FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

FEUP

Real-Time Sensing of Traffic Information in Twitter Messages

Sara Filipa Lemos de Carvalho

Master in Informatics and Computing Engineering

Supervisor: Rosaldo Rossetti (Ph.D)

Second Supervisor: Luís Sarmento (M.Sc.)

28^{th} June, 2010

Real-Time Sensing of Traffic Information in Twitter Messages

Sara Filipa Lemos de Carvalho

Master in Informatics and Computing Engineering

Approved in oral examination by the committee:

Chair: Luís Paulo Reis (Ph.D)
External Examiner: Paulo Cortez (Ph.D)
Supervisor: Rosaldo Rossetti (Ph.D)

27^{st} July, 2010

Abstract

Traffic issues affect the mobility of many people and the dynamics of the big urban centers. The study of the traffic urban networks is of great importance for the improvement routes and traffic flow.

This document intends to introduce a new source of information that can be helpful in the traffic scene analysis: microblogging messages. This new source surpasses some of the disadvantages of traditional traffic sensors, like area coverage and the costs of installation and maintenance.

The problem in focus in this study is to investigate whether information can be found, in microblogging messages, that is relevant to the traffic study. The microblogging platform used to collect the messages was Twitter and the objective was to retrieve messages shared by individual users, in opposite to official sources like news agencies.

The approach followed to solve the problem was to address it as a text classification problem using SVMs. The solution was divided into two main iterations of built and improvement of two classification models (bootstrapping strategy) to capture traffic messages. In each phase, for each model, the results achieved were registered and the evaluation measures were calculated and compared. Also, the improvements made from one phase to the other were registered.

In the end, the two classification models were able to capture a generic traffic message with a precision of more than 80% and a traffic message shared by an individual user with a precision of approximately 50%.

In conclusion, the objectives were achieved and the results were considered satisfactory in capturing messages from individual users, although in the capture of general traffic messages the results were a success.

Resumo

Problemas de trânsito afectam a mobilidade de muitas pessoas e a dinâmica dos grandes centros urbanos. O estudo das redes de tráfego urbano são, portanto, muito importantes para o melhoramento das estruturas e da eficiência de escoamento de trânsito.

Este documento pretende apresentar uma nova fonte de informação que pode ser útil para a análise e identificação de situações de trânsito: microblogging. A característica mais relevante desta fonte é o facto de conseguir ultrapassar alguns dos problemas dos sensores de trânsito tradicionais como por exemplo a limitada área de cobertura e os elevados custos de instalação e manutenção.

O problema abordado neste estudo é a verificar se é possível extrair informação relevante para o estudo do trânsito, de plataformas de microblogging. A plataforma escolhida para este estudo foi o Twitter e o objectivo é capturar as mensagens de trânsito que são partilhadas por utilizadores individuais, em oposição a mensagens partilhadas por agências noticiosas.

O problema foi abordado com um problema de classificaçâo de texto, usando SVMs. A soluçâo foi desenhada em duas iterações diferentes de construção e melhoramento de dois modelos de classificação diferentes (estratégia bootstrapping). Em cada fase, para cada modelo, foram recolhidos os resultados obtidos e calculados e comparados os valores de avaliação. Foram também assinalados os melhoramentos conseguidos de uma fase para outra.

No final, ambos os modelos conseguiram uma precisão na captura de mensagens de trânsito acima dos 80% e um precisão de 50% na captura de mensagens partilhadas por utilizadores individuais. Em conclusão, os objectivos do estudo foram alcançados e os resultados foram considerados satisfatórios no caso da captura de mensagem de trânsito partilhadas por utilizadores individuais, embora o resultado na captura de mensagem genéricas de trânsito seja considerado um sucesso.

Acknowledgements

In this page, it is a pleasure to address my thanks to the many people who made this thesis possible.

Firstly, I would like to thank my supervisors, Rosaldo Rossetti and Luís Sarmento, for the orientation and supervision of my work in the project and their many contributions, from the beginning to the end of the project, which made the work each day more interesting and captivating.

I would also like to express my gratitude to the institution this project was developed, Faculdade de Engenharia da Universidade do Porto, and the people who make part of this institution, specially the many lecturers whose classes I have attended.

I am grateful, as well, to the many people, teachers and others, who have influenced and encouraged me throughout the different school stages and made the learning process a fun and stimulating one.

A very special thanks is addressed to the friends made throughout the years for the endless moments of fun, support and caring. All these years would not have been the same without you.

Finally, and most importantly, I would like to thank all my family, for the loving, supporting and happy environment I grew up in and for always being the firsts to motivate, advise and guide me in the right direction. To you I dedicate this dissertation.

Sara Carvalho

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

Abbreviations

API	Application Programming Interface
FN	False Negative
FP	False Positive
FPR	False Positive Rate
ROC	Receiver Operator Characteristic
SVM	Support Vector Machines
TN	True Negatives
TP	True Positives
TPR	True Positive Rate
URL	Uniform Resource Locator

Chapter 1

Introduction

Traffic issues affect the mobility of many people and the dynamics in the big urban centers. The study of traffic is of paramount importance for the improvement of facilities and the generation of efficient routes.

This document presents a study that introduces a new information source to aid the traffic characterization problem: microblogging messages, a powerful new Social Media form that as revolutionized the communication between Internet users.

In this introductory section, the background and context of the problem are presented, along with the motivation, objectives and the methodological approach.

1.1 Context and Motivation

Traffic issues affect the mobility of many people and the dynamics in the big urban centers. Therefore, the study of the urban networks and its characteristics has been the focus of many studies for a long time. To study an urban network, several types of sensors and devices are placed along the network to gather data. This data is used to calculate such measures as the traffic density and flow, that allow a more clear understanding of the network at a given moment. This information is then utilized to improve the traffic flow and organization of the urban center.

The sensors and devices used to collect data include cameras, inductive loops and radars. Even though they are effective in their respective functions, these sensors have some limitations: their expensive to buy and repair, suffer damage easily (in the case of inductive loops), the covered area is constrained to some points in an network and the information they collect property of one organization/company. Moreover, the data they collect is usually specific, in a sense that it only collects data of one type, like vehicle count.

This document presents a study that introduces a new information source that can be relevant to the traffic characterization problem and surpasses some of the limitations of the traditional sensors described earlier: microblogging messages, a powerful new Social Media form that as revolutionized the communication between Internet users.

Microblogging is one of the forms of Social Media that has revolutionized the interaction between Internet users. It is a simple, agile form of communication that provides light-weight information updates and exchange of several sources. These messages can come from various communication channels such as cell phones, instant messaging, e-mail and the Web. The most famous microblogging platforms include Twitter[Twi], Tumblr[Tum], Plurk[Plu] and Jaiku[Jai].

Twitter is one of the applications that has experienced a remarkable popularity between Internet users and has transformed the way many people communicate around the world. As a result the contents of the messages shared between its users has become the focus of many studies: whether to determine the users intentions [JSFT07] and users' social interactions[HRW08] or, more recently, sentiment analysis and outcome predictions [AH10][TSSW10].

1.2 Objectives

The objective of the study addressed in this paper is the identification of messages that are relevant to the traffic characterization problem in a continuous flow of messages open to any subject. The focus is the capture of the user generated messages, in oppose to the messages broadcast by official sources like news agencies and traffic reporting agencies. These official sources report the traffic events that the traditional sensors capture and, therefore, are not the target of this investigation. The user generated messages represent a percentage of less than 0.05% in the Twitter universe of messages shared in Portuguese.

The capture of traffic messages from Twitter can be very useful for the traffic characterization purposes because Twitter overcomes some of the problems that exist with traditional traffic sensors. First, there are virtually no costs involved: these are software sensors that require practically no maintenance, since information is voluntarily communicated by users. Second, Twitter-based sensors allow obtaining information from potentially every point of the network, even those that are far from the main traffic axis. Third, Twitter users can describe a wide-range of traffic related events, which go beyond simple frequency counts. Finally, the information gathered is open to the public and can be accessed freely.

Additionally, the number of cars increases every year, and in 2007, in Portugal, there were 413 passenger cars for each 1000 inhabitants [Eur09]. The number of Internet users has also been growing every year and, in 2008, there were about 1.71 million users with mobile Internet access and around 1.58 million fixed Internet users, in Portugal. This was

the first year that the number of mobile broadband users surpassed the number of fixed broadband costumers [ANA08]. These two facts support the idea that information can be found in the Web that is relevant for the problem. More importantly, Keshav[Kes05] presents a study on how cell phones will dominate the future Internet, which is one more reason why more traffic news should become more frequent in the future.

1.3 Methodological Approach

Given a certain message from Twitter, a *tweet*, in the Twitter universe of messages in Portuguese, the objective is to identify the messages has being relevant or not. This is, therefore, a binary classification problem. To build the classifier, SVM[CV95][Vap95] was used, a powerful binary text classification technique that has proven to be very effective in problems with large feature spaces like this one[Joa98].

As the traffic related, user generated messages exist in a very small percentage in the Twitter universe ($> 0.05\%$), collect an appropriate corpus of positive examples could be a very time-consuming task. Therefore, for the initial positive training corpus to be fed to the classifier, official sources were used. These sources encompass messages that are broadcast from news agencies or traffic reporting agencies previously identified. The idea is for the classifier to capture the messages from non official sources using the official messages as a model.

1.4 Structure of the Document

This document is composed by five more Chapters. In Chapter 2 a bibliographic review will be presented on the areas that are relevant for this study and the areas for which this work is relevant. The next chapter, Chapter 3, will present the problem itself and its expected contributions. In Chapter 4 the approach and solution followed to solve the problem will be described and in Chapter 5 the results achieved with the implementation of the solution will be presented. In this chapter an comparative analysis of the results will also be made. In the last chapter, Chapter 6, the final achievements will be described, along with the further developments to be made to the project and the future work.

Chapter 2

Bibliographic Review

This Chapter presents a bibliographic review on the areas of interest and contribution of this study and it is mainly divided in three areas. The first one is the area of microblogging and social media, where special focus was given to Twitter and related work. The second area discussed is text classification, one of its techniques (Support Vector Machines) , evaluation measures (Precision-Recall, ROC) and improvement techniques (*Bootstrapping*). Finally, a review on traffic characterization is presented.

2.1 Microblogging

The World Wide Web is a fast-growing, global information center for news, advertisements, education and information services. The Web 2.0 phenomenon changed the Web paradigm and brought along the concepts of community and collaboration to applications. This transformed the regular website visitor into an application user that is encouraged to participate and contribute. Internet-based applications that embrace this philosophy are known as Social Media. The variety of Social Media websites include not only blogs, social networking and collaborative wikis, but also interest/hobby sharing applications and entertainment. Any of these applications provide a constant source of information and opinion gathering.

Microblogging is one of the most recent phenomena of the Web. Microblogging platforms allow users to share and broadcast light-weight information in the form of short written messages. The messages can be broadcast from several channels of communication, like short messages from cell phones, instant messaging, e-mail and the Web. The most famous microblogging platforms include Twitter[Twi], Tumblr[Tum], Plurk[Plu] and Jaiku[Jai].

2.1.1 Twitter

Twitter is one of the most popular microblogging services and was created in 2006 by Evan Williams, Biz Stone and Jack Dorsey. The philosophy behind the platform is simple: the user is invited to share with the community anything he wishes in less than 140 characters long. These messages are called *tweets*. Each user has a group of followers that receive all the *tweets* they share and receives all the *tweets* from the users they decide to follow. The plataform also allows its users to send direct responses to *tweets* from other users, through the use of *tags*, and group discussions under a certain topic, using *hashtags*. The *tags* are the representation of the Twitter name of an user and it usually starts with an '@'. A *hashtag* is usually a word that represents a certain topic and it is used within the message on that topic, being normally represented by the character '#' followed by the word or expression.

The Twitter community has been growing at a considerable pace since its creation [Pon] and its role in society has become relevant to many people and businesses [Joh]. In fact, Twitter has been helping the dissemination of information in a vast number of areas. In foreign political events, for example, Twitter was used in the conflicts that followed the Iranian Presidential election in 2009 and considered one of the most important means of comunication with the outside world, after several network services were restricted in the area [Vei][Mus]. In emergency situations, Twitter was used by the victims of the fires in California [Blob], in 2007, and Victoria [You], Australia, in 2009, to get accurate real-time information out and help other victims.

2.1.2 Related Work

Twitter is a very popular social media application, and has been so for the past three years. However, there have not been many published studies on the extraction of knowledge from this source in particular. In fact, it is even rarer to find studies that explore the user-generated content on a determined topic, as specific as traffic.

Huberman et al. [HRW08] presented a study on the social interactions between Twitter users, to conclude that its usage is driven by a sparse and hidden network of connections underlying the groups of friends and followers. Java and others [JSFT07] state that the main intentions for the use of microblogging are reporting daily activities and information exchange. It goes on to show that users with similar intentions connect with each other, by analyzing user intentions at a community level.

Sakaki and others [SOM10] present a system that uses Twitter as a *social sensor* for the real-time sensing of messages that report earthquake related events. This system is able to detect, with high probability, most of the earthquakes, with an seismic intensity scale greater than two, reported by the Japan Meteorological Agency, just by monitoring Twitter messages.

Asur and Huberman [AH10] demonstrate how social media, in particular Twitter, can be use to predict real-world outcomes. The study focuses in the prediction of box-office revenues for out-coming movies. In the end they conclude that a simple model that senses *tweets* on a particular topic can outperform some market-based predictors, therefore proving the forecasting power of social media. A similar study was done by Tumasjan and others [TSSW10], focusing in the predictions of Elections. The conclude that the mere number of *tweets* reflect voter preferences and comes close to election poles, and that the *tweets* are not only about spreading political opinions, but also to discuss these opinions with other users.

The finding of high-quality content in social media was studied by Agichtien et al. [ACD+08]. In the paper they investigate methods that allow the identification of high-quality content automatically by exploiting community feedback, such as links between items and explicit quality ratings from members of the community.

2.2 Text Classification

Text classification, which is also called text categorization, is a process of text analysis used to classify documents into a number of predefined categories. This is achieved through machine learning from examples which perform automate category assignments [HK06].

The documents are usually represented by vectors of attribute values, numeric or boolean. Each distinct word or group of words, in the document is a different feature and has associated to it the number of times it occurs in the document. This generates vectors with a very large dimension. To avoid this problem, some adjustments can be made to the feature identification, such as, only consider a feature words that appear more than 3 times or do not identify as feature grammatical conjunctions (like “and”, “or”, etc..) [Joa98] [Lew90].

The process of text classification can be seen as the following sequence of actions: feeding the classifier with pre-classified training data in order to build the classification scheme; refine and improve the classifier with a testing phase; and, finally, use the built scheme to classify new documents [HK06]. In the next sections it will be revised some of the basic techniques for text classification, that are relevant to the problem, such as Support Vector Machines and Naïve Bayes Classifiers.

2.3 Support Vector Machines

Support Vector Machines (SVMs) are a technique of classification and regression of both linear and nonlinear data, using supervised machine learning methods. It is a technique

that derived and belongs to the family of statistical methods for text classification [HK06]. It was originally introduced by Vapnik and Cortes [CV95][Vap95].

SVMs are binary classifiers, which means that they classify each document as belonging to one of two categories. The process of classification using SVMs works as the following. Primarily, a set of documents, represented in the form of feature vectors described in 2.3, are pre-classified as positive or negative examples. A positive example is an example of a document that we wish to identify, because it contains a certain number of features. A negative example is an example of any other type of information. These pre-classified examples are, then, fed to the SVM and during this time, it builds its classification model. After the training phase, the model is them prepared to classify new documents based on their features.

More formally, SVM construct a linear optimal hyperplane, during the training phase, separating the two categories. A visual examples can be found in Figure 2.1. The hyperplane contemplates the larger margin possible, in order to be more accurate. The hyperplane can be defined as

$$W.X + b = 0$$

where W represents the weight vector with n attributes, X is the respective values of the attributes in vector W, and b is a scalar, also referred to as bias [HK06]. For example[HK06], consider a vector of two attributes, $W = \{w_1, w_2\}$, and training tuples with the form $X = \{x_1, x_2\}$. Considering b as an additional weight w_0, the separating hyperplane is defined as

$$w_0 + w_1 x_1 + w_2 x_2 = 0.$$

Being y the class to which an examples belongs, the training data satisfies the following constrains:

$$w_0 + w_1 x_1 + w_2 x_2 \geq 1 \text{ for } y = 1$$

$$w_0 + w_1 x_1 + w_2 x_2 \leq -1 \text{ for } y = -1.$$

Joachims [Joa98] states that the reason SVMs work well for text categorization is because they acknowledge the properties of text in the documents. Namely, its large feature spaces, the dense concept vectors, with very few irrelevant features and the sparse instance vectors. It also shows the good performance of SVMs by comparison to other machine learning methods used for text categorization, like Naive Bayes[Zha04] and the Rocchio Algorithm[Roc71].

2.4 Evaluation Measures

In the field of machine learning and information retrieval, the notions of precision and recall are used to evaluate the model built, in addition to the accuracy results. In fact,

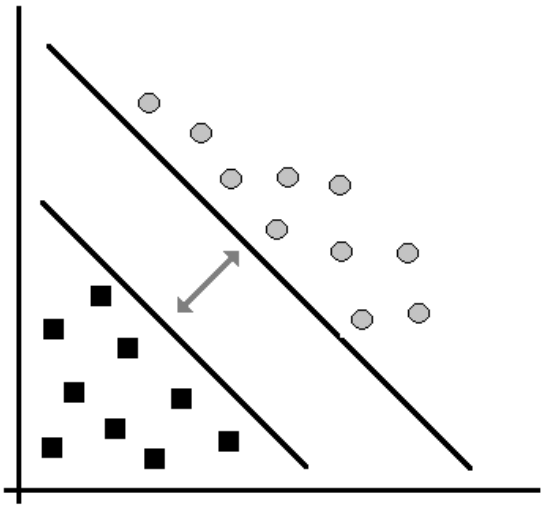

Figure 2.1: Support Vectors and the maximum separating hyperplane.

Provost et al.[PFK97] state that accuracy results are not enough for the evaluation and comparison of classification models. In this section, an overview of the notions of precision and recall will be presented. In binary text classification problem, each example is labeled by the classifier as positive or negative. The results achieved can be represented by the confusion matrix, which contemplates four different categories: the true positives (TP), the examples correctly classified as positives; the false positives (FP), the examples that are incorrectly classified as positives; the false negatives (FN), the examples that are incorrectly classified as negatives; and, finally, the true negatives (TN), the examples that are correctly classified as negatives [DG06][MS99]. In table 2.1, one can find a visual representation of the matrix.

Table 2.1: Confusion matrix in a binary classification problem

	actual positive	actual negative
predicted positive	TP	FP
predicted negative	FN	TN

Precision is the percentage of positive examples that the binary model classified correctly and is defined as

$$Precision = \frac{TP}{TP+FP}.$$

Recall is the percentage of positive examples existent in the test sample that the classifier identified as positives and is defined as

$$Recall = \frac{TP}{TP+FN}.$$

In most cases the relationship between Precision and Recall is inverse and, therefore, the cost of increasing one of these values results in the decrease of the other. The balance of the two measures of performance depends on the specific objectives of the study is to retrieve the maximum number of positive documents (maximum recall), at the expense of retrieving a large number of false positives, or to retrieve the maximum number of relevant results (maximum precision), at the expense of retrieving a small number of detected positives.

Another measure of performance often used in binary classification are the Receiver Operator Characteristic (ROC) curves, which present the proportion between the number of correctly classified positive examples (TP) and the incorrectly classified negative examples (FP) [DG06]. An ROC space is represented by the relationship between the true positive rate (TPR) and the false positive rate (FPR), which are defined as

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Davis and Goadrich [DG06] show that there is a deep connection between the ROC space and the Precision-Recall in such a way that a curve dominates in ROC space if and only if it dominated in Precision-Recall space.

2.5 *Bootstrapping*

In machine learning, the term bootstrapping refers to the technique of building a sampling distribution for a statistic by resampling from the data available. The technique translates into the iterative train and evaluation of a classifier in order to improve its performance. A known type of bootstrapping associated with text categorization is the co-training. Co-traning is the process of inducing a classifier using a small set of labeled data and a much larger set of unlabeled data (Blum & Mitchell[BM98], Nigam et al.[NMTM99]).

2.6 Traffic Characterization

Traffic can be defined as the movement of pedestrians, animals or vehicles in a route. In the 21th century, its study is more active in the resolution of the amount of traffic/capacity of the route problem [MS03]. More specifically, the vehicular traffic is described, by B. Kerner [Ker09], as a complex dynamic process associated with interactions between three major dynamic processes in a traffic network: (i) travel decision behavior, which determines travel demand, (ii) routing of vehicles, which is associated with traffic supply, and (iii) traffic congestion occurrences, which restricts free flow traffic.

Traffic engineering is the field of engineering that is responsible for such studies as route planning and design for the safe and efficient movement of pedestrians and vehicles. In *Traffic engineering design: principles and practice* [MS03]:

> "Traffic Engineering is used to either improve an existing situation or, in the case of a new facility, to ensure that the facility is correctly and safely designed and adequate for the demands that will be placed on it."

In the case of an existing situation, the study is done to collect information about the current situation of the traffic demands in the area. In the case of a new facility, the study is made to collect information about the future need of the facility and the needs of the surrounding area.

To collect the information about the needs of a certain area, in terms of routes, surveys are made. Traffic surveys gather all the processes used to collect data that provide and objective measure for those needs [MS03]. The measures only concern a certain moment in time and, therefore, there is the a constant need for surveys that analyse the traffic situation along the day, the week and, even the year. There are many types of surveys that can be made. From traffic counting methods, like volume counts or speed counts, to video collection systems or questionnaires and travel diaries, all can be considered traffic surveys. Its use depends on the type of data that is requires for the situation, but also the cost of taking such a study [JPR09].

To perform automatic studies in some given area, sensors are used. These include inductive loops, radars, video streams and infra-red sensors. In recent years there has been a grown focus in the use of video cameras. This last resource is particularly interesting because it allows to simultaneously record information of many types: traffic flow, turning movements, speeds and congestion and delays [MS03] [JPR09]. In 1994, Huang et al. [HKM+94], presented a prototype that successfully combined a vision-based surveillance system with a dynamic belief network dedicated to analysing traffic scenes. Later, Zhu and others [ZXY+00] have developed a cost-effective, real-time system for traffic monitoring, using 2D spacio-temporal image analysis. In 2004, Tai and others [TTLS04] present an image tracking system and its performance in traffic monitoring and accident detection at road intersections.

2.7 Summary

In this section a review of the state of the art for this study was presented. It started with a review on microblogging with a special focus on the Twitter platform. The section also presented a review of the work done around this platform, whether to study its characteristics and its users or to make use of the user generated content. The area of text classification is also presented in the next subsections. In particular, the technique

of classification, SVM, the most commonly used evaluation measures for classifiers and the performance improvement technique known as *bootstrapping*. In the end, there is a review on the traffic characterization topic. In the next Chapter, the problem and its contribution will be discussed, with reference to the investigation presented in this Chapter.

Chapter 3

Problem Statement

In this Chapter, the problem will be described and justified, with reference to the bibliographic review presented in Chapter 2. The first section includes a more detailed description of the study and its objectives, while Section 3.2 states the expected contributions of the study.

3.1 Problem Description

The main questions to be answered in this study are whether there is information present in microblogging messages that is relevant to the traffic characterization problem and to which degree efficiency can these messages be sensed. The aim is the identification of these messages in a continuous flow of messages, open to any subject.

As far as messages that contain traffic information are concerned, in this problem, they were divided into two categories:

- **Official traffic messages:** the messages that are shared by official sources of information, like news agencies or traffic forecasting agencies. These messages represent an estimated percentage of 0.02% of the Twitter universe of messages in Portuguese. The information these messages contain are a result of the information gathered by the traditional sensors and, therefore, suffer from the same limitations referred in Section 1.1. As a result these messages are not the focus of the investigation, but represent a model of the messages that contain relevant traffic information. An example of official traffic messages can be seen next:

  08:53, 29-04-2010, IP4, km 97, acidente circule com precaução, sentido Amarante.

  #diarioeconomico Trânsito muito condicionado no final do IC19 devido a acidente http://ow.ly/174EN1.

- **User generated traffic messages:** the messages that are shared by individual, independent users of Twitter. These messages represent an estimated percentage of 0.05% of the Twitter universe of messages in Portuguese. The information these messages contain can refer to any point of the traffic network and be about any type of traffic related event and, more importantly, the information comes directly from the users of the network. These messages are the focus of this problem and are the type of messages that the study aims to identify. An example of these messages can be seen next:

  O Eixo N-S está completamente parado no sentido norte.

  Mas que fila na a3, sentido porto braga.

As it can be seen in the messages above, there are considerable differences between the official and user generated traffic messages. The first group of messages presents a set of characteristics that allows an easy identification. These characteristics include the use of an *hashtag* in the messages or an specific *url* that identifies the source by its name, obeys to a certain structure and is usually written in a very neutral and objective way. The user generated messages, on the other hand, do not obey to a structure and present a much more free style of writing.

The Twitter platform was chosen by its popularity between Internet users and the importance it has achieved in the past few years as a new way of communication, as reviewed in Section 2.1.1. Also, it has been proven to be a usefull *sensor* for other eventś detection as reviewed in Section 2.1.2.

3.2 Expected Contributions

This scientific contribute of this study is mainly inserted into to two main areas: the traffic study problem and the social media content study.

The real-time sensing of traffic information in social networks is helpful for the study of the urban mobility, as it provides data that is relevant to the problem given by its users. This information can be helpful at identifying problematic areas that are not covered by the sensors network or identifying sporadic traffic issues, like stalled vehicles and accidents, or unusual high volume of cars in a certain point of the network.

Moreover, it can also overcome some of the disadvantages of conventional sensors. More specifically, the information is available for free and is open, as oppose to the information that is collected by most sensors, the area that is covered is not constraint but spread along the network, as opposed to the area covered by traditional sensors, and the financial investment that has to be made is considerably smaller than the one required

with most traditional sensors. In sum, the aim, in the traffic studies area, is to present microblogging messages as another source of information that can be relevant to the study of urban mobility.

As far as the social media content study is concerned, there has been a growing interest in it with the bloom of the social media applications in the past decade. This study is relevant to this area as it is applied to the automatic detection of traffic information, a topic that is not known as being common or popular in social-networks[Bloa], and therefore expected to have fewer contributions. However, as discussed in chapter one, there is also a growing potential for this subject in microblogging messages, as the number of drivers is each year bigger [Eur09] and mobile phones with Internet access will, according to some studies, play a big and important part in Internet accesses[Kes05].

3.3 Summary

In this Chapter the problem was described and its objectives stated, along with the expected contributions of the study. The problem of finding information relevant to the traffic characterization problem, in messages shared by Twitter users, is an innovative use for this user content generator. In the traffic scene analysis and characterization this presents itself as a source of information that overcomes some of the disadvantages of the traditional sensors. In the next Chapter, the approach and solution followed to solve the problem will be described.

Chapter 4

The Proposed Solution

This section presents description of the solution followed for the problem described in Chapter 3. The short summary of the methodological approach for this problem has already been presented in Section 1.3. In this section, the approach to the problem will be explained as well as the solution architecture drawn to solve the problem.

The aim of this study is the identification of messages that contain traffic related information in a continuous flow of messages. This means that given a certain messages, the objective is to determine whether the messages are traffic related, or not. Therefore, this is a binary text classification problem.

There are several algorithms for text classification, which, given the appropriate training set, achieve relatively high results. In this problem we opted for using SVM, which is a binary classification technique, as reviewed in Section 2.3, that acknowledges the properties of text and, therefore, if very efficient when dealing with large feature spaces and sparse attribute vectors.

4.1 Dataset Definition

As referred in Section 3.1, one of the main challenges of the problem is generating a representative dataset for such an unbalanced problem (the percentage of relevant messages in the Twitter stream is less than 0.05%). For this reason, the manual selection of positive examples for the problem among the messages available for training was considered an unfeasible task.

However, as explain in Section 3.1 there are several official sources that report traffic related events into the Twitosphere. These messages present a strict format and language, much unlike the user-generated messages, which tend to be very loose in terms of grammatical structure and usually contain many spelling mistakes, non-standard punctuation,

emoticons among many other characteristics. Nevertheless, the traffic messages from official sources provide good examples of the standard vocabulary for traffic related events (e.g. names of routes and critical locations in the traffic network) and can be very easily identified in the Twitosphere. For these reasons, the training strategy will use these easily identified messages as positive examples in the first iteration, together with a balanced number of negative examples chosen randomly from a large collection of messages in the Twitosphere.

4.2 Bootstrapping Strategy

The training and construction of the classification models consisted in a two iteration bootstrapping strategy.

The first iteration consisted in training the classifier with the official traffic messages, as positive examples, and randomly picked messages from the Twitosphere, as negative examples. After the training, the first version of the classification model is obtained.

Following the bootstrapping approach, this version of the classifier just described is used to find additional positive and negative examples that can be used to enrich the initial training set and build a better classification model after the second training iteration. The messages classified as relevant by the first version of the classifier were, then, manually divided into three categories:

1. Relevant traffic messages generated by human, anonymous users, which are the focus of the problem;

2. Traffic related messages that are generated by an official source, which are not the focus of the problem;

3. Messages that are not traffic related and have been incorrectly classified as positive by the classification model.

To build the second iteration training set, (1) were added as additional positive examples, while (3) were used as negative examples. The messages in (2) were not included in the training set because they can neither be considered a positive example, as they are not the focus of the problem, or a negative example, as they are traffic related. With this new training set, the second version of the classification model is build and is expected to achieve better performance than the first one.

The bootstrapping process is represented in Figure 4.1

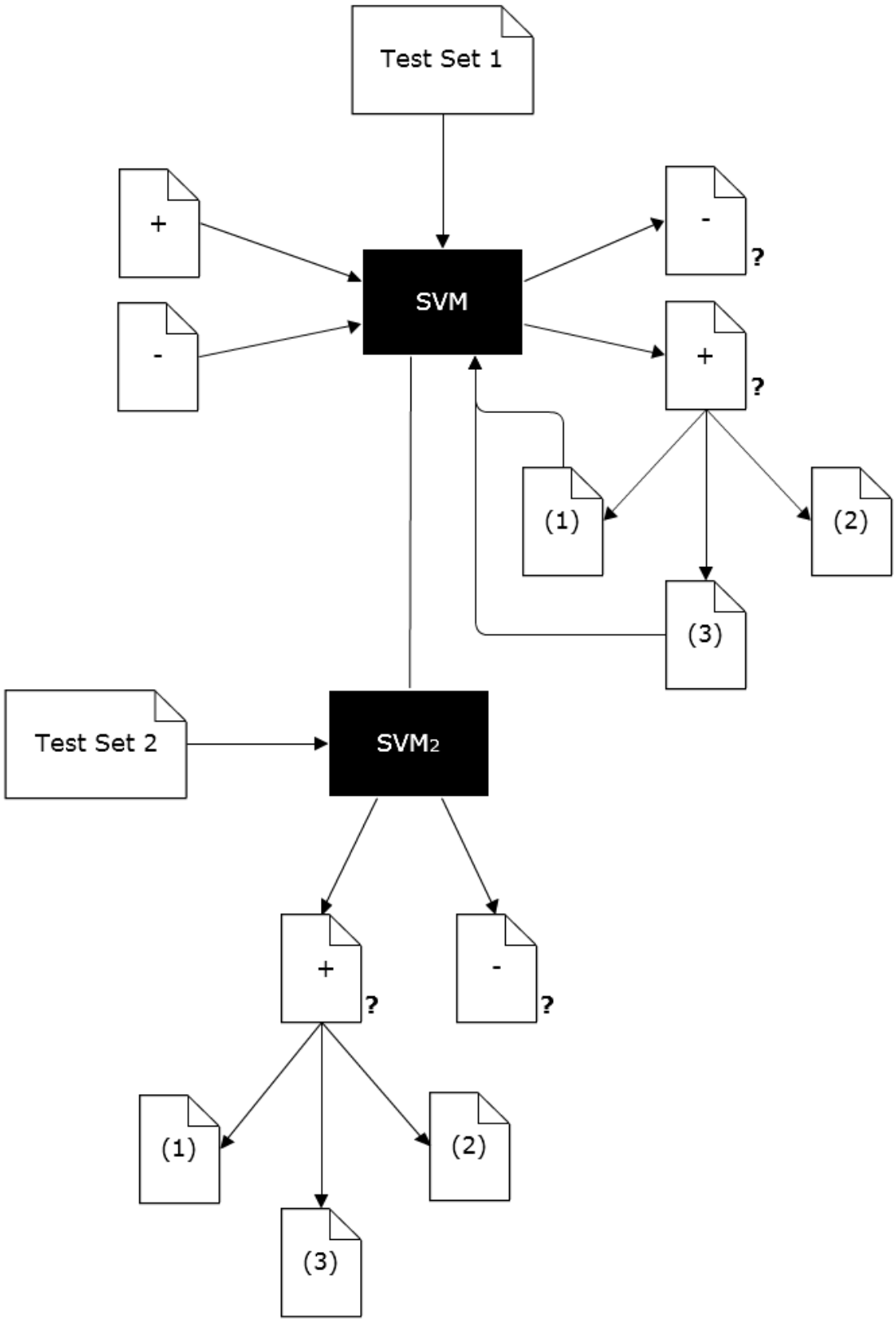

Figure 4.1: Bootstrapping strategy followed (two iterations).

4.3 Experimental Set-Up

In this section, it will be described the set-up in which the experiment occur in terms of the datasets used, the classification set-up, the features chosen for the problem and evaluation measures used to assess the models.

4.3.1 Datasets

The experiments were performed over a dataset from Twitter (*tweets*) of approximately 565,000 messages, collected between March and April of 2010. The messages were mostly written in Portuguese and were provided by Sapo[1].

In the initial dataset, the positive examples consisted in approximately 3,300 manually chosen messages from previously identified official sources. The negative examples consisted on a random sample of approximately 41,000 messages from the initial collection of messages from the total dataset. This corresponded to a ratio of approximately one positive example for 12 negative examples. This configuration was thought to represent in some way the unbalanced situation in the case of traffic-relates messages.

4.3.2 Classification Set-Up

As mentioned above, the classification technique chosen for the experiment was SVM. The SVM software used was provided by the Weka[HFH+09] toolkit, together with the LibSVM[CL01] library. The model of the classifier built worked as a black-box, which means that it is only possible to control the inputs and receive the outputs.

In the beginning of the project, some preliminary experiments were performed to decide which SVM algorithm to use. In the end, the results shown that the most adequate was the linear kernel function because, although polynomial and radial kernels obtained similar results, this approach saved considerable computational time.

4.3.3 Feature Definition

For these experiments, two types of features were used to vectorize the messages:

- **Uni-grams:** where each distinct word was a feature.
- **Bi-grams:** where each two consecutive distinct words were considered a feature.

Before the messages were fed to the classification model, they went through a pre-processing state. This process involved removing the *stop words* (i.e. words with less and four characters written in lowercase) and punctuation marks. Also, to reduce scarcity, all the tokens in the message were lowercased.

[1] *www.sapo.pt*, the main Portuguese ISP

4.3.4 Evaluation

Most of the experiments in the field of text classification use the k-fold cross-validation process to evaluate the performance of a classification model. However, in this case, the messages used as positive examples in the training set (i.e. traffic messages from official sources) are not the messages in which this study is focuses (i.e. the human, anonymous users traffic related messages). Therefore, this type of evaluation is not the most suitable for this problem.

Instead, the evaluation process followed consisted in having the classification model process an certain number of unlabeled data and then proceed to the manual verification of the results. The messages, from the large collection of unlabeled data, that were classified as positive by the classification model were, in the manual verification process, divided in the three categories defined in Section 4.2. According to that definition the messages were divided into true positives (TP) and false positives (FP) as follows:

- TP_{ALL} = (1) + (2)
- TP_{HMN} = (1)
- FP_{ALL} = (3)
- FP_{HMN} = (2) + (3)

The number of false negatives in both scenarios (FN_{ALL} and FN_{HMN}) were estimated based on the size of the dataset and the percentage of their existence in the Twitosphere.

With this values it becomes possible to calculate the evaluation measures of the classification model performance, Precision(P) and Recall(R):

$$P_{ALL} = \frac{TP_{ALL}}{TP_{ALL} + FP_{ALL}}$$

$$R_{ALL} = \frac{TP_{ALL}}{TP_{ALL} + FN_{ALL}}$$

$$P_{HMN} = \frac{TP_{HMN}}{TP_{HMN} + FP_{HMN}}$$

$$R_{HMN} = \frac{TP_{HMN}}{TP_{HMN} + FN_{HMN}}$$

In every experiment, in order to decrease the number of false negatives, it was imposed a minimum threshold, th_{min}, for the positive decision. For every experiment, it was tested at least more than one value for this parameter.

4.4 Summary

In this section, the approach chosen to solve the problem presented in Chapter 3 was described along with the solution architecture and the experimental set-up to study the results of the experiment. The approach consists in a bootstrapping strategy with two iterations to train a binary text classification model that uses SVM. In the next chapter, Chapter 5, the results of the solution presented will be described.

Chapter 5

Results

This section presents the results of the experiment described in Chapter 4, as well as the experiments that determined the difficulty of the problem. The results will be presented according to the evaluation measures defined in Section 4.3.4.

5.1 Problem Complexity

This experiment consisted in collecting approximately 4,000 randomly picked messages from the total dataset of messages and, then, manually classify them to fit one of the three categories described in Section 4.2. The results of the manual verification allowed the estimation of the percentage of existence of messages from each kind in the Twitosphere. The percentages are presented in Table 5.1.

Table 5.1: Traffic *tweets*, manually classified, on a randomly picked sample.

Messages	(1) + (2)	(1)
4092	0.07%	0.05%

The percentage of messages that contain traffic information, written in Portuguese, is estimated to be 0.07% and 0.05% is the percentage of traffic messages sent by human anonymous sources, which are the focus of the problem.

5.2 Iteration One

In this section, the results concerning the first iteration of the bootstrapping strategy will be presented. For this iteration, which used the messages shared by official sources as positive examples, the results will be presented for several values of the decision threshold, th_{min}, using a test sample of approximately 260,000 unlabeled messages from Twitter.

5.2.1 Uni-grams based classification model

Tables 5.2 and 5.3 show the results achieved by the uni-grams based model in iteration one.

Table 5.2: Absolute results for the uni-grams based classification model in iteration one.

th_{min}	$TP+FP$	TP_{ALL}	TP_{HMN}
0.60	62	20	13
0.70	50	15	13

Table 5.3: Evaluation results for the uni-grams based classification model in iteration one.

th_{min}	P_{ALL}	R_{ALL}	P_{HMN}	R_{HMN}
0.60	32.26%	10.99%	20.97%	10.00%
0.70	30.00%	8.24%	26.00%	10.00%
Average	31.13%	9.62%	23.49%	10.00%

5.2.2 Bi-grams based classification model

Tables 5.4 and 5.5 show the results achieved by the bi-grams based model in iteration one.

Table 5.4: Absolute results for the bi-grams based classification model in iteration one.

th_{min}	$TP+FP$	TP_{ALL}	TP_{HMN}
0.60	54	12	5
0.70	47	12	5

Table 5.5: Evaluation results for the bi-grams based classification model in iteration one.

th_{min}	P_{ALL}	R_{ALL}	P_{HMN}	R_{HMN}
0.60	22.22%	6.59%	9.26%	3.85%
0.70	25.53%	6.59%	10.64%	3.85%
Average	23.88%	6.59%	9.13%	3.85%

5.3 Iteration Two

In this section, the results concerning the second iteration of the bootstrapping strategy will be presented. For this iteration, that used the messages labeled by the classification model build in iteration one to enlarge the training dataset, the results will be presented for the same values of the decision threshold, th_{min}, in the previous section using a test sample

of approximately 260,000 unlabeled messages from Twitter (distinct from the ones used in iteration one).

5.3.1 Uni-grams based classification model

Tables 5.6 and 5.7 show the results achieved by the uni-grams based model in iteration two.

Table 5.6: Absolute results for the uni-grams based classification model in iteration two.

th_{min}	$TP+FP$	TP_{ALL}	TP_{HMN}
0.60	47	41	21
0.70	37	35	20

Table 5.7: Evaluation results for the uni-grams based classification model in iteration two.

th_{min}	P_{ALL}	R_{ALL}	P_{HMN}	R_{HMN}
0.60	87.23%	22.56%	44.68%	16.18%
0.70	94.59%	19.26%	54.05%	15.41%
Average	90.91%	20.91%	49.37%	15.79%

5.3.2 Bi-grams based classification model

Tables 5.8 and 5.9 show the results achieved by the bi-grams based model in iteration two.

Table 5.8: Absolute results for the bi-grams based classification model in iteration two.

th_{min}	$TP+FP$	TP_{ALL}	TP_{HMN}
0.60	30	24	13
0.70	23	19	13

Table 5.9: Evaluation results for the bi-grams based classification model in iteration two.

th_{min}	P_{ALL}	R_{ALL}	P_{HMN}	R_{HMN}
0.60	80.00%	13.20%	43.33%	10.01%
0.70	82.61%	10.45%	56.52%	10.01%
Average	81.30%	11.83%	49.93%	10.01%

5.4 Overview

This section compares and evaluates the improvements made in the classifiers built in iteration one and iteration two of the bootstrapping strategy. When looking at the results,

it is important not to forget that the percentage of existence of traffic related messages shared by anonymous, independent users is very low (> 0.05%) in the universe of Twitter messages being processed.

In general, the improvement is significant from one iteration to the other and the precision and recall values increase in every case. The additional labeled data used to build the second version of the classifiers in iteration two consisted in 15 positive examples - traffic related messages from independent, anonymous users - and 75 negative examples - misclassified messages. This increase is, after analysis, more justified by the addition of the negative examples rather than by the addition of the negative examples. These negative examples allowed the classifier to become more robust to certain misleading messages that were not traffic related but contained a certain number of keywords in common with the traffic vocabulary.

Figure 5.1 shows the evaluation measures of the uni-grams based model in both iterations. One can see that precision, in the case of the detection of traffic related messages from independent and anonymous users, achieves nearly 50% and the correspondent value of recall nearly doubles from iteration one to iteration two.

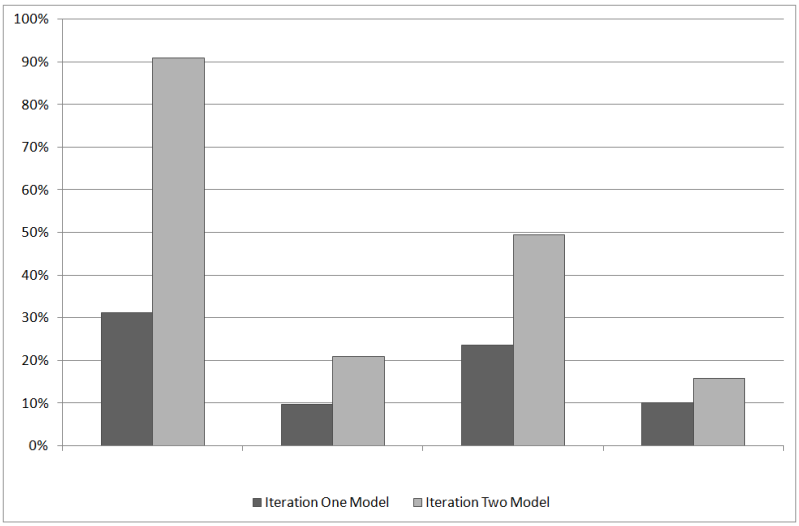

Figure 5.1: Precision-Recall values of the uni-grams model in phase one and two.

Figure 5.2 shows the evaluation measures of the bi-grams based model in both iterations. The results are similar to the ones achieved by the uni-grams based model.

In Figure 5.3, there is a comparison between the uni-grams and bi-grams based classification models, in the end of iteration two. The overall results of the uni-grams based model are better than the results of the bi-grams based model.

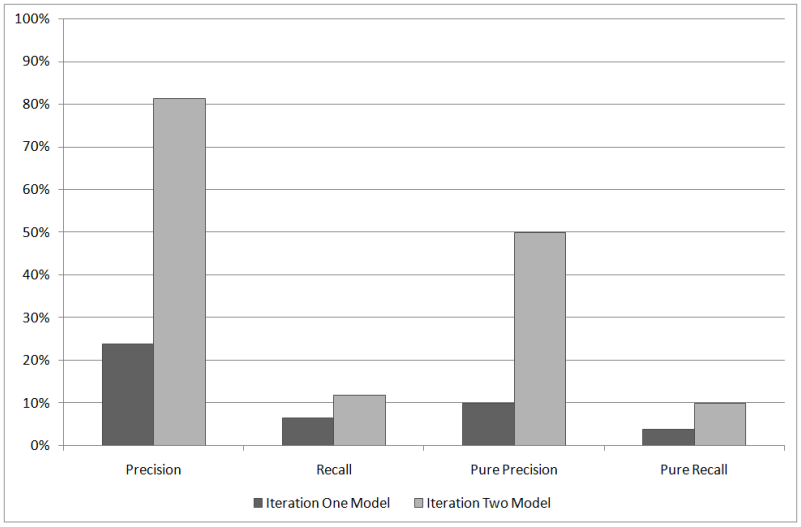

Figure 5.2: Precision-Recall values of the bi-grams model in phase one and two.

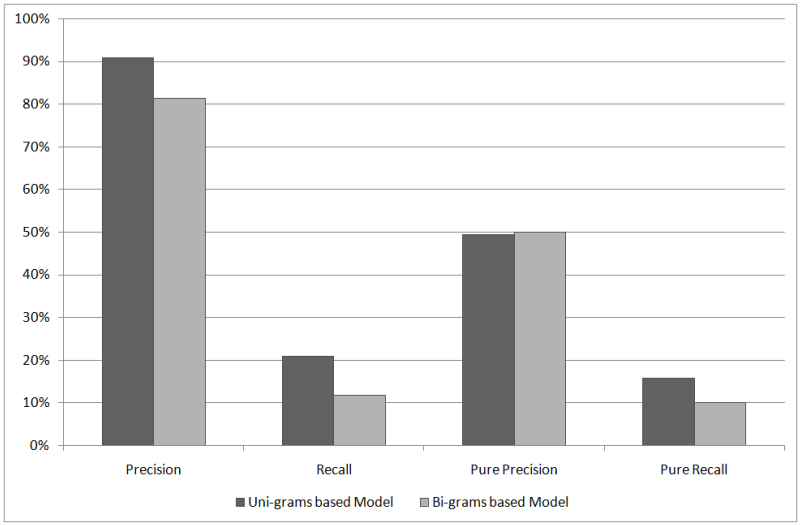

Figure 5.3: Precision-Recall values of the uni-grams and bi-grams model in phase two.

5.5 Summary

In this Chapter the results of the implementation described in Chapter 4 are presented. In each Section, a different iteration of the project is presented. In iteration one and iteration two, both classification models are assessed in terms of absolute results (number of *tweets* captured) and evaluation measures (precision and recall for the capture of traffic messages and *pure* traffic messages).

Chapter 6

Conclusion

The problem described in chapter 3 has been solved according to the solution presented in chapter 4 and the results achieved from this approach have been presented in chapter 5. In this last chapter, in Section 6.1, some last remarks will be made about this study, as well as a summary of the main achievements and results. In Section 6.2 the further improvements to the project will be presented and in Section 6.3 the future works will be described.

6.1 Final Remarks

As mentioned in the Section 1, the characterization of a traffic network is of great importance for the improvement of traffic flow and an efficient mobility plan within the urban area. The information shared by users through microblogging represents an additional source of information that can the relevant to this problem. Moreover, this innovative *traffic sensor*, is not constraint to certain areas of the network and can contribute with information for traffic scene analysis and the identification of new problematic areas.

The problem of sensing messages, shared by Twitter users, as been solved has a text classification problem using SVMs. Two models were built, the first used as feature, uni-grams, while the other used bi-grams. The main objective was to sense the traffic messages from individual, anonymous users, using the offical traffic messages as a model for training the classifier.

The construction of the classification models was divided into two iteration of a bootstrapping strategy and in the end of iteration two the model achieved satisfactory results. The classification model that used bi-grams as features, reached a precision of approximately 80%, in the case of traffic *tweets* and a precision of 50%, in the case of the traffic *tweets* from independent users. The classification model that used uni-grams, presented

similar results, reaching a precision of approximately 90% in the case of traffic results and a precision of nearly 50% in the case of traffic *tweets* from independent users. The values of recall have also increased, in fact doubled in some cases, but are still under the 25% margin. Unlike must binary classification problems, with this problem it was decided not to level the values of precision and recall and the reason for this was the focus given to precision. Since traffic is such a specific topic, leveling the two values would result in many false positives, which was something to avoid.

The resolution of this problem proved that information related to traffic can be found in microblogging messages that is relevant to the traffic characterization problem. The results are especially important if one keeps in mind that the *tweets* with traffic information, shared by individual users, represent an estimate 0.05% of the universe of *tweets* shared in Portuguese.

6.2 Further Developments

The main improvements and additions that can be made to the final work presented are divided into 3 groups: the improvement of the classification model, the addition of an application for the capture of messages from Twitter and connection with the classifier and the automatic detection of the true positives, by feature extraction.

The first group is related with the improvement of the classification models. Two types of changes can be made to the classifiers, that could improve its performance. The first is to perform another iteration of bootstrapping following iteration two of the solution, using another training file collected from the previous testing phase, to increase the precision and recall of the *pure* traffic messages. The second improvement is related with training examples. The classification models received a much higher percentage of negative examples than positive ones and, therefore, the model is *pessimist*, meaning that is more likely that an example is classified as negative, than positive. To revert some of the *pessimism*, in another retraining phase, the model should be fed with more positive examples never, however, making the number of positives even with the negative examples, because the number of positives in the Twitter universe is much smaller than the negatives.

The second group of improvements is the addition of an application that would collect the messages from Twitter over frequent intervals, using Twitter Search Api[1], and, then, fed them to the model and presented the results of the classification process.

The last group of improvements can be seen as an addition to be used over the results of the classification process to automatically determine which of the classified *tweets* actually contain relevant information, for example the extraction of a location. This process is known as feature extraction and selection for text classification[Lew92].

[1]Twitter Search Api, http://apiwiki.twitter.com/Twitter-API-Documentation

6.3 Future Works

In this section it will be discussed the projects that can be derived from this study as well as the applications and domains that can be dealt using the same methodological approach.

The methodological approach was elaborated to the automatic capture messages that do exist in a small percentage in the microblogging universe. Therefore, it is natural to say that this approach will work with most subjects that are in the same conditions, i.e. the information that is intended to be captured exists in a very low percentage in the microblogging universe, provided there is a function for the data collected.

In the future, this project could make part of a much widespread project of capturing messages from microblogging that report events. The objective of such a project would be to capture information on where large volumes of people are located, and it would encompass several smaller text classification projects for different types of events. One of them would be traffic, which associated, for example, with a project that captures public events, could predict where an unusual amount of people would be located at one moment.

Conclusion

References

[ACD+08] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 183–194, New York, NY, USA, 2008. ACM.

[AH10] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. Mar 2010.

[ANA08] ANACOM. Number of mobile broadband users surpasses number of fixed broadband users for first time. ANACOM, 2008.

[Bloa] Twitter Blog. Top twitter trends of 2009. http://blog.twitter.com/2009/12/top-twitter-trends-of-2009.html. Last visited: 26-06-2010.

[Blob] Andy Bloxham. Facebook more effective than emergency services in a disaster. The Daily Telegraph, December 20, 2008.

[BM98] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, New York, NY, USA, 1998. ACM.

[CL01] Chih C. Chang and Chih J. Lin. *LIBSVM: a library for support vector machines*, 2001.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[DG06] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *In ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM Press, 2006.

[Eur09] Eurostat. Energy, transport and environment indicators, 2009 edition. European Comission, 2009.

[HFH+09] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[HK06] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, 2nd edition, 2006.

[HKM+94] T. Huang, D. Koller, J. Malik, G. Ogasawara, B. Rao, S. Russell, and J. Weber. Automatic symbolic traffic scene analysis using belief networks. In *PROCEEDINGS 12TH NATIONAL CONFERENCE IN AI*, pages 966–972. AAAI Press, 1994.

[HRW08] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. 2008.

[Jai] Jaiku. http://www.jaiku.com/. Last visited: 26-06-2010.

[Joa98] Thorsten Joachims. Text categorization with support vector machines: learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[Joh] Steve Johnson. How twitter will change the way we live. Time, June 05, 2009.

[JPR09] Brian Slack Jean-Paul Rodrigue, Claude Comtois. *The Geography of Transport Systems*, chapter 4. Routledge, 2009.

[JSFT07] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM.

[Ker09] Boris S. Kerner. *Introduction to Modern Traffic Flow Theory and Control: The Long Road to Three-Phase Traffic Theory*, chapter 1. Springer, 2009.

[Kes05] S. Keshav. Why cell phones will dominate the future internet. *SIGCOMM Comput. Commun. Rev.*, 35(2):83–86, 2005.

[Lew90] David D. Lewis. Representation quality in text classification: An introduction and experiment. In *In: Proceedings of Workshop on Speech and Natural Language. Hidden*, pages 288–295. Morgan Kaufmann, 1990.

[Lew92] David D. Lewis. Feature selection and feature extraction for text categorization. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 212–217, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

[MS99] Christopher D. Manning and Hinrich Schuetze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1 edition, June 1999.

[MS03] Peter Guest Mike Slinn, Paul Matthews. *Traffic Engineering Design: Principles and Practice*, chapter 1 and 2. Elsevier Butterworth-Heinemann, 2nd edition, 2003.

[Mus] Mike Musgrove. Twitter freedom's only link in iran. The Washington Post, July 09, 2009.

[NMTM99] Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. In *Machine Learning*, pages 103–134, 1999.

[PFK97] Foster Provost, Tom Fawcett, and Ron Kohavi. The case against accuracy estimation for comparing induction algorithms. In *In Proceedings of the Fifteenth International Conference on Machine Learning*, pages 445–453. Morgan Kaufmann, 1997.

[Plu] Plurk. http://www.plurk.com/. Last visited: 26-06-2010.

[Pon] Jason Pontin. From many tweets, one loud voice on the internet. The New York Times, April 22, 2007.

[Roc71] J. Rocchio. *Relevance Feedback in Information Retrieval*, pages 313–323. 1971.

[SOM10] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 851–860, New York, NY, USA, 2010. ACM.

[TSSW10] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *AAAI-10: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Atlanta, USA, 2010. AAAI.

[TTLS04] Jen-Chao Tai, Shung-Tsang Tseng, Ching-Po Lin, and Kai-Tai Song. Real-time image tracking for automatic traffic monitoring and enforcement applications. *Image and Vision Computing*, 22(6):485 – 501, 2004.

[Tum] Tumblr. http://www.tumblr.com/. Last visited: 26-06-2010.

[Twi] Twitter. http://twitter.com/. Last visited: 26-06-2010.

[Vap95] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

[Vei] E. Veiszadeh. Twitter freedom's only link in iran. The Australian, July 16, 2009.

[You] Emma Young. Crisis puts a new face on a social networking. The Sydney Morning Herald, June 07, 2009.

[Zha04] Harry Zhang. The optimality of naive bayes. In Valerie Barr, Zdravko Markov, Valerie Barr, and Zdravko Markov, editors, *FLAIRS Conference*. AAAI Press, 2004.

[ZXY+00] Zhigang Zhu, Guangyou Xu, Bo Yang, Dingji Shi, and Xueyin Lin. Visatram: a real-time vision system for automatic traffic monitoring. *Image and Vision Computing*, 18(10):781 – 794, 2000.

REFERENCES

Appendix A

Publications

The following papper was accepted in the ATSS Workshop, IEEE ITSC 2010.

Real-Time Sensing of Traffic Information in Twitter Messages

Sara Carvalho Luís Sarmento Rosaldo J. F. Rossetti, *Member, IEEE*

***Abstract*— This paper presents an initial attempt to use micro-blogging messages posted on Twitter (by users in transit) to perform real-time sensing of traffic-related information. We propose a text classification approach to the problem: we wish to automatically identify traffic-related messages posted on Twitter, among the millions of unrelated messages posted by users. Given that the fraction of relevant traffic messages on Twitter is extremely low ($< 0.05\%$), the main challenges involved at this stage are (i) creating a suitable training set for setting up the classifier, and (ii) driving the classifier to a reasonable level of precision in identifying relevant messages. We opted for a dual-stage bootstrapping strategy for tackling both these problems simultaneously. First, we used short message reports that are automatically posted on Twitter by certain official news source to compile an initial set of training messages that are comparable in contents to the user-generated messages we wish to identify. Then, using a classifier trained on such robot-sent messages, we process a large collection of Twitter messages to identify traffic-related messages set by users, which are then added to the training set and are used to train a second version of the classifier. Results show that, despite the highly-unbalanced example distribution, we are able to almost double the performance of the classifier from the first iteration to the second, and that F-measures above 23% in identifying relevant traffic-related Twitter messages can be achieved, with little human effort involved in creating a training set.**

I. INTRODUCTION

The study of an urban network involves the placement of several types of sensors and devices along the network to gather data. Such data is used to calculate measures, such as the *traffic density* and *flow*, which allow a more clear understanding of the network at a given moment. This information is then used for improving the traffic flow and organization of the urban center.

The sensors and devices used to collect data include hardware such as cameras, inductive loops and radars. Even though they are effective in their respective functions, these sensors have some limitations. First, they are relatively expensive, and require constant maintenance. Second, they are not mobile and cover only a restricted area of the network. Third, these sensors tend to be very specialized, i.e., they only collect data of one particular type (e.g., vehicle count). Finally, the information they collect is usually property of one organization or company, and it is usually difficult to obtain access to it, even for the purpose of research. We thus propose exploring an alternative information source that can be useful to the traffic characterization problem, and, simultaneously, overcomes some of the limitations of the traditional sensors described earlier: Twitter messages.

Twitter is one of the applications that has experienced an increasing popularity among Internet users and has deeply transformed the way people communicate around the world. The Twitter community has been growing at a considerable pace since its creation [1] and its role has become very important in many social, economic and even political contexts [2]. For example, Twitter was used to report the conflicts that followed the Iranian Presidential election in 2009, despite the active censorship performed by the Iranian authorities over more traditional media [3][4]. Twitter was also used by the victims of the fires in California [5], in 2007, and Victoria [6], Australia, in 2009, to report accurate information in real-time and to help other victims.

Likewise, due to their real-time and ubiquous nature, we believe that Twitter messages can be used for a variety of other purposes, such as, for example, as a complementary source of relevant and up-to-date traffic information. In fact, anonymous users sometimes post messages that describe relevant traffic information, such as, for example: *This traffic jam doesn't seem to have an ending. I'm still stuck in EN12.*

Therefore, the identification of this type of messages may be very useful for traffic characterization purposes, especially because Twitter allows overcoming some of the problems that exist with other sensors. First, there are virtually no costs involved: these are software sensors that require practically no maintenance, since information is voluntarily communicated by users. Second, Twitter-based sensors allow obtaining information from potentially every point of the network, even those that are far from the main traffic axis. Third, Twitter users can describe a wide-range of traffic related events, which go beyond simple frequency counts. Finally, the information gathered is open to the public and can be accessed freely.

Official traffic information sources also use Twitter to broadcast information. Several Twitter "users" are in fact robots from news agencies the inject messages containing traffic information in the Twitosphere, such as: *08:53, 29-04-2010, IP4, km 97, direction west, accident, drive safely.*

Obviously, these robot-sent messages are not interesting from the point of view of sensing traffic information (they report information previously sensed and analyzed). The problem we address in this paper is the identification of messages relevant to the traffic characterization sent by *anonymous users* (i.e. not by robots).

There are the two main challenges in this work. First, from the text classification point of view, this problem is highly *unbalanced*. By manually sampling a large collection

Faculdade de Engenharia da Universidade do Porto, Departamento de Engenharia Informática, Laboratório de Inteligência Artificial e Ciência de Computadores, Rua Dr. Roberto Frias, s/n 4200-465, Porto, PORTUGAL sara.carvalho.l@gmail.com las@fe.up.pt rossetti@fe.up.pt

of Twitter messages, we were able to estimate that the percentage of traffic messages sent by anonymous users (i.e. excluding messages automatically posted by official traffic agencies) correspond to less than 0.05% of the messages at stake. This means, that it should be extremely difficult to achieve high precision in the identification of relevant Twitter messages. Secondly, as a consequence of this highly unbalanced distribution, we face the additional challenge of creating an appropriate and representative dataset for training the classifiers. With such low ratio of positive cases, manual annotation of a balanced corpus becomes unfeasible, so alternative strategies have to be devised. In this paper, we focus on these two challenges.

II. RELATED WORK

Twitter is a very popular social media application, and has been so for the past three years. However, there have not been many studies exploring the user-generated content on issues related to traffic or mobility.

Sakaki and others [7] present a system that uses Twitter as a *social sensor* for the real-time sensing of messages that report earthquake related events. This system is able to detect, with high probability, most of the earthquakes, with a seismic intensity scale greater than two, reported by the Japan Meteorological Agency, just by monitoring Twitter messages.

Asur and Huberman [8] demonstrate how social media, in particular Twitter, can be use to predict real-world outcomes. The study focuses on the prediction of box-office revenues for out-coming movies. In the end they conclude that a simple model that senses *tweets* on a particular topic can outperform some market-based predictors, therefore proving the forecasting power of social media. A similar study was done by Tumasjan and others [9], focusing in the predictions of Elections. They conclude that the mere number of *tweets* reflect voter preferences and comes close to election polls, and that the *tweets* are not only about spreading political opinions, but also to discuss these opinions with other users.

The finding of high-quality content in social media was studied by Agichtien et al. [10]. In the paper they investigate methods that allow the identification of high-quality content automatically by exploiting community feedback, such as links between items and explicit quality ratings from members of the community.

III. TEXT CLASSIFICATION APPROACH

Given a stream of messages from Twitter, i.e., *tweets*, our goal is to identify the messages that contain relevant information for traffic characterization. This can, therefore, be seen as a binary text classification problem. There are multiple algorithms for performing text classification, which, given the appropriate training set, achieve relatively high performances. One of the main challenges in this setting is precisely that of generating a representative training set, since percentage relevant messages in the Twitter stream is extremely low ($< 0.05\%$). Thus, manually selecting positive examples among the messages available for training becomes an unfeasibly laborious task.

However, as explained before, there are several official agencies injecting valid traffic messages in the Twitosphere, using robot users. Robot users are quite simple to identify since their messages have very strict formatting. On the other hand, Twitter messages posted by human users tend to be very loose in terms of grammatical structure, and usually contain many *spelling mistakes*, *non-standard punctuation*, emoticons among other idiosyncrasies. Nevertheless, robot-sent messages provide good examples of *vocabulary* related to the description of traffic-related events (e.g. names of routes or critical locations in the network), which should be shared by many of the human-generated messages. Therefore, our strategy consists in using robot-sent messages – which are relatively easy to identify – for gathering an *initial* set of positive examples needed for training a classification model. Given the very small fraction of Twitters related to traffic, a balanced number of *negative* examples can be randomly chosen from the entire collection of Twitter messages.

After training the classifier with robot-sent messages (as positive examples) and messages randomly chosen from a collection of Twitter messages (as negative examples), we obtain a first version of the classifier. We then proceed by following a boot-strapping approach: we use the first version of the classifier (just described) to find additional positive and negative examples that can be used enrich the initial training set for training a better classifier. Among the examples classified as relevant traffic messages by the first classifier we should have:

1) Messages generated by human users that are, in fact, relevant traffic messages, and which are interesting from the point of view of traffic sensing;
2) Message about traffic event, but which are sent by robot-users and, thus, not interesting for traffic sensing purposes;
3) Messages that are not related to traffic, i.e., they have been incorrectly classified in terms of topic.

We add (1) as additional positive examples to the training set, while (3) are used as negative training examples. The messages in (2) are not included in the expanded training because they can neither be considered negative examples (they are traffic-related), nor they can be considered positive examples (they do not represent the messages we want to capture). Using such expanded training set, we train a second version of the classifier, which is expected to be more robust and to achieve higher precision in identifying relevant traffic messages posted by human users.

IV. EXPERIMENTAL SET-UP

A. Data Sets

All the experiments were performed over a data set of approximately 565,000 Twitter messages crawled from the

Portuguese Twitosphere[1] between March and April of 2010. Most of the messages are written in Portuguese but some contain excerpts of other languages, such as English.

For building the *initial* training set, we manually identified several robot users sending traffic-related messages. We collected 3,300 messages of such robot users to be used as *positive* examples. Then, we randomly picked 41,000 messages from the entire collection to be used as *negative* examples. We opted for generating an unbalanced data set (ratio of 1 positive example to approximately 12 negative examples) in order to create a *pessimist* classification model.

B. Classification Set-Up

For the classification task at hand, we opted for using Support Vector Machines (SVM) [11][12]. SVM is a powerful binary classification algorithm that has proven to be effective in many text classification settings [13]. We used the LibSVM library [14] available through the Weka [15], a software toolkit.

After some preliminary experiments, we decided to configure the SVM algorithm to use a *linear* kernel function. Simple tests allowed us to check that the performance obtained using polynomial and radial kernels was similar, and thus, did not compensate the extra computational burden. All the remainder parameters were kept in their default values.

Messages are vectorized using a *unigram* bag-of-words approach. However, words with less than 4 characters written in lowercase were considered to be stop words and were removed. We also removed all *punctuation* characters. For reducing sparsity in the end all tokens are lowercased.

C. Evaluation

Most of the times, the performance of a classifier is evaluated over a test set, usually under a k-fold cross-validation scheme. However, in our case, the messages we wish to capture, i.e. traffic-related messages posted by *human users*, are not part of the initial training set, which is composed of traffic-related messages sent by robot users. Therefore, traditional k-fold cross-validation schemes cannot be applied in our setting.

Instead, we opted for *manually* assessing the performance of the classifier: we use the classifier to process a large set of unlabeled data, and we manually verify the results. Messages can be evaluated according to three possible cases, already defined in Section III ((i) traffic-related and posted by humans; (ii) traffic-related but posted by robots; and (iii) not related to traffic. We will evaluate the classifier for the task of finding traffic-related messages posted by humans (i.e. case (i)). Traffic related messages posted by robot users will *not* be considered valid. Thus, true positives (TP) and false positives (FP) are given by:

- $TP_{HMN} = (1)$
- $FP_{HMN} = (2) + (3)$

The number of false negatives FN_{HMN}, i.e., the number of valid messages in the collection that were not identified by the classifier, is given as a function of estimate on the total number of traffic messages posted by (human) users, which was found to be 0.05% of all messages.

We can now compute on Precision, Recall and F-score:

$$P = \frac{TP_{HMN}}{TP_{HMN} + FP_{HMN}}$$

$$R = \frac{TP_{HMN}}{TP_{HMN} + FN_{HMN}}$$

$$F = 2 \cdot \frac{P \cdot R}{P + R}$$

The SVM classifier produces a continuous classification decision value ranging from -1 to 1. Negative decision values correspond to messages that the classifier considered *not* to be related to traffic, while positive values correspond to messages that are considered to be related to traffic. We can impose a minimum threshold on the value for positive decision th_{min} to reduce the number of false positives. We can then compute all the performance measures previously presented for various values of th_{min} and, hence, test the sensitivity of the classifier to this parameter.

V. RESULTS

In this section, we will present the results obtained in both stages of the bootstrapping approach. Table I and Table II show the absolute and evaluation results achieved by the classification model in the first and second iterations.

In the first iteration (i.e., using the classifier trained only with robot-sent messages as positive examples), we present the results for several values of the decision threshold, th_{min}, using as test sample of approximately 260,000 unlabeled Twitter messages.

TABLE I

ABSOLUTE RESULTS ACHIEVED IN THE SECOND STAGE OF THE BOOTSTRAPPING PROCESS.

	1^{st} Iteration		2^{nd} Iteration	
th_{min}	$TP + FP$	TP_{HMN}	$TP + FP$	TP_{HMN}
0.60	62	13	47	21
0.70	50	13	37	20

TABLE II

PRECISION, RECALL AND F-SCORE RESULTS IN THE FIRST AND SECOND ITERATION OF THE BOOTSTRAPPING TRAINING STRATEGY

	1^{st} Iteration			2^{nd} Iteration		
th_{min}	P_1	R_1	F_1	P_2	R_2	F_2
0.6	21.0%	10.0%	13.5%	44.7%	16.2%	23.8%
0.7	26.0%	10.0%	14.4%	54.1%	15.4%	24.0%
AVG	23.5%	10.0%	**14.0%**	49.4%	15.8%	**23.9%**

When looking at these results, it is important to keep in mind that the percentage of *tweets* containing traffic information sent by users was estimated to be less than 0.05% in the universe of *tweets* being processed (0.07 if we account for the robot sent messages). The precision figures we are achieving are much larger than this baseline value.

[1]This data set was provided by Sapo (http://www.sapo.pt), the main Portuguese ISP.

Tables I and II also show the results achieved using the second version of the classification model and another test corpus. The additional labeled data used to build the second version of the classifier was composed by 15 human-generated messages, correctly found in the previous iteration - as positive examples - and 75 messages previously misclassified - as negative examples. The test corpus was composed of approximately the same size as the first one (about 260,000 messages) but with distinct messages.

As it can be seen, the bootstrapping strategy we used led to a significant improvement in the performance of the classifier both in terms of precision and recall. In Table II we can see that the value of average F-score after bootstrapping was approximately 2 times higher. In other words, we were able to almost double the average value of precision and recall simultaneously.

Notably, this increase in performance was obtained by adding only a few additional positive (15) an negative (75) examples to the initial training set (3,300 positive examples and 41,000 negative examples). After analysis, we verified that this increase in performance was mostly due to the addition of the negative examples, which allowed building a classification model that was more robust to messages of certain topics. Among these, the most frequent cases were related to Twitter messages about political matters, which tend to share some specific keywords (e.g. "right" and "left"), and also make frequent references to certain locations.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a two-stage bootstrapping strategy to train a text classifier capable of identifying traffic-related messages posted by users on Twitter. The main challenges at stake were related to the relatively low percentage of relevant messages in the Twitter stream ($< 0.05\%$), which was problematic both for compiling appropriate training sets and for achieving useful classification performance. We trained a first version of the classifier using traffic-related Twitter messages automatically sent by a few official news sources (which were identified manually), and we used this classifier to compile additional positive and negative examples from a large collection of Twitter messages. By expanding the training set with these newly found examples, we were able to train a second version of the classifier, which achieved F-measure of approximately 23% (i.e. an 85% increase in performance in terms of F-measure in relation to the previous classifier). In other words, we show that it is possible to train a relatively accurate text classifier for traffic-related Twitter messages (given the percentage of relevant messages in the universe at stake) with almost no annotation effort.

Given the very large number of Twitter messages being posted hourly, we believe that our experiments show that exploring traffic-related information from this media can become an interesting option for complementing information gathered by more traditional sensors (which tend to be limited by several factors such as cost, mobility, availability, etc.).

In the future, our goal is to improve classification performance even further. More specifically, we wish to check how much improvement it is possible to achieve by executing additional iterations of the bootstrapping process. Also, we will focus on trying to resolve the geographic reference explicitly or implicitly related to the traffic events described in Twitter messages, in order to match the information with an exact location.

ACKNOWLEDGMENTS

This work was partially supported by grant SFRH/BD/23590/2005 from FCT (Portuguese research funding agency), and by a special grant from the Master Program in Informatics and Computing Engineering, Faculdade de Engenharia da Universidade do Porto, Portugal. We also wish to thank Sapo.pt for providing access to the Twitter data set used in this work.

REFERENCES

[1] J. Pontin, "From many tweets, one loud voice on the internet," The New York Times, April 22, 2007.

[2] S. Johnson, "How twitter will change the way we live," Time, June 05, 2009.

[3] E. Veiszadeh, "Twitter freedom's only link in iran," The Australian, July 16, 2009.

[4] M. Musgrove, "Twitter is a player in iran's drama," The Washington Post, July 09, 2009.

[5] A. Bloxham, "Facebook more effective than emergency services in a disaster," The Daily Telegraph, December 20, 2008.

[6] E. Young, "Crisis puts a new face on a social networking," The Sydney Morning Herald, June 07, 2009.

[7] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *WWW '10: Proceedings of the 19th international conference on World wide web*. New York, NY, USA: ACM, 2010, pp. 851–860.

[8] S. Asur and B. A. Huberman, "Predicting the future with social media," Mar 2010. [Online]. Available: http://arxiv.org/abs/1003.5699

[9] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *AAAI-10: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. Atlanta, USA: AAAI, 2010.

[10] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *WSDM '08: Proceedings of the international conference on Web search and web data mining*. New York, NY, USA: ACM, 2008, pp. 183–194.

[11] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine Learning*, 1995, pp. 273–297.

[12] V. N. Vapnik, *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

[13] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in *Proceedings of ECML-98, 10th European Conference on Machine Learning*, C. Nédellec and C. Rouveirol, Eds., no. 1398. Chemnitz, DE: Springer Verlag, Heidelberg, DE, 1998, pp. 137–142. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.6982

[14] C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines*, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/˜cjlin/libsvm

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009. [Online]. Available: http://dx.doi.org/10.1145/1656274.1656278