Media Placement: Using Sentiment Analysis in Brand Reputation Maintaining*

Dmitry Kuznetsov[1], Ilya Gavrilov[1], Nikita Benkovich[1], Tatyana Charnetskaya[1], and Rostislav Yavorskiy[2]

[1] Higher School of Economics
Moscow, 101000, Russia
E-mail: dskuznetsov_5@edu.hse.ru
[2] Surgut State University
Surgut, 628403, Russia
E-mail: javorski_re@surgu.ru

Abstract. Sentiment analysis is a class of text analysis methods in Natural Language Processing designed to automatically identify emotionally coloured vocabulary and emphatic evaluation of authors regarding objects in the text. Businesses are using such methods to analyze the competitive environment and monitor the customers' opinion. In this paper, we present Media Placement news analysis tool targeted at measurement of brand recognition and media coverage in online media.

Keywords: Natural Language Processing · Sentiment Analysis · Machine Learning

1 Introduction

The world economy is increasingly determined by industries with services and intangible assets put in spot light. Nowadays, intangible assets such as brand recognition, experience and tacit knowledge of employees, relationships and informal commitments bring for most companies a larger share of their total value than tangible assets such as equipment and infrastructure. Intangible assets and their effective management are the key to success in the long term, and many research papers have been devoted to this subject, see e.g. [3, 10].

Brand reputation of a company is affected by various factors including the product performance [11], employee behaviors [4], quality of customer service [6] etc. Our research is focused on analyzing of the company's image on the basis of open media data. Namely, local news media. Automatic analysis of news pieces seems to be beneficial for two reasons. Firstly, they are "hot" information and feedback, and secondly, the way they are delivered is meant to be objective. The process of news mining involves the following steps:

– parsing of news websites to obtain the text to be analyzed;

– recognition of companies' names including abbreviations and different spellings;
– detection of lexical homonymy and resolving lexical ambiguity;
– sentiment analysis of news related to a certain company.

As a pilot project, there has been conducted automatic news analysis of Krasnodar region (Russia) local media. As a result of this research we have estimated brand reputation profiles of major corporations which are operating in southern Russia.

2 Technology of sentiment analysis in business

Our overview is based on paper [5] and [7, 9] There are 4 main approaches most vendors use today for text sentiment analysis.

1. **Rule-based approach.** Such an approach consists of a certain set of rules on the basis of which the system draws a conclusion about the tonality of a text piece. This approach may deliver good results with a large set of rules. At the same time, drawing up a large set of rules is a very time consuming process. The rules are often tied to a specific subject area; and this approach is not suitable in case of noisiness of data due to of errors in the text.
2. **Sentiment lexicon approach.** In this case, each word in the dictionary is assigned a tonality value. To obtain the final tonality value, one has to take the arithmetic average or calculate the sum of the tonality values of all the words in a document. A similar, yet more complex way is to train a classifier (for example, a neural network) on a specially selected labeled dataset which takes into account the peculiarities of a field dealt with. Examples of such dictionaries for the English language are SentiWordNet [2], ANEW [8], etc. The main restriction of this method the need to come up with a new set of vocabulary for each new subject area.
3. **Unsupervised learning.** The difference from the previous method is that hidden patterns and relations between objects which define the sentiment are detected from unlabeled data (or labeled data is taken, but labling is ignored). One should take into account that there could be low accuracy compared to the supervised learning.
4. **Supervised learning.** The algorithm for implementing this approach can be briefly described as follows:
   (a) Before proceeding with the algorithm, one is expected to decide how many classes and which type of classification will be used. When putting forward a plain classification, it is very difficult to achieve good results. Research shows [1] that hierarchical classification gives much better results.
   (b) Firstly, one needs to collect a set of documents; on its basis the classifier is to be trained. The bigger the set is, the better. There is no point in collecting less than ‘10.000 items.

(c) Each document should be presented in the form of a n-dimensional feature vector. The quality of the classification directly depends on which set of characteristics will be used. The most common ways of presenting documents are either in the form of a so-called bag of words or in the form of n-grams. In order to compose a vector, it is necessary to assign a weight to each of its attributes. A common method for estimating weight is the TF-IDF measure. The bottom line is to put more weight on words that have either an obvious positive or negative tone.
(d) Each document should be assigned with a correct type of tonality. It is usually done manually which is the most time and effort consuming part of the whole process.
(e) Applying the resulting model employing the chosen classification algorithm and method for training the classifier.

All the above methods have been used for business purposes. We see the forth approach as the most efficient and promising.

3 The Media Placement Assignment

Media Placement Startup based in Moscow, Russia, ventures to offer a complex sentiment analysis tool to big companies so that they get an idea of where they are in terms of brand recognition and media coverage compared with their competitors.

In this paper we present our pilot project, which is based on news data from Russian open new media sources such as RBC[4], Krasnodarmedia[5], Kommersant Yug, Delovaya Gazeta.

3.1 Company name recognition

For company recognition we used a simple but effective approach based on dictionaries and rules because there aren't a lot of ways to mention a company in the text compared to a more general problem called entity recognition with using more complicated methods such as neural networks.

Firstly, we collected a dictionary of company names. To do that, we parsed a website of top 500 Russian companies[6]. Then we applied our rules to detect potential company names, obtained the top of the most frequently run across company names in our data, excluding companies which are in the dictionary, verified them manually and added to the dictionary. We split our dictionary into two parts due to the fact that some companies have names which do not represent a unique word, i.e. a word has a meaning and used in context not a as company name, or some companies have the same names as another companies (but in a different domain), projects, funds and etc. First part is a dictionary

[4] http://rbc.ru
[5] https://krasnodarmedia.su/
[6] https://ru.investinrussia.com/russia-200

which contains “strong” names that are unique and used only as a company name, for instance, “Sberbank”. The second part is a dictionary with “weak” names which will be used not only as a company name but also as a regular word or name of a project or fund etc, for example, “Saturn” or “Vozrozhdenie”. We are not sure in names from the second dictionary, so we applied some verification rules which will be described later.

As mentioned above, we used simple rules to detect potential companies. We split rules into two types: “strong” rules which we are sure in, and “weak” ones which we aren't sure in and it's just a possible company's name to be verified manually. The strong rule description assumes it's a word in quotas which are followed by special word determined type of organisation, for instance “OAO”, “PAO” and etc. It is very strict rule and we've never seen false positive detection, but it captures very limited number of companies. The second rule is weak and we determine this as words in quotas and there is some special word in a window of this words such as “company”, “bank”, “fabric” and etc. Verifying rules for a weak dictionary are almost the same: if the name is followed by one of the special word such as “OAO”, or there is a special word in a window around such as “company” or “enterprise”. Then, we introduced some minor additional rules, such as verifying that the first letter is capitalized, a potential company's name (weak rule) is less than 5 words to avoid detecting some quotes which may include some special words around, if there are more open quotas than close (and obviously in this case some of them are nested), that first close quota close two last open, checking that detected company names aren't nested and if they are, we'll remove the shortest.

As a result, with these two simple approaches we covered a big part of company names which would do for the MVP stage. Finally, we decided that existing datasets for NER consist a lot more wise definition of organisation than just companies, so it's not suitable way to estimate perfomance of our algo. As a result, we marked our parsed data manually and estimated the algo. We got 87% of accuracy, 80% of recal and precision equals 93%.

3.2 Sentiment analysis

As a first step of analyzing news we picked binary classification for sentiment analysis. For that we used open dataset for sentiment analysis from linis source, trained model on this and then applied to our news data. Learning dataset includes 5 classes from negative to positive, we treated them as all negative (-2, -1 classes) as negative and the rest (including neutral) as a positive, because we have hypothethis that if company is mentioned even in a neutral context it's a good, because people will remember the company and when they want to buy some goods they will think about this company firstly neither about their competitors. Then we tried two approach to build final model.

First approach is a simple baseline, we use TF-IDF encoding over normalized words (with pymorphy2) and logistic regression over the encoding. Quite a tricky task was to tune the parameters such as regularization, type of optimization, type of a kernel etc. We split learning dataset into train and test and estimated

model quality on test part by AUC-ROC, then chose the best and the simplest approach: logistic regression with a linear kernel and next hyperparameters: penalty = l2, C = 1, solver = newton-cg, final score of the model is 0.74. Then we applied the model on our news data and manually read some news and scored. Finally we got approximately 0.85 AUC-ROC score.

As for the second approach, we used the pretrained fastText model for word embedding and full text vectorization. To vectorize the text, we comupted coordinate-wise average, maximum and minimum of all words and concatenate this statistics into one vector. After that we apply logistic reggresion and chose hyperparameters as in previous approach. Our setting is logistic regression with penalty = l2 , C = 10, newton = cg optimizer. That brought about the improvment of our final AUC-ROC score from 0.85 to 0.89. This approach perfoms much better and seems good enough to move further to do more comlicated analysis such as splitting sentiment into parts and analyse which of them if more improtant and etc.

Model	ROC-AUC	Final ROC-AUC
LogReg + Tf-idf	0.74	0.85
LogReg + FastText	0.75	0.89

Table 1. Results of models

The results of our experiments are presented on table 1. Figure 4 shows that FastText performs better, then tf-idf approach in terms of precision-recall ratio.

4 Conclusion

The presented project is currently under development. We also consider other areas of application for the techniques described above. One possible direction is analysis of medical records.

We would like to thank Dr. Oleg Lavrov for his participation and input, and very valuable comments on this work.

References

1. Babbar, R., Partalas, I., Gaussier, E., Amini, M.R.: On flat versus hierarchical classification in large-scale taxonomies. In: Advances in neural information processing systems. pp. 1824–1832 (2013)
2. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Lrec. vol. 10, pp. 2200–2204 (2010)
3. Barth, M.E., Clement, M.B., Foster, G., Kasznik, R.: Brand values and capital market valuation. Review of accounting studies **3**(1-2), 41–68 (1998)

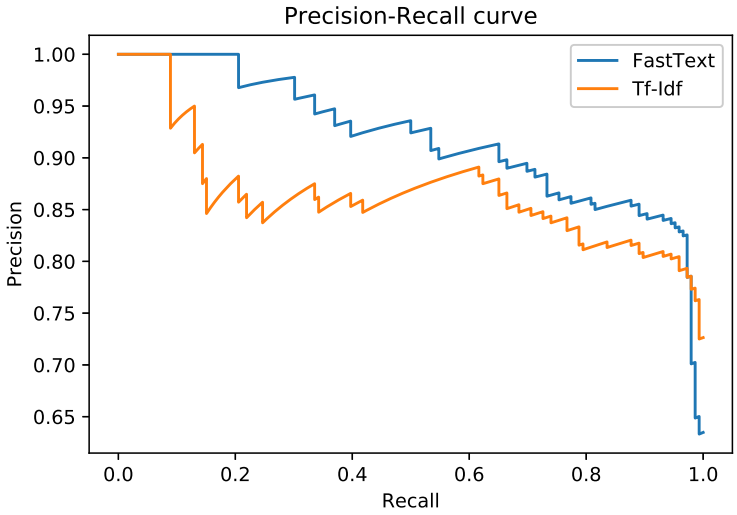

Fig. 1. PR plot. On this figure we see that FastText performs better, then tf-idf approach in terms of precision-recall ratio

4. Biedenbach, G., Bengtsson, M., Wincent, J.: Brand equity in the professional service context: Analyzing the impact of employee role behavior and customer–employee rapport. Industrial Marketing Management **40**(7), 1093–1102 (2011)
5. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
6. Cretu, A.E., Brodie, R.J.: The influence of brand image and company reputation where manufacturers market to small firms: A customer value perspective. Industrial marketing management **36**(2), 230–240 (2007)
7. Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies **5**(1), 1–167 (2012)
8. Nielsen, F.Å.: A new anew: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903 (2011)
9. Pang, B., Lee, L., et al.: Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval **2**(1–2), 1–135 (2008)
10. Salinas, G.: The International Brand Valuation Manual: A complete overview and analysis of brand valuation techniques, methodologies and applications. John Wiley & Sons (2011)
11. Selnes, F.: An examination of the effect of product performance on brand reputation, satisfaction and loyalty. European Journal of marketing **27**(9), 19–35 (1993)