

Instance-Based Learning for Tweet Monitoring and Categorization

Julien Gobeill^{1,2(✉)}, Arnaud Gaudinat¹, and Patrick Ruch^{1,2}

¹ BiTeM Group, HEG/HES-SO, University of Applied Sciences,
7 rte de Drize 1227, Carouge, Switzerland
{julien.gobeill, arnaud.gaudinat, patrick.ruch}@hesge.ch

² SIBtex Group, SIB Swiss Institute of Bioinformatics,

1 rue Michel-Servet 1206, Genève, Switzerland
{julien.gobeill, patrick.ruch}@hesge.ch

Abstract. The CLEF RepLab 2014 Track was the occasion to investigate the robustness of instance-based learning in a complete system for tweet monitoring and categorization based. The algorithm we implemented was a k -Nearest Neighbors. Dealing with the domain (automotive or banking) and the language (English or Spanish), the experiments showed that the categorizer was not affected by the choice of representation: even with all learning tweets merged into one single Knowledge Base (KB), the observed performances were close to those with dedicated KBs. Interestingly, English training data in addition to the sparse Spanish data were useful for Spanish categorization (+14% for accuracy for automotive, +26% for banking). Yet, performances suffered from an overprediction of the most prevalent category. The algorithm showed the defects of its virtues: it was very robust, but not easy to improve. BiTeM/SIBtex tools for tweet monitoring are available within the DrugsListener Project page of the BiTeM website (<http://bitem.hesge.ch/>).

1 Introduction

BiTeM/SIBtex has a long tradition of participating in large evaluation campaigns, such as TREC, NTCIR or CLEF [1-4]. The CLEF RepLab 2014 Track was the occasion to integrate several local tools into a complete system, and to evaluate a simple and robust statistical approach for tweet classification in competition. The goal of the first task was to perform text categorization on Twitter, i.e. to design a system able to assign a predefined category to a tweet. This category was one out of eight related to companies' reputations. All tweets dealt with entities from the automotive (20 entities) or the banking (11 entities) domain, and were in English (93%) or in Spanish (7%). For training and/or learning purposes, participants were provided with approximately 15,000 tweets labeled by human experts (the training set). Then, the systems had to predict the good categories for 32,000 unlabeled tweets (the test set).

In this task, the main difficulty was to efficiently preprocess the text, as standard Natural Language Processing strategies can fail to deal with the short, noisy, and strongly contextualised nature of the tweets. Another difficulty was to efficiently

learn from unbalanced classes. Finally, this was a multilingual task, but the language distribution also was unbalanced, with less than 10% Spanish learning instances. We applied a simple and robust statistical approach in order to design our system, based on instance-based learning for categorization purposes.

Two particular questions were investigated during this study. Q_1 : is it better to build one Knowledge Base (KB) for each domain, or to merge automotive and banking into the same KB ? Q_2 : is it better to build one KB for each language, or to merge English and Spanish into the same KB ?

2 Methods

2.1 Overall Architecture of the System

Figure 1 illustrates the overall architecture of our system. The workflow is divided into two steps: the training phase (offline), and the test phase (online). Three independent components act cooperatively to preprocess data (component 1), to build the knowledge base (component 2) and to classify tweets (component 3).

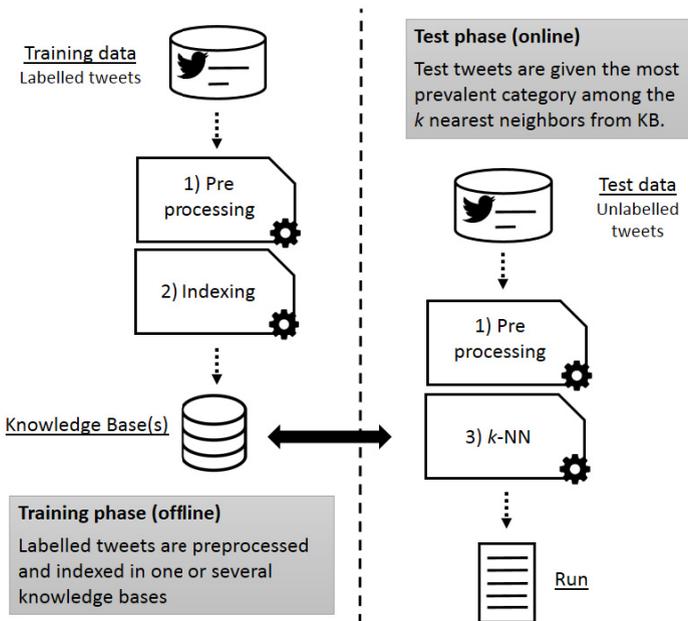


Fig. 1. Overall architecture of the system

During the training phase, all tweets belonging to the training set were preprocessed by component 1. Component 1 merges several standard Natural Language Processing treatments, along with a language detector. Then, they were indexed in

one or several indexes by component 2, in order to make the KB. Component 2 is an Information Retrieval platform, which builds indexes for related documents retrieval.

During the test phase, all tweets belonging to the test set also were preprocessed by component 1. Then, for a given test tweet, the component 3 (k -NN) exploited the KB in order to retrieve the most similar tweets seen in the training data, and to infer a predicted category. Official runs were computed with the whole test set.

Tweets often contain metadata within tags, the most frequent being hyperlinks (<a>) and emphasis (). Moreover, they often don't have proper punctuation.

2.2 Preprocessing

The goal of the component 1 was to preprocess the tweets in order to have proper and efficient instances to index (for the training phase) or search (for the test phase). For this purpose, a set of basic rules was applied. Tags were first discarded. Contents within an emphasis tag () were repeated in order to be overweighted. Contents within a hyperlink tag (<a>) also were repeated, and were preceded by the "HREF" mention.

For language detection purposes, we performed a simple N-Gram-Based Text Categorization, based on the Cavnar and Trenkle works [5]. This approach aims at comparing n-grams frequency profiles in a given text, with profiles observed in large English and Spanish corpus. This simple approach is reported to have an accuracy in the range of 92% to 99%. N-grams profiles were taken from [6].

2.3 Indexing

The goal of the component 2 was to build one or several indexes from the training data, in order to obtain a related documents search engine. For this purpose, we used the Terrier platform [7]. We used default stemming, stop words and a Poisson weighting scheme (PL2).

Dealing with Q_1 and Q_2 , we investigated several strategies and built several indexes, mixing tweets from the cars or banks domains, and tweets in English or Spanish.

2.4 k -NN

The goal of the component 3 was to categorize tweets from the test set. For this purpose, we used a k -Nearest Neighbors, a remarkably simple algorithm which assigns to a new text the categories that are the most prevalent among the k most similar tweets contained in the KB [8]. Similar tweets were retrieved thanks to component 2. Then, a score computer inferred the category from the k most similar instances, following this formula:

$$predcat = \arg \max_{c \in \{c_1, c_2, \dots, c_m\}} \sum_{x_i \in K} E(x_i, c) \times RSV(x_i)$$

where $predcat$ is the predicted category for a test tweet, c_1, c_2, \dots, c_m are the possible categories, K is the set of the k nearest neighbors of the test tweet, $RSV(x_i)$ is the retrieval status value given by the component 2 (i.e. the similarity score) for the neighbor x_i , and $E(x_i, c)$ is 1 when x_i is of category c , 0 otherwise.

3 Results and Discussions

The Q_1 and Q_2 issues were addressed with the training data, thanks to a ten-fold cross validation strategy.

3.1 Q_1 : Is It Better to Build One KB for Each Domain, or to Merge Automotive and Banking into the Same KB ?

First, we investigated Q_1 , by exploiting KB with only bank tweets (*banks* index), only automotive tweets (*car* index), or both (*all* index). English and Spanish were merged into the same KB. Figure 2 shows the performances of the system for the banks test set, for different values of k .

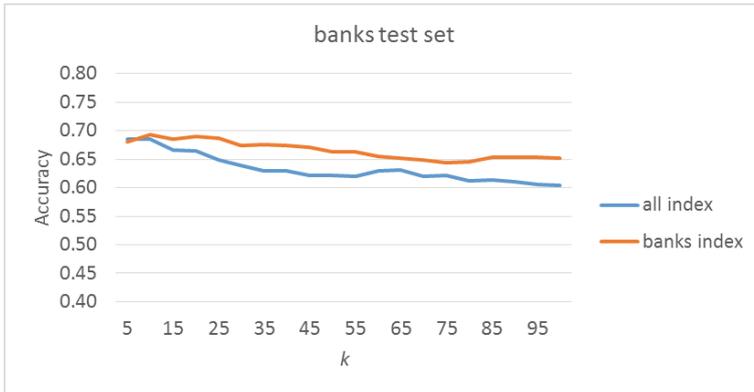


Fig. 2. Performances for the banks test set, using the *all* index (all training data merged) or the specific *banks* index (only banks training data), for different values of k .

Experiments showed that the optimal k for these data was around 10. They also showed that throughout the curves, it was better to use specific indexes (orange curve) versus a unique merged index (blue curve). Yet, the difference between best performances is not significant, with an accuracy of 0.69 for the *all* and the *banks* indexes for banks tweets (at $k=10$), and accuracies of 0.77 versus 0.76 for the *cars* index and the *all* index. We can say that, for categorizing tweets from a given domain, data from the other domain do not provide useful information, but do not degrade the optimal performances, thanks to the k -NN robustness.

3.2 Q_2 : Is It Better to Build one KB for Each Language, or to Merge English and Spanish into the Same KB ?

Then, we investigated Q_2 , especially for the Spanish language that represented less than 7% of the training data. We exploited the *cars*, *banks*, *cars_es* and *banks_es* indexes. Figure 3 shows the performances of the system for the cars test set, for different values of k .

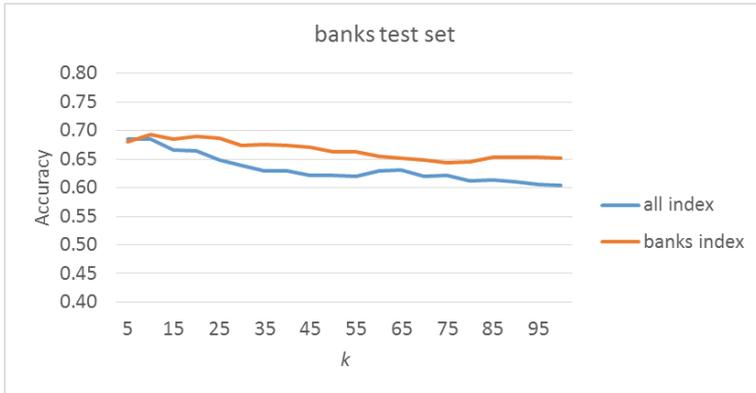


Fig. 3. Performances for the cars - Spanish test set, using the cars index (English and Spanish merged) or the specific cars - Spanish index (only Spanish data), for different values of k .

Experiments showed that the optimal k for Spanish data was around 30, significantly higher than the general case. This could be explained by the smaller set of Spanish instances. They also showed that it was better to use both languages indexes (orange curves) versus a Spanish-specific index (blue curves). We can say that, for categorizing tweets from Spanish, an additional amount of English data provides useful information and increases the top accuracy (from 0.69 to 0.79 for cars, from 0.57 to 0.72 for banks).

The same experiments with the English language showed no significant differences between the merged and the English-specific indexes. We have not tried any cross-language strategy [10,11].

4 Conclusion

We designed a complete system for tweet categorization according to predefined reputational categories. Dealing with the domain (automotive or banking) and the language (English or Spanish), the experiments showed that the k -NN was not very affected by the kind of representations: even with all data merged into one single KB, the observed performances are close to those observed with dedicated KB. Moreover, English training data were useful for Spanish categorization (+14% for accuracy for automotive, +26% for banking). Yet, the unbalanced labels make the k -NN to predict

the most prevalent category more often than necessary; this issue needs to be investigated in future works. The k -NN showed the defects of its virtues: it was robust, but not easy to improve. BiTeM/SIBTex tools for tweet monitoring are available within the DrugsListener Project page of the BiTeM website [9].

References

1. Gobeill, J., Teodoro, D., Pasche, E., Ruch, P.: Report on the trec 2009 experiments: chemical IR track. In: The Eighteenth Text REtrieval Conference (2009)
2. Gobeill, J., Pasche, E., Teodoro, D., Ruch, P.: Simple pre and post processing strategies for patent searching in CLEF intellectual property track. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (eds.) CLEF 2009. LNCS, vol. 6241, pp. 444–451. Springer, Heidelberg (2010)
3. Teodoro, D., Gobeill, J., Pasche, E., Ruch, P., Vishnyakova, D., Lovis, C.: Automatic IPC encoding and novelty tracking for effective patent mining. In: The 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies, Tokyo, Japan, pp. 309–317 (2010)
4. Vishnyakova, D., Pasche, E., Ruch, P.: Selection of relevant articles for curation for the comparative toxicogenomic database. In: BioCreative Workshop [Internet], pp. 31–38 (2012)
5. Cavnar, W., Trenkle, J.: N-gram-based text categorization. In: Proceedings of SDAIR-1994, 3rd Annual Symposium on Document Analysis and Information Retrieval (1994)
6. Practical cryptography. <http://practicalcryptography.com/>
7. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A high performance and scalable information retrieval platform. In: Proceedings of ACM SIGIR 2006 Workshop on Open Source Information Retrieval (2006)
8. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
9. BiTeM website. <http://bitem.hesge.ch/>
10. Müller, H., Geissbühler, A., Ruch, P.: ImageCLEF 2004: combining image and multi-lingual search for medical image retrieval. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 718–727. Springer, Heidelberg (2005)
11. Müller, H., Geissbühler, A., Marty, J., Lovis, C., Ruch, P.: The use of medGIFT and easyIR for imageCLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 724–732. Springer, Heidelberg (2006)