Research Proposal

Title of research: Multi-lingual Text Classification of Financial Documents

Name of student: Halvard Skogsrud

Proposed supervisor: Dr Rafael A. Calvo

Keywords: information retrieval, document classification, machine learning

Introduction and Synopsis

One of the biggest challenges facing financial research and trading organizations is how to well exploit unstructured financial announcements such as textual announcements. The automatic classification of this type of data poses many challenges for learning systems because the feature vector used to represent a document must capture some of the complex semantics of natural language. The proposed research will focus on classification of financial news in a multi-lingual context. Several types of machine learning techniques will be studied and their performances measured. Using a combination of several such techniques will also be studied, as this could lead to performance improvement. One of the goals is to produce a system for the classification of financial news that can be applied to a data set of any language.

Financial Documents

Financial analysts, traders and government agencies regulating the financial markets must process thousands of daily announcements. These announcements are sometimes relevant to their particular area of interest, have market-wide impact, or will have an impact in the value or a particular share. Thus, analysts and traders are often searching for the proverbial needle in the haystack, reading through corporate announcements, press releases and even searching for potentially useful “gossip” in discussion forums and chat rooms in the World Wide Web. Meanwhile, government agencies are trying to find signs of market manipulation, such as individuals and organizations trying to change the value of a share by releasing false information. They are also searching for signs of insider trading, the trading of public shares based on non-public information. The agencies are also searching for uncertified individuals giving financial guidance. Automatic document classification is appropriate in all of the scenarios mentioned above. These classification techniques are described by several authors, see [12, 4, 2]. A successful application of some of these techniques is shown in [1].

Vector Models

Vector models have been successful in information retrieval and text classification, see [7, 8]. Vector models are based on the assumption that a document can be represented as a vector, dismissing the order of words and other grammatical issues, and that this representation is able to retain enough useful information to enable text classification.

In order to reduce the number of distinct terms in the database, a list of stopwords is removed from the documents. This technique is common for most languages where some words have low information content. In English and many European languages, articles and prepositions are among the words considered stopwords. Removing 100-300 stopwords from English documents reduces the document length by 30–40

Several weighting schemes can be used in the vector representation for the values of terms (words). Two common weighting schemes are *Term Frequency* and *Inverse Document Frequency* weighting. These schemes are described in more detail in [7].

The dimensions of these vector spaces is proportional to the number of terms remaining after stopword removal and stemming. Even for moderate-size text collections this could be tens or hundreds of thousands of terms. This is prohibitively high for some algorithms, so dimensionality reduction techniques are needed. The χ^2 statistic is one such technique. The underlying assumption in this technique is that features whose appearance in a document is highly correlated to a class membership will be useful for measuring class membership. In [12], different dimensionality reduction techniques were studied, and it was found that *Information Gain* and χ^2 were the most effective.

Machine Learning Techniques

Different statistical approaches can be used as classifiers, these are usually machine learning techniques. Available techniques include k-nearest neighbours (kNN), linear least square fit (LLSF), support vector machines (SVM), Naive Bayes (NB) and neural networks (NNet). In [11], an evaluation of these approaches, applied to document classification, was undertaken. Neural networks was used as the classifier in experiments performed in [1]. An introduction to neural networks can be found in [3, 6]. Part of the research will look into comparing classifiers for the document sets, as well as evaluating the use of averaging over a number of simultaneous learning methods. For measuring the performance of various learning techniques, *recall* and *precision* are commonly used performance measures. Another common performance measure is the F-measure.

When dealing with documents from multiple classes, there are two possible ways of averaging these measures, namely macro-average and micro-average.

In macro-averaging, one result table is used for each class, the performance measures are computed on each of them and averaged. In micro-averaging, only one result table is used, and an average of all the classes is computed for each cell in this table. The macro-average weights equally all the classes, regardless of how many documents belong to it. The micro-average weights all documents equally, thus favoring the performance on common classes.

Different classifiers will perform different in common and rare categories. Learning algorithms that are trained more often on more populated classes thus risk local overfitting.

Research problems, methods and approaches

This Master of Engineering thesis project will consist in implementing new document classification models, such as Neural Networks and K-Nearest Neighbours. Their performance will be measured over the corpus of the Australian Stock Exchange (ASX) announcements and the Oslo Stock Exchange press releases and announcements. The implementation will be preceded by a system design of a whole document classification application.

The expected outcomes of the project are:

- **N**eural Network classifier. Its performance will be compared with baseline results obtained by other authors on the Reuters corpus [11, 1].
- **K**-Nearest Neighbours classifier.Its performance will be compared with baseline results obtained by other authors on the Reuters[11, 1].
- **E**nglish documents classification. A new corpus of announcements from the Australian Stock Exchange will be used measuring the performance of the above two methods
- **N**orweigian documents classification. A new corpus of announcements from the Oslo Stock Exchange will be used, and the performance of the Neural Network and kNN classifier will be compared.

References

[1] R. A. Calvo and H. A. Ceccatto. Intelligent document classification. *Intelligent Data Analysis*, 4(5), 2000.

[2] W. B. Croft, editor. *Advances in Information Retrieval.* Kluwer Academic, 2000.

[3] J. Hertz, A. Krogh, and R. Palmer. *Introduction to the theory of neural computation*. Addison-Wesley, Redwood, CA, 1991.

[4] C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[5] W. G. Ng and K. Low. Feature selection, perceptron learning, and usability case study for text categorization. In *Conference on Research and Development in Information Retrieval (SIGIR'97)*, pages 67–73, 1997.

[6] B. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, Cambridge, 1996.

[7] G. Salton. *Automatic text processing: The transformation, analysis and retrieval of information by computer*. Addison-Wesley, Reading, MA, 1989.

[8] G. Salton. Developments in automatic text retrieval. *Science*, 253:974–979, 1991.

[9] H. Schutze, D. Hull, and P. J. O. A comparison of classifiers and document representations for the routing problem. In *Conference on Research and Development in Information Retrieval*, pages 22–34, 1995.

[10] J. P. Wiener and A. Weigend. A neural network approach to topic spotting. In *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95)*, 1995.

[11] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Conference on Research and Development in Information Retrieval (SIGIR'99)*, 1999.

[12] Y. Yang and J. Pedersen. Feature selection in statistical learning of text categorization. In *The Fourteenth International Conference on Machine Learning*, pages 412–420, 1997.