Defining Semantic Meta-Hashtags for Twitter Classification

Joana Costa[1,2], Catarina Silva[1,2], Mário Antunes[1,3], and Bernardete Ribeiro[2]

[1] School of Technology and Management, Computer Science Communication and Research Centre
Polytechnic Institute of Leiria; Portugal
{joana.costa,catarina,mario.antunes}@ipleiria.pt
[2] Department of Informatics Engineering, Center for Informatics and Systems of the University of Coimbra (CISUC); Portugal
{joanamc,catarina,bribeiro}@dei.uc.pt
[3] Center for Research in Advanced Computing Systems (CRACS); Portugal
mantunes@dcc.fc.up.pt

Abstract. Given the wide spread of social networks, research efforts to retrieve information using tagging from social networks communications have increased. In particular, in Twitter social network, *hashtags* are widely used to define a shared context for events or topics. While this is a common practice often the *hashtags* freely introduced by the user become easily biased. In this paper, we propose to deal with this bias defining semantic meta-hashtags by clustering similar messages to improve the classification. First, we use the user-defined *hashtags* as the Twitter message class labels. Then, we apply the meta-hashtag approach to boost the performance of the message classification.
The meta-hashtag approach is tested in a Twitter-based dataset constructed by requesting public *tweets* to the Twitter API. The experimental results yielded by comparing a baseline model based on user-defined *hashtags* with the clustered meta-hashtag approach show that the overall classification is improved. It is concluded that by incorporating semantics in the meta-hashtag model can have impact in different applications, e.g. recommendation systems, event detection or crowdsourcing.

Keywords: Meta-hashtags, Semantic, Text Classification, Twitter

1 Introduction

Twitter is a social media platform that provides a microblogging service where users are able to post text-based messages of up to 140 characters, also known as *tweets*. It can also be considered an online social network, as users can link themselves by defining others to follow, and consequently have their own followers. The underlying concept of Twitter is to share the everyday activities with friends and family in a simple way. However, *tweets* may contain information of broad interest [1] and have a wide range of applications and uses, like event detection [2–5], academic tool [6–8], news media [2, 9] or mining political opinion [10, 11].

Twitter provides the possibility of including an *hashtag*, a single word started with the symbol "#" , in order to classify the content of a message and improve search capabilities. This can be particularly important considering the amount of data produced in Twitter social network. Besides improving search capabilities, *hashtags* have been identified as having multiple and relevant potentialities, like promoting the phenomenon described in [12] as *micro-meme*, i.e. an idea, behavior or style that spreads from person to person within a culture [13]. By tagging a message with a trending topic hashtag, a user expands the audience of the message, compelling more users to express their feelings about the subject [14].

Considering the importance of the hashtag in Twitter, it is relevant to study the possibility of evaluating message contents in order to predict its hashtag. If we can classify a message based on a set of *hashtags*, we are able to suggest an hashtag for a given *tweet*, bringing a wider audience into discussion [15], spreading an idea [16], get affiliated with a community [17], or bringing together other Internet resources [18].

We propose an approach to deal with the bias resulting from the freely user-defined *hashtags*, by defining semantic meta-hashtags to identify clusters of similar messages, in order to improve their classification. First, we use the user-defined *hashtags* as the Twitter message class labels. Then, we define meta-hashtags by grouping the most used *hashtags* and their related *hashtags* into a meta-class and applied the meta-hashtag approach to boost the performance of the initial message classification. Both the initial model and the meta-hashtag model were tested with the initial user defined *hashtags* and the results are presented by comparing the classification performances obtained.

The rest of the paper is organized as follows. We start in Section 2 by describing the related work regarding social networks and meta-class approaches. We then proceed in Section 4 to explain the experimental setup, including the dataset description, the pre-processing methods, learning and evaluation approaches. In Section 5 we present and analyse the results obtained. Finally, in Section 6 we present the most relevant conclusions and delineate some directions for future work.

2 Related Work

Social networks have gained significant importance and are being widely studied in many fields in the last years. Modern challenges in social networks involve not only computer science matters but also social, political, business, and economical sciences. In computer science, and considering our focus on Twitter, recent works comprise event detection [3, 4], information spreading [19], community mining [20], crowdsourcing [21] and sentiment analysis [11].

Regarding Twitter *hashtags*, and particularly hashtag recommendation, we have identified the recent study presented in [22], where an approach for hashtag recommendation is introduced. This approach computes a similarity measure between *tweets* and uses a ranking system to recommend *hashtags* to new *tweets*.

A different approach is proposed in [23], where an event detection method is described to cluster Twitter *hashtags* based on semantic similarities between the *hashtags*. Two methods for *tweet* vector generation are proposed and their performance evaluated on clustering and event detection in comparison to word-based vector generation methods. This work is in line with our work except for the fact that the semantic similarities are computed based on the message content similarities rather than being based on semantic hashtag similarities.

The foundation of our proposal is the use of meta-classes to boost the performance of Twitter messages. Although the application on a Twitter classification problem is a novel contribution, the use of meta-classes has been studied in other classification contexts. In [24] the use of meta-classes is proposed to improve the performance of classifiers in an handwritten character recognition problem. The use of meta-classes is promoted in this study based on the complex boundaries between classes, the classes overlapping and the lack of sufficient number of samples for some classes.

The related work presented so far sheds light on the importance of social networks in the scientific community, specially the recently explored niche of Twitter *hashtags*, that can have multiple applications like recommendation systems or improvement of search capabilities. In the next section we will detail our proposed approach in order to settle our contribution in the field.

3 Proposed Approach

This section describes the proposed approach to define meta-hashtags and to use them in a classification application to improve the overall classification obtained. Our approach is twofold, resulting in two final models, the baseline model, that considers the user-defined *hashtags*, and the meta-hashtags model, that considers the clusters of similar messages grouped by a single meta-hashtag. In Fig. 1 we depict the proposed framework.

The baseline model is constructed and trained with labelled examples that use the user-defined *hashtags* as a "one-against-all" two-class problem. In the meta-hashtags model, semantic meta-hashtags were heuristically defined, by clustering similar in a meta class. Related classes are then relabelled according to the new defined meta-hashtags and a similar training process occur in order to construct the new proposed model.

The underpinning idea behind the use of meta-hashtags is to combine the class label of similar messages in order to mitigate the effects of the bias introduced by freely user-defined *hashtags*, and thus improving the overall classification of Twitter messages according to their *hashtags*.

4 Experimental Setup

In this section we start by describing the built data set for the purpose of testing our approach. We also characterize the methodology for documents representation. We then proceed dealing with the pre-processing method and finally, we

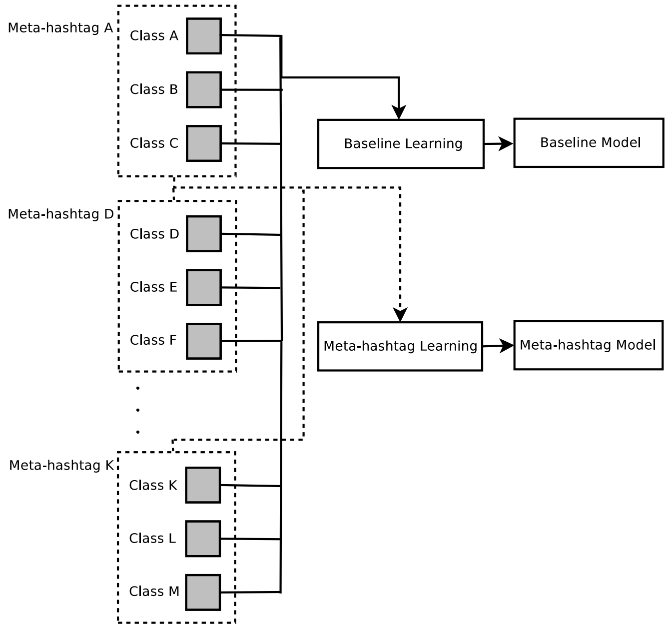

Fig. 1: Proposed approach.

conclude by introducing the performance metrics used to evaluate the proposed approach.

4.1 Dataset

The dataset was constructed by requesting public *tweets* to the Twitter API[4]. We have collected more than 230.000 messages during four days and, considering the worldwide usage of Twitter, *tweets* were only considered if the user language was defined as English. All the messages that did not have at least one hashtag were discarded, as the hashtag are assumed as the message classification. Finally, *tweets* containing no message content besides *hashtags* were also discarded and all the *hashtags* are removed from remaining *tweets*. From the 230.000 collected messages, we reach 10.000 *tweets* that have a body part and at least one hashtag.

As users are able to define their own *hashtags*, a high number of different *hashtags* is present in the requested *tweets*. In order to narrow the number of classes, we have only considered the most used *hashtags*. Finally, we used the crowdsourcing platform *http://tagdef.com/* to discover the hashtag meaning and the related *hashtags*, and considered the most used ones, which have at least two or more related *hashtags*, so a meta-hashtag class may empirically be defined. A total number of 15 *hashtags* were found to match this presumption and a

[4] https://dev.Twitter.com/

total number of 1.230 *tweets* were considered as being labelled and suited for classification purposes. The individual *hashtags* were semantically clustered in 5 meta-hashtag classes, as depicted in Table 1.

The *tweets* were then split into two equally sized and disjoint sets: training and testing. The training data set is used to build classification learning models, and the testing data set to evaluate performance.

Table 1 describes the positive documents of each class and the corresponding meta-class of the data set. As can be observed from the Table 1, there is an heterogeneous distribution of *hashtags* in the dataset. For example, class TEAMFOLLOWBACK has 274 documents, while class NOWFOLLOWING has only 17 documents. The amount of positive documents in the training and testing data sets is balanced because it is obtained by the equally split of training and test sets.

	Training	Testing
NP	138	139
NOWPLAYING	72	68
meta-hashtag NP	**209**	**207**
SEX	70	54
PORN	65	56
XXX	23	19
HOT	14	6
meta-hashtag SEX	**104**	**76**
JOB	32	36
JOBS	41	37
meta-hashtag JOB	**59**	**61**
NW	32	32
NOWWATCHING	7	4
meta-hashtag NW	**39**	**36**
TEAMFOLLOW	17	18
TEAMFOLLOWBACK	126	148
FOLLOWBACK	14	24
NF	58	57
NOWFOLLOWING	7	10
meta-hashtag NF	**207**	**238**

Table 1: Amount of positive documents in the training and testing phases.

4.2 Pre-processing Methods

A *tweet* is represented as one of the most successful and commonly used document representation, which is the vector space model, also known as *Bag of Words*. The collection of features is built as the dictionary of unique terms present in the documents collections. Each document of the document collection is indexed with the *bag* of the terms occurring in it, i.e., a vector with one element for each term occurring in the whole collection.

High dimensional space can cause computational problems in text-classification problems where a vector with one element for each occurring term in the whole connection is used to represent a document. Also, overfitting can easily occur which can prevent the classifier to generalize and thus the prediction ability becomes poor. In order to reduce feature space pre-processing methods are often applied. These techniques aim at reducing the size of the document representation and prevent the mislead classification as some words, such as articles, prepositions and conjunctions, called *stopwords*, are non-informative words, and occur more frequently than informative ones. These words could also mislead correlations between documents so *stopword* removal technique was applied. *Stemming* method was also applied. This method consists in removing case and inflection information of a word, reducing it to the word stem. Stemming does not alter significantly the information included, but it does avoid feature expansion.

4.3 Learning and Evaluation

The evaluation of our approach was done by the dataset with the Support Vector Machine (SVM) method. This machine learning method was introduced by Vapnik [25], based on his Statistical Learning Theory and Structural Risk Minimization Principle. The idea behind the use of SVM for classification consists on finding the optimal separating hyperplane between the positive and negative examples. Once this hyperplane is found, new examples can be classified simply by determining which side of the hyperplane they are on. SVM constitute currently the best of breed kernel-based technique, exhibiting state-of-the-art performance in text classification problems [26–28]. SVM were used in our experiments to construct the model with user-defined *hashtags* and the meta-hastags model.

In order to evaluate the binary decision task of the proposed models we defined several measures based on the possible outcomes of the classification, such as, error rate ($\frac{FP+FN}{TP+FP+TN+FN}$), recall ($R = \frac{TP}{TP+FN}$), and precision ($P = \frac{TP}{TP+FP}$), as well as combined measures, such as, the van Rijsbergen F_β measure [29], which combines recall and precision in a single score.

F_β is one of the best suited measures for text classification used with $\beta = 1$, i.e. F_1 ($F_1 = \frac{2*P*R}{P+R}$), an harmonic average between precision and recall.

5 Experimental Results and Analysis

In this Section we evaluate the performance obtained on the Twitter data set using the two approaches described in Section 3, namely the baseline approach considering the 15 initial *hashtags* and the meta-hashtag approach. Table 2 summarises the performance results obtained by classifying the datasets.

	Baseline Model			Meta-hashtags Model		
	Precision	Recall	F1	Precision	Recall	F1
NP	45.73%	53.96%	49.50%	35.44%	92.81%	51.29%
NOWPLAYING	23.96%	33.82%	28.05%	15.11%	80.88%	25.46%
meta-hashtag NP				**50.55%**	**88.89%**	**64.45%**
SEX	70.18%	74.07%	72.07%	38.69%	98.15%	55.50%
PORN	80.00%	71.43%	75.47%	40.88%	100.00%	58.03%
XXX	21.05%	21.05%	21.05%	11.45%	100.00%	20.54%
HOT	7.14%	16.67%	10.00%	3.61%	100.00%	6.98%
meta-hashtag SEX				**69.23%**	**94.74%**	**80.00%**
JOB	53.33%	66.67%	59.26%	34.31%	97.22%	50.72%
JOBS	50.79%	86.49%	64.00%	33.33%	91.89%	48.92%
meta-hashtag JOB				**56.86%**	**95.08%**	**71.17%**
NW	17.65%	9.38%	12.24%	15.38%	12.50%	13.79%
NOWWATCHING	0.00%	0.00%	-	3.85%	25.00%	6.67%
meta-hashtag NW				**19.23%**	**13.89%**	**16.13%**
TEAMFOLLOW	100.00%	100.00%	100.00%	26.09%	100.00%	41.38%
TEAMFOLLOWBACK	40.80%	55.41%	46.99%	32.91%	87.84%	47.88%
FOLLOWBACK	0.00%	0.00%	-	5.57%	91.67%	10.50%
NF	25.00%	5.26%	8.70%	13.92%	96.49%	24.34%
NOWFOLLOWING	-	0.00%	-	2.53%	100.00%	4.94%
meta-hashtag NF				**54.94%**	**91.18%**	**68.56%**

Table 2: Comparative results.

Analysing the table we can observe that the use of a meta-hashtag outperforms the overall classification of the initial *hashtags* when they are considered individually. For example, the class NP has a F1 measure of 49.50% in the baseline model, in the class NOWPLAYING the corresponding value is 28.05% and with the use of a meta-hashtag the new proposed model presents a F1 value for the meta-hashtag NP class of 64.45%. This might be related to the fact that the content being classified in the meta-hashtag dataset is different from the initial dataset, thus misleading the classifier. Theses improvements on the results obtained may also be observed for other cases, like in class JOB with F1 of 59.26% and class JOBS with 64.00%, while the corresponding meta-hashtag JOB has 71.17%.

With the use of a meta-hashtag we unify the labelling process by grouping similar messages and placing them in the same classification class, thus boosting the performance of the overall classifier. This analysis is in line with previous results in [23] where from a set of different methods proposed, a pseudo-meta-hashtag approach was presented to be beneficial in a clustering problem.

Other noteworthy results are obtained when using the meta-hashtag model to classify the initial classes. Although the F1 measure decreases considering the baseline model, the recall increases in a higher proportion. As an example, the class XXX classified by the baseline model presents a precision and a recall of 21.05%. Classified by the meta-hashtag model, the precision falls to 11.45% and the recall raises up to 100.00%, which means that precision proportionally decreases less than the increase of recall. This is due to the false positive increase being less than the increase of true positives. The increase of false positives, and thus the decrease of the F1 measure was expected, as we mislead the classifier by training it with the meta-hashtags examples, which means we used as positive examples not only the initial class messages, but also the related messages that belong to semantically similar *hashtags*.

In the baseline model, classes like NOWFOLLOWING or FOLLOWBACK have no F1 measure. This occurs because the classifier did not identified any true positive document, probably due to the lack of information in the training phase, so precision and recall are 0.00% and F1 can not be calculated. In these classes the use of the meta-hashtags approach, more than increasing the classifier performance, permits the identification of these classes documents.

6 Conclusions and Future Work

In this paper we have presented a meta-hashtag approach in order to deal with the bias produced by freely user-defined hashtag in Twitter social network. The main idea of defining semantic meta-hashtags that cluster similar messages is to boost the classification performance of messages, by avoiding the mislead classification problem raised by having multiple classes to identical features.

For that purpose, we have constructed a dataset by requesting public *tweets* to the Twitter API and conducted a set of experiments by comparing a baseline model based on user-defined *hashtags* with our approach based on meta-hashtags.

The preliminary results are very promising. It is possible to observe that the proposed approach outperforms the F1 measure of each initial class included in its composition, with the exception of the initial class TEAMFOLLOW that is already correctly classified in the initial approach. It is also important to note the overall improvement of the recall metric when using the meta-hashtag model to classify the initial classes. This improvement sustains the use of meta-hashtags and makes it possible to infer that, the more information we give to the classifier in the training process, the better for the identification of true positive documents in the testing phase, as the increase of recall is due to the increase of true positives.

Our future work will give a formalization of the meta-hastags model including *hashtags* similarities in order to evaluate the clustering quality. Moreover, evaluation of dynamically created Twitter meta-hashtags will be performed possibly for enrichment of applications in real-time.

Acknowledgment

This work is financed by the ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project «PEst-C/EEI/LA0014/2011»

References

1. M. Efron, "Hashtag retrieval in a microblogging environment," in *SIGIR*, 2010, pp. 787–788.
2. F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Semantic enrichment of twitter posts for user profile construction on the social web," in *ESWC (2)*, 2011, pp. 375–389.
3. T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 851–860.
4. A.-M. Popescu and M. Pennacchiotti, "Detecting controversial events from twitter," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, ser. CIKM '10, 2010, pp. 1873–1876.
5. H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on twitter," in *ICWSM*, 2011.
6. S. Ovadia, "Exploring the Potential of Twitter as a Research Tool," *Behavioral & Social Sciences Librarian*, vol. 28, no. 4, pp. 202–205, Oct. 2009.
7. M. Rowe and M. Stankovic, "Mapping tweets to conference talks: A goldmine for semantics," *Workshop on Social Data on the Web (SDoW)*, 2010.
8. K. Weller and C. Puschmann, *Twitter for Scientific Communication: How Can Citations/References be Identified and Measured?*, 2011, pp. 1–4.
9. F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Analyzing user modeling on twitter for personalized news recommendations," in *Proceedings of the 19th international conference on User modeling, adaption, and personalization*, 2011, pp. 1–12.
10. A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010, pp. 178–185.
11. B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," in *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010.

12. J. Huang, K. M. Thornton, and E. N. Efthimiadis, "Conversational tagging in twitter," in *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, 2010, pp. 173–178.
13. (2012, Oct) Merriam-webster's dictionary.
14. M. Zappavigna, "Ambient affiliation: A linguistic perspective on Twitter," *New Media & Society*, vol. 13, no. 5, pp. 788–806, 2011.
15. S. Johnson, "How twitter will change the way we live," *Time Magazine*, vol. 173, pp. 23–32, 2009.
16. O. Tsur and A. Rappoport, "What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities," in *Proceedings of the fifth ACM international conference on Web search and data mining*, ser. WSDM '12, 2012, pp. 643–652.
17. L. Yang, T. Sun, M. Zhang, and Q. Mei, "We know what @you #tag: does the dual role affect hashtag adoption?" in *Proceedings of the 21st international conference on World Wide Web*, ser. WWW '12, 2012, pp. 261–270.
18. H.-C. Chang, "A new perspective on twitter hashtag use: diffusion of innovation theory," in *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, ser. ASIS&T '10, 2010, pp. 85:1–85:4.
19. B. Doerr, M. Fouz, and T. Friedrich, "Why rumors spread so quickly in social networks," *Commun. ACM*, vol. 55, no. 6, pp. 70–75, 2012.
20. C. Tantipathananandh and T. Y. Berger-Wolf, "Finding Communities in Dynamic Social Networks," in *2011 IEEE 11th International Conference on Data Mining*, 2011, pp. 1236–1241.
21. M. Treiber, D. Schall, S. Dustdar, and C. Scherling, "Tweetflows: flexible workflows with twitter," in *Proceedings of the 3rd International Workshop on Principles of Engineering Service-Oriented Systems*, 2011, pp. 1–7.
22. E. Zangerle, W. Gassler, and G. Specht, "Recommending #-tags in twitter," in *Proceedings of the Workshop on Semantic Adaptive Social Web 2011 in connection with the 19th International Conference on User Modeling, Adaptation and Personalization, UMAP 2011*, 2011, pp. 67–78.
23. H. O. O Ozdikis, P Senkul, "Semantic Expansion of Hashtags for Enhanced Event Detection in Twitter," in *The First International Workshop on Online Social Systems (WOSS)*, 2012.
24. A. Koerich, "Improving classification performance using metaclasses," in *IEEE International Conference on Systems, Man and Cybernetics*, 2003, pp. 717–722.
25. V. Vapnik, *The Nature of Statistical Learning Theory.* Springer, 1999.
26. T. Joachims, *Learning Text Classifiers with Support Vector Machines.* Kluwer Academic Publishers, Dordrecht, NL, 2002.
27. S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2002.
28. J. Costa, C. Silva, M. Antunes, and B. Ribeiro, "On using crowdsourcing and active learning to improve classification performance," in *Proceedings of 11th International Conference on Intelligent Systems Design and Applications (ISDA)*, 2011, pp. 469–474.
29. C. van Rijsbergen, *Information Retrieval.* Butterworths Ed., 1979.