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