# POPSTAR at RepLab 2013: Name ambiguity resolution on Twitter

Pedro Saleiro, Luís Rei, Arian Pasquali, Carlos Soares, Jorge Teixeira, Fábio Pinto, Mohammad Nozari, Catarina Félix, Pedro Strecht

DEI-FEUP, Labs Sapo UP, INESC TEC
University of Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal
{pssc, csoares}@fe.up.pt

**Abstract.** Filtering tweets relevant to a given entity is an important task for online reputation management systems. This contributes to a reliable analysis of opinions and trends regarding a given entity. In this paper we describe our participation at the Filtering Task of RepLab 2013. The goal of the competition is to classify a tweet as relevant or not relevant to a given entity. To address this task we studied a large set of features that can be generated to describe the relationship between an entity and a tweet. We explored different learning algorithms as well as, different types of features: text, keyword similarity scores between entities metadata and tweets, Freebase entity graph and Wikipedia. The test set of the competition comprises more than 90000 tweets of 61 entities of four distinct categories: automotive, banking, universities and music. Results show that our approach is able to achieve a Reliability of 0.72 and a Sensitivity of 0.45 on the test set, corresponding to an F-measure of 0.48 and an Accuracy of 0.908.

**Keywords:** Online Reputation Management, Word Sense Disambiguation

## 1 Introduction

The relationship between people and public entities has changed with the rise of social media. Online users of social networks, blogs and micro-blogs are able to directly express and spread opinions about public entities, such as politicians, artists, companies or products. Online Reputation Management aims to automatically process online information about public entities. Some of the common tasks within Online Reputation Management consist in collecting, processing and aggregating social network messages to extract opinion trends about such entities .

Twitter, one of the most used online social networks, provides a search system that allows users to query for tweets containing a set of keywords. Online Reputation Management systems often use Twitter as a source of information

when monitoring a given entity. However, search results are not necessarily relevant to that entity because keywords can be ambiguous. For instance, a tweet containing the word "columbia" can be related with several entities, such as a federal state, a city or a university. Furthermore, tweets are short which results in a reduced context for entity disambiguation. When monitoring the reputation of a given entity on Twitter, it is first necessary to guarantee that all tweets are relevant to that entity. Consequently, other processing tasks, such as sentiment analysis will benefit from filtering out noise in the data stream.

In this work, we tackle the aforementioned problem by applying a supervised learning approach. We studied a large set of features that can be generated to describe the relationship between an entity and a tweet and different learning algorithms. Concerning features, we used meta-data, tweet postings represented with TF-IDF, similarity between tweets and Wikipedia, Freebase entities disambiguation, feature selection of terms based on frequency and transformation of content representation using SVD. The algorithms tested include Naive Bayes, SVM, Random Forests, Decision trees and Neural networks.

The resulting classifier participated in the Filtering task of RepLab 2013 [1]. The corpus used for the competition consisted of a collection of tweets both in English and Spanish, possibly relevant to 61 entities from four domains: automotive, banking, universities and music.

The reminder of this paper consists in the overview of the Filtering task followed by the explanation of our methodology in Section 3. Experimental setup and results are described in Section 4 and 5, respectively, followed by the conclusion.

## 2 Task Overview

RepLab 2013 [1] focus on monitoring the online reputation of entities on Twitter. The Filtering task consists in determining which tweets are relevant to each entity. The corpus consists of a collection of tweets obtained by querying the Twitter Search API with 61 entity names during the period from the June 2012 until the December 2012. The corpus contain tweets both in English and Spanish. The balance between both languages varies for each entity. Tweets were manually annotated as "Related" or "Unrelated" to the respective target entity.

The data provided to participants consists in tweets and a list of 61 entities. For each tweet in the corpus we have the target entity id, the language of the tweet, the timestamp and the tweet id. The content of each URL in the tweets is also provided. Due to Twitter's terms of service, the participants were responsible to download the tweets using the respective id. The data related with entities contain the query used to collect the tweets (e.g. "BMW"), the official name of the entity (e.g. "Bayerische Motoren Werke AG"), the category of the entity (e.g. "automotive"), the content of its homepage and both Wikipedia articles in English and Spanish.

## 3 Methodology

The task we are tackling consists in building a relevance classifier: given an entity $e_i$ and a tweet $t_j$ we want to classify $t_j$ as *Related* or *Unrelated* to $e_i$. We use a supervised learning approach to address this problem. In this section, we describe our approach which comprises pre-processing of raw tweets and selecting the most appropriate feature representation of the relationship between entities and tweets.

### 3.1 Pre-processing

Contrary to other type of online texts (e.g. news or blog posts) tweets contain informal and non-standard language containing emoticons, spelling errors, wrong letter casing, unusual punctuation and abbreviations. Therefore, we apply some pre-processing techniques for text normalization. We use a tokenizer [2] optimized for segmenting words in tweets. After tokenization we apply the following procedure:

1. extract user mentions and URLs.
2. convert hashtags to words by removing the hash symbol.
3. remove all punctuation.
4. convert text to lower case.
5. remove accents and convert non-ASCII characters to their ASCII equivalent.
6. remove stopwords based on the list of stopwords for English and Spanish of NLTK.

We apply the same normalization process to metadata about the entities, namely query and entity name.

### 3.2 Features

We are interested in exploring the best combination of features to optimize relevance classification. We investigate several types of features: TF-IDF of n-grams, keyword similarities between tweets and entities as well as external resources projections.

**RepLab metadata:** we use entity's category, query and the language of tweets as features.

**TF-IDF:** we calculate TF-IDF of uni-grams, bi-grams and tri-grams using the normalized text of tweets.

**Text probability** : we encapsulate text in a single feature to avoid high dimensionality issues when adding other features. We use the TF-IDF of uni-grams, bi-grams and tri-grams for training a text classifier which calculates the probability of a tweet being related to the expected entity. We use the output probabilities of the classifier as a feature by applying a scheme of cross folds to train and classify within the training set. Regarding the test set, we use all tweets of the training set as training of the text classifier.

**Keyword similarity:** we calculate similarity scores between Replab metada and the tweets, by calculating the ratio of the number of common terms in the tweet and the terms of query and entity name. We also calculate similarities at character level in order to include possible spelling errors in the tweet. We apply the same procedure for user mentions and hashtags.

**Web similarity:** we calculate the similarity between the tweet text and the normalized content of the entity's homepage and normalized Wikipedia articles. The similarity value is the number of common terms multiplied by logarithm of the number of terms in tweet.

**Freebase:** For each keyword of the entity's query present in the tweet we create two bi-grams, containing the keyword and the previous/subsequent word. We submit these bi-grams to the Freebase Search API and compare the list of retrieved entities with the id of the target entity on Freebase. We calculate a Freebase score by using the inverse position of the target entity in the list of results retrieved. If the target entity is the first result, the score is 1, if it is the second, the score is 0.5, and so on. If the target entity is not in the results list, the score is zero. The feature corresponds to the maximum score of the extracted bi-grams of each tweet.

**Category classifier:** We create a sentence category classifier using the Wikipedia articles of each entity. We annotate each sentence of the Wikipedia articles with the category of the corresponding entity. We calculate TF-IDF for unigrams, bi-grams and tri-grams and train a multi-class classifier (SVM). We classify each tweet using this classifier. We use as feature the probability of the tweet being relevant to its target class.

**Twitter metadata:** We use URL domains, hashtags and user mentions as features.

## 4 Experimental Set-up

The dataset provided by the RepLab 2013 organization is divided in training, test and background. The test dataset is not labeled and it is used to create submissions for the competition. We discarded the background tweets which were also not labeled. The text and metadata of tweets was collected using a script provided by the organization. The training set consists in a total of 45671 tweets from which we were able to download 43582. Approximately 75% of tweets in the training set are labeled as "Related" as depicted in Table 1.

We split the training dataset into a development set and a validation set, containing 80% and 20% of the original, respectively. We adopted a randomly stratified split approach per entity, i.e., we group tweets of each entity and randomly split them preserving the balance of "Related"/"Unrelated" tweets. The submission dataset consists of 90356 tweets from which we were able to download 88934.

We used the development set for trying new features and test algorithms. We divided the development set in 10 folds generated with the randomly stratified approach. The validation set remained untouched until near the submission

| Dataset | Related | Unrelated | Total |
|---|---|---|---|
| Training | 33193 | 10389 | 43582 |
| Development | 26534 | 8307 | 34841 |
| Validation | 6659 | 2082 | 8741 |
| Test | - | - | 90356 |

**Table 1.** Dataset description.

deadline. At this time, we used the validation set to validate the results obtained in the development set. The purpose of this validation step is to evaluate how well our classifier generalizes from its training data to the validation data and thus estimate how well it will generalize to the test set. It allows us to spot overfitting. After validation, our submissions were trained using all of the data in the training dataset.

## 5 Results

We tried to create different submissions using different algorithms, features and we also tried to create entity specific models as explained in Table 2. We applied selection of features based on frequency and transformation of content representation using SVD. The algorithms tested include Naive Bayes, SVM, Random Forests, Decision trees and Neural networks. The evaluation measures used are accuracy and the official metric of the competition, F-measure which is the harmonic mean of Reliability and Sensitivity [3]. We submitted a total of 10 submissions to the RepLab competition though, we only present the top 4 submissions regarding the F-measure.

| Submission | Algorithm | Features | No. of models |
|---|---|---|---|
| popstar_filtering_2 | Random Forests | No TF-IDF and no Twitter metadata | 1, global |
| popstar_filtering_3 | Logistic Regression | Both TF-IDF and Twitter metadata | 1, global |
| popstar_filtering_7 | SVM | No TF-IDF and no Twitter metadata | 1, global |
| popstar_filtering_8 | Random Forests | No TF-IDF and no Twitter metadata | 61, 1 per entity |

**Table 2.** Submissions description.

Table 3 shows the results of our top submissions and the official baseline of the competition. This baseline classifies each tweet with the label of the most similar tweet of target entity in the training set using Jaccard similarity coefficient. The baseline results were obtained using 99.5% of the test set.

Based on the results achieved we are able to conclude that the models of our classifier are able to generalize successfully. Results obtained in the validation

| Submission | Accuracy (Val. Set) | Accuracy | Reliability | Sensitivity | F-measure |
|---|---|---|---|---|---|
| popstar_filtering_2 | 0.945 | 0.908 | 0.729 | 0.451 | 0.488 |
| popstar_filtering_3 | 0.947 | 0.907 | 0.765 | 0.440 | 0.480 |
| popstar_filtering_7 | 0.944 | 0.906 | 0.759 | 0.428 | 0.470 |
| popstar_filtering_8 | 0.948 | 0.902 | 0.589 | 0.444 | 0.448 |
| Official Baseline | - | 0.8714 | 0.4902 | 0.3199 | 0.3255 |

**Table 3.** Official results for each submission plus our validation set accuracy.

set are similar to those obtained in the test set. During development, solutions based on one model per entity were consistently outperformed by solutions based on global models. We also noticed during development that language specific models did not exhibit improvements in global accuracy, therefore we opted to use language as a feature. Results show that the best submission uses the Random Forests classifier with 500 estimators for training a global model and it does not contain the TF-IDF feature. Though, the Text Probabilities feature encapsulates text by using a specific model trained just with TF-IDF of n-grams of tweets.

## 6 Conclusion

In this paper we have described the POPSTAR participation at the Filtering task of RepLab 2013. The main goal of this task was to classify tweets as relevant or not to a given target entity. We have explored several types of features, namely similarity between keywords, TF-IDF of n-grams and we have also explored external resources such as Freebase and Wikipedia. Results show that it is possible to achieve an Accuracy over 0.90 and an F-measure of 0.48 in a test set containing more than 90000 tweets of 61 entities.

## References


1. E. Amigó, J. Carrillo de Albornoz, I. Chugur, A. Corujo, J. Gonzalo, T. Martin, E. Meij, M. de Rijke, and D. Spina, "Overview of replab 2013 evaluating online reputation monitoring systems," in *Fourth International Conference of the CLEF initiative, CLEF 2013, Valencia, Spain. Proceedings*, Springer LNCS, September 2013.
2. G. Laboreiro, L. Sarmento, J. Teixeira, and E. Oliveira, "Tokenizing micro-blogging messages using a text classification approach," in *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, AND 10, (New York, NY, USA), pp. 81–88, ACM, 2010.
3. E. Amigó, J. Gonzalo, and F. Verdejo, "A general evaluation measure for document organization tasks," in *Proceedings SIGIR 2013*, July.