Chapter 78
Natural Language Processing as Feature Extraction Method for Building Better Predictive Models

Goran Klepac
Raiffeisen Bank Austria d.d., Croatia

Marko Velić
University of Zagreb, Croatia

ABSTRACT

This chapter covers natural language processing techniques and their application in predicitve models development. Two case studies are presented. First case describes a project where textual descriptions of various situations in call center of one telecommunication company were processed in order to predict churn. Second case describes sentiment analysis of business news and describes practical and testing issues in text mining projects. Both case studies depict different approaches and are implemented in different tools. Language of the texts processed in these projects is Croatian which belongs to the Slavic group of languages with more complex morphologies and grammar rules than English. Chapter concludes with several points on the future research possible in this domain.

INTRODUCTION

In big data era, predictive models development should not be based on internal data from structured relation databases as the only disposable data sources for model development. Growing trend of unstructured data gives us opportunity to use potentials from unstructured data sources.

This does not mean that traditional methodology for predictive model development should be neglected; it means that it should be improved with patterns from unstructured data for better performance of the models. For the business problems solving, like churn prediction, fraud detection or other predictive

DOI: 10.4018/978-1-5225-1759-7.ch078

model development in business, introducing elements (patterns) found by natural language processing into predictive business model development introduces gains on model reliability and efficiency.

Traditional approach to predictive model development does not consider textual data as valuable data source for model constructions. Textual data sources like customer comments in call centers or similar data sources are excluded from model development sample, even if it could contain valuable information in domain of churn understanding/ prediction, fraud understanding/ prediction, customer needs for the next best offer modeling etc. Main reason for that is unclear methodology and idea how to use it, beside common attitude that this type of data is useless for predictive business statistical model development based on Bayesian networks, logistic regression, neural networks or similar.

Croatian language is a member of the Slavic group of languages together with Bosnian, Slovenian, Serbian, Macedonian, Russian, Czech, Polish, Ukrainian etc. Altogether Slavic group counts 18 different languages and is spoken by more than 200 million people.

Slavic languages are similar in the roots of the many words and different in grammar rules. Considering natural language processing techniques that try to mitigate problems with grammatical and different morphological word features it is reasonable to assume that it is worth experimenting with models on different languages. More on this will be covered in final sections of this chapter.

Chapter will give solutions on how unstructured data (different kind of text data) with natural language processing could be used as the elements for building better business predictive models. This will be illustrated with two cases.

First case will describe a scenario where a telecom company wants to develop churn predictive model. Case will show how this company used textual data from call center (customer comments written by operators in call center). Collected textual data contains variety of information, questions, and comments from customers entered into textual fields by operators. It contains questions about new services/ products, notifications about equipment failure, questions about bills etc. Natural language processing showed some patterns within textual data, which showed strong impact on churn commitment. Recognized textual pattern leads company to conclusion about churn nature and causes. Characteristic of this case is relying on internal data sources – structured and unstructured, where recognized textual pattern could be joined to unique customer, which is important when we want to make predictive business data mining models on customer level.

Second case will show different scenario – developing predictive models for stock market. In this case, public text data will be used for predictive model developing purposes. Stock market predictions are often based on previous price trends (technical analysis) and company's financial reports (fundamental analysis). There are systems that include collaborative filtering methods (also known as Wisdom of the Crowds) where many users rate stocks, similar to rating movies or books on popular online systems. In addition, there are advancements in sentiment analysis where stock market news or social network messages are being processed to identify possible future trends. This section will show one case where technical analysis, fundamental analysis and collaborative filtering are already in use on one Croatian stock market web portal. In addition, chapter will present development of the sentiment analysis module for mining business news. In the effort to collect annotated dataset that would allow for sentiment analysis, experts (brokers) are asked to annotate more than 500 business news i.e. RSS abstracts of the news collected by the portal's web parsers.

RSS news include headings, abstracts, date and the link to the source. Experts' inputs are companies that news relates to and the overall sentiment for the particular company. Sentiment can be positive, neutral or negative. Relation news-stock-sentiment is important since the same news can be positive for