hierarchical clustering and certain density based methods (e.g. DBSCAN).

The k-means clustering is a partitioning-based algorithm [***33***]. It relies on the iterative relocation of data points between clusters and divides the data set into non-overlapping groups [***33***]. In this approach, clusters are described by their mean vectors; therefore, it can also be said that this technique follows a centroid model [***33***]. In mathematical terms the k-means methods can be described as an approximation of a normal mixture model with an estimation of the mixtures by maximum likelyhoods, where all the mixture components (clusters) are assumed to have spherical covariance matrices and equal sampling probabilities [***33***]. K-means algorithms are actually easy to implement and this fact, together with their good computational efficiency and low memory consumption, has made these methods immensely popular [***33***]. Throughout time numerous variations of the k-means approach have been developed. For example, the Jenks optimization method [***34***] is a one-dimensional version of k-means clustering, where the main idea is the minimization of variance within groups and the maximization of variance between groups. Another interesting variation is the spherical k-means algorithm, which uses cosine similarity in order to minimize the mean-squared error [***35***]. Other well-known variations are, for example, the k-medians clustering [***36***], the X-means clustering [***37***] or the Minkowski-weighted k-means [***38***].

Mean shift is a nonparametric clustering method [***39***]. Unlike k-means clustering, it does not require prior knowledge of the number of clusters nor does it constrain their shape [***39***]. Therefore this method is suitable for a wide range of data analysis applications even where there is little or no information available about the data set. The main idea behind mean shift clustering is the use of kernel density estimation in order to locate local maxima (or modes) of the data distribution and to identify clusters

[*40*]. Mean shift is a robust technique, however, it generates excessively high computational costs as the space dimension increases, therefore it cannot be applied on high dimensional data sets without alterations [*39*].

Hierarchical clustering is a flexible non-parametric method, which requires very little a priori knowledge or constraints over the data set [*41*]. As the name suggests, it builds a hierarchy of clusters, where partitions are nested into each other. The cluster hierarchy is produced by recursive partitioning of the instances in the data set, which can be performed in an agglomerative or divisive manner [*42*]. The nested grouping created by the algorithm is represented by a dendrogram, which can be cut at different similarity level to obtain different clusterings [*32*]. There are many different algorithms that implement hierarchical clustering; however, it can be observed that most of them are some sort of variants of the single-link, complete-link or minimum variance algorithms [*32*]. These approaches differ in the way they measure similarity distance between clusters: single-link methods assume that the distance between two clusters equals to the shortest distance between all pairs of objects taken from the two clusters, complete-link methods suppose it to be the longest pairwise distance and minimum variance (also known as average-link) methods consider the average distance between members of the two clusters to be the distance of the clusters [*32*]. Hierarchical clustering methods have become largely popular due to their versatility; however, they also have certain weaknesses: for example, they do not scale well with larger data sets as their time complexity is not linear, but polynomial (at least $O(m^2)$, where m is total number of instances) and they have no back-tracking capabilities [*42*].

Density based methods treat the data space as a set of spatially diffused points and identify dense point regions as clusters. This approach is especially well suited to handle spatial data with arbitrary cluster shapes [*43*]. It also has a good efficiency on large databases and requires minimal domain knowledge to determine the input parameters [*43*]. As opposed to approaches such as k-means clustering, which partitions every instance of the data set into clusters, it also has the capability to recognize data points that do not actually belong to any clusters, so it is able to deal with noisy data. One of the most widely known and commonly used density-based clustering algorithms is DBSCAN, which was introduced by Ester et al. in 1996 [*43*]. It identifies dense regions by classifying data points into core points, density reachable points and outliers with the help of two parameters: Ɛ (the neighbourhood radius) and MinPts (the minimum required number of points to form a dense region). A core point can directly reach at least MinPts within Ɛ distance and a point is density reachable from a core point if there is a path between them that consists of core points only and it forms all points of the path are directly reachable from the point preceding them [*43*]. Outliers are points that do not belong to any of the clusters (cannot be reached from other points) and therefore are considered as noise [*43*]. The DBSCAN algorithm is able to perform with good efficiency on large and noisy spatial databases; however, it also has a few shortcomings: for example, it can only provide a flat partitioning of data objects based on a single global threshold, which poses the risk of inaccurate characterization of data sets with very diverse density or nested clusters [*44*].

2.3.2. Clustering in traffic data analysis

Clustering is a widely used technique in many fields. Traffic data analysis is one of the areas where clustering can be helpful. Previous research has demonstrated that clustering algorithms can be used effectively to automatically detect freeway incidents [*45*], monitor road traffic state [*46*] or to disseminate congestion [*47*], for example. This section presents some of the most interesting projects.

Automatic incident detection and characterization is an important issue in traffic control as it enables timely reaction and problem resolution (e.g. dissemination or prevention of congestion). Numerous studies have proposed possible solutions for this issue, all of them suggesting different approaches. One of these approaches is the use of clustering techniques as it can be seen in the study presented by Sheu [*45*], for example. He developed a method based on fuzzy clustering theories in order to identify and characterize freeway incidents and traffic conditions associated with them [*45*]. The algorithm uses raw traffic data obtained from upstream and downstream detectors in order to identify traffic conditions [*45*]. It detects incidents based on predefined decision variables and irregular traffic patterns and then determines their location and end time [*45*]. Off-line evaluation tests with simulation data yielded good results and showed that the method has the potential of detecting and characterizing incidents real-time.

Another interesting research performed by Sohn and Lee [*48*] investigated the efficacy of clustering techniques in road accident classification. The study examined data fusion, ensemble and clustering algorithms in order to assess how they can improve classification accuracy when grading the severity of road incidents [*48*]. The research applied k-means clustering to road traffic accident data and found that compared with the other methods it performed significantly better on data with large variation [*48*].

Jiang et al. [*46*] studied the possibility of dynamic road state identification by using fuzzy clustering approach. The aim of the research was to provide a solution for determining traffic state from limited sensor data. The analysis method suggested by the authors is based on fuzzy k-means clustering and, according to the test results, has the capability to convert microwave sensor data into traffic state concepts [*46*]. The case study demonstrated that information obtained this way might be highly useful for local traffic managers: however, it is important to highlight that the translation of quantitative data into traffic states is not always straightforward and there is no universal solution as people in different cities and countries may perceive traffic conditions differently.

2.4. Advanced Traveller Information Systems

The enhanced focus on real-time traffic event detection and its benefits for traffic operators and authorities as well as travellers resulted in the development of numerous advanced traveller information systems. By definition, an Advanced Traveller Information System (ATIS) is a system that "assists travellers with pre-tip and en route travel information to improve the convenience, safety and efficiency of travel" [***49***]. This definition leaves room for broad interpretation in regards to the actual implementation of an ATIS. It can manifest in the form of a website or mobile application, for example, and the data used by the system can also be obtained from different sources such as local authorities, official databases, news agents or the travellers themselves. The latter case is called crowd-sensing. This section will focus on ATIS using crowd-sensing.

2.4.1. Real-life examples

Nowadays advanced traveller information systems are being used in numerous places. Many of these systems rely on crowd-sourced data, which can be obtained by either using a dedicated platform (e.g. specific websites, communication channels or mobile applications) or by mining social media streams. The following part of this section is going to list some real-life examples of crowd-sensing-based advanced traveller information systems.

Waze is one of the most significant and largest community-based traffic and navigation applications in the world [***50***]. In 2013 it counted approximately fifty million users [***51***]. It is basically a map application, but the data collected from its users enables it to dynamically adapt to traffic conditions and to inform users about congestions or other incidents. Users can contribute to the map in two ways: firstly, the application collects sensor data from their devices in order to calculate their speed and mobility from which traffic conditions can be inferred, secondly, users can also report irregular events and these reports can be seen by other users as well. Waze has demonstrated high efficiency in several locations; however, as it relies solely on user-generated data, its performance in areas with a low number of users is questionable.

Roadify, a real-time transit information application used mainly in the United States of America, offers a common information interface combining data received from official authorities and users [***52***]. The application provides real-time arrival and schedule information about subways, buses, trains, ferries, bike sharing and car2go services, as well as maps with directions and service alerts from official agency sources and other travellers. Users contribute to the application by sharing their mobility data (coming from their device sensors), which helps to calculate real-time arrival information, and by submitting reports about transit problems. Compared to other applications, the main advantage of Roadify is that it provides a unified transit information interface, so travellers do not need to search for information regarding different transit services from different applications and portals. However, it needs a critical number of users in order to be the most efficient and it has not been achieved yet in several locations.

Tiramisu is a project started by the Carnegie Mellon University, which became a relatively widely used application in the Pittsburgh area [*53*]. The application provides real-time information about buses based on the transit service provider's own schedule and the GPS traces obtained from users [*54*]. It also offers users the possibility to report problems perceived during their travel, which does not only inform other users about issues on certain vehicles, but it also alerts local authorities of acute problems which need timely intervention or improvement, thus promoting co-design. The concept of co-design is what distinguishes Tiramisu from other transit information applications: it is not just a simple information interface like many other similar systems, but also a platform for travellers to voice their opinion about transit services, thus offering them the possibility to take part in service planning. However, the application faces the same problem as the previously mentioned systems: it needs a critical amount of users in order to be efficient and useful.

SMARTY was a smart transport project implemented in Tuscany, Italy [*55*]. It involved developing a unified transport service platform which provides real-time transit and traffic information based on sensor data (obtained from external sensors installed in different points of the city as well as user device sensors), user feedbacks, local municipal databases and social media mining [*56*]. In addition to traffic information and route recommendations, SMARTY also offers a wide range of services such as mobile payments (mainly for transport ticket purchases), parking spot reservation, bike sharing and carpooling. Therefore SMARTY is not a simple information system, but rather a complex platform, which can be considered as an intelligent transportation system.

Smart City is a web application developed as part of an initiative endorsed by the government of Jakarta [*57*]. It provides different types of information on Jakarta: in addition to traffic information and bus schedules, it also features weather information, danger alerts, programme recommendations (e.g. museum exhibitions, concerts etc.) and reviews of local businesses such as restaurants, bars etc. [*58*]. The data used by the site is obtained from different sources such as information provided by the government and city authorities, external sources (e.g. companies) and user reports collected from different applications such as Twitter, Google Maps, Waze and Qlue (a local mobile application for collecting problem reports from users). Because of the wide range of data sources used, a fair level of efficiency and accuracy is achievable even with a low number of users, thus the system can be considered more reliable than solely crowd-sensing-based information systems.

UbiBus is an on-going project, which aims to develop an Advanced Public Transportation System to support Brazilian passengers [*59*]. Part of the project was the development of several transit information applications: Your City On Time provides real-time arrival information about buses based on their location, speed, route and traffic condition information; UbibusRoute recommends routes to users based on the transit information extracted from Twitter and EPITrans collects traffic condition information from Facebook posts. Preliminary experiments produced promising results, however, it is difficult to make further deductions about the performance and efficiency of the proposed system as it is still under development.

2.5. Summary

This section has presented the technical background of the research topic as well as some real-life examples of existing systems, which were designed to solve similar problems. The approaches described above have their advantages and disadvantages and they demonstrate different levels of efficiency in different settings, which are all important to consider when designing the proposed solution. The selected methods and the underlying reasons for their choice will be explained in further detail in the next chapter.

3. Research methods

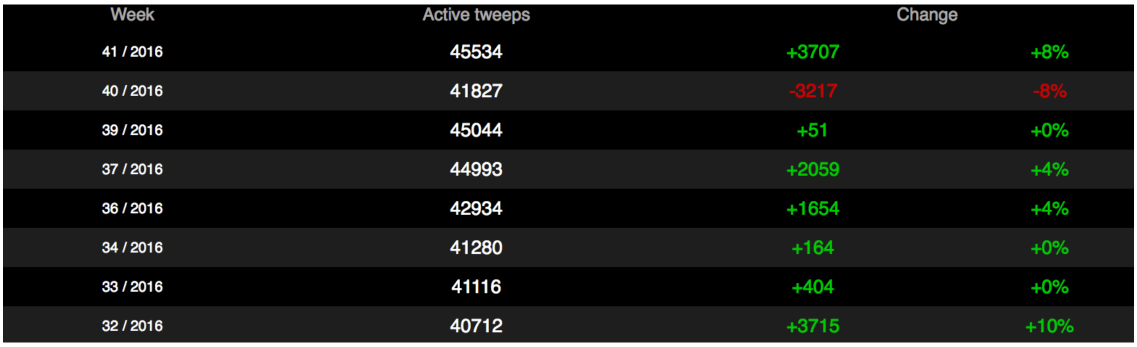
This chapter is going to describe the methods and approaches selected to solve the research problem as well as give a detailed presentation of the proposed traffic event detection and classification system.

Several factors had to be considered during the planning of the proposed system and analysis methods used in it. The complexity of Finnish language and the relatively small population of the city of Tampere were among the most important ones, which had a significant impact on the design. The design decisions made during this research are going to be explained in the next subsection.

3.1. Design decisions

Certain design decisions had to be made during this research in order to facilitate data analysis and to enhance efficiency.

One of the most important decisions was to limit the scope to social media data acquired from Twitter. The reason behind this restriction is that Twitter – unlike the other popular social media platforms – provides short and concise data entries by limiting the length of messages in 140 characters. Many of the posts submitted by its users are also public and therefore easily accessible, whereas content published on other social networking sites such as Facebook often has restricted visibility, which makes data mining notably more difficult. Moreover, Twitter has well-defined APIs to provide access to its public streams, which facilitates data collection significantly. The number of active Finnish users also shows some potential: according to SuomiTwitter (a website presenting statistical data of Finnish Twitter users), the number of such users is estimated to be around 438 000 [*60*]. These users demonstrate a relatively high activity, submitting an estimated average of 3000 tweets per hour [*60*]. These statistics show that Twitter can be a valuable data source for social media based research.



Week	Active tweeps	Change	
41 / 2016	45534	+3707	+8%
40 / 2016	41827	-3217	-8%
39 / 2016	45044	+51	+0%
37 / 2016	44993	+2059	+4%
36 / 2016	42934	+1654	+4%
34 / 2016	41280	+164	+0%
33 / 2016	41116	+404	+0%
32 / 2016	40712	+3715	+10%

Figure 3.1 Number of active Finnish Twitter users per week (Source: Suomi-Twitter, 2016)

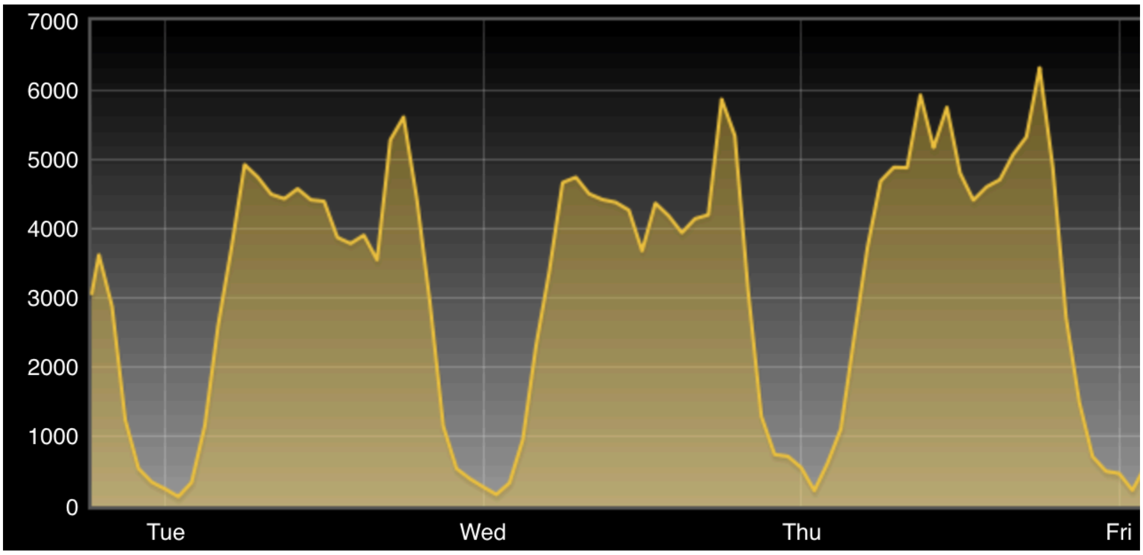


Figure 3.2 Finnish tweets per hour between 11-14th October, 2016 (Source: Suomi-Twitter, 2016)

Even though the number of Finnish Twitter users is relatively high in proportion to the size of the population, the amount of traffic related content produced by them is often quite low, especially when restricting the geographical scope to the Tampere area.

Therefore official sources should be included in the monitoring process as well. It was also mentioned by the research partner (the transportation department of the city of Tampere) that some of the unexpected traffic problems were caused by utility companies performing previously unannounced road works; therefore special attention should be paid to them too.

Encouraging Twitter users to report traffic related events is also an important task in order to increase the amount of obtainable data. In addition to active advertising, providing an easy reporting method can also positively impact user activity. Therefore it seems reasonable in the initial phase to develop a simple mobile application, through which it is extremely effortless to submit informative posts about traffic problems. The use of such application would also have the added benefit of adhering to pre-defined reporting formats and including geographic coordinates in tweets. As the application is Twitter-based, meaning that it uses Twitter accounts and it publishes user posts on Twitter, its introduction would not compromise the original concept of acquiring data solely from social media.

Although the main focus of the research is on the analysis of Finnish texts, inclusion of other languages can be considered too in order to expand the data pool. As Finland has an increasing number of non-Finnish-speaking immigrants, inclusion other languages can be a reasonable decision. It is not in the scope of this research to add support for all the foreign languages spoken by residents of Finland, however, as it would increase analysis complexity significantly with little added benefit. Including English at least seems to be a good choice though, as it is widely used and it has a relatively simple grammatical structure; therefore its addition does not require any significant additional effort in terms of analysis. The inclusion of Swedish – the second official language of Finland –, on the other hand, does not seem necessary in the case of Tampere: according to the 2013 information of Statistics Finland, the percentage of Swedish speaking inhabitants in the city was around 0.5% [***61***].

The decision of including another language besides Finnish entails the requirement for the system to support multilingualism. This needs to be incorporated in the user interface of the applications as well as in the analysis logic. Creating applications with multilingual user interfaces is relatively simple on most platforms; therefore it does not require much effort. Handling multilingualism in the textual analysis might be a more difficult task; however, there are already some existing frameworks that support several languages and perform with satisfactory efficiency. One example is the Grammatical Framework, which has already been discussed in the literature review section.

3.2. Overview of the traffic event detection system

With the above-described decisions considered, a social media-based traffic event detection and classification system was designed as a proposed solution to the research problem. The system monitors Twitter streams to collect and analyse traffic related social media messages submitted from the Tampere area. Information deemed reliable is forwarded to the users, either automatically by the system itself or by the person in charge of the administration of posts (e.g. a traffic operator).

User reports are submitted via Twitter, either through the service's own application or through the mobile application designed for simplifying the reporting process. The mobile application is available only for Android platform at the moment, but developing an iOS version in the future is also considered. In ordered to present the acquired and classified data to the traffic operators in a comprehensible and transparent way, a web application was developed too. The application was built using the MeteorJS framework, which is an open-source web development framework using JavaScript and HTML5.

A simplified illustration of the interactions between the different parts of the system can be seen below:

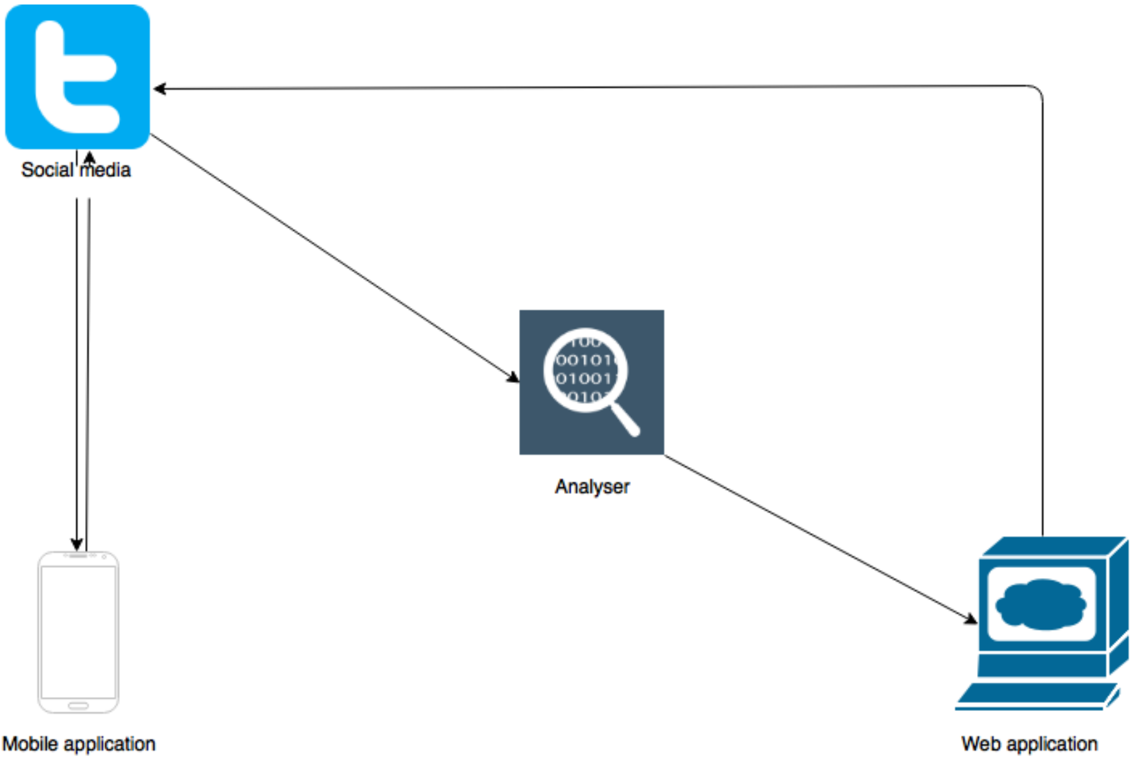


Figure 3.3 An overview of the system

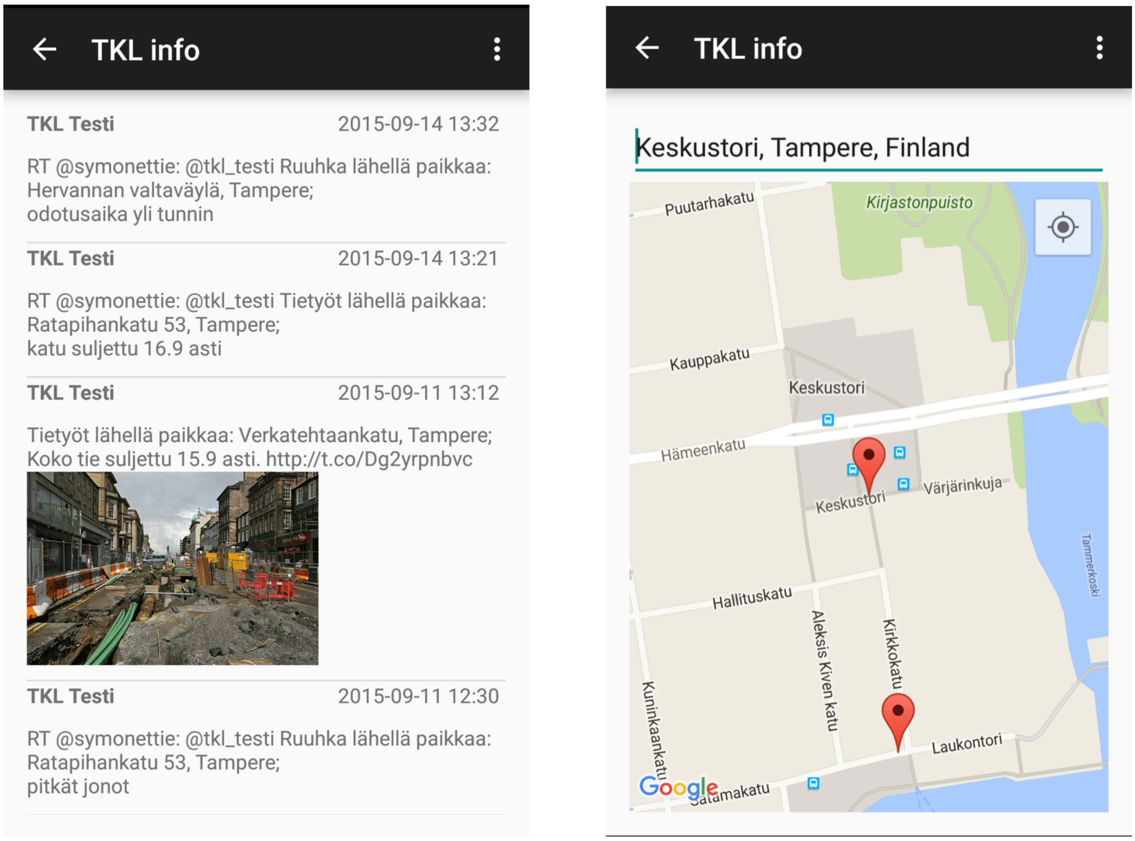
The following subsections are going to discuss the different parts of the system in further detail.

3.2.1. The mobile application

As mentioned above, the main purpose of the mobile application is to simplify the reporting process and thus increase user activity. In addition, it also provides a more comprehensible and more visual interface for browsing officially published posts (tweets submitted through the transportation department's Twitter account). Therefore the application has two main functions: the display of official tweets collected from the traffic operator's Twitter account and the facilitation of submitting user reports.

Official posts are presented in two different ways in the application: the home view displays all the posts collected from the official account in a list and the map view shows tweets which have geographical information associated with them placed on a map. The latter gives users a better visual understanding of traffic conditions throughout the city and helps them to identify problematic spots.

The images below present the list and the map view of official posts:



TKL info

TKL Testi 2015-09-14 13:32

RT @symonettie: @tkl_testi Ruuhka lähellä paikkaa: Hervannan valtaväylä, Tampere;
odotusaika yli tunnin

TKL Testi 2015-09-14 13:21

RT @symonettie: @tkl_testi Tietyöt lähellä paikkaa: Ratapihankatu 53, Tampere;
katu suljettu 16.9 asti

TKL Testi 2015-09-11 13:12

Tietyöt lähellä paikkaa: Verkatehtaankatu, Tampere; Koko tie suljettu 15.9 asti. http://t.co/Dg2yrpnbvc

TKL Testi 2015-09-11 12:30

RT @symonettie: @tkl_testi Ruuhka lähellä paikkaa: Ratapihankatu 53, Tampere;
pitkät jonot

TKL info

Keskustori, Tampere, Finland

Figure 3.4 List view and map view of official posts in the mobile application

The reporting page has a simple user interface in order to make submitting information effortless and thus encourage users to be more active. Reports can be generated easily by selecting the type of event from a list of predefined categories and adding geographical information by using the application's geolocator. Thus, in the most basic use case, an informative report can be created and submitted at the press of three buttons. Additionally, the application also allows users to attach images to their posts and to add their own comments as well.

A screen shot of the reporting page can be seen below:

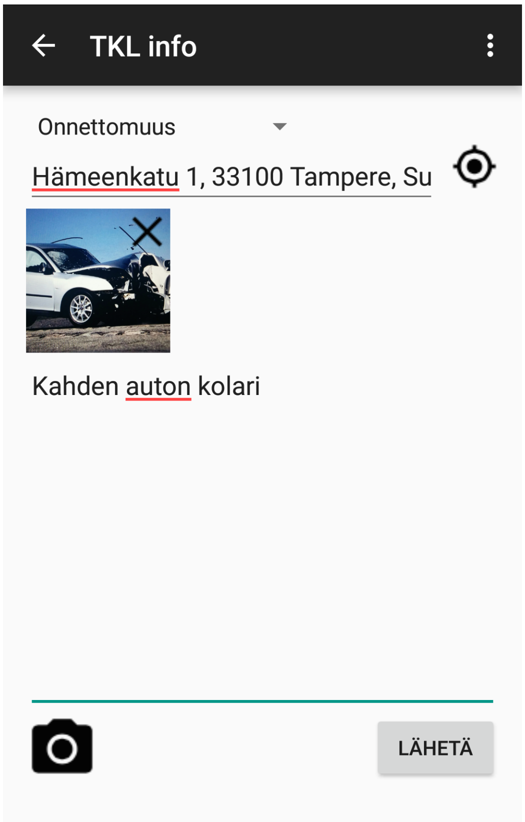


Figure 3.5 The reporting page of the mobile application

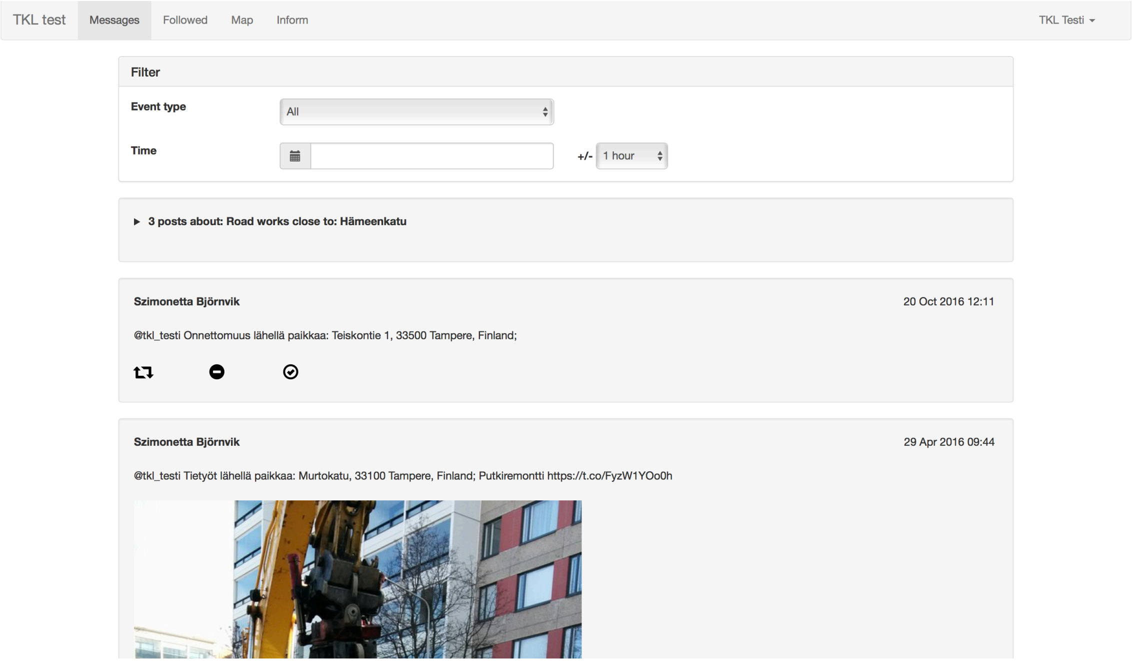
The application was developed for Android platform and is available on Google Play Store, the official digital store of Android applications. As mentioned before, it relies strongly on Twitter. Users need to log in through Twitter in order to be able to use the application as it receives data from the social media service and it also forwards the user reports there. The Twitter integration was implemented using Fabric SDK, the official mobile development platform of Twitter, which provides a simple and reliable way to access the service’s API.

3.2.2. The web application

The web application was mainly designed for traffic operators to use, but it also provides a simpler, more restricted interface to simple users, which offers the same functionalities as the mobile application. The administrative interface (used by officials) displays all the user reports and posts detected by the monitoring system and enables administrators to forward or enter relevant information, manage posts, block misbehaving users and follow other channels. In order to present collected tweets in an easily comprehensible way, the application categorises and clusters them. This is performed by the built-in analyser, which uses a combined approach of grammatical and word list-based analysis.

As mentioned before, the main added feature of the administrator view is the display of all posts detected by the system (simple users can see posts published by the official Twitter account only). In order to enhance transparency, tweets referring to the same event are grouped together and it is also possible to set filters on the posts. The administrator can manually forward ("retweet") messages deemed accurate and informative, but the system also has an automatic forwarding mechanism. This feature relieves the burden of continuous monitoring from operators and ensures that information is forwarded to users even when there is no human administrator working on the task. This also enables to create a real-time information system.

The layout of the incoming messages page can be seen below:



Filter

Event type: All

Time: +/- 1 hour

▸ 3 posts about: Road works close to: Hämeenkatu

Szimonetta Björnvik 20 Oct 2016 12:11

@tkl_testi Onnettomuus lähellä paikkaa: Teiskontie 1, 33500 Tampere, Finland;

Szimonetta Björnvik 29 Apr 2016 09:44

@tkl_testi Tietyöt lähellä paikkaa: Murtokatu, 33100 Tampere, Finland; Putkiremontti https://t.co/FyzW1YOo0h

Figure 3.6 Messages page of the web application

In other functionalities the web application is quite similar to the mobile application. Just like the Android application, it also has strong Twitter integration as user authentication and general data flow are handled through the social media service.

3.2.3. The analyser

The analyser is one of the most important parts of the system. It is responsible for the classification and clustering of social media messages as well as for determining what information should be forwarded to the users.

As mentioned above, the analysis logic is part of the web application as it is integrated into its server-side logic. Tweets collected from Twitter are automatically analysed, classified and clustered by analyser. If a certain piece of information is deemed reliable and useful by the system, it is also forwarded to users. Thus the analyser's main purpose is to help or even substitute the work of human operators.

The analyser uses a combined approach of grammatical and word list-based analysis, which will be discussed in detail in later sections. The necessity of using more than one analysis method can be explained by the fact that not all tweets follow standard grammatical structures (e.g. because of use of slang and simplified internet language); however, often more information can be extracted from grammatically correct messages with a sufficiently accurate grammatical analysis framework and therefore solely examining the occurrence of certain words (word list-based analysis) is not enough.

The exact structure and mechanism of the analyser as well as the acquisition of data to be analysed are going to be explained in further detail in the following sections.

3.3. Collection of data

As mentioned in the previous sections, the system collects data solely from Twitter. Public tweets are relatively easy to obtain through the REST and Streaming APIs of Twitter [***62***]. The REST APIs provide access to read and write Twitter data and can be used, for example, to perform elaborate searches or to collect tweets belonging to certain users. The Streaming APIs, on the other hand, offer insight into the on-going data flow of Twitter and allow for real-time detection of new posts.

Although the Twitter APIs provide access to a wide range of data and operations, they have certain limitations, which should not be ignored. The most significant restriction concerns the amount of data exposed through the APIs: without specifically purchased permission, only 1% of the public stream can be accessed. Although this might seem rather restrictive, it is possible to increase the number of relevant tweets collected with the effective use of filtering as filters are applied on the whole stream and not the limited data set. Therefore it is crucial to define adequate search criteria.

As the aim is to collect traffic information about the Tampere area, it can be considered reasonable to search for tweets containing traffic related keywords posted from the Tampere area. As Twitter allows for both keyword and location based filtering, it seems like a feasible task. However, it is important to note that only a very low percentage of tweets have geographical coordinates associated with them, therefore location based filtering would inordinately decrease the amount of posts collected. Broadening the scope to all Finnish social media messages mentioning traffic conditions, on the other hand, would include too many irrelevant posts as well.

One possible solution to increase data collection efficiency is to introduce dedicated channels. For example, tweets posted to a certain accounts (e.g. official accounts of transportation authorities) or containing certain “hashtags” (special keywords included in social media posts, which also function as searchable links) could be monitored. For this purpose, a test Twitter account (@tkl_testi) and a specific hashtag (#TKLinfo) were created. Tweets mentioning the test user or using the special hashtag are automatically detected and collected by the system. Additionally, posts published by certain users of interest such as official traffic information channels (e.g. @vtrafficSuomi) or local utility companies (e.g. Tampereen Sähkölaitos) are also monitored.

In addition to the tweets collected from specific channels, traffic related posts written in Finnish language were also collected for testing and observation purposes. These tweets are used to test and calibrate the linguistic analyser and are not utilised by the actual system. The reason behind the collection of additional data was to increase the volume and variety of testing data set. As the social media activity of the residents of Tampere is still relatively low, such measure can be deemed necessary.

3.4. Analysis of acquired data

The main aim of the analysis is to extract sufficient information from the social media messages to be able to properly classify and cluster them as well as determine what should be forwarded to users. Tweets following standard grammatical patterns are subjected to linguistic analysis in order to increase information extraction efficiency. In the case of non-grammatical tweets, a simpler word list-based analysis is performed to obtain certain basic information.

In order to correctly categorise and cluster tweets, it is important to determine the type of event they describe as well as the location and the timeframe of the incident. As Twitter messages are automatically time-stamped by the site, it is relatively easy to infer the time of the event, provided that it is assumed that users post about everything they witness immediately. Determining the place of the event is not always straightforward though: as not all tweets are "geo-tagged" (having geographical coordinates associated with them,) the location often has to be inferred from the textual content and the success of geographical information extraction highly depends on the quality of the text (e.g. amount of information present, spelling, use of language etc.) and the capabilities of the analyser. In some cases, location is not even mentioned at all; however, these posts will be generally ignored as attempting to deduce spatial information based on other social media data would be a too complex task (albeit not entirely impossible) with little added benefit.

The following subsections are going to discuss data analysis process. The different textual analysis approaches used for information extraction and classification are going to be explained in detail as well as the methods used for clustering and determining reliability and relevance.

3.4.1. Word list-based analysis

In order to identify the event described by the tweets collected, a word list-based analysis is performed on them. The analysis method follows a similar approach to that of discussed by Wanichayapong et al. [***9***], which uses categorised vocabularies in order to determine the type of the incident reported. The word lists used by the analyser were compiled from glossaries published by transportation companies and keywords extracted from traffic related tweets. As the system focuses mainly on identifying events such as accidents, traffic jams or road works, the vocabularies are also divided into these three categories.

Due to the complexity of Finnish language, certain modifications have to be made to the original method mentioned above, which was developed for posts written in Thai, a language with a considerably simpler grammatical structure. For example, the texts analysed have to be pre-processed first, which involves morphological stemming among others (e.g. elimination of hyperlinks and irregular characters, lowercase conversion etc.). Also, the exact same location extraction method defined in the original study – which is based on the occurrence of location names and certain prepositions – cannot be applied in this system as Finnish language expresses locative cases with declensions instead of separate prepositional words and the inspection of such cases with mere word lists is rather difficult. Moreover, any suffixes carrying locative information get eliminated during the pre-processing. Therefore, if the place of the event has to be determined solely from the textual content using word list-based analysis – which occurs only in the case of non-grammatical tweets –, some simplifying assumptions have to be made: for example, if the tweet contains only one location name, it is assumed that it is the exact spot of the incident and if the post mentions two locations, the place of the event is considered to be at their midpoint.

The logic behind the word list-based analysis is rather simple: the analyser checks for the occurrence of certain key words and determines the type of the event described based on which word list the extracted key words belong to. As tweets are rather short due to the official length restrictions, it can be assumed that even one identified key word can be sufficient to infer the overall message of the post. In some cases, however, it is not completely obvious how to classify the post in question, as there might be keywords from two different categories present. Such is the case, for example, when an accident resulting in huge traffic jams (due to road sections being closed off for police investigation) is reported. As the system allows for only one category per event in order to avoid redundancy, a primary event category has to be determined in these situations. In the example mentioned above this would be the accident, as it is considered to be a more important piece of information (and also, the traffic jam can be seen as its logical consequence, which almost always occurs).

The main advantage of the word list-based analysis is that it can help to obtain information in cases where grammatical analysis cannot be performed. In the case of tweets following standard grammatical patterns, however, the linguistic analysis approach is preferred as it allows for deeper and more accurate analysis. The next subsection is going to discuss the grammatical analysis framework used in this project.

3.4.2. Grammatical analysis

As mentioned above, grammatical analysis is performed on tweets that follow standard grammatical patterns. Besides information extraction and classification, it also helps to gain a deeper understanding of the relationships between different parts of the text and thus perform more thorough analyses.

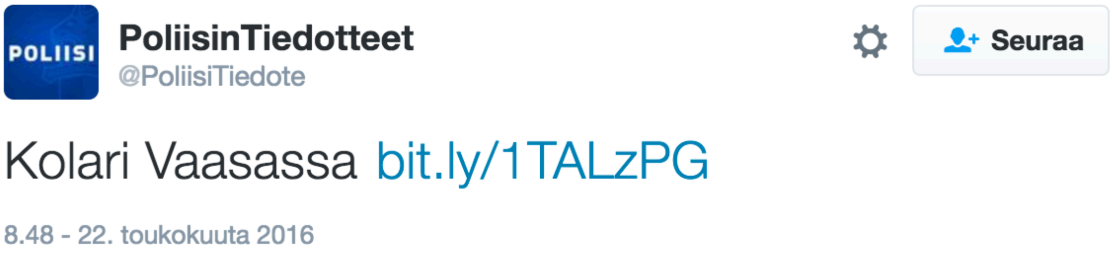
The grammar is built on the Grammatical Framework already described earlier and it also utilises the extensive Finnish and English resource grammar libraries developed for the framework. The resource grammars are used to provide syntax rules and lexical paradigms, which are utilised in the custom grammar. The custom grammar was designed to support traffic related messages following typical formats and includes structural rules and specific lexicons.

The underlying idea of the specific traffic grammar is that reporting messages are most likely to contain the following elements: the event that occurred, its location and some additional descriptive adjectives. Therefore, variations and permutations of these elements can cover a relatively large part of all the possible tweet formats. In order to build an effective analyser, it has to support a sufficiently large amount of different structures and use amply extensive lexicons. It should be borne in mind, however, that it is practically impossible to prepare the analyser for all the possibilities; therefore an optimal degree of comprehensiveness should be determined.

The traffic specific lexicon used in the grammatical analyser was compiled using external glossaries and common words extracted from traffic related reports and tweets. It also includes a list of all official street names found Tampere; the list was obtained from a web site of the municipality [***63***]. The explicit inclusion of street names does not only facilitate the textual analysis, but it also helps to identify the actual geographic location being referred to. As the scope of the research is restricted to one town with a finite and relatively small set of possible street names, the addition of an exhaustive list is feasible and requires no overbearing effort.

The syntax rules created for the traffic grammar were formed in a way to support most patterns detected in grammatically correct traffic related messages. It can be observed that most of these messages have similar grammatical structures as they were submitted by official or authorative sources. Therefore, even a small set of well-formed rules can achieve a satisfactory coverage and enable the analyser to process a large part of traffic related tweets. In order to maximise efficacy, however, rules for recognising less common (but grammatically correct) message formats were also added.

An example of one of the most basic reporting formats, which can actually be observed in some messages, is the following: “onnettomuus Tampereella” (tr. “accident in Tampere”). This can be easily modelled as a function of an event and a location, both of them being nouns. Other common message formats often extend this basic form e.g. by adding qualitative adjectives, verbs or relational clauses. All these adhere to standard grammatical rules and therefore can be easily implemented using the Grammatical Framework.



PoliisinTiedotteet
@PoliisiTiedote

Kolari Vaasassa bit.ly/1TALzPG

8.48 - 22. toukokuuta 2016

Figure 3.7 Example of a tweet following a standard reporting format

The implemented textual analyser, which combines the above described grammatical analyser and the word list-based classifier, is able to categorise most traffic related messages correctly as well as extract some additional information. Besides accurate classification, effective clustering of tweets referring to the same event is another important task. The next subsection is going to discuss the clustering approach used in this project.

3.4.3. Cluster analysis

Clustering tweets referring to the same event helps to assess the importance and impact of certain incidents. Grouping messages together also enables to present information in a more transparent and comprehensible way, which facilitates the work of human operators. Therefore it is important to find an adequate clustering method, which can effectively identify relationships between social media data entries.

In order to be able to decide whether two separate tweets are referring to the same event, appropriate decision criteria have to be determined. If certain posts describe the same type of incident and were posted from the same location at approximately the same time, it is highly likely that they are in fact referring to the same event. Therefore, a decision rule could be to ensure that the tweets belong to the same classification category and were posted within a certain radius and within a given timeframe. The optimal limit of the spatial and temporal distance may vary between event categories, as the size of their impact area and their duration often differ (e.g. road construction works often last for months whereas traffic jams usually dissolve within an hour).

As the spatial aspect is a very important factor, the grouping task could be treated as a density based clustering problem. However, as temporal proximity and identical event categories are also decision criteria, it is not enough to examine the location only. Therefore an additional temporal dimension needs to be included in the analysis alongside with constraints on the type of the incident.

Considering existing clustering solutions, a DBSCAN style approach seems most suitable for this grouping problem, as it examines spatial characteristics and is able to handle arbitrarily shaped clusters and noise. However, the original method cannot be applied without further modifications as it focuses on the spatial dimension only and defines the same ε radius for all the clusters. As mentioned above, temporal dimension has to be considered as well alongside with event category constraints and the size of the distance radii may also vary. Therefore these modifications need to be included in the clustering method in order to perform effective analyses.

Based on the definitions of DBSCAN and the additional modifications, the following rules can be defined:

- *Category*: Let $C(p) = c$ define the category of a p data point (message), where $c \in \{a\!: accident, t\!: traffic\ jam, r\!: road\ work\}$.
- *Maximum distances*: $\varepsilon_{sc}, c \in \{a, t, r\}$ denotes the category specific maximum temporal distance allowed between two points to be considered neighbours and $\varepsilon_{tc}, c \in \{a, t, r\}$ denotes the category specific maximum temporal distance.
- *Distance*:
  $$dist_s(p,q) = 2R \arcsin\left(\sqrt{sin^2\left(\frac{q_x - p_x}{2}\right) + cos(p_x)cos(q_x)sin^2\left(\frac{q_y - p_y}{2}\right)}\right)$$
  defines the spatial (geodetic) distance between points p and q, while $dist_t(p,q) = |p_t - q_t|$ denotes the temporal distance of the points, given in hours.
- *Neighbourhood*: A given q point is in the neighbourhood of p if $C(q) = C(p)$ and $dist_s(p,q) \leq \varepsilon_{sc}$ and $dist_t(p,q) \leq \varepsilon_{tc}$.
- *Directly density-reachable*: A point q is directly density-reachable from point p if q is a neighbour of p and the neighbourhood of p contains at least *MinPts* number of points, where *MinPts* is a the minimum required number of points to form a dense region.
- *Density-reachable*: A point q is density-reachable from point p if there is a $x_1 \ldots x_n$ path between the points, $x_1 = p$ and $x_n = q$, in which p_{i+1} is directly density reachable from p_i.
- *Density-connected*: A point q is density-connected to point p if there is a point *o* from which both q and p are density-reachable.
- *Cluster*: A cluster *C* is a non-empty subset of the database where all members are density-connected and if a point p is a member of *C*, then all points that are density reachable from p are members too.
- *Noise*: In a given database *D*, noise is defined as a set of points that do not belong to any of the clusters in *D*.

With the definitions described above, the optimal values for MinPts and the category specific spatial and temporal distances should be determined in order to be able to apply the clustering approach to practice. As the social media activity in Tampere in regards to reporting traffic conditions is relatively low at the moment, the expected size of clusters is quite small. Therefore the value of MinPts should not be too big either. Based on the social media data collected so far from Twitter, the average number of Finnish tweets referring to the same traffic event seems to be around three, therefore it may be a sensible decision to set the value of MinPts to this.

Setting appropriate values for the category specific distances is crucial for achieving a satisfactory accuracy; however, it is not always a simple task, especially in the case of time intervals. For example, road works can range from one-day repair jobs to large construction projects lasting for years. Therefore a too small ε_{tr} value can result in ignoring prolonged events, while a too big value can lead to mistakenly identifying isolated incidents as being the same. With all these considered, the final decision was to set the distances to average values. Based on historical observations [***64***], the average duration of road construction works is 6,42 months, whereas traffic jams last for 8 minutes on average [***65***]. Differences in the size of the impact area are less outstanding, however: the average length of road sections under construction is 3,86 km, for example, and the length of queues caused by traffic jams is only slightly shorter, ranging from 1 km to 3 km.

Clustering tweets referring to the same event does not only allow for presenting data in a simpler and more comprehensible way, but it also helps with deciding what information should be forwarded to users. The following subsection is going to discuss the main considerations in regards to tweet verification and automatic forwarding.

3.4.4. Verification and automatic forwarding

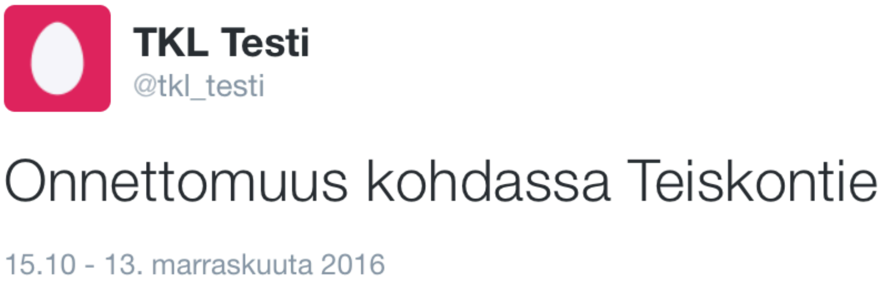
One purpose of the system is to inform passengers about unexpected traffic events and conditions in real-time, even when a human operator is not present to execute the task. For this reason, an automatic forwarding mechanism needs to be added. However, not all the reports received should be forwarded, as they may contain false or redundant information. Therefore the system needs to be able to determine the verity of messages and to decide what content should be published to users.

Verifying tweets is not always a simple task, especially in the case of localities with small population and relatively low online activity. For example, verification methods such as comparison to external news sources might not be applicable, as minor events are not always officially reported. Requiring a critical level of social media coverage might not be an optimal solution either, as it might result in ignoring events that actually occurred, but were mentioned only by a few users.

Certain messages, however, can be deemed reliable with a high certainty. Such is the case when an event is reported by authorities or official news channels. According to past observations, official and authorative Twitter accounts generate most of the traffic related social media activity in Finland, therefore such situations can be expected to occur quite frequently. In these cases, there is no actual need for further inspections and the reports can be directly forwarded to users.

User submitted messages cannot be treated with the same confidence as official announcements as some individuals may post with malicious intent. Irrelevant spam is immediately filtered out by the textual analyser, but a solution for recognising false reports following valid formats should also be found. A good initial approach might be to require a minimum number of unique users to report the same event or to examine the number of “likes” and “retweets” a certain post receives in order to assess the verity of the message. However, as the current social media activity in the area regarding traffic events is not really high, the threshold number in the initial phase cannot be too big. At the moment this number is set to three, which is the same as the minimum number of neighbouring data points required in the clustering method in order to form dense regions.

After identifying true and relevant events, appropriate information should be forwarded to the users. Simply reposting one of the user submitted reports might not always be an optimal solution though as the content or the language of the post might not adhere to official standards. However, as the most important pieces of information are the type of the event occurred and its location, which are automatically extracted by the analyser, an official announcement with a standard and simple format can easily be compiled.



TKL Testi
@tkl_testi

Onnettomuus kohdassa Teiskontie

15.10 - 13. marraskuuta 2016

Figure 3.8 A simple official announcement posted from the test account

The verification and automatic forwarding mechanism supports and complements the work of human operators. This means that administrators have full control over the system and can make the final decision regarding the content to be shared publicly and the automatic posting feature is mostly intended to use at times when a human operator cannot be present to monitor the system (e.g. outside of working hours). However, with some further improvements and development, it might be possible in the future for the automatic system to completely replace human work.

3.5. Summary

This chapter has thoroughly explained the approaches and methods used for developing an advanced traffic information system, which is the suggested solution to the research problem. The system monitors Twitter for relevant information and benefits of a combined word list-based and grammatical analyser to classify messages. A modified DBSCAN-style clustering is performed in order to group reports referring to the same event together, which – besides allowing for simpler and more comprehensible data presentation – helps also with verification and decision making regarding automatic forwarding.

The implemented system has the potential to effectively identify, analyse and forward relevant and informative messages describing traffic conditions in the Tampere area. The next chapter is going to present and evaluate the results of the research.

4. Results

This chapter is going to present the main results of the research project as well as the tests performed in order to evaluate the efficacy and the accuracy of the proposed solution.

4.1. Overview

The main goal of the research was to find a solution to identify and classify traffic events occurring in the Tampere area based on Finnish social media data. For this purpose a social media-based traffic information system was developed. The system monitors Twitter in order to detect traffic related messages and analyses them to determine the event they describe and to assess their relevancy and verity. The classification method relies on the help of a textual analyser, which combines grammatical and word list-based analysis methods. The initial presumption is that this method is able to correctly classify most traffic related tweets, however, evaluation tests should be performed to confirm this. As clustering posts referring the same event was also an important part of the research, the efficacy and accuracy of the proposed clustering approach should also be evaluated.

The following subsection is going to discuss the tests performed to assess the performance of the system. As textual analysis and clustering are considered to be the core of the system, special attention will be paid to their evaluation.

4.2. Testing and evaluation

In order to evaluate the performance of the system, it is important to see how it processes social media data and draw the necessary conclusions. As the accuracy of textual analysis and clustering has the most impact on the overall efficacy, special focus should be placed on the testing of those components. The following subsections are going to present the results of the tests performed on the textual analyser and the clusterer.

4.2.1. Testing the textual analyser

The main task of the textual analyser is to determine what type of event the detected messages describe; therefore it is important to assess classification accuracy. As the analyser combines grammatical and word list-based analysis methods, both approaches will be tested.

For the general assessment of classification accuracy, traffic related tweets written in Finnish were collected using the Twitter Search API. The data was collected over the span of six months and the resulting data set consists of 188 unique tweets describing traffic events or conditions. In addition, a total of 120 test user reports were generated to test some system specific use cases.

The performance of the analyser is assessed from the test results by inspecting the number of correctly classified messages (true positives and true negatives) in relation to the number of incorrectly classified data entries (false positives and false negatives). Based on these numbers, evaluation measures such as precision, recall and accuracy are calculated. The precision value indicates the fraction of retrieved data that is relevant (not stating whether all relevant data entries were identified), while recall denotes the fraction of relevant instances recognised. Accuracy refers to the number of correctly classified items out of all the entries analysed.

The precision, recall and accuracy values are calculated based on the definitions of Olson and Delen [***66***]:

- Precision: $P = \frac{TP}{TP+FP}$
- Recall: $R = \frac{TP}{TP+FN}$
- Accuracy: $A = \frac{TP+TN}{TP+TN+FP+FN}$

Where TP stands for true positive, i.e. the number of messages correctly classified by the analyser; TN denotes true negative, i.e. the number of irrelevant posts correctly ignored by the classifier and FP and FN indicate the false positives and false negatives respectively.

4.2.1.1 Evaluation of the grammatical analysis

The grammatical analyser was tested on the above-mentioned data set consisting of 188 generally collected tweets and 120 test user reports, which had been previously manually categorised. The test results were compared to the manual classification in order to calculate the appropriate evaluation measures.

The analyser successfully parsed and classified 53 out of the 188 general tweets, which corresponds to the amount of grammatically correct posts present in the data set.

However, the test results yielded better results, with 118 out of 120 messages correctly parsed and the remaining two tweets properly recognised as irrelevant.

The following tables summarise the test results as well as present the appropriate evaluative values:

Data set	TP	TN	FP	FN
General	53	14	0	121
Test reports	118	2	0	0
Sum	171	16	0	121

Table 4.1 Test results of the grammatical analyser

Data set	Precision	Recall	Accuracy
General	100%	30,46%	35,64%
Test reports	100%	100%	100%
Combined	100%	58,56%	60,71%

Table 4.2 Evaluation results of the grammatical analyser

The results above demonstrate the viability of the grammatical analyser. As a relatively large part of tweets submitted by individual users does not adhere to rules of standard grammar, 100% accuracy is not expected. However, as the traffic grammar used in the system has highly specific structural rules and vocabulary, it is highly likely that messages identified and parsed by the analyser are correctly classified. As the results of the tests on user reports show, accuracy can be improved with the use of the mobile application and by concentrating on specific channels. Therefore better results are expected from actual use of the system.

4.2.1.2 Evaluation of the word list-based analysis

The word list-based analyser was tested on the same data set as the grammatical analyser. According to the test results, the analyser successfully identified all the relevant traffic related messages; however, there were also some “false positives” in the case of messages, which contained certain traffic specific keywords, but did not describe actual road incidents. Such posts were, for example, tweets reporting car race accidents.

The following tables present the test results as well as the values calculated for the evaluation measures:

Data set	TP	TN	FP	FN
General	174	0	14	0
Test reports	118	2	0	0
Sum	292	2	14	0

Table 4.3 Test results of the word list-based analyser

Data set	Precision	Recall	Accuracy
General	92,55%	100%	92,55%
Test reports	100%	100%	100%
Combined	95,42%	100%	95,45%

Table 4.4 Evaluation results of the word list-based analyser

The results demonstrate that the analyser is able to operate with high precision and recall and thus achieve a good overall accuracy. The word list-based analyser therefore can compensate for the possible deficiencies of the grammatical analyser by being able to correctly classify even those messages that cannot be parsed with the traffic specific grammar. One drawback of the word list-based analysis compared to the grammatical analysis is that is more prone to producing false positives as it is less strict regarding the textual content and format, however, as the system is expected to receive mostly relevant messages, it might not have a significant effect on the overall accuracy. It is also a future development goal to improve the precision of the analyser, therefore the risk of incorrect classification is hoped to be minimised.

4.2.2. Evaluation of the clustering

The clustering method was mainly tested on the test user reports as well as on a set of 77 posts selected from the general traffic tweets. Other entries of the data set were excluded from the tests, as the exact location could not be inferred from their textual content often due to references to broader geographic areas or vague mentions of place names.

The clustering algorithm performed well on both, assigning the correct cluster to almost all the entries, with the exception of a few posts referring to separate incidents occurring in close temporal and geographical proximity. Thus the cluster analysis demonstrated an average accuracy of 99%.

For more extensive validation, evaluation metrics such as precision, recall and Rand Index (accuracy) were calculated, following the same formulae as the ones defined in the previous section. The number of negative and positive decisions, however, is counted by data point pairs instead of individual instances. In this context, a true positive decision, for example, refers to similar data points that were correctly placed in the same cluster, while a true negative decision is made when two dissimilar items are assigned different clusters, as expected from an accurate method. The correct identification of noise is also treated as a true negative classification.

The following tables summarise the test and evaluation results:

Data set	TP	TN	FP	FN
General	26	2899	1	0
Test reports	146	6989	1	4
Sum	172	9888	2	4

Table 4.5 Clustering test results

Data set	Precision	Recall	Rand Index
General	96,29%	100%	99,96%
Test reports	99,32%	97,33%	99,93%
Combined	98,85%	97,73%	99,94%

Table 4.6 Clustering evaluation results

The results above clearly demonstrate the efficacy of the selected clustering method. Although the achieved accuracy can already be considered satisfactory, further improvements might be possible attain by careful selection of the category specific distances. However, their optimal value is often difficult to determine as the spatial and temporal scope of events of the same category may vary greatly. In order to find the appropriate distance values for maximising accuracy, more observations are needed.

4.3. Summary

This chapter has presented the results of the tests performed on the textual and cluster analysers. The evaluation of the selected analysis methods has determined that they can achieve a satisfactory accuracy on traffic related social media data, which can be improved with further developments. The actual use of the system is also expected to yield better results.

5. Conclusions

This section is going to summarise and discuss the main findings of the research as well as present plans and ideas for future research and development.

5.1. Summary of the findings

The main purpose of this research was to develop a real-time social media-based traffic event detection system for the city of Tampere. As no similar research has been yet performed in a locality with such a small population and therefore relatively low social media activity, it was also a research task to determine whether it is feasible to build such a system under these circumstances. The complexity of Finnish language also presented additional challenges, which needed to be tackled too.

The research focused mainly on Twitter as information source as it provides a larger amount of publicly available data than its main competitors. The short and concise nature of Twitter messages and the well defined and well documented APIs provided by the service also facilitate data collection and analysis significantly. The Finnish user base of the application can also be considered rather large in proportion to the overall population and, based on the data samples collected, their activity can already be deemed sufficient to build a social media-based traffic information system on it. The use of social media is also expected to increase in the next few years and this, paired with an intensive advertisement campaign for the appropriate traffic problem reporting channels, can result in significantly more user submitted reports and thus better results.

As Twitter provides access to only 1% of the whole public stream, it is important to invent effective data collection methods. Defining appropriate search filters can improve retrieval efficacy, as filters are applied on the whole public data set before returning a limited number of tweets. Although the optimal solution would be to scan for traffic related messages submitted from the Tampere area, filtering by location is often difficult as most posts do not have geographical information associated with them. Therefore one design decision during the development of the traffic information system was to concentrate mainly on specific channels and official accounts. As official and authorative sources account for the majority of traffic related tweets at the moment, a satisfactory performance can be achieved by focusing mainly on these accounts. However, increased user activity directed at specific channels is also expected in the future, which will improve efficacy and accuracy.

Although the complexity of Finnish language might seem intimidating at first in terms of textual analysis and classification, there are some grammatical frameworks, which can provide significant help. For example, the Grammatical Framework has a comprehensive resource grammar library for Finnish, which enables the implementation of complex topic specific grammars. The traffic grammar defined in this research was found to be highly effective in parsing texts following standard grammatical patterns; however, a relatively large part of social media messages do not adhere to classical grammar rules. In order to be able to classify non-grammatical posts as well, a word list-based analyser was also added to the system, which demonstrated exceptionally high classification accuracy. However, performing word list-based analysis only on the whole data set was not considered an optimal solution as the grammatical analyser allows for more thorough inspections, which can be particularly useful for future developments.

Another important research task was the effective clustering of tweets referring to the same event. As the spatial aspect of posts is considered very important in the data analysis, a density-based clustering approach such as DBSCAN seemed the optimal choice. However, as temporality and classification category are also important factors in the decision-making, the original method had to be modified to include those aspects as well. The customised DBSCAN demonstrated an accuracy of 99% on the test data set, which proves the efficacy of this method.

All in all, the foundations have been created for a real-time social media-based traffic event detection and advanced traveller information system, which is able to achieve good performance even in small localities such as Tampere and in a complex linguistic environment such as Finnish. With increased user activity and additional future developments, this system is expected to yield even better results and become an essential part of traffic management and operator-passenger communication.

5.2. Relationship and contribution to earlier studies

As mentioned earlier, real-time traffic event detection and social media mining are prevalent issues nowadays and have inspired a vast amount of research projects and studies. This research has certainly benefited from the knowledge accumulated by other researchers, which also shows in some of the approaches selected to solve the research problem.

Certain studies on topics such as social media mining, traffic event detection and textual and cluster analysis have helped significantly during the development of the traffic information system implemented as part of this research project. For example, the classification methods suggested by Wanichayapong et al. [***9***] and Kosala et al. [***8***] provided inspiration for the keyword and word list-based textual analysis, while previous works and studies in relation to the Grammatical Framework [***30***] [***31***] have helped significantly with the grammatical analysis. In addition, the density based approach proposed by Ester et al. [***43***] presented certain core concepts and ideas, which were successfully utilised in the construction of the tweet clustering method defined in this study.

Although this research has gained much inspiration from previous studies, it also features some novel elements. For example, no similar research has addressed the problem of social mining in small localities, for which this paper presents a possible solution. The proven feasibility of such research may also inspire others to implement similar projects in other places with small population, which can be highly beneficial for the local inhabitants.

The textual analysis and classification of social media messages written in a highly complex language such as Finnish has also been a rather uncharted territory, especially in connection with the topic of traffic event detection. This research has examined the issue thoroughly and has successfully provided a possible solution, which could be used by others seeking to solve similar problems as well.

5.3. Limitations of the research

Although this research has successfully lay the foundation for further studies and practical applications, it has had certain limitations, which should be addressed in the future.

The low social media activity in the area, especially in regards to reporting traffic problems, was one major concern when attempting to build an effective social media-based event detection system. Although the online content submitted by official and authorative sources can already be sufficient to obtain near real-time information about most incidents, a more active involvement of private users can significantly improve efficacy as they might notice more problems than what is officially reported to operators and they might also gain knowledge about certain events earlier. In addition, having a larger data set, of which a significant part is generated by normal individuals, can also help to conduct more thorough observations and to identify prevalent trends and patterns. This problem was partly mitigated by expanding the search scope to include all Finnish tweets describing traffic events and conditions regardless of location and the results already provided some information; however, more insight could be gained from a sizeable location specific data set.

Certain limitations applied to the analysis of data as well. For example, the textual analyser is not able to cope with colloquial Finnish or slang. As these dialects are often used in social contexts, it poses the risk of ignoring a considerable portion of relevant messages. It is, however, difficult to include colloquial Finnish and slang in the textual analysis, as they change dynamically and rapidly and they also vary greatly among different regions, which also results in the lack of proper documentation. In addition, texts written in these dialects at times follow extraordinary grammar rules, which can present certain challenges in terms of grammatical analysis. The Finnish Resource Grammar Library of the Grammatical Framework, which is used to parse standard, grammatically correct texts, does not support colloquial language or slang at the moment. Based on current data samples, the majority of traffic related social media messages uses standard Finnish language, however, this might change in the future with the increased involvement of private users.

In addition to non-standard dialects, spelling mistakes can also affect classification accuracy. At the moment the textual analyser has no coping strategy defined to mitigate this potential problem. Eliminating the impact of spelling mistakes is actually a very difficult task, which most classifiers fail to perform. This has been realised by other researchers as well such as Batool et al. [*25*]. Although in some instances pre-processing and the use of an extensive dictionary as reference [*12*] can help to correct most spelling mistakes, at times confusion might arise in the case of several words having very similar written forms. This phenomenon is relatively common in the Finnish language; therefore further investigation is needed to solve the problem of wrongly spelt words.

Although the above-mentioned limitations had no significant effect on the research because of the current circumstances, they should be considered in the future in order to be able to build an effective and robust system. The next subsection is going to list ideas and plans for further research and development.

5.4. Ideas for future research and development

As mentioned earlier, low social media activity and lack of users are problems that should be addressed in the future. One possible solution is to launch intensive and large-scale advertisement and social media campaigns in order to encourage inhabitants to report problems through the specific channels established in this research. In addition, possible solutions to increase user engagement within the reporting application could also be investigated. For example, gamification or the addition of a reputation system of some sort could motivate users to be more active, which could increase the amount of available data. Enabling rating of user-submitted reports could also help with verification, which would improve reliability and accuracy.

Improvement possibilities of the analyser should be examined as well. As mentioned above, the appearance of colloquial language or slang in texts can present certain problems, which should be dealt with. Although adding support for such non-standard and rapidly changing dialects is challenging, it might not be entirely impossible. Therefore a careful investigation of possible solutions and a thorough study of different variations of slang and colloquial language are planned as future research. Other possible improvements such as defining coping mechanisms for spelling mistakes are also being considered.

Identification of the location a post was submitted from can often be problematic, as most tweets do not have geographical coordinates associated with them. Although the textual analyser has the capability to extract location information in certain cases, it is unable to cope with ambiguous names or the lack of explicit place references. In these cases, alternative solutions to determine the exact geographic position of the user could be investigated. The inspection of options such as inferring current location based on the textual context or past user activity is among future plans.

As automatisation and artificial intelligence are trending topics in information technology right now, it would be an interesting task to explore the numerous possibilities the inclusion of machine learning methods could provide. The utility of machine learning in data mining and classification has already been demonstrated in several studies [*6*] [*24*] [*25*] and the system developed in this research could also benefit from the use of this approach. Therefore an extensive study of theoretical background and existing solutions as well as the implementation of a custom learning method are among the future development plans.

Acknowledgement

In this section, I would like to express my gratitude to all the people whose help and support enabled me to write this thesis.

First and foremost, I would like to thank my supervisor, Jyrki Nummenmaa, for granting me the opportunity to work on this exciting research project and for providing continuous help and support throughout the whole research and thesis writing process. His active contribution and guidance I would not have been able to conduct this research and write this thesis.

I would also like to thank the transportation department of the city of Tampere for providing me with such an interesting real-life problem and for their financial support, which has allowed me to work full-time on this research as an intern at the University of Tampere. The valuable insight I gained during my site visits and our meetings has also helped me significantly.

I would also like to express my gratitude to all those colleagues and classmates who provided me with ideas, suggestions, advice and feedback in regards to this thesis. Their help has allowed me to identify and correct mistakes as well as introduce further improvements.

Last but not least, I would like to thank my friends and family whose continuous emotional support has helped me through many difficulties and allowed me to go on. Without them, all this would not have been possible.

References

1] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Washington DC, USA, 2010, pp. 178-185.

2] Arjumand Younus et al., "What do the Average Twitterers Say: a Twitter Model for Public Opinion Analysis in the Face of Major Political Events," in *2011 International Conference on Advances in Social Networks Analysis and Mining*, Kaohsiung, Taiwan, 2011, pp. 618 - 623.

3] Mohamed M Mostafa, "More than words: Social networks' text mining for consumer brand sentiments ," *Expert Systems with Applications* , vol. 40, pp. 4241–4251, 2013.

4] Jianshu Weng, Yuxia Yao, Erwin Leonardi, and Bu-Sung Lee, "Event Detection in Twitter," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, 2011, pp. 401-408.

5] Adrien Guille and Cécile Favre, "Mention-anomaly-based Event Detection and Tracking in Twitter," in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Beijing, China, 2014, pp. 375-382.

6] Maximilian Walther and Michael Kaisser, "Geo-spatial Event Detection in the Twitter Stream," in *Proceedings of the 35th European Conference on Advances in Information Retrieval*, Moscow, Russia, 2013, pp. 356-367.

7] Eric Mai and Rob Hranac, "Twitter Interactions as a Data Source for Transportation Incidents," in *Transportation Research Board 92nd Annual Meeting Compendium of Papers*, Washington DC, USA, 2013, p. 11.

8] Raymondus Kosala, Erwin Adi, and Steven, "Harvesting Real Time Traffic Information from Twitter," *Procedia Engineering*, p. 12, January 2012.

9] Napong Wanichayapong, Wasawat Pruthipunyaskul, Wasan Pattara-Atikom, and Pimwadee Chaovalit, "Social-based Traffic Information Extraction and Classification," in *2011 11th International Conference on ITS Telecommunications (ITST)*, St. Petersburg, Russia, 2011, pp. 107-112.

10] Enrico Steiger, Timothy Ellersiek, and Alexander Zipf, "Explorative Public Transport Flow Analysis from Uncertain Social Media Data," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, Dallas, USA, 2014, pp. 1-7.

Alexander Pak and Patrick Paroubek, "Twitter as a Corpus for Sentiment

11] Analysis and Opinion Mining," in *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, Valletta, Malta, 2010, p. 7.

12] Farhan Hassan Khan, Saba Bashir, and Usman Qamar, "TOM: Twitter Opinion Mining Framework Using Hybrid Classification Scheme," *Decision Support Systems*, vol. 57, pp. 245-257, January 2014.

13] Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades, and Nick Bassiliades, "Ontology-based Sentiment Analysis of Twitter Posts ," *Expert Systems with Applications*, vol. 40, no. 10, pp. 4065-4074, 2013.

14] Manoochehr Ghiassi, J Skinner, and David Zimbra, "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network," *Expert Systems with Applications* , vol. 40, pp. 6266–6282 , 2013.

15] Kazufumi Watanabe, Masanao Ochi, Makoto Okabe, and Rikio Onai, "Jasmine: A Real-time Local-event Detection System based on Geolocation Information Propagated to Microblogs," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, Glasgow, UK, 2011, pp. 2541-2544.

16] Rui Li, Kinh Hou Lei, Ravi Khadiwala, and Kevin Chen-Chuan Chang, "TEDAS: a Twitter-based Event Detection and Analysis System," in *2012 IEEE 28th International Conference on Data Engineering (ICDE)*, Washington DC, USA, 2012, pp. 1273-1276.

17] Dennis Thom, Harald Bosch, Steffen Koch, Michael Wörner, and Thomas Ertl, "Spatiotemporal Anomaly Detection through Visual Analysis of Geolocated Twitter Messages," in *2012 IEEE Pacific Visualization Symposium (PacificVis)*, Songdo, South Korea, 2012, pp. 41-48.

18] Chenliang Li, Aixin Sun, and Anwitaman Datta, "Twevent: Segment-based Event Detection from Tweets," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, Maui, USA, 2012, pp. 155-164.

19] Richard McCreadie, Craig Macdonald, Iadh Ounis, Miles Osborne, and Sasa Petrovic, "Scalable Distributed Event Detection for Twitter," in *2013 IEEE International Conference on Big Data*, Silicon Valley, USA, 2013, pp. 543-549.

20] Sri Krisna Endarnoto, Sonny Pradipta, Anto Satriyo Nugroho, and James Purnama, "Traffic Condition Information Extraction & Visualization from Social Media Twitter for Android Mobile Application," in *2011 International Conference on Electrical Engineering and Informatics (ICEEI)*, Bandung, Indonesia, 2011, pp. 1-4.

IBM, "Mining Urban Traffic Events and Anomalies," Dublin, Ireland, 21] Research Report 2012.

22] Elizabeth M Daly, Freddy Lecue, and Veli Bicer, "Westland Row Why So Slow? Fusing Social Media and Linked Data Sources for Understanding Real-Time Traffic Conditions," in *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, Santa Monica, USA, 2013, pp. 203-212.

23] Bharath Sriram, David Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas, "Short Text Classification in Twitter to Improve Information Filtering," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Geneva, Switzerland, 2010, pp. 841-842.

24] Kyosuke Nishida, Ko Fujimura, Ryohei Banno, and Takahide Hoshide, "Tweet Classification by Data Compression," in *Proceedings of the 2011 International Workshop on Detecting and Exploiting Cultural Diversity on the Social Web*, Glasgow, Scotland, 2011, pp. 29-34.

25] Rabia Batool, Asad Masood Khattak, Jahanzeb Maqbool, and Sungyoung Lee, "Precise Tweet Classification and Sentiment Analysis," in *2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS)*, Niigata, Japan, 2013, pp. 461-466.

26] Alchemy API Inc. (2015) Alchemy API | Powering the New AI Economy. [Online]. http://www.alchemyapi.com

27] Axel Schulz and Frederik Janssen, "What Is Good for One City May Not Be Good for Another One: Evaluating Generalization for Tweet Classification Based on Semantic Abstraction," in *S4SC'14 Proceedings of the Fifth International Conference on Semantics for Smarter Cities*, vol. 1280, Riva del Garda, Italy, 2014, pp. 53-67.

28] Tommi A Pirinen, "Modularisation of Finnish Finite-State Language Description—Towards Wide Collaboration in Open Source Development of Morphological Analyser," in *NODALIDA 2011 Conference Proceedings*, Riga, Latvia, 2011, pp. 299-302.

29] Krister Lindén, Erik Axelson, Sam Hardwick, Tommi A Pirinen, and Miikka Silfverberg, "HFST — Framework for Compiling and Applying Morphologies," in *Systems and Frameworks for Computational Morphology: Second International Workshop, SFCM 2011, Zurich, Switzerland, August 26, 2011. Proceedings*, Zurich, Switzerland, 2011, pp. 67-85.

30] Aarne Ranta, "The GF Resource Grammar Library," *Linguistic Issues in Language Technology*, vol. 2, no. 2, pp. 1-63, December 2009.

Aarne Ranta, "Grammatical Framework," *Journal of Functional*

31] *Programming*, vol. 14, no. 2, pp. 145-189, January 2004.

32] A K Jain, M N Murty, and P J Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, September 1999.

33] Laurence Morissette and Sylvain Chartier, "The k-means clustering technique: General considerations and implementation in Mathematica," *Tutorials in Quantitative Methods for Psychology*, vol. 9, no. 1, pp. 15-24, February 2013.

34] George F Jenks, "The data model concept in statistical mapping," *International Yearbook of Cartography*, vol. 7, no. 1, pp. 186-190, 1967.

35] Shi Zhong, "Efficient Online Spherical K-means Clustering," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005*, vol. 5, Montreal, Canada, 2005, pp. 3180-3185.

36] Alfons Juan and Enrique Vidal, "Fast K-means-like clustering in metric spaces," *Pattern Recognition Letters*, vol. 15, no. 1, pp. 19-25, January 1994.

37] Dan Pelleg and Andrew Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters," in *ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning*, Stanford, CA, USA, 2000, pp. 727-734.

38] Renato Cordeiro de Amorim and Boris Mirkin, "Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering," *Pattern Recognition*, vol. 45, no. 3, pp. 1061–1075, March 2012.

39] Bogdan Georgescu, Ilan Shimshoni, and Peter Meer, "Mean Shift Based Clustering in High Dimensions: A Texture Classification Example," in *Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV'03)*, vol. 1, Nice, France, 2003, pp. 456-463.

40] Dorin Comaniciu and Peter Meer, "Mean Shift: A Robust Approach Toward Feature Space Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence* , vol. 24, no. 5, pp. 603-619, May 2002.

41] F Murtagh, "A Survey of Recent Advances in Hierarchical Clustering Algorithms," *The Computer Journal*, vol. 26, no. 4, pp. 354-359, 1983.

42] Lior Rokach and Oded Maimon, "Clustering Methods," in *Data mining and knowledge discovery handbook*, Lior Rokach and Oded Maimon, Eds.: Springer US, 2005, pp. 321-352.

43] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the Second Knowledge Discovery and Data Mining Conference*, vol. 96, Portland, OR, USA, 1996, pp. 226-231.

Ricardo J G B Campello, Davoud Moulavi, and Jörg Sander, "Density-Based

44] Clustering Based on Hierarchical Density Estimates," in *Advances in Knowledge Discovery and Data Mining*, vol. 2, Gold Coast, Australia, 2013, pp. 160-172.

45] Jiuh-Biing Sheu, "A fuzzy clustering-based approach to automatic freeway incident detection and characterization," *Fuzzy Sets and Systems*, vol. 128, no. 3, pp. 377–388, June 2002.

46] Gui-yan Jiang, Jiang-feng Wang, Xiao-dong Zhang, and Long-hui Gang, "The Study on the Application of Fuzzy Clustering Analysis in the Dynamic Identification of Road Traffic State ," in *Intelligent Transportation Systems, 2003. Proceedings*, vol. 1, Shanghai, China, 2003, pp. 408-411.

47] Sandor Dornbush and Anupam Joshi, "StreetSmart Traffic: Discovering and Disseminating Automobile Congestion Using VANET's," in *2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring*, Dublin, Ireland, 2007, pp. 11-15.

48] So Young Sohn and Sung Ho Lee, "Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea," *Safety Science*, vol. 41, no. 1, pp. 1-14, February 2003.

49] Chun-Hsin Wu et al., "An Advanced Traveler Information System with Emerging Network Technologies," in *Proceedings of the 6th Asia-Pacific Intelligent Transportation Systems Forum*, Taipei, Taiwan, 2003, pp. 230-231.

50] Waze. (2009) Waze. [Online]. https://www.waze.com

51] Vindu Goel, "Maps That Live and Breathe With Data," *New York Times*, p. B1, June 2013.

52] Roadify. (2010) Roadify. [Online]. http://www.roadify.com

53] Tiramisu Transit LLC. (2011) Tiramisu: the real-time bus tracker. [Online]. http://www.tiramisutransit.com/

54] John Zimmerman et al., "Field Trial of Tiramisu: Crowd-sourcing Bus Arrival Times to Spur Co-design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, Canada, 2011, pp. 1677-1686.

55] SMARTY. (2007) SMARTY. [Online]. http://www.smarty.toscana.it/

56] Giuseppe Anastasi et al., "Urban and Social Sensing for Sustainable Mobility in Smart Cities," in *Sustainable Internet and ICT for Sustainability 2013*, Palermo, Italy, 2013, pp. 1-4.

57] Jakarta Smart City. (2015) Jakarta Smart City. [Online]. http://smartcity.jakarta.go.id/

58] Dewanti A Wardhani, "Jakarta launches Smart City program," *The Jakarta Post*, p. 9, December 2014.

59] Vaninha Vieira et al., "The UbiBus Project: Using Context and Ubiquitous Computing to build Advanced Public Transportation Systems to Support Bus Passengers," Project Report 2011.

60] Toni Nummela. (2016, October) Suomi-Twitter. [Online]. http://www.toninummela.com/suomi-twitter/

61] Statistics Finland. (2016, October) Statistics Finland. [Online]. http://tilastokeskus.fi/tup/kunnat/kuntatiedot/837.html

62] Twitter Inc. (2016) Twitter Developer Documentation. [Online]. https://dev.twitter.com/rest/public

63] City of Tampere. (2016) Puhdistussuunitelmat. [Online]. http://www.puhdistussuunnitelmat.fi/tampere/kadut.htm

64] Liikennevirasto. (2016, April) Likkennevirasto. [Online]. http://alknet.tiehallinto.fi/alk/tietyot/tietyo_maak_14.html

65] Karoliina Lehtonen. (2016, September) Tamperelainen. [Online]. http://www.tamperelainen.fi/artikkeli/434877-lauantaivieras-tampereella-viihdytaan-sujuvasti

66] David L Olson and Dursun Delen, *Advanced Data Mining Techniques*, 1st ed.: Springer, February 2008.