2013

A Combined Approach for Extracting Financial Instrument-Specific Investor Sentiment from Weblogs

Achim Klein
University of Hohenheim, Information Systems 2, Stuttgart, Germany, achim.klein@uni-hohenheim.de

Olena Altuntas
University of Hohenheim, Information Systems 2, Stuttgart, Germany, olena.altuntas@uni-hohenheim.de

Martin Riekert
University of Hohenheim, Information Systems 2, Stuttgart, Germany, martin.riekert@uni-hohenheim.de

Velizar Dinev
University of Hohenheim, Information Systems 2, Stuttgart, Germany, velizar_dinev@uni-hohenheim.de

A Combined Approach for Extracting Financial Instrument-Specific Investor Sentiment from Weblogs

Achim Klein, Olena Altuntas, Martin Riekert, and Velizar Dinev

University of Hohenheim, Information Systems 2, Stuttgart, Germany
{achim.klein, olena.altuntas, martin.riekert, velizar_dinev}
@uni-hohenheim.de

Abstract. Investor sentiment about future returns of financial instruments is a highly relevant information source for investment managers and other stakeholders in the financial industry. Investor sentiments are abundant in financial blog texts. Making use of these sentiments constitutes a massive information management challenge when considering the millions of blog articles with ever-changing and growing amounts of information that need to be acquired and interpreted. We propose a novel approach for investor sentiment extraction from blogs by combining machine-learning on the document-level and knowledge-based information extraction on the sentence-level. The proposed artifact is a financial instrument-specific investor sentiment extraction method, which we apply to a set of blog articles. The evaluation suggests that the combined approach achieves a higher precision compared to a standalone knowledge-based approach.

Keywords: Financial information management, investor sentiment, financial weblogs, machine learning classification, knowledge-based web information extraction

1 Introduction

Financial markets and the financial industry are information-driven domains. Information is the key for decision making of professional and individual investors [1]. Making optimal investment decisions critically depends on acquiring, filtering, and interpreting the relevant information with respect to financial instruments (e.g., stock indices, stocks).

Structured information such as price and economic time series can easily be accessed through financial information systems (e.g., Bloomberg Terminal) and integrated into investment decision models. However, unstructured textual information is less integrated and cannot be directly used in an automatic way. This limitation is in particular critical, because the Web provides a huge amount of relevant unstructured information. Specifically, blogs have become a prime means for market participants to communicate opinions, investments analysis, trade ideas, and rumors. These kinds of unstructured information are subjective and referred to as investor sentiment. The

11th International Conference on Wirtschaftsinformatik,
27th February – 01st March 2013, Leipzig, Germany

literature provides evidence that investor sentiments are relevant information for investors; in particular, the demand for risky assets by noise traders is significantly affected by their sentiments [3]. In addition, investor sentiment can effectively be used for predicting stock returns [3]. Recently, it was shown how a sentiment-based trading strategy that exploits the sentiment found in weblog articles consistently generates favorable returns [4]. However, making investor sentiment from weblogs available for financial decision makers is an unsolved information extraction (IE) problem. Whereas IE has made great advances, the literature reports very few approaches that specifically concern financial weblogs. The majority of IE methods in finance are concerned with other media such as corporate disclosures [5], news articles [6], and Twitter messages [7]. Compared to these media, the automatic assessment of full-length web documents is more difficult due to noise and high ambiguity [8].

Current approaches to financial sentiment extraction are typically based on supervised machine-learning. Machine learning is a domain-independent technique for document classification. Since financial blogs often include investor sentiments on more than one financial instrument [9], document classification alone is not sufficient. Thus, a heuristic approach for selecting financial-instrument specific text parts for separate applications of machine-learning classification has been proposed [9]. In contrast to this simple approach, knowledge-based IE applies domain-specific and linguistic knowledge for text analysis and can directly work on (sub-) sentence-level with respect to specific financial instruments. However, it lacks the inherent optimization capabilities of machine-learning methods. Thus, the objectives of our research are to: (1) develop a combined investor sentiment extraction method that enriches machine learning by a knowledge-based financial instrument-specific text selection and pre-classification and (2) apply this method to a set of blog articles to demonstrate its usefulness by determining precision and recall. We hypothesize the combined method to perform better than the standalone knowledge-based approach. The *contribution* of this research is an extraction method that combines strengths of machine learning on the document level and knowledge-based IE on the sentence level.

The remainder of this paper is organized as follows. In section 2, we discuss the approaches for investor sentiment extraction and compare our approach with the relevant literature. In section 3, we define a basic model for investor sentiment in blogs and formally specify the extraction problem. In section 4, we present the proposed extraction method. Section 5 reports the experimental evaluation. Section 6 concludes the paper and outlines future work.

2 Related Work

Sentiment extraction from web documents is a subfield of opinion mining and sentiment analysis [10]. Sentiment extraction has become a widely adopted research topic and has gained also adoption by practitioners. A major stream of research focuses on sentiment with respect to consumer products (e.g., books, movies) and reviews of such products on the web [11]. In recent years, investor sentiment has attracted specific research. Most approaches in the financial domain assume only one financial in-

strument in a document and perform classification on the document-level. Next, we review two groups of approaches: (1) document-level approaches, and (2) object-level approaches.

2.1 Document-level Approaches

Supervised machine-learning is a widely utilized approach for classifying documents in the financial domain. Supervised machine-learning is a statistical technique that creates a classification model bottom-up in a data-driven way. That is, it creates a mapping of a numerical representation of a document to the classification of the document by means of labeled examples of texts. Different machine-learning methods such as Support Vector Machines (SVM) (e.g., in [5]), Naïve Bayes (NB) (e.g., in [9]) and Artificial Neural Networks (ANN) (e.g., [12]) have been used for this task. SVMs are widely used for text classification as according [13], they are well suited for this task as they achieved highest accuracy of all compared classifiers (i.e., Naïve Bayes, Rocchio, k-nearest neighbor (KNN), decision tree learner), are robust, fully automatic, and can cope well with large amounts of input machine-learning features generated from texts. If not complemented with other methods, document-level text classification using machine-learning approaches is not specific to an object, i.e., a financial instrument.

The survey of classification approaches for corporate *news* [14] reports that news are labeled (positive/negative) with regard to the post-publication price reaction of the respective stocks. For instance, Groth and Muntermann classify corporate disclosures with respect to short term future price volatility using SVM, NB, KNN, and ANN approaches [5]. They find SVM to perform best in this application context. Schumaker et al. predict stock prices on a 20 minute horizon based on a text representation that includes the document-level sentiment polarity of financial news articles [15].

Sentiment classification is primarily concerned with determining the sentiment but not directly with predicting financial variables. Thus, a manually labeled corpus is required. Antweiler & Frank [16] use a manually labeled set of 1000 messages from *stock message boards* in classes positive, negative, and neutral. Using this corpus they train a classifier using SVM and Naïve Bayes methods. Using this classifier, they propose a multiple-document aggregate measure of “bullishness” that can be interpreted as a positive/negative sentiment score [16]. The measure neglects neutral (hold) messages as they were found to be dominated by noise [16]. Also, neutral is not required for testing market reactions. The measure significantly predicts stock price volatility [16]. Das and Chen [8] also classify messages from stock message boards using a majority-voting among various machine-learning methods and a sentiment word count approach. The manually classified corpus consists of 913 messages. Das and Chen report classification accuracy of only 40.6% on a large test sample due to high ambiguity [8]. With a set of selected texts of low ambiguity, they achieve 66.9% accuracy [8].

With respect to *blogs*, Gilbert and Karahalios [17] use user-provided document-level tags that express their mood to train a classifier for detecting anxious posts in the

LiveJournal website. With an aggregate anxiety index, they predict next day returns of the S&P 500 stock index and find a Granger-causal relationship. The content Gilbert and Karahalios analyze is not finance-specific, analysis is on document-level, the sentiment is narrowed to anxiety, and it also is not specific to financial objects.

The most recent research is concerned with *microblogs*. For instance, Bollen et al. [7] propose a dictionary-based approach for classifying positive/negative sentiment and also 6 mood states in Twitter messages. These messages do not explicitly refer to stocks. However, they find a predictive relationship to the Dow Jones Industrial Average prices. Classifying sentiment in full-length articles in weblogs is more difficult due to higher expressiveness, multiple objects and ambiguity.

The approaches of the extant literature typically assign a polarity classification on the document-level to texts. This kind of sentiment analysis is rather coarse as it usually does not refer to specific financial instruments. Since web texts contain lots of ambiguity and noise, the reported classification performance is rather low. For creating a classifier, a labeled corpus is required in any case for supervised machine-learning. The corpus size is typically less than 1000 documents due to the high amount of effort required for manual annotation.

2.2 Object-level Approaches

The approaches for object-level sentiment analysis in web documents can be segmented in the following groups: (1) machine-learning based approaches that integrate a method for selecting text parts referring to a specific object, (2) dictionary-based, and (3) linguistic or knowledge-based approaches that use formalized linguistic or domain knowledge. We first review approaches in the financial domain with financial instruments being the relevant objects.

Concerning *machine-learning approaches*, O'Hare et al. [9] propose an approach using SVM and NB methods to train a classifier for sentiment in financial blogs with regard to stocks on a 979 document corpus. This approach extracts stocks and respective companies from blog texts and uses the surrounding *n* (a) words, (b) sentences, and (c) paragraphs to train a stock-specific classifier. With 25 words, the NB achieves an accuracy of 75% and SVM achieves 74% accuracy. This approach comes close to ours as it also performs financial instrument-specific sentiment classification in financial weblogs. However, assuming that (all) surrounding text parts relate to a specific financial instrument is a heuristic that might fail for (1) sentences that actually do not contain a sentiment but just mention a financial instrument or (2) sentences that contain multiple sentiments with respect to different financial instruments.

Concerning *dictionary-based approaches*, Zhang and Skiena [4] use co-referenced occurrences of positive/negative sentiment words from a dictionary and company named entities to extract sentiment with respect to companies on sentence level from Twitter messages, blog articles, and news. The exact approach is described in Godbole et al. [18] and assigns sentiment of lexicon words to companies "juxtaposed" (co-occurring) in the same sentence. A ratio computed from the number of positive and negative sentiments serves as document-level sentiment and is used in a trading strategy that yields better than benchmark results [4]. Neither [4] nor [18] report on

the accuracy of their classification approach. However, Pang et al. [19] have shown that machine-learning sentiment classification approaches perform better than a simple lexical approach. This observation provides a strong indication for the superiority of the approach of O'Hare et al.

Concerning *knowledge-based approaches,* Klein et al. [20] propose a method for sentiment classification of financial weblog documents. They formalize correlations between economic indicators and future returns in an ontology [20]. Thus, in contrast to O'Hare et al [9], this approach allows for a detailed and thorough analysis of sentiment in text that refers to the feature "future returns" of a financial instrument instead of using assumptions and heuristics. Klein et al. also formalize linguistic patterns by means of regular-expression-based rules. They find their knowledge-based approach to be superior in terms of classification accuracy with respect to baseline machine-learning approaches.

As the literature in the financial domain only provides few examples regarding the identified different approaches, we also report on approaches that are not specific to this domain. Typically, these approaches are referred to as topic-specific.

Thet et al. [21] use a *linguistic approach* for classifying the sentiment with respect to multiple aspects (e.g., storyline, music) of a movie at sub-sentence level of discussion board messages. This approach combines grammatical relations in dependency trees constructed by a parser with lexicons for contextual sentiment classification. They apply their approach to a movie dataset and achieve an accuracy of 81%. Nasukawa and Yi [22] analyze sentence-level opinions about products. Their approach also uses a syntactic parser and uses its output in rules together with a lexicon to classify sentiment. They extract only sentiment that relates semantically to the subject, i.e., a product. This relationship is ensured by manually defined rules and sentiment words. They achieve a high precision (up to 95%) but a low recall (up to 28.6%). The rather deep analysis that is part of these two approaches comes typically at high computational cost and time cost for parsing of documents (cf. [23]). As timely decisions are crucial in the financial domain, we neglect such approaches.

Yi et al. [24] propose a *natural language processing* (NLP) approach for (1) extracting product aspects from customer reviews and general web documents at a sub-sentence level and (2) classification of the sentiment with respect to these aspects. Their approach is based on modified lexicons and "sentiment patterns" (for describing grammatical relations between semantically orientated words and extracted aspects). They report an accuracy of 85.6%. As the corpus is not publicly available and sentiment patterns are not described comprehensively, the approach is not replicable.

The extant literature indicates that simple dictionary-based approaches perform worse in terms of precision and accuracy compared to machine-learning approaches [19]. To this respect, O'Hare et al. have proposed a simple heuristic for selecting text parts that refer to a specific object, i.e., a stock, to be used as input for machine-learning. Since more elaborated linguistic and knowledge-based approaches can better detect which sentences actually contain sentiment and also distinguish diverging sentiments on multiple objects in one sentence, we hypothesize that the text selection process for machine-learning can be improved by such methods. However, as deep analysis approaches such as [21] require time consuming parsing, we propose to uti-

lize the approach of Klein et al. [20] for object-specific text selection. This ontology and rule-based classifier will be further used for generating additional machine-learning features by sentiment polarity classification at the sentence level. Further, using [20] for text selection also incorporates indirect sentiments by means of a domain ontology of indicators for future returns which is beyond the other approaches reviewed. This type of knowledge-based feature generation is also proposed by Gabrilovich and Markovitch [25] who find it to improve classification performance. Further, in the proposed combined approach machine-learning can exploit more information (e.g., bag of words) than a standalone knowledge-based approach and can optimize the weighting of sentence-level sentiments in the document-level classification. Thus, we hypothesize the proposed combined approach to provide higher precision for document-level classifications than a standalone knowledge-based approach.

3 Basic Model and Problem Specification

3.1 Basic Model

The basic model comprises sentiments and documents. It is derived from the domain-unspecific sentiment analysis framework of Liu [26]. A sentiment on a sentiment object is a positive or a negative view, attitude, emotion or appraisal [26]. We are interested only in financial instruments as sentiment objects and define formally:

Definition (sentiment): Sentiment is a tuple $s_l=(fi, so)$ on sentiment level $l \in \{sentence, document\}$, using the following elements:

- $\boldsymbol{fi \in FI}$: A financial instrument, i.e., the object to which the sentiment is expressed. We assume a sentiment expressed with respect to a financial instrument to refer to its future returns.
- $\boldsymbol{so \in}$ {*positive*, *negative*}: The *sentiment orientation.*

A sentiment orientation *so* is expressed by an *orientation term ot* $\in$ *OT* with *positive* or *negative* semantic orientation *so*. We omit the neutral orientation (implying "hold" with respect to a financial instrument) because of the following reasons. O'Hare et al. [9] found that reducing the number of sentiment orientations to two (positive/negative) significantly improves human annotator agreement and also the machine classifier performance in terms of accuracy. This is presumably due to less ambiguity. Furthermore, Antweiler and Frank [16] argue that in "hold"-documents there is a dominating amount of noise. They conclude that the developed bullishness measures perform significantly better at predicting returns, volatility, and trading volumes without consideration of "hold". The bullishness measure is a ratio of positive and negative messages that can be directly used for decision making which does not require "hold". This view is, e.g., supported by Schumaker et al. [15], who define the classification of sentiment in financial news as a two-class (positive/negative) problem and define a respective trading model.

A sentiment orientation *so* can also be expressed by an (economic) indicator $i \in I$ such as GDP growth. The sets *OT*, *FI* and *I* are defined in section 4.1.

If a sentiment with respect to a financial instrument is expressed indirectly via an economic indicator, we assume that the sentiment orientation referring to it can be obtained by the **correlation coefficient** $c(i,fi)$: $I \times FI \rightarrow \{-1,+1\}$, i.e., a function that provides a positive or negative correlation for an indicator *i* that refers to the future returns of a financial instrument *fi*. Using the correlation coefficient and the semantic orientation expressed directly towards the indicator, the sentiment orientation for the financial instrument can be inferred. The following list provides some examples:

- Orientation term, "*rise*": has a positive orientation, while "*drop*" has a negative orientation.
- Directly expressed sentiment, "*I expect the FTSE 100 to rise.*": The orientation term with positive orientation "rise" provides a sentiment orientation, expressed directly with respect to (the future returns of) the FTSE 100 stock index.
- Sentiment expressed by an indicator, "*Earnings of IBM are on the rise.*": The orientation term "rise" refers to the indicator "earnings", which has a positive correlation to the future returns of a stock or a stock index. Thus, the sentiment orientation of the IBM stock with respect to its future returns is positive.

Next, we define the document model following [20]:

Definition (document): A document $d = (P, ST, T, PH)$ is based on these model elements:

- $\boldsymbol{d \in D}$**:** A *document* consists of sets of paragraphs, sentences, tokens, and phrases.
- $\boldsymbol{p \in P}$**:** A document *d* consists of a finite number of *paragraphs*. A paragraph is a sequence of sentences of finite length.
- $\boldsymbol{st \in ST}$**:** Each *sentence* is a finite length sequence of tokens.
- $\boldsymbol{t \in T}$**:** A *token* is a finite length sequence of characters which can be of type word, number, symbol, punctuation, or space.
- $\boldsymbol{ph \in PH}$**:** A *phrase* can contain sentiment(s) and is a sequence of tokens that focus around a head element.

3.2 Problem Specification

The problem is to classify the sentiment orientation $so \in \{positive, negative\}$ of a sentiment document tuple $s_{document} = (fi, so)$ with respect to a financial instrument *fi* and its future returns in a document *d*. Each document can contain multiple sentiment document tuples. To detect relevant text parts for financial instrument-specific classification, first all $s_{sentence} = (fi, so)$ need to be extracted and the sentiment orientation *so* has to be classified. We assume this pre-classification on sentence-level to improve the document-level classification when used as additional input.

4 Investor Sentiment Extraction Approach

Our approach is a combination of knowledge-based and machine learning techniques for investor sentiment extraction. This approach performs *financial instrument-*

specific extraction on the document-level. The extracted sentiment is defined as $s_{document}$=(*fi, so*). As shown in the data flow diagram below (Figure 1), the approach consists of three successive steps: (1) preprocessing extracts financial instruments, (2) ontology- and rule-based extraction of text parts that refer to specific financial instruments, and (3) machine-learning based classification of overall investor sentiment on document level, and a domain ontology.

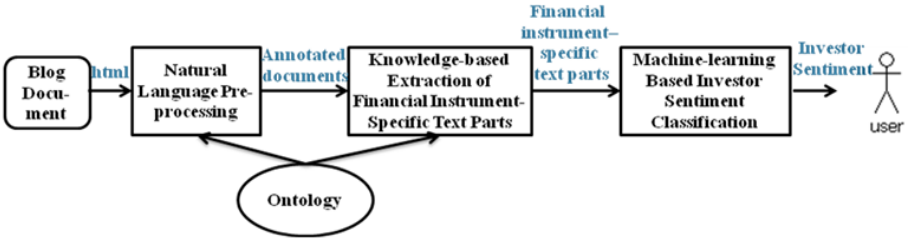

Fig.1. Combined approach for investor sentiment classification

4.1 Domain Ontology and Lexical Resources

There are two resources that are being used in the sentiment extraction approaches: (1) a domain ontology, and (2) a lexicon of semantically oriented words.

Ontology. We develop an ontology (extending from [20]) consisting of the concepts *FI*, *I*, and *OT* defined in the basic model and the formalized relations between the concepts. The concept *FI* has instances of stocks and stock indexes.

In finance, a person rarely expresses a sentiment about future returns of a financial instrument directly. He or she often rather would express a sentiment about a factor or indicator that influences returns such as company earnings. In investment analysis two major theories prevail: fundamental analysis [27] and technical analysis [28]. We formalize knowledge on economic indicators *I* for future returns from both theories in the ontology, capturing various classes of indicators and their correlation to stock returns. The correlation is assumed static and either positive (+1) or negative (−1). 14 fundamental indicators and their correlations are derived from [27], [29], [30], and [31]. 11 technical indicators and their correlations are derived from Lo et al. [28]. We refer to [20] for details. Labels (i.e., textual representations) for instances of instruments and indicators are provided by manual expert text annotations (cf., section 5.1).

We use the General Inquirer [32] (http://www.wjh.harvard.edu/~inquirer/) **lexicon of orientation terms** (*OT*). Only the words tagged as positive/negative after clearing duplicates are considered, leaving 1791 and 2198 words respectively. As General Inquirer is not specific to the financial domain, we also use a modified version of this lexicon. The modification was carried out by an undergraduate student without knowledge of the golden standard corpus or the sentiment extraction method in this work. 41 words were added (e.g., high, low, large, small), the polarity of 89 words was changed (e.g., arrest was moved to negative) and 360 words were deleted (e.g., company, share, thank). The resulting modified lexicon consists of 1575 positive and 2020 negative words.

4.2 Natural Language Pre-processing

The natural language pre-processing step uses GATE's ANNIE information extraction system (Maynard et al. [33]). The pre-processing includes tokenization, sentence splitting, part of speech (POS)-tagging, morphological analysis (for lemmatization), noun and verb chunking, and identification of ontology concepts defined above. The output of this step is an annotated document represented by a list of tokens, sentences, and ontology-based entities (financial instruments, indicators, and orientation terms).

4.3 Knowledge-based Extraction of Financial Instrument-specific Text Parts

The extraction of investor sentiment contained in single sentences of a weblog document is performed based on the ontology and rules, incorporating financial domain expertise and linguistic knowledge. The extraction of financial instrument-specific text parts consists of the following consecutive steps: (1) identification of relevant sentences, (2) extraction of sentiment sentences, and (3) sentiment classification on the sentence-level.

Identification of relevant sentences. A relevant sentence potentially contains sentiment on a given financial instrument *fi*. Table 1 contains heuristic rules for identifying relevant sentences according [20]. Additionally, we employ a co-reference recognition that exploits the topology of concepts in the domain ontology. For example, the instance "IBM" of the concept "stock" is recognized by extracting "stock" in a sentence for which the *fi* "IBM" has been already assigned by the rules (e.g., if it occurs in proximity, for instance in the first sentence of the paragraph). The subsequent extraction steps are carried out only on relevant sentences.

Table 1. Rules for identifying relevant sentences (following [20])

No	Rules
1	Sentence contains a financial instrument *fi.*
2	Sentence contains a macro fundamental indicator *i* (referring to the economy or financial markets in general) as this implicitly refers to any *fi.*
3	Sentence after a sentence that contains a *fi.*
4	All sentences in the paragraph that begins with a sentence that contains a *fi.*

Sentiment sentence extraction. A sentiment sentence refers to a financial instrument *fi* explicitly (e.g., S&P 500) or implicitly via an indicator *i* (e.g., interest rates). It must contain an orientation term *ot* to infer the sentiment orientation. The *fi* is not required to occur in a sentiment sentence. The *fi* can be heuristically inferred by rules from Table 1. Table 2 presents the rules for extracting sentiment sentences using the following elements: "Adj" (adjective), "Adv" (adverb), "N" (noun), "V" (verb) and "Prep" (preposition) are tokens differentiated by part of speech (POS). All other tokens are denoted as "to". Orientation terms are specified with their POS in subscript. A "?" indicates that a token sequence has arbitrary length including zero. Ontology concepts are denoted as FI (financial instrument), I (indicator), and OT (orientation term).

Table 2. Extraction rules for sentiment sentences (extending [20])

No	Rules	Example
1	OT_{Adj} (FI|I)	*"positive S&P500"*
2	(FI|I) OT_N	*"market crash", "index decline"*
3	(FI|I) (to)? Prep (to)? (OT_N)	*"stock market in decline"*
4	(OT_N|OT_{Adj}) (to)? Prep (to)? (FI|I)	*"run-up in S&P500"*
5	(FI|I) (Adj|Adv)? V (Adj|N)? OT_N	*" stock market makes new highs"*
6	(FI|I) (Adv)? OT_V	*"oil prices decrease"*
7	(FI|I) (Adj|Adv)? V OT_{Adj}	*"unemployment remains high"*

Sentiment sentence classification. For extracted sentiment sentences the sentiment orientation *so* is classified as follows. If the sentiment sentence contains the financial instrument *fi* (or if it was inferred by Table 2 rules), the sentiment orientation *so* is given by the lexicon's classification of the orientation term *ot*. If the sentiment sentence contains an indicator *i*, the sentiment orientation *so* depends in addition on the correlation coefficient $c(i,fi)$ of the indicator modeled in the domain ontology. If the correlation coefficient is $c(i,fi)=1$ (positive), then the *so* is given by *ot*'s classification. If $c(i,fi)=-1$ (negative), the sentiment orientation given by the *ot* is inverted. Example: "$high^{(ot,positive)}$ $unemployment^{(i,-1)}$ *rate*" will be classified with *so=negative*. In any case, if a negation (e.g., "*no*", "*not*", "*never*") occurs in the sentiment sentence, the sentiment orientation *so* is inverted.

The output of this step is a set of classified (positive/negative) financial instrument-specific text parts that are used as input for machine-learning document level classification in the next step.

4.4 Machine Learning-based Investor Sentiment Classification

For investor sentiment classification on document-level the optimization capabilities of machine-learning techniques are used. Document-level classification is performed by the linear kernel SVM as it has been shown to perform well for text classification tasks [13] in comparison to other methods such as Naïve Bayes, Rocchio, k-nearest neighbor (KNN), and decision tree learner. Further, Groth and Muntermann [5] have found SVM to be particularly well-suited for classifying financial texts and their finance domain-specific application context. We use the one-against-another classification method which means only one binary classifier is defined. The features used by SVM are unigrams represented as bag of words (i.e., a vector containing the number of occurrences of each word in a document) such as in O'Hare et al. [9]. Pang et al. [19] have shown unigrams to be a good language model for sentiment classification with SVM. We use SVM with default parameters, without allowing for a soft margin (i.e., cost=1). No optimization of SVM parameters was conducted.

As each document may contain sentiments with respect to multiple financial instruments *fi*, the overall document sentiment orientation *so* is analyzed with respect to each distinct financial instrument separately, delivering a set of document-level sen-

timent tuples $s_{document}$=(*fi*, *so*). For each distinct financial instrument a sub-document is created. This sub-document consists only of text parts referring to this financial instrument. All tokens (normalized by lemmas) included in this sub-document and the sentiment orientation of sentiment sentences serve as machine-learning features. Using all the machine-learning features obtained, the SVM classifier obtains the document-level sentiment classification that refers to a specific financial instrument *fi*.

5 Evaluation

In this section, we report evaluation results of our proposed combined investor sentiment extraction method with respect to a baseline method. A set of manually labeled documents that comprise the gold standard corpus serves as basis for evaluation.

5.1 Gold Standard Corpus

We reuse and extend the corpus of Klein et al. [20]. The extension results in a total of 528 financial instrument-specific document-level investor sentiment annotations in 409 distinct documents that stem from the following sources:

1. 165 blog documents classified on the document level as positive or negative (from Klein et. al. [20]). These documents were classified bi-polar by three graduate students independently of each other. None of them is an author of [20]. Overall classification was derived by majority vote. All sentiments refer to future returns of the S&P 500 stock index.
2. Further, 161 unique blog documents were annotated with 217 financial instrument-specific sentiment annotations by 4 finance industry professionals, none of whom is an author of this work. A fuzzy sentiment classification with 5 levels of degrees of membership for the classes positive and negative were used respectively with assigned labels: no amount (0), a small amount (0.25), a medium amount (0.5), a large amount (0.75), and a maximum amount (1). Each annotator assigned degrees of membership for positive and negative. We subtract the negative from the positive degree of membership for each annotator and use the median of these values to obtain the aggregate score. An aggregate score >0 results in a positive label, negative otherwise.
3. The remaining 83 unique blog documents with 164 financial instrument-specific investor sentiment annotations were split randomly in 3 sets. Each set was annotated by one finance industry professional each. The sentiment annotation schema is identical to (2).

For the last two sources, the annotated investor sentiments refer to specific U.S. or EU stocks, both large and small caps. Further, in all three sources the annotators also annotated textual representations of economic indicators for future returns of stocks to be used as labels for ontology instances. 58.2% of the overall corpus documents are classified as positive and 41.8% as negative.

5.2 Cross-Validation Methodology

For comparing automatic to manual classifications, we utilize stratified ten-fold cross-validation as suggested by Kohavi [34] who finds indications for this approach to be better than leave-one-out cross-validation.

Ten-fold cross-validation divides the corpus into ten subsets of approximately equal size. Each subset is stratified, i.e., every subset contains approximately the same proportion of positive and negative investor sentiment annotations as the whole corpus. In one fold, one subset is used as test set and the others as training set. In ten folds, each of the subsets becomes the test set once. For our combined approach, labeled documents in the training set are used to train a machine-learning classifier. Further, for the knowledge-based approach (that is part of the combined approach, see section 4.3), the annotations of economic indicators for future returns of stocks in the documents in the training set are used to dynamically create the labels (textual representations) for the indicator instances in the ontology used by this approach for extraction and inference. The resulting classifiers are then applied for automatically classifying the documents in the test set. Comparing the classifications of the classifier vs. the human-provided classifications in all ten test sets that together comprise the whole corpus, we derive the standard information retrieval metrics [35].

5.3 Classification Accuracy Results

We evaluate the financial instrument-specific sentiment classification of our combined approach on the golden standard corpus using the described cross-validation methodology. We compare the combined classifier results to the also financial instrument-specific knowledge-based approach for sentence-level classification described in section 4.3 which provides the basis for the combined classifier. To provide a document level sentiment classification, the sentences classified by the standalone knowledge-based method are aggregated as the net of positive and negative sentiment sentences referring to the same financial instrument [20]. Table 3 (below) summarizes metric results [35]. All metrics have been micro-averaged over the classes positive and negative according to Yang [35]. We report results for each classifier using (1) the General Inquirer (GI) lexical resource and (2) our version (GI mod.) modified with respect to the financial domain as described in section 4.1.

Table 3. Classifier performance of our combined approach vs. knowledge-based.

Classifier approach	Precision	Recall	F1-Measure	Accuracy
Knowledge-based (GI)	62.5%	56.3%	59.2%	61.3%
Knowledge-based (GI mod.)	68.8%	**62.1%**	65.3%	67%
Combined (GI)	**73.3%**	59.8%	**65.9%**	**69%**
Combined (GI mod.)	71.4%	58.7%	64.4%	67.6%

5.4 Discussion of Results

Results indicate that with respect to the classification of sentiment regarding specific financial instruments, the proposed combined approach outperforms the knowledge-based approach with respect to almost all metrics. We assume this due to the optimization capabilities of the machine-learning method that can weigh its inputs and also has more information (e.g., bag of words). Note, that the financial instrument-specific classification results in terms of recall, f1-measure and accuracy are harmed by cases in which the approaches were not able to extract a sentiment, thus increasing the false negative rate.

Another finding is that our modified version of the General Inquirer lexical resource of sentiment words helped to achieve improved results. The knowledge-based classifier increased accuracy from 61.3% to 67%. Counter-intuitively, this does not hold for the combined method. Investigation of this issue remains future work.

One could argue accuracies of the approaches to be quite low. However, we have to consider that sentiment classification of full-length web documents is a complex task due to noise and ambiguity [8]. Das and Chen [8] achieve 66.9% accuracy for a document-level classifier approach. As blog documents often discuss multiple financial instruments, a financial instrument-specific classification is required which is more complex and cannot be tackled by standalone machine-learning methods. Thus, the accuracy of 69% for our combined financial instrument-specific classifier method can be considered a fair result considering the more complex problem as the one in [8].

Our corpus of 409 documents could be considered small. However, it is substantially larger than the one of Klein et al. [20] and is also comparable to sizes of corpora of related work, e.g., a set of 440 messages [8] or 423 news [5]. As human annotation is a huge effort and there is no large publicly available corpus for blogs, we consider the size of our corpus reasonable.

6 Conclusion

We presented a combined approach for automatic extraction and crisp classification of investor sentiment from blogs. The classification is with regard to specific financial instruments. This fine-grained analysis is enabled by the underlying knowledge-based sentiment analysis approach which integrates domain knowledge of economic indicators for returns and linguistic knowledge formalized as rules and works on the sentence-level. In contrast to this, most related work deals only with document-level classification or uses heuristic assumptions for chunking a text in financial instrument-specific parts. Evaluating whether our approach outperforms such a simple heuristic approach is subject to future work.

We showed the combined approach to substantially improve on the classification performance of the standalone knowledge-based approach. The performance is comparable to document-level classification results of other authors (e.g., Das and Chen [8]) but solves the more difficult problem of financial instrument-specific classification. In terms of absolute numbers, we consider the performance to be fair as web

docment classification is a complex task, even for humans, because of high ambiguity [8].

In future work, we aim to improve precision and accuracy of our approach and plan to add additional machine-learning features such as ontology concepts, e.g., orientation terms and economic indicators. We also plan to extend the corpus.

Acknowledgements. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) within the context of the project FIRST (Large scale information extraction and integration infrastructure for supporting financial decision making) under grant agreement no. 257928. The work presented in this paper was also partly funded by the German Federal Ministry of Education and Research (BMBF) under the project SCRMI (FKZ 01IS10044D).

References

1. Admati, A., Pfleiderer, P.: Selling and Trading on Information in Financial Markets. Am Econ Rev, 96-103 (1988)
2. Shleifer, A., Summers, L.H.: The Noise Trader Approach to Finance. Journal of Economic Perspectives 4, 19-33 (1990)
3. Brown, G. W., Cliff, M.T.: Investor Sentiment and Asset Valuation. Journal of Busisness 78, 405-440 (2005)
4. Zhang, W., Skiena, S.: Trading Strategies to Exploit Blog and News Sentiment. In: 4th International AAAI Conference on Weblogs and Social Media, 375-378 (2010)
5. Groth,S., Muntermann, J.: An intraday market risk management approach based on textual analysis. Decision Support Systems 50, 680-691 (2011)
6. Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S.: More than Words: Quantifying Language to Measure Firms' Fundamentals. Journal of Finance 63, 1437–1467 (2008)
7. Bollen, J., Mao, H., and Zeng, X.-J. Twitter mood predicts the stock market. Journal of Computational Science 2, 1–8 (2011)
8. Das, S., Chen, M.: Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. Management Science 53, 1375-1388 (2007)
9. O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., et al.: Topic-Dependent Sentiment Analysis of Financial Blogs. In: 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, 9-16 (2009)
10. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2, 1–135 (2008)
11. Tang, H., Tan, S., Cheng, X.: A survey on sentiment detection of reviews. Expert Syst Appl 36, 10760-10773 (2009)
12. Moraes, R., Valiati, J., Neto, W.: Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Syst Appl 40, 621 - 633 (2013)
13. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. LNCS 1398, 137–142 (1998)
14. Mittermayer, M.-A., Knolmayer, G.F.: Text Mining Systems for Market Response to News: A Survey. University of Bern, Bern (2006)
15. Schumaker, R.P., Zhang, Y., Huang, C.-N., Chen, H.: Evaluating Sentiment in Financial News Articles. Decision Support Systems 53, 458-464 (2012)
16. Antweiler, W., Frank, M.Z.: Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. Journal of Finance 59, 1259–1294 (2004)

17. Gilbert, E., Karahalios, K.: Widespread Worry and the Stock Market. In:4th International AAAI Conference on Weblogs and Social Media, 58-65 (2010)
18. Godbole, N., Srinivasaiah, M., Skiena, S.: Large-scale sentiment analysis for news and blogs. In: Proceedings of the international conference on weblogs and social media, ICWSM'07 (2007)
19. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 79–86 (2002)
20. Klein, A., Altuntas, O., Häusser, T., Kessler, W.: Extracting Investor Sentiment from Weblog Texts: A Knowledge-based Approach. In: IEEE Conference on Commerce and Enterprise Computing, 1-9 (2011)
21. Thet, T., Na, J., Khoo, C., Shakthikumar, S.: Sentiment analysis of movie reviews on discussion boards using a linguistic approach. In: Proceeding of the international CIKM workshop on topic-sentiment analysis for mass opinion measurement,TSA'09, 81-84 (2009)
22. Nasakuva, T., Yi, J.: Sentiment Analysis: Capturing Favorability using Natural Language Processing. K-CAP'03, 70-77 (2003)
23. Banko, M., Cafarella, M., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In International Joint Conference on Artificial Intelligence, 2670-2676 (2007)
24. Yi, J., Nasukawa. T., Bunescu, R., Niblack, W.: Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In: Proceedings of the IEEE international conference on data mining, ICDM'03, 427-434 (2003)
25. Gabrilovich, E., Markovitch, S.: Feature Generation for Text Categorization Using World Knowledge. Proceedings of the 19th International Joint Conference for Artificial Intelligence, 1048–1053 (2005)
26. Liu, B.: Sentiment Analysis and Subjectivity. In: Indurkhya, N., Damerau, F.(eds.): Handbook of Natural Language Processing, 2nd ed, Chapman and Hall/CRC Press, Boca Raton, FL, U.S.A., 627-666 (2010)
27. Lev, B., Thiagarahan, S. R.: Fundamental Information Analysis. Journal of. Accounting Research31, 190-215 (1993)
28. Lo, A.W., Mamaysky, H., Wang, J.: Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation. Journal of Financ 55, 1705-1765 (2000)
29. Chen, N., Roll, R., Ross, S.A.: Economic Forces and the Stock Market. Journal of Business 59, 383-403 (1986)
30. Cheung, Y-W., Lilian, K. N.: International evidence on the stock market and aggregate economic activity. Journal of Empirical Finance 5, 281-296 (1998)
31. Humpe, A., Macmillan, P.: Can macroeconomic variables explain long-term stock market movements? A comparison of the US and Japan. Appl Financ Econ 19, 111–119 (2009)
32. Stone, P.J., Dunphy, D.C., Smith, M.S., Ogilvie, D.M.: The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Cambridge, MA (1966)
33. Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K., Yorick W.: Architectural elements of language engineering robustness. Natural Language Engineering 8, 257-274 (2002)
34. Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In:14th international joint conference on Artificial intelligence, Vol. 2, 1137-1143 (1995)
35. Yang, Y.: An evaluation of statistical approaches to text categorization. Inform Retrieval 1, 69-90 (1999)