Media Monitoring using News Recommenders

Francesco Barile
Free University of Bozen-Bolzano
francesco.barile@unibz.it

Francesco Ricci
Free University of Bozen-Bolzano
francesco.ricci@unibz.it

Marko Tkalcic
Free University of Bozen-Bolzano
marko.tkalcic@unibz.it

Bernardo Magnini
Fondazione Bruno Kessler
magnini@fbk.eu

Roberto Zanoli
Fondazione Bruno Kessler
zanoli@fbk.eu

Alberto Lavelli
Fondazione Bruno Kessler
lavelli@fbk.eu

Manuela Speranza
Fondazione Bruno Kessler
manspera@fbk.eu

KEYWORDS

Media Monitoring, News Recommender Systems, Content Analysis

1 INTRODUCTION

Media monitoring (MM) services, also referred to as clipping services, provide their customers with a daily selection of media content that is of interest to them. Such content, here referred to as documents, can be obtained from any kind of media, such as newspapers and other print media, video and audio services, and web and social media [5]. This monitoring service is used by companies to analyse special topics of interest, in order to determine the impact on the market and the value of their brands, but also to monitor competitors and to protect their reputation and plan the company's policies [6, 8]. There are several MM service providers, but very few papers in the literature describe the adopted technological solutions. Meltwater [1] is a monitoring service tracking keywords and phrases on more then 300,000 online sources, offering a personalized dashboard that allows customers to perform different analyses on the retrieved documents. Cision [2] and News Exposure [3] provide monitoring on different kinds of media: online monitoring on internet, broadcast monitoring on TV and radio transmissions, print monitoring on newspapers and social monitoring that analyze the social networks, also providing analytic tools. Mention [4] is a social media monitoring, hence it focuses on web and social media contents, providing tools to monitor in real-time customers' mentions and allowing also the tracking of competitors.

Even though these tools use data mining techniques to analyze the documents and support the customers with reports and statistics, the selection of the documents related to the customers is mostly based on keyword matching techniques by using keywords that reflect the customers' name, products and competitors. However, keyword-based techniques are not precise and it is necessary to manually inspect the keyword-filtered documents to remove false positives: a time-consuming process. Therefore, the work-flow in the customer company typically involves a human editor who inspects, on a daily basis, each document provided by the MM service and decides whether it is really relevant or not. In order to minimize the daily work done by human editors, we have developed a recommender system (RS) that operates after the keyword-filtering step and before the final check of human editors. Our RS uses automatic classification techniques in order to label the keyword-filtered documents either as relevant or non-relevant and to ease the work of the editors.

The work presented here was done in conjunction with the Italian company *Euregio*. The company provides the *Infojuice system* (IJ), a MM service providing a tool for the qualitative and quantitative analysis of the customers' level of representativity on the media. The product is able to collect documents in Italian, German and English, from a range of different sources. More details on these results can be found in [1].

2 THE PROPOSED SOLUTION

The proposed solution is a recommender system (RS) that identifies document recommendations by using one classifier for each customer. The RS is used by the IJ system to support the editors during their daily activity. Several factors have been taken into consideration in the design of the RS. In particular, since the number of customers is small and their interests quite diverse, collaborative filtering was not applicable. Furthermore, the system collected feedback of the editors is restricted to the actions of removing documents from the lists produced by the keyword-based filtering queries; hence, we have at disposal only negative feedback. Therefore, we decided to implement a solution optimized differently for each customer by using automatic classification techniques that leverage data generated by the editors' actions in order to distinguish relevant from non-relevant documents.

We decided to perform the classification tasks with two classical approaches: Multinomial Naïve Bayes (MNB) and Support Vector Machines (SVM) [4]. In particular, regarding the SVM classifier, we decided to use a linear kernel (LIN SVM) and an exponential one, the radial basis function kernel (RBF SVM) [3]. Regarding the features, extracted from the documents and used by the classifiers, they represent information derived from the *Title*, *Text* and *Source* of the documents. The document's Source (for instance a magazine)

[1] Meltwater Official Website. https://www.meltwater.com/uk/ [Accessed 02.09.2019]
[2] Cision Official Website. https://www.cision.com/us/ [Accessed 02.09.2019]
[3] News Exposure Official Website. https://newsexposure.com/ [Accessed 02.09.2019]
[4] Mention Official Website. https://mention.com/en/ [Accessed 02.09.2019]

IIR 2019, September 16–18, 2019, Padova, Italy

Table 1: Aggregate performance results (P, R and F1) obtained using the best features configuration for three classifiers (MNB, LIN SVM and RBF SVM) when the hyper-parameters are set to optimize either precision or F1. KBS is the keywords based filtering system.

Optimization	MNB			LIN SVM			RBF SVM			KBS
Score	*Prec.*	*Recall*	F_1	*Prec.*	*Recall*	F_1	*Prec.*	*Recall*	F_1	*Prec.*
Precision	0.853	0.703	0.782	0.880	0.597	0.729	0.989	0.140	0.204	0.747
F_1	0.801	0.918	0.850	0.841	0.894	0.866	0.841	0.891	0.860	0.747

is represented as a binary vector, with a binary feature associated with each one of the available sources. For the textual components of the documents, i.e., the *Title* and the *Text*, we use a Bag of Words Model (*tf-idf*) and a Word Embeddings Model (*Fasttext* [2]).

3 SYSTEM EVALUATION

We evaluated the proposed RS in a set of offline experiments using a dataset containing 365,430 Italian documents and the editors' actions performed over a time window of six months. We considered five reference customers, whose editors performed removing actions in all the six months considered. We evaluated several different configurations specified by the classifier, hyper-parameters optimization criteria (P and F1) and used feature sets combination. The evaluation of a configuration requires two steps: (1) the selection of the hyper-parameters for each classifier and optimization criteria, and (2) the evaluation of the configurations by using the best hyper-parameters determined at the previous step. Since for many customers the number of relevant and not relevant documents was substantially different in each test month, we used resampling strategies for balancing these numbers [7]. We found the optimal values of the classifiers hyper-parameters by using grid search. Our data is time stamped, so, random cross-validation is not appropriate; validation data must be posterior to training data. We defined two time correct train-validation splits and searched for the hyper-parameter values that maximise the averaged obtained scores (P or F1) over these two splits. In the first split, the first two months were used as training and the third month as validation; then the first three months were used as training and the fourth month as validation. Once we have identified the best hyper-parameter combination for a given configuration, we evaluated the performance of that configuration using the fifth and the sixth month as test set, referred to as test month 5 and test month 6 respectively. In the first case, we trained the classifier using the first four months as a training set, while in the second case we used the first five months as a training set.

4 RESULTS ANALYSIS

As shown in Table 1, all the proposed algorithms outperform (improve) the precision of the current Keyword-Based System (KBS), which is the ratio of the number of documents the editor considered as relevant over the total number of documents selected by KBS. Moreover, high *Precision* (equal or close to 1.000) can be obtained if this is used in hyper-parameters selection step. This means that the system can find a set of relevant documents with few false positives (i.e., documents identified as relevant are mostly truly relevant). We also measured the computational time needed to train the classifiers and to perform the recommendations on the real data of customers and documents managed by the IJ system. We have observed that the proposed approach can be executed on off-the-shelf hardware in reasonable time and does not require to alter the current workflow.

5 CONCLUSIONS

We have presented a system that reduces the daily work of an editor for generating press release selections from a large data set of documents managed by a media monitoring (MM) system. In the future we plan to improve the developed system. One line of research will study techniques to identify the documents that are clearly non-relevant, hence shrinking even more the grey area of documents that the editor needs to inspect. Additionally, we plan to evaluate the system's performance in a multilingual scenario by using documents of different languages (Italian, German, and English) and language-independent features. Furthermore, we plan to collect additional feedback from the editors and also from end users of the system in order to generate personalised press releases. Finally, an online evaluation will be performed to validate the offline results in a real scenario.

ACKNOWLEDGEMENT

This work was supported by the autonomous province of Bolzano-Bozen (Alto Adige-Südtirol) under the EUCLIP_RES project (EUregio CrossLinguistic REcommender System).

REFERENCES

[1] Francesco Barile, Francesco Ricci, Marko Tkalcic, Bernardo Magnini, Roberto Zanoli, Alberto Lavelli, and Manuela Speranza. 2019. A News Recommender System for Media Monitoring. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI '19)*. ACM.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[3] Aurélien Géron. 2017. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.* " O'Reilly Media, Inc.".

[4] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Text classification and Naïve Bayes. *Introduction to information retrieval* 1, 6 (2008).

[5] Stephen D Rappaport. 2010. Listening solutions: A marketer's guide to software and services. *Journal of Advertising Research* 50, 2 (2010), 197–213.

[6] Ioannis Stavrakantonakis, Andreea-Elena Gagiu, Harriet Kasper, Ioan Toma, and Andreas Thalhammer. 2012. An approach for evaluation of social media monitoring tools. *Common Value Management* 52, 1 (2012), 52–64.

[7] Aixin Sun, Ee-Peng Lim, and Ying Liu. 2009. On strategies for imbalanced text classification using SVM: A comparative study. *Decision Support Systems* 48, 1 (2009), 191–201.

[8] Boyang Zhang and Marita Vos. 2014. Social media monitoring: aims, methods, and challenges for international companies. *Corporate Communications: An International Journal* 19, 4 (2014), 371–383.