Computational Content Analysis of European Central Bank Statements

Viorel Milea[1], Rui J. Almeida[1], Nurfadhlina Mohd Sharef[2], Uzay Kaymak[1,3], Flavius Frasincar[1]

[1]Econometric Institute,
Erasmus School of Economics,
Erasmus University Rotterdam,
P.O. Box 1738, 3000 DR Rotterdam, the Netherlands
{*milea, almeida, frasincar*}@*ese.eur.nl*

[2]Intelligent Computation Group,
Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia, Malaysia
fadhlina@fsktm.upm.edu.my

[3]Information Systems Group,
School of Industrial Engineering,
Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven, the Netherlands
u.kaymak@ieee.org

Abstract: **In this paper we present a framework for the computational content analysis of European Central Bank (ECB) statements. Based on this framework, we provide two approaches that can be used in a practical context. Both approaches use the content of ECB statements to predict upward and downward movement in the MSCI EURO index. General Inquirer (GI) is used for the quantification of the content of the statements. In the first approach, we rely on the frequency of adjectives in the text of the ECB statements in relation to the content categories they represent. The second approach uses fuzzy grammar fragments composed of economic terms and content categories. Our results indicate that the two proposed approaches perform better than a random classifier for predicting upward or downward movement of the MSCI EURO index.**
Keywords: index prediction, fuzzy inference system, fuzzy grammar, General Inquirer, MSCI EURO.

I. Introduction

For a large part, corporate as well as government communication consists of free text. Although accessible to the human mind, such information fragments are difficult to process in an automated way. When faced with high volumes of such information, one would find it desirable to use machines for the processing, interpretation and aggregation of this knowledge. Ideally, such a process should lead to an advice in the form of a recommended decision that follows from the text that is being considered.

This issue is especially relevant for financial investors, who are often faced with high volumes of information and are under time pressure to incorporate this into their decision-making, in a time short enough to provide a competitive advantage over other market participants. The sources of information can be very diverse, ranging from formal means of communication to social media. Indeed, different studies show that European Central Bank (ECB) statements hold predictive power over financial markets [15, 16], and that the general mood on Twitter can be used in predicting upward or downward movement in the Dow Jones Industrial Average (DJIA) index [5].

Most approaches to content and sentiment analysis in economic text are currently focused on different isolated problems rather than providing frameworks for this problem. In this paper, we aim at providing a general framework towards the automated analysis of ECB statements, with the goal of aiding decision-makers with investment decisions. We focus the analysis on the content of fragments of text, based on the General Inquirer service [29], which employs the Harvard-IV-4 and Lasswell content dictionaries.

We provide two approaches for our proposed framework, that can be used in a practical context. The first focuses on the frequency of content categories as encountered in text. The second one takes a more sophisticated approach in that, rather than focusing on word frequencies, it focuses on fragments of text containing both an economic term as well as a word denoting some content category. Again, the frequency of such fragments is measured in text.

The information source we choose consists of European Central Bank (ECB) communication. The statements we consider have appeared monthly starting at the end of 1998. In addition to discussing the levels of the key interest rates in the European Union, these statements provide an overview of the

economy in the past month, as well as an economic outlook for the period succeeding the issued statement. Given the importance of these statements to financial markets, the anticipation with which they are received by market participants, and the extensive information they contain, we deem these statements relevant for the price forming mechanism of European assets. Hence, we hypothesize that automated content extraction of the ECB statements can help predict a Europe-wide financial market index, such as the MSCI EURO. For modeling the levels of the index, we rely on a Fuzzy Inference System (FIS), mainly due to the interpretability of such models, which provides us with some insight into the relationship between the content of the ECB statements and the movement of the index.

We validate our framework by measuring the accuracy of the proposed approaches in terms of correctly predicting upward or downward movement of the index. We find that both approaches show a performance that exceeds the accuracy of a random classifier, thus validating our general framework fpr automated financial decision support based on economic text. While one approach gives a better accuracy on the test set, the other one helps at reducing the number of features used for prediction and gives more stable models.

The outline of the paper is as follows. In Section II we present studies related to the extraction of content and sentiment from text. Section III presents our general framework for the content analysis of ECB statements. Two approaches based on our framework are described in Section IV. We present the fuzzy model that we use for the analysis in Section V. The experimental setup and the results are presented in Section VI. Our conclusions and suggestions for further work are described in Section VII.

II. Content and Sentiment Analysis

The first attempt at content analysis in an economic context is presented in [10]. Here, the authors investigate the relationship between a focus on wealth and wealth-related words in the speeches of the German Emperor and the state of the economy over the period 1870-1914. They find a strong relationship between the focus on wealth and the state of the German economy. More recent research, such as [27], relies on the GI dictionary for explaining the market prices and the trading volumes. The author finds that a relationship exists between a daily Wall Street Journal column, 'Abreast of the Market', and the market prices and trading volumes of that day for the stocks discussed in the column. In [19] the authors develop a method for the automated extraction of basis expressions that indicate economic trends. They are able to classify positive and negative expressions which hold predictive power over economic trends, without the help of a dictionary.

Other research has focused on the extraction of sentiment from free text in an economic context. In [28], the authors focus on eight dimensions of sentiment: joy, sadness, trust, disgust, fear, anger, surprise, and anticipation. They are able to provide visualizations of how these eight sentiments evolve over time for some concept, e.g., Iraq, based on news messages. The results are validated against ratings of human reviewers of the news messages. The method performs satisfactorily in visualizing the evolution of these sentiments.
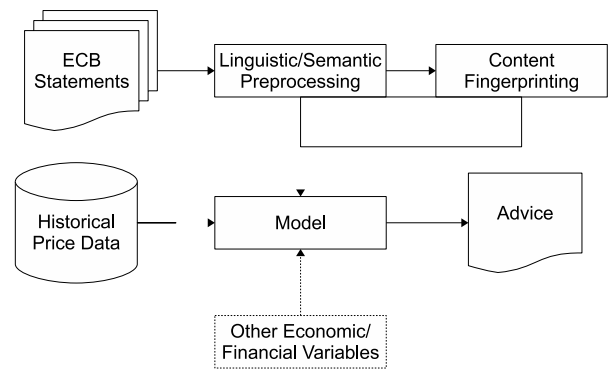
Figure. 1: ECB statements analysis framework.

In [7], the authors discuss a sentiment mining approach related to the extraction of term subjectivity and orientation from text. The approach starts with two training sets consisting of *positive* and *negative* words, respectively. It extends these two sets with WordNet synonyms and antonyms of the words in the sets. Then, a binary classifier is built by a supervised learner that is able to categorize vectorized representations of terms and classify them as *positive* or *negative*. In another approach in [1] extraction of fuzzy sentiment is done, where the authors are able to assign a fuzzy membership of *positive* or *negative* to a set of words using the so-called Sentiment Tag Extraction Program (STEP).

The first approach presented in this paper differs from the above approaches in that it relies on selected content categories from GI, and employs a fuzzy model for the prediction of movements in the MSCI EURO index. Rather than focusing on sentiment, we select a total of thirteen categories from GI and employ the percentages of words that fall under those categories as document fingerprints for the individual ECB statements. By using a fuzzy model, we are able to investigate how each category impacts the index, and draw economic conclusions. Contrary to the approach in [27], we do not aggregate all content categories into one single indicator, which would lead to losing the ability to question the impact of the different content categories on the explanandum.

In the second approach, we focus on the use of fuzzy grammars that are learnt from the text. Central to our work is the approach described in [12], where the evolution of fuzzy grammar fragments is studied for matching strings originating in free text. The basis of our approach are the methods described in [21, 22] for learning and extracting such fuzzy grammar fragments from text.

III. Content Analysis Framework

In this section, we introduce the framework that we propose for the automated content analysis of ECB statements. An overview of the architecture that we propose is given in Figure 1. In the remaining part of this section, we discuss the different modules and the reasons for including them in the architecture, as well as different approaches for concretizing each of these modules.

Our framework consists of three main modules: the Linguisting/Semantic Preprocessing module, the Content Fingerprinting module, and the module responsible for creating

the model based on historical data and the content of the text documents. We discuss each of these three modules in detail, together with the inputs they require and the output they generate.

Given a collection of ECB statements, some linguistic and/or semantic processing is required before analyzing the content of such documents. In the linguistic/semantic preprocessing module one can envision transformations such as stemming, stopword removal, part-of-speech tagging, and/or more complex operations such as semantic analysis of the concepts presented in text, possibly based on a domain-ontology or economic thesaurus such as [33]. The output of this module consists of a text that is at least transformed in such a way that makes syntactic comparison across documents possible. If a semantic approach is considered, then the comparison across documents can move from the syntactic level to an analysis of how different concepts are incorporated in different documents, i.e., a comparison of the content is enabled at a deeper level.

The preprocessed documents can then be analyzed in terms of the content they present. While the previous step mainly concerns linguistic and semantic analysis, further processing of the preprocessed documents can be purely quantitative. Here, we envision the generation of content fingerprints that make a quantitative comparison between documents possible. The content fingerprints can be generated based on the frequency of (some) words and the content they denote. A more complex approach can incorporate an analysis of predefined concepts within the text, an ontology-based approach, or an analysis of recurring structures/patterns as encountered in the text documents, in terms of the link between the concepts and content that describe these patterns. The content analysis can be based on content dictionaries such as the General Inquirer (GI) [29]. The GI service provides over 180 content categories, each of them described by a set of words that fall under that category. Regardless of the approach being considered, the output of this step consists of a quantification of the content of the text document, a content fingerprint, which describes the document being considered in terms of the content it describes. This approach makes a comparison between documents possible, and, simultaneously, enables the mapping of the content fingerprints to some numerical variable that is influenced by the content of the economic texts, e.g., stock prices for some company's shares, levels of stock indexes, etc.

Provided that one has access to historical prices of an asset or the index being considered, and that these prices can be matched with the time when the economic text documents have been made public, the price variable can be modeled based on the content fingerprints of the documents being considered. In this step, one can also consider other economic/financial variables that are relevant in the price formation of the object being considered. Different approaches can be envisioned in this step, such as Fuzzy Inference Systems (FIS) when interpretability of the model is desired, or Neural Networks (NN) for capturing the possible non-linearity of the relationship between the content fingerprints and the numerical variable being predicted without a focus on interpretability. Other approaches could also be considered.

The output of the modeling step is a forecast of the level of the asset value being considered, which can be translated to a recommendation with regard to some portfolio. For example, a price projection higher than the current price can be regarded as an advice of buying/increasing the weight of that asset within the portfolio, while a lower price projection can be interpreted as a (short) selling recommendation. In this paper, we use the forecasts of our system for the evaluation of the approach.

IV. Two Approaches to Computational Content Analysis

In this section we provide two approaches that show how the generalized framework that we present can be applied in a practical context. The text documents that we consider are the monthly statements of the European Central Bank (ECB). We consider these statements due to their comparable structure over the years, and the fact that they are issued regularly at predictable moments. In addition to the level of key interest rates, the ECB statements focus on the current state of the economy, and discuss likely economic developments for the short-, medium-, and long-term in the European Union (EU). Being received with much anticipation by the financial markets, we consider these statements highly relevant in the price formation of assets across the EU. Since the individual assets are affected to various extents by the considerations in ECB statements, we consider an aggregate index (MSCI EURO index) as the measure of performance for both approaches that we present in this paper. This index is a measure of the equity market performance of developed European markets, and currently considers sixteen countries: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom. For the quantification of the content of the ECB statements we use GI, while the model used is a FIS, chosen based on its ability to capture non-linear relationships and its interpretability.

The two approaches that we present quantify the content of ECB statements at different levels of complexity. The first approach that we consider, the Adjective Frequency Approach (AFA) presented in Section IV-A, looks at the frequency of adjectives within the text in relation to the content category they describe. The fingerprints thus generated are mapped onto levels of the MSCI EURO index. The second approach that we consider, the Fuzzy Grammar Fragments Approach (FGFA) presented in Section IV-B, looks at the frequency of grammar fragments composed of at least one economic term and a word belonging to a content category. Again, the fingerprints that we generate are mapped onto the levels of the MSCI EURO index.

A. The Adjective Frequency Approach (AFA)

In this approach, we require data from two different sources. On the one hand, we use ECB statements available from the ECB press website [30]. On the other hand, we use the MSCI EURO index, available from the Thomson One Banker website [32].

An ECB statement consists of different parts. The first part deals with the key ECB interest rates and their levels for the coming months. The following four parts deal with the eco-

nomic and monetary analysis, as well as the fiscal policies and structural reforms. These first five parts are considered relevant for our purpose. Finally, approximately the second half of an ECB statement consists of questions and answers from the press directed towards the president of the ECB. For the current scope, we consider the Q&A part of an ECB statement relevant only indirectly, and only focus on the part describing the current and expected future state of the economy.

The relevant parts of the ECB statements for the selected period are extracted by using an HTML wrapper from the ECB press website. Upon successful extraction, each statement is annotated for parts of speech using the Stanford POS Tagger [25, 26]. Based on the part of speech annotation, we extract only the adjectives from the text. It should be noted that all ECB statements, at least in the part we consider relevant for the current purpose, follow a similar structure. For this reason, we believe that the adjectives in the text could provide a good discrimination among the different statements.

For each ECB statement from the relevant period, the set of all adjectives contained in the text is fed to the General Inquirer web service. Based on this input, GI generates a document fingerprint consisting of the percentages of words from the document that fall under each category supported by GI. GI currently supports over 180 content categories, but for our current purpose we focus only on 13 of them, namely [29]:

- *positiv*, consisting of 1045 positive words, such as harmony, improve, and resolve;

- *negativ*, made up of 1160 negative words, such as adversity, grief, and quit;

- *strong*, consisting of 1902 words implying strength, such as apprehension, constrain, and fought;

- *weak*, containing 755 words implying weakness, such as defect, flee, and pitiful;

- *ovrst*, consisting of 696 words indicating overstatement, such as chronic, hopeless, and ridiculous;

- *undrst*, containing 319 referring to understatement, such as careful, hesitant, and light;

- *need*, made up of 76 words related to the expression of need or intent, such as famine, intent, and prefer;

- *goal*, consisting of 53 words referring to end-states towards which muscular or mental striving is directed, such as innovation, purposeful, and solution;

- *try*, containing 70 words indicating activities taken to reach a goal, such as compete, redeem, and seek;

- *means*, made up of 244 words denoting what is utilized in attaining goals, such as budget, debt, and necessity;

- *persist*, 64 words indicating endurance, such as always, invariable, and unfailing;

- *complet*, consisting of 81 words indicating the achievement of goals, such as enable, recover, and sustain;
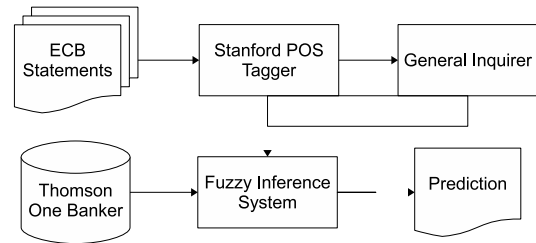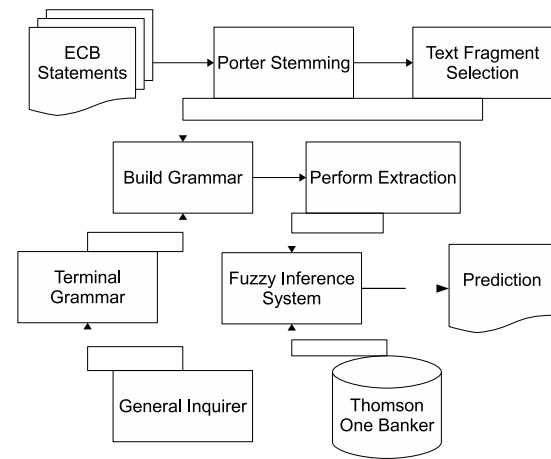

Figure. 2: AFA steps.

Figure. 3: FGFA steps.

- *fail*, which consists of 137 words that indicate that goals have not been achieved, such as bankrupt, forfeit, and ineffective.

By feeding the adjectives from each relevant ECB statement to GI, we obtain a matrix of percentages that indicate for each document, for each content category, the percentage of words in that document that fall under that category. Upon generating this matrix, we normalize it using min-max normalization across each content category.

Finally, we obtain the data on the MSCI EURO index from Thomson One Banker (T1B). We extract monthly, end-of-month data for the period January 1st 1999 until December 31st 2009. An overview of AFA is provided in Figure 2.

B. *The Fuzzy Grammar Fragments Approach (FGFA)*

The fuzzy grammar fragments approach focuses on the recognition and extraction of fragments from text. These fragments are defined and parsed based on a terminal grammar. In addition, the matching of text fragments to grammar fragments is achieved through a fuzzy parsing procedure.

For the purpose of extracting the fuzzy grammar fragments, we focus on a subset of 33 ECB statements. This is done in order to test the generalizability of the the approach and reduce the computational time. These statements are selected such that they are uniformly spread over the dataset: for each year from 1999 to 2009 we select 3 statements, from March, June, and September, respectively. We employ the same content categories from GI as in the AFA. These statements are then processed according to the flow in Figure 3.

1) Terminal grammar

For the purpose of information extraction, we begin by defining a terminal grammar around which the fuzzy grammar fragments are built. The complete terminal grammar employed for the current purpose is presented in the Appendix. The terminal grammar is centered around `<EconomicTerm>` and `<ContentCategory>` as the current focus is on extracting combinations of the two from the text of the ECB statements.

2) Porter stemming

In order to be able to identify text fragments that are identical, one must be able to abstract beyond dissimilarities between the words and the dissimilarities that may relate to things like the tense of verb, plural vs. singular, etc. For this reason, both the terminal grammar as well as the text of the ECB statements are reduced to a root form through the Porter stemming algorithm [18].

3) Text fragment selection

The topic of interest in the current case consists of the words contained in `<Economic Term>`. For the purpose of building a grammar for ECB statements, text fragments consisting of 5 words preceding an economic term and 5 words succeeding an economic term are automatically selected from the text of the statements. In order to preserve the meaning of the selected text fragments, we focus only on words that are included in the same sentence. Thus, if an economic term is the first word in a sentence, no predecessors will be selected, and if an economic term is the second word in a sentence, only one word (the word preceding the economic term) will be selected as predecessor. The same applies in the case of successors. It should be noted that the text fragments that are automatically extracted have a length of maximum five, i.e., predecessors and successors are never considered together.

4) Building the grammar

Once all text fragments related to an economic term have been extracted, the process can proceed towards building a grammar for the ECB statements. For this purpose, all selected text fragments are transformed into grammar fragments, based on the terminal grammar presented in the Appendix. An example is presented next, where given a selected text fragment T1 and the terminal grammar as presented in the Appendix, T1 would be translated into a grammar fragment F1 as follows, with `<aw>` denoting any word that is not included in the terminal grammar but is present in the text fragment:

```
T1:     earli upward pressur price

F1:     <aw><PositivCat><StrongCat><EconomicTerm>
```

Once all text fragments have been translated to grammar fragments, we proceed to building the ECB statements grammar as described in [21].
The focus of this research is on combinations of words from `<Economic Term>` and words from `<ContentCategory>`. For this reason, we only focus on fuzzy grammar fragments that contain at least one `<EconomicTerm>` and at least one `<ContentCategory>`, regardless of the number of `<aw>`. For example, the fuzzy grammar fragment F1 would be selected, while F2 and F3 would be removed from the grammar:

```
F1:     <aw><PositivCat><StrongCat><EconomicTerm>

F2:     <EconomicTerm><aw><EconomicTerm>

F3:     <aw><PositivCat><aw><StrongCat>
```

Furthermore, in order to simplify the fuzzy grammar fragments we obtain, all trailing and preceding `<aw>` are removed from the fragments. For example, fuzzy grammar fragment F1 would become the fragment eF1:

```
F1:     <aw><PositivCat><StrongCat><EconomicTerm>

eF1:    <PositivCat><StrongCat><EconomicTerm>
```

Finally, we group all the resulting fuzzy grammar fragments according to the `<ContentCategory>` they describe. When a fragment contains more than one `<ContentCategory>`, we classify this fragment under each of the `<ContentCategory>` elements it contains. For example, fragment F4 would be classified as strong, and fragment eF1 as both strong and positive:

```
F4:     <StrongCat><EconomicTerm>

eF1:    <PositivCat><StrongCat><EconomicTerm>
```

This classification is important when we employ the fuzzy inference system for predicting the MSCI EURO index. There, we rely on the frequencies of each group of grammar fragments, i.e., positive, negative, strong, etc., for the prediction of upward or downward movement in the index.
A final note that should be made regarding the building of the ECB grammar is that some words may fall under multiple categories, such as for the example the word *growth*, that falls both under `<EconomicTerm>`, as well as `<StrongCat>`. For this reason, we impose the following preference ordering over the grammar presented in the Appendix.

1. `<EconomicTerm>`
2. `<PositivCat>`
3. `<NegativCat>`
4. `<StrongCat>`
5. `<WeakCat>`
6. `<OvrstCat>`
7. `<UndrstCat>`
8. `<MeansCat>`

9. `<CompletCat>`

10. `<FailCat>`

11. `<NeedCat>`

12. `<PersistCat>`

13. `<GoalCat>`

Following this ordering, the word *growth*, that falls both under `<EconomicTerm>` as well as `<StrongCat>`, will be considered under `<EconomicTerm>`.

5) The extraction

After having built the grammar for the ECB statements, we proceed to the extraction of strings that can be parsed by the ECB grammar as described in [22]. The extraction from our set of documents is focused around the groups of 13 content categories as described in the Appendix. We count the number of strings that can be parsed by the grammar fragments under each category, for each ECB statement.

The output of this step consists of a matrix of frequencies of strings parsed by fuzzy grammar fragments under each of the 13 GI content categories. These frequencies are reported for each ECB statement that is available.

After the extraction process, no fuzzy grammar fragments have been found for the following content categories in combination with an `<EconomicTerm>`:

- `<Need>`;

- `<Complet>`;

- `<Fail>`.

In addition, the content category `<Goal>` is only seldomly encountered in the documents. For this reason we remove this content category from the results list. This reduces the number of content categories available for experiments to 9, namely:

- `<Means>`;

- `<Negativ>`;

- `<Ovrst>`;

- `<Persist>`;

- `<Positiv>`;

- `<Strong>`;

- `<Try>`;

- `<Undrst>`;

- `<Weak>`.

V. The Fuzzy Model

In this section we outline the basics of the adopted fuzzy model for the prediction of the MSCI EURO index based on the content of ECB statements.

Several techniques can be used in fuzzy identification. One possibility is to use identification by product-space clustering to approximate a nonlinear problem by decomposing it into several [2, 8] subproblems. The information regarding the distribution of data can be captured by the fuzzy clusters, which can be used to identify relations between various variables regarding the modeled system.

Let us consider an n-dimensional classification problem for which N patterns $\mathbf{x}_p = (x_p^1, \ldots, x_p^n)$, $p = 1, 2, \ldots, N$ are given from κ classes $C_1, C_2, \ldots, C_\kappa$. The task of a pattern classifier is to assign a given pattern $\mathbf{x}_p$ to one of the κ possible classes based on its feature values. Thus, a classification task can be represented as a mapping $\psi : X \subset \mathbb{R}^n \to \{0, 1\}^\kappa$ where $\psi(\mathbf{x}) = \mathbf{c} = (c_1, \ldots, c_\kappa)$ such that $c_l = 1$ and $c_j = 0$ $(j = 1, \ldots, \kappa, j \neq l)$. When the classification problem is binary, regression models can also be used as classifiers. In this approach, the regression model computes a score , e.g. probability of belonging to a class c_l, for each pattern. By applying a threshold to the score values at a suitable cutoff value, the class that a data pattern belongs to can be determined.

Takagi and Sugeno (TS) [24] fuzzy models are suitable for identification of nonlinear systems and regression models. A TS model with affine linear consequents can be interpreted in terms of changes of the model parameters with respect to the antecedent variables, as well as in terms of local linear models of the system.

One of the most simple forms of TS models contains rules with consequents in the affine linear form:

$$R^k : \text{ If } \mathbf{x} \text{ is } A^k \text{ then } y^k = (\mathbf{a}^k)^T \mathbf{x} + b^k, \quad (1)$$

where R^k is the kth rule, A^k is the rule antecedent, $\vec{a}^k$ is a parameter vector and b^k is a scalar offset. The consequents of the affine TS model are hyperplanes in the product space of the inputs and the output.

To form the fuzzy system model from the data set with N data samples, given by $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T$, $Y = [y_1, y_2, \ldots, y_N]^T$ where each data sample has a dimension of n $(N >> n)$, the model structure is first determined. Afterwards, the parameters of the model are identified. The number of rules characterizes the structure of a fuzzy system. For the models used in this work, the number of rules will be the same as the number of clusters. Fuzzy clustering in the Cartesian product-space $X \times Y$ is applied to partition the training data. The partitions correspond to the characteristic regions where the systems' behavior is approximated by local linear models in the multidimensional space. Given the training data X_T and the number of clusters K, a suitable clustering algorithm is applied.

In this work, we use the fuzzy c-means (FCM) [4] algorithm. As result of the clustering process, we obtain a fuzzy partition matrix $U = [\mu_p^k]$. The fuzzy sets in the antecedent of the rules are identified by means of the matrix U that have dimensions $[N \times K]$. One dimensional fuzzy sets A_i^k, $i = 1, \ldots, n$ are obtained from the multidimensional fuzzy sets by projections onto the space of the input variables x^i.

This is expressed by the point-wise projection operator of the form

$$\mu_{A_i^k}(x_p^i) = \text{proj}_i(\mu_p^k), \qquad (2)$$

after which the pointwise projections are approximated by Gaussian membership functions.

When computing the degree of fulfillment $\beta^k(\mathbf{x})$ of the k-th rule, the original cluster in the antecedent product space is reconstructed by applying the intersection operator in the cartesian product space of the antecedent variables: $\beta^k(x) = \mu_{A_1^k}(x^1) \wedge \mu_{A_2^k}(x^2) \wedge \ldots \wedge \mu_{A_p^k}(x^n)$. Other t-norms, such as the product, can be used instead of the minimum operator. The consequent parameters for each rule are obtained by means of least square estimation, which concludes the identification of the classification system.

After the generation of the fuzzy system, rule base simplification and model reduction could be used [20], but we did not consider this step in our current study.

We proceed as follows to generate the class labels. With the exception of the first observation from the dataset, all output values are set to 1 if the predicted value for the index in period $t + 1$ is higher than or equal to the predicted value of the index in period t, and to 0 if the predicted value of the index is lower in period $t + 1$ compared to the same value in period t. The same procedure is applied to the actual values of the index.

VI. Experiments and Results

In this section we outline the experiments that we perform and the obtained results. After first describing the experimental setup in Section VI-A, we present the results of the AFA approach in Section VI-B. The results obtained by using the FGFA approach are described in Section VI-C. The section ends with a discussion of our results in Section VI-D.

A. Experimental Setup

The dataset we used consisted of ECB statements and monthly closing values of the MSCI EURO index in the period January 1st, 1999 to December 31st, 2010. The index data is shown in Figure 4. We use 70% of the data for training the model and leave the remaining 30% for testing. For the training dataset, we generate a random permutation of indexes of the data points covering 70% of the complete dataset. In this way, every run of the system will be using different, randomly selected data. We do this in order to test the accuracy of the system regardless of economic cycles, as training the system on the first 70% of the data cannot account for the economic crisis from 2008 onwards. By using multiple runs on randomly selected data points we aim at reducing this effect. Furthermore, the model is then less likely to be influenced by any trend information that may be present in the data.

We run 100 experiments, and for each experiment the data are randomly drawn again from the dataset. For all 100 experiments, we maintain 70% of the dataset for training and 30% of the dataset for testing. Although different types of fuzzy systems have been tested, the best results have been obtained with a Takagi-Sugeno fuzzy system based on fuzzy c-means clustering [4]. We tried several numbers of clusters, and obtained the best results when using three clusters.
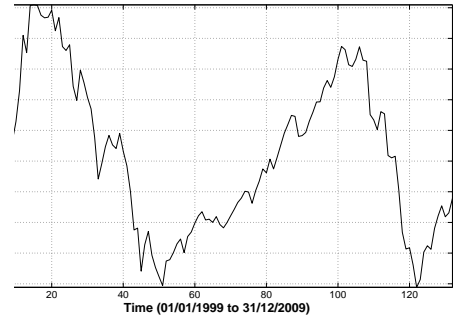

Figure. 4: The MSCI EURO index.

The fuzzy model is developed to predict the actual level of the MSCI EURO index in the month of the statement that is considered. In the final analysis, however, we are interested in the upwards or downwards movements of the index. Thus, the prediction of the index value by the FIS is used to determine whether the index will move up or down in the month of the respective statement. The accuracy of the fuzzy system is measured as the percentage of times that the system is able to correctly predict whether the index will move up or down. The formula for accuracy is presented in (3), where M^+ stands for the number of datapoints correctly predicted as upward movement, M^- stands for the number of datapoints correctly predicted as downward movement. D stands for the total number of datapoints.

$$ACC = (M^+/D + M^-/D) * 100\% \qquad (3)$$

B. AFA Results

In Table 1 we present an overview of the results of 100 experiments on the data described in the previous paragraphs. For the 100 runs, for both the training set as well as the testing set, we provide an overview of the minimum, maximum, and the mean accuracy obtained. The standard deviation of the accuracy is shown between parentheses.

Table 1: AFA – Results of 100 experiments

	Min (%)	Max (%)	Mean (%)
Training	58.82	77.65	69.18
			(4.01)
Testing	44.44	80.56	63.03
			(7.88)

On the training set, the average accuracy ranges between 58.82% and 77.65%, while having a mean of 69.18% with a standard deviation of 4.01. A small standard deviation indicates consistent results. The average accuracy shows that in about 2/3 cases, the system is able to correctly identify an increase or decrease in the MSCI EURO index. The average accuracy goes down over the 100 experiments for the test set, but only slightly to 63.03%, indicating that some overfitting occurs. However, the standard deviation nearly doubles to 7.88, which can also be observed in the much wider range between the minimum and the maximum accuracy. Having a minimum accuracy as low as 44.44% on the test data might indicate that periods are present in the test set when the model does a very poor job at predicting the change in

the index, such as when the movement of the index is solely determined by a crisis period.

In Table 2 we present the average confusion matrix for 100 fuzzy inference systems that we generate. The rows indicate the predicted movement direction of the index, while the columns indicate the true change in the index value.

Table 2: AFA – Confusion matrix for 100 experiments

	True Up	True Down
Pred. Up	34.28%	16.72%
Pred. Down	20.25%	28.75%

As it can be seen from Table 2, a slight bias can be observed between true positives and true negatives. The system seems to be able to better predict upward movement rather than downward movement. In terms of misclassifications, the same can be stated about the false positives and the false negatives. In Table 2 we also show the standard deviation for all mean values between parentheses.

In Figure 5 we provide an overview of the FIS output surface for selected pairs of inputs. The values of the MSCI EURO index in this figure have been obtained by min-max normalization. Therefore, the values for the index range between 0 and 1. From this figure, one can notice the high non-linearity of the relations, such as for example in the case of *positiv* vs. *negativ* selected inputs pair. The presence of nonlinear relations supports our choice for a fuzzy inference system to model the relationship between the content of ECB statements and the MSCI EURO index. It can also be observed that small parts of the results are sometimes counterintuitive, as in the case of the *positiv - negativ* plot. For example, one could observe that there is a positive correlation between *negativ* and the MSCI EURO index for very small values of the *negativ* variable. We consider this a spurious effect, and a direct result of the limited amount of data that is available for training, as well as testing, the model. For higher values of the *negativ* variable, the relation is as expected: higher values of this variable result in lower values for the index.

C. FGFA Results

In this section we report the results obtained from 100 experiments for FGFA. An overview hereof is provided in Table 3. For both the training and the testing set we report the minimum, maximum, and mean accuracy. Additionally, we report the standard deviation of the accuracy between parentheses.

Table 3: FGFA – Results of 100 experiments

	Min (%)	Max (%)	Mean (%)
Training	52.94	70.59	61.65 (3.34)
Testing	47.22	77.78	61.36 (6.43)

In Table 4 we present the average confusion matrix for 100 fuzzy inference systems that we generate. The rows indicate the predicted movement direction of the index, while the columns indicate the true change in the index value. It can be seen that the confusion matrix obtained is similar to the confusion matrix obtained from AFA.

Table 4: FGFA – Confusion matrix for 100 experiments

	True Up	True Down
Pred. Up	33.72%	17.64%
Pred. Down	21.00%	27.64%

In Figure 6 we provide a few surface plots for pairs of selected inputs, for one of the fuzzy models generated by the system. All pairs of inputs are plotted against the output, which consists of the normalized levels of the MSCI EURO index. Again, we notice the non-linearity describing the relations between our content input variables and the values of the index. The results indicate that the *negativ* variable is inversely related to the values of the index, while the *positiv* category positively influences the index. The *ovrst* content variable also results in higher values for the index, when this category is present to a greater extent in the text of the ECB statements.

D. Discussion

The performance of a random classifier is expected to be roughly equal to 50% because the classes up and down movement of the MSCI EURO index are equally distributed in our dataset. For the selected period, we can conclude that both AFA, as well as FGFA, provide superior performance, at least in terms of the mean accuracy of prediction, when compared to a random investment strategy. Hence, both approaches to computational content analysis of the ECB statements have predictive power over the MSCI EURO index. Hence, the general framework that we propose is useful for the computational content analysis of ECB statements. Such analysis can form the basis for the aggregation of multiple documents in a way that is more accessible to decision makers. In addition, such analysis can form the input to models that take such economic information into account and stand at the basis of (semi-) automated investment strategies that might be used, for example, in algorithmic trading.

Finally, we note that the relation between the content of ECB statements and the MSCI EURO index appears to be non-linear, both in the case of AFA, and in the case of FGFA. Although both approaches show this type of relation, further investigation is needed into the extent of the non-linearity before we can conclude that the contents of economic text is always non-linearly related to the variable being forecasted.

VII. Conclusions and Future Work

In this paper we present a general framework for the computational analysis of ECB statements. The application of this framework is illustrated by means of two concrete approaches. One is based on the frequency of adjectives in the text in relation to the content categories as outlined by GI. The other is focused on the frequency of fuzzy grammar fragments in relation to the economic terms and content categories they describe, again based on GI. The documents being considered are the monthly statements of the ECB, and they are used for the prediction of the upward or the downward movement in the MSCI EURO index. Our results indicate that, in both approaches, the movement of the index can be predicted with a higher accuracy than when a random
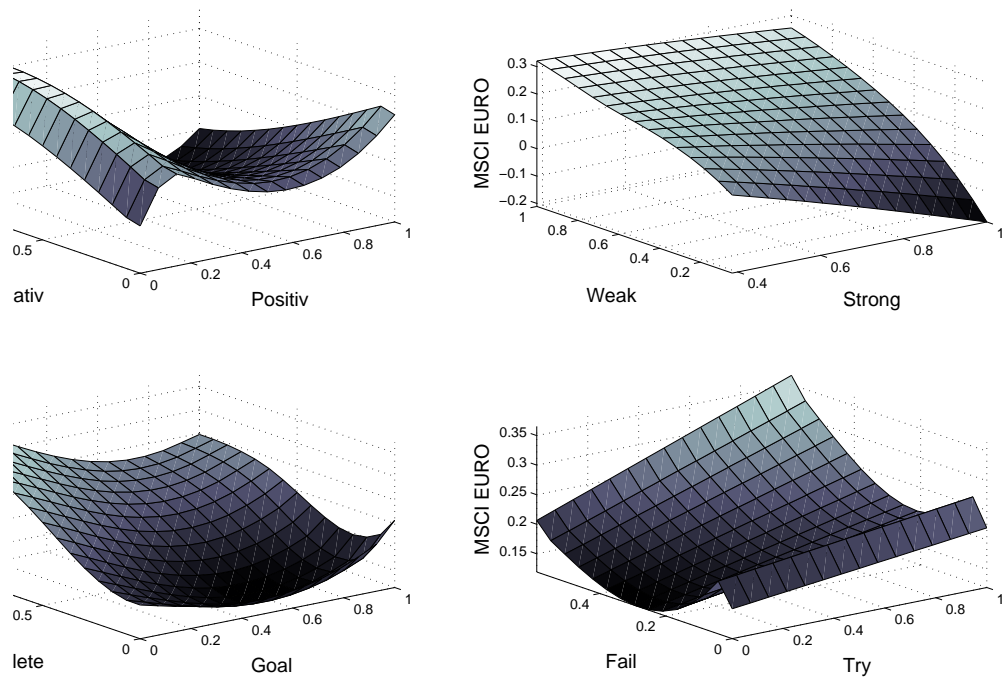
Figure. 5: Fuzzy inference system output surface for selected pairs of inputs (AFA)
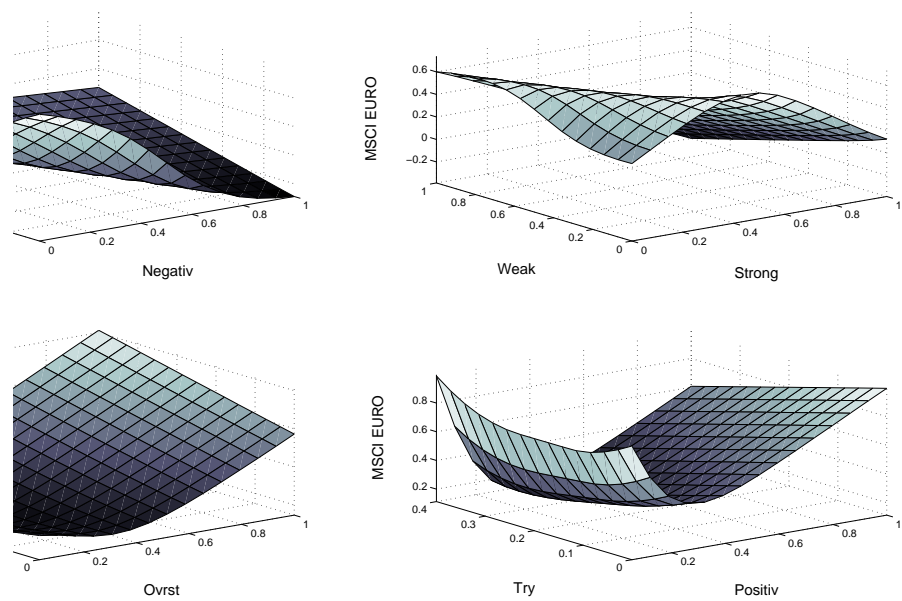
Figure. 6: Fuzzy inference system output surface for selected pairs of inputs (FGFA)

classifier is used. We use these results to validate the ability of our proposed general framework to analyze the content ECB statements.

Note that our approach does not consider deep knowledge about the semantics of the text. It can be expected that the results could improve if the semantics of the text are taken into account explicitly. Ontology-based approaches based on state-of-the-art languages such as the Web Ontology Language (OWL) [3] in static contexts or tOWL [13, 14] in time-varying contexts is an interesting direction for further research.

Acknowledgments

References

[1] Andreevskaia, A. and Bergler, S.: *Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses*, Proceedings the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006), pp. 209-216, 2006

[2] Babuska, R.: Fuzzy modeling for control, Kluwer Academic Publishers Norwell, MA, USA, 1998

[3] Bechhofer, S., Van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., and Stein, L.A.: OWL Web Ontology Language Reference, W3C Recommendation, 2004

[4] Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, Boston, USA, 1981

[5] Bollen, J., Mao, H., and Zeng, X.: Twitter mood predicts the stock market, Journal of Computational Science, **2**(1) 1–8, 2011

[6] Dunphy, D.C., Bullard, C.G., and Crossing, E.E.M.: *Validation of the General Inquirer Harvard IV Dictionary*, 1974 Pisa Conference on Content Analysis, 1974

[7] Esuli, A. and Sebastiani, F.: *Determining term subjectivity and term orientation for opinion mining*, Proceedings the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006), pp. 193-200, 2006

[8] Kaymak, U. and Babuska, R.: Compatible Cluster Merging for Fuzzy Modelling, Proceedings of 1995 IEEE International Conference on Fuzzy Systems, 897–904, 1995

[9] Kelly, E.F. and Stone, P.J.: *Computer Recongnition of English Word Senses*, Amsterdam: Noord-Holland, 1975

[10] Klingemann, H.D., Mohler, P.P., and Weber, R.P.: *Das Reichtumsthema in den Thronreden des Kaisers und die Okonomische Entwicklung in Deutschland 1871-1914*, Computerunterstutzte Inhaltsanalyse in der empirischen Sozialforschung, Kronberg: Athenaum, 1982

[11] Laswell, H.D. and Namenwirth, J.Z.: *The Laswell Value Dictionary*, vols. 1-3, New Haven: Yale University Press, 1968

[12] Martin, T., Shen, Y., and Azvine, B.: Incremental Evolution of Fuzzy Grammar Fragments to Enhance Instance Matching and Text Mining. IEEE Transactions on Fuzzy Systems **16**(6), 1425–1438, 2008

[13] Milea, V., Frasincar, F., and Kaymak, U.: tOWL: A Temporal Web Ontology Language. IEEE Transactions on Systems, Man, and Cybernetics: Part B, 2011, To appear

[14] Milea, V., Frasincar, F., Kaymak, U., and Houben, G.J.: Temporal Optimizations and Temporal Cardinality in the tOWL Language. International Journal of Web Engineering and Technology, 2011, To appear

[15] Milea, V., Almeida, R.J., Kaymak, U., and Frasincar, F.: A Fuzzy Model of the MSCI EURO Index Based on Content Analysis of European Central Bank Statements. Proceedings of the 2010 IEEE World Congress on Computational Intelligence (WCCI 2010), pp. 154–160, 2010

[16] Milea, V., Sharef, N.M., Almeida, R.J., Kaymak, U., and Frasincar, F.: Prediction of the MSCI EURO Index Based on Fuzzy Grammar Fragments Extracted from European Central Bank Statements, The 2010 International Conference of Soft Computing and Pattern Recognition (SoCPaR 2010), pp. 231–236, IEEE, 2010

[17] Namenwirth, J.Z. and Weber, R.P.: *The Laswell Value Dictionary*, 1974 Pisa Conference on Content Analysis, 1974

[18] Porter, M.F.: An Algorithm for Suffix Stripping. Program **14**(3), 130–137, 1980

[19] Sakaji, H., Sakai, H., and Masuyama, S.: *Automatic extraction of basis expressions that indicate economic trends*, Advances in Knowledge Discovery and Data Mining, pp. 977-984, 2008

[20] Setnes, M., Babuska, R., Kaymak, U., and Nauta Lemke, H.R.: Similarity Measures in Fuzzy Rule Base Simplification, IEEE Transactions on Systems, Man, and Cybernetics, Part B, **28**(3) 376–386

[21] Sharef, N.M., Martin, T., and Shen, Y.: Minimal Combination for Incremental Grammar Fragment Learning. Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference (IFSA-EUSFLAT 2009), pp. 909–914, 2009

[22] Sharef, N.M. and Shen, Y.: Text Fragment Extraction Using Incremental Evolving Fuzzy Grammar Fragments Learner. Proceedings of the 2010 IEEE World Congress

on Computational Intelligence (WCCI 2010), pp. 3026–3033 2010

[23] Stone, P.J., Dunphy, D.C., Smith, M.S., and Ogilvie, D.M.: The General Inquirer: A Computer Approach to Content Analysis. MIT Press Cambridge, 1966

[24] Takagi, T. and Sugeno, M.: *Fuzzy Identification of Systems and its Applications to Modeling and Control*, *IEEE Trans. Systems*, *Man*, and *Cybernetics*, vol. 15, no. 1, pp. 116-132, 1985

[25] Toutanova, K. and Manning, C.D.: *Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger*, Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70, 2000

[26] Toutanova, K., Klein, D., Manning, C., and Singer, Y.: *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network*, Proceedings of HLT-NAACL 2003, pp. 252-259, 2003

[27] Tetlock, P.C.: *Giving content to investor sentiment: The role of media in the stock market*, Journal of Finance, **62**3, 1139–1168, 2007

[28] Zhang, J., Kawai, Y., Tumamoto, T., and Tanaka, K.: *A Novel Visualization Method for Distinction of Web News Sentiment*, 10th International Conference on Web Information Systems Engineering (WISE 2009), pp. 181-194, 2009

[29] `http://www.wjh.harvard.edu/~inquirer/`. Last accessed: August 2nd, 2011

[30] `http://www.ecb.int/press/pressconf`. Last accessed: August 2nd, 2011

[31] `http://www.mscibarra.com/products/indices/international_equity_indices/euro/`. Last accessed: August 2nd, 2011

[32] `http://banker.thomsonib.com/`. Last accessed: August 2nd, 2011

[33] `http://zbw.eu/stw/versions/latest/about`. Last accessed: August 2nd, 2011

Appendix – Terminal Grammar

```
<ContentCategory> ::= <Positiv> | <Negativ> | <Strong> | <Weak> |
                      <Ovrst> | <Undrst> | <Need> | <Goal> | <Try> |
                      <Means> |<Persist> | <Complet> | <Fail>
        <Positiv> ::= All words contained in the Positiv category
                      by the General Inquirer.
        <Negativ> ::= All words contained in the Negativ category
                      by the General Inquirer.
         <Strong> ::= All words contained in the Strong category
                      by the General Inquirer.
           <Weak> ::= All words contained in the Weak category
                      by the General Inquirer.
          <Ovrst> ::= All words contained in the Ovrst category
                      by the General Inquirer.
          <Undrst> ::= All words contained in the Undrst category
                      by the General Inquirer.
           <Need> ::= All words contained in the Need category
                      by the General Inquirer.
           <Goal> ::= All words contained in the Goal category
                      by the General Inquirer.
            <Try> ::= All words contained in the Try category
                      by the General Inquirer.
          <Means> ::= All words contained in the Means category
                      by the General Inquirer.
        <Persist> ::= All words contained in the Persist category
                      by the General Inquirer.
        <Complet> ::= All words contained in the Complet category
                      by the General Inquirer.
           <Fail> ::= All words contained in the Fail category
                      by the General Inquirer.
   <EconomicTerm> ::= <CoreTerms> | <MonetaryTerm> | <Commodity> |
                      <Institution> | <Person>
      <CoreTerms> ::= {bank, cost, corporation, crisis, credit, debt,
                      economy, employment, euro, export, fund, growth,
                      GDP, import, inflation, investment, labour, liquidity,
                      loan, market, policy, price, protectionist, rate, risk,
                      sector, spread, tax, taxation, trade, volatility, wage,
                      yield, balance sheet, financial market, foreign trade
                      financial sector, global economy, interest rate
                      inflation rate, labour market, macro-economic
                      financial corporation, policy measure, opportunity cost
                      yield curve}
   <MonetaryTerm> ::= {monetary, M0, MB, M1, M2, M3, MZM}
      <Commodity> ::= {commodity, oil, gold, energy}
    <Institution> ::= {European Central Bank, ECB, International Monetary
                      Fund, IMF, Governing Council}
         <Person> ::= {President, Vice-President, Commissioner}
```

Author Biographies

Viorel Milea obtained the M.Sc. degree in Informatics & Economics from Erasmus University Rotterdam, the Netherlands, in 2006. Currently, he is working towards his PhD degree at the Erasmus University Rotterdam, the Netherlands. The focus of his PhD is on employing Semantic Web technologies for enhancing the current state-of-the-art in automated trading with a focus on processing information contained in economic news messages and assessing its impact on stock prices. His research interests cover areas such as Semantic Web theory and applications, intelligent systems in finance, and nature-inspired classification and optimization techniques.

Rui Jorge Almeida graduated from the five year program in Mechanical Engineering in 2005 and received his M.Sc. degree in Mechanical Engineering in 2006. Both titles were obtained from Instituto Superior Técnico, Technical University of Lisbon, Portugal. He is currently a PhD Candidate at the Department of Econometrics of the Erasmus School of Economics, Erasmus University Rotterdam, the Netherlands. His research interests include fuzzy decision making, combining fuzzy modeling techniques and statistical methods as well as data mining in finance.

Nurfadhlina Mohd Sharef obtained the B.IT degree in Science and System Management from the Universiti Kebangsaan Malaysia, Malaysia and M.Sc. in Software Engineering from Universiti Putra Malaysia, Malaysia. Then she pursued her PhD in University of Bristol. She is currently attached to the Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia.

She is also a member of the Intelligent Computing Group, Database Group and Applied Informatics Group in her workplace. Her research interests include text mining, semantics, knowledge management, and intelligent computing.

Uzay Kaymak received the M.Sc. degree in electrical engineering, the Degree of Chartered Designer in information technology, and the Ph.D. degree in control engineering from the Delft University of Technology, Delft, The Netherlands, in 1992, 1995, and 1998, respectively. From 1997 to 2000, he was a Reservoir Engineer with Shell International Exploration and Production. He is currently professor of intelligence and computation in economics with the Econometric Institute, Erasmus University Rotterdam, the Netherlands and holds the chair of information systems in the healthcare at the School of Industrial Engineering, Eindhoven University of Technology, the Netherlands. Prof. Kaymak is a member of the editorial board of several international journals, such as Fuzzy Sets and Systems, and Soft Computing.

Flavius Frasincar obtained his M.Sc. in computer science from Politehnica University Bucharest, Romania, in 1998. In 2000, he received the professional doctorate degree in software engineering from Eindhoven University of Technology, the Netherlands. He got the PhD degree in computer science from Eindhoven University of Technology, the Netherlands, in 2005. Since 2005, he is assistant professor in information systems at Erasmus University Rotterdam, the Netherlands. He has published in numerous conferences and journals in the areas of databases, Web information systems, personalization, and the Semantic Web. He is a member of the editorial board of the International Journal of Web Engineering and Technology.