International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-018-0805-x

ORIGINAL ARTICLE

FineNews: fine-grained semantic sentiment analysis on financial microblogs and news

Amna Dridi[1] · Mattia Atzeni[1] · Diego Reforgiato Recupero[1]

Received: 1 March 2017 / Accepted: 7 March 2018

Abstract
In this paper, a fine-grained supervised approach is proposed to identify bullish and bearish sentiments associated with companies and stocks, by predicting a real-valued score between − 1 and + 1. We propose a supervised approach learned by using several feature sets, consisting of lexical features, semantic features and a combination of lexical and semantic features. Our study reveals that semantic features, most notably BabelNet synsets and semantic frames, can be successfully applied for Sentiment Analysis within the financial domain to achieve better results. Moreover, a comparative study has been conducted between our supervised approach and unsupervised approaches. The obtained experimental results show how our approach outperforms the others.

Keywords Sentiment analysis · Financial domain · Microblogs · News

1 Introduction

User-generated data in blogs and social networks have recently become a valuable resource for mining user sentiments to the end of capturing the "pulse" of stock markets [1]. Therefore, Sentiment Analysis in the financial domain is becoming more and more a big concern for businesses, organizations and marketing researchers, mainly due to the high subjectivity of this content, as users express freely their opinions, contrary to news articles which are known by their objectivity and implicit opinions [2].

Both lexicon-based [1, 3] and machine learning methods [2, 4] have been used in Sentiment Analysis within the financial domain. Most of lexicon-based methods have focused on the coarse-grained analysis of sentiment expressed in the text. However, coarse-grained methods are insufficient for the detection and polarity classification of sentiment expressed about companies in financial news text, as not all expressions of sentiment are related to the company of interest [5]. To tackle this problem, machine learning techniques have been recently proposed [2, 5, 6], mainly investigating a fine-grained schema to pinpoint the particular phrases which express sentiment and analyze these sentiment expressions in a fine-grained manner.

Both approaches of research in Sentiment Analysis in the financial domain are still too much focused on word occurrence methods and they seldom even use *WordNet* [7], ignoring consequently advancements of techniques in semantics. However, semantics is crucial for text classification problems. From this perspective, this work lies at the intersection of NLP, Semantic Web and Sentiment Analysis which are recently being increasingly researched for many emerging needs, such as the financial one. There have been some early-stage efforts to integrate a semantic abstraction layer in the financial domain [8]. However, no previous study has focused on investigating Semantic Web in Sentiment Analysis within the financial domain. In this research work, we aim to fill this gap. We believe that by grasping commonsense knowledge bases and semantic networks, this study adds a deep understanding of sentiments and opinions from natural language expressed by means of user-generated data. By using *Framester* [9], a wide coverage hub of linguistic linked data standardized using frame semantics, this work also adds breadth to the debate on the strengths of using

✉ Diego Reforgiato Recupero
diego.reforgiato@unica.it

Amna Dridi
amna@unica.it

Mattia Atzeni
m.atzeni38@studenti.unica.it

[1] Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124 Cagliari, Italy

Published online: 10 March 2018

semantics for Sentiment Analysis in the financial domain. Additionally, by focusing on user-generated texts, this work enriches the knowledge base of financial user-generated data.

The proposed approach has been evaluated on two datasets which have been proposed as training data for SemEval 2017 (task 5).[1] The task that needed to be fulfilled, and that we target in our paper, is the following: given a text instance, predict the sentiment score for each of the companies/stocks mentioned in the message. Sentiment values need to be floating point values in the range between − 1 (very negative/bearish) and 1 (very positive/bullish), with 0 designating neutral sentiment. To fulfill this task, we introduce in this paper a supervised approach by training five machine learning classifiers. By boosting the training model with the employment of semantics through replacement and augmentation this research shows that the accuracy of fine-grained sentiment detection in financial domain when using semantic features is better in terms of cosine similarity score compared to the baselines in the microblogs dataset specifically. We also compared the performances of our proposed supervised approach based on Framester semantic features with existing unsupervised approaches using SentiWordNet [10] features combined with *K*-means and Latent Dirichlet Allocation [11]. Results indicate that our proposed approach outperforms the others. To the best of our knowledge, the proposed approach represents the first attempt exploiting Semantic Web in Sentiment Analysis within the financial domain. The major contributions of this work are therefore the following:

- we conduct some experiments showing how semantic features, most notably BabelNet synsets and semantic frames, can be exploited to perform Sentiment Analysis tasks in the financial domain;
- we show that simple fine-grained approaches, based on specific parts of a message, are particularly useful in the financial domain, since financial messages often reveal both positive and negative trends, so that it becomes crucial to determine the entity which the sentiment score relates to;
- all the experiments are replicable and the proposed system developed on top of Apache Spark is freely available on GitHub under GPLv3 license: https://github.com/UnicaSSA/FineNews.

The remainder of this paper is organized as follows: Sect. 2 presents a summary of the existing Sentiment Analysis approaches within the financial domain. Section 3 introduces the material we used for supporting the whole task of Sentiment Analysis on financial user-generated data. Section 4 discusses the financial data we have used and how they have been prepared. Section 5 presents the method used to conduct the Sentiment Analysis. Section 6 presents how we employed five machine learning algorithms for automated Sentiment Analysis and present results. We further compare performances of these methods for our two datasets with unsupervised approaches based on SentiWordNet. Section 7 concludes the paper.

2 Related work

Investigating hidden patterns in consumer's attitudes towards brands, capturing their opinions and understanding their preferences have long been a major concern for researchers studying the impact of media on the stock markets. While most of existing work in Sentiment Analysis within the financial domain has used the traditional finance news [5], tweets and blogs, as more subjective sources of information, are increasingly attracting attention of marketing researchers in the past decade since they are readily available for free. Also, they overcome the problem of the subjectivity of news data. For instance, blogs and tweets are highly subjective as authors frequently use opinionated sentences to freely express their evaluations and projections towards products, companies and stocks.

Sentiment Analysis within the financial domain has been applied for a wide range of economic and financial fields [12], such as market prediction [8, 12, 13], box office prediction for movies [14], analyzing consumer's attitudes towards certain brands [1, 3], determining the financial blogger's sentiment towards companies and their stock [2] and detecting crisis [15]. Both lexicon-based [1, 3] and machine learning methods [2, 4] have been used. Authors in [1], for instance, have used an *expert-predefined lexicon* including around 6800 seed adjectives with known orientation to conduct the analysis of consumer brand sentiments. They have shown that their study added breadth and depth to the debate over attitudes towards cosmopolitan brands. In the same context, Ghiassi et al. [3] have developed a *Twitter-specific lexicon* for Sentiment Analysis and augmented it with brand-specific terms for brand-related tweets in order to perform Twitter brand Sentiment Analysis. They have shown that the reduced lexicon set, while significantly smaller (only 187 features), reduces modeling complexity, maintains a high degree of coverage over their Twitter corpus, and yields improved sentiment classification accuracy. On the other hand, Ferguson et al. [4] have explored the use of paragraph-level and document-level annotations, examining how additional information from paragraph-level annotations can be used to increase the accuracy of document-level sentiment classification. Similarly, O'Hare et al. [2] have proposed

[1] http://alt.qcri.org/semeval2017/task5/.

and evaluated simple text-extraction approaches to extract most relative segments of a document with respect to a given topic. Then, they have trained and tested sentiment classifiers on the extracted sub-document representation (word-, sentence-, and paragraph-text extraction).

More recently, Li et al. [16] have further investigated the applications of sentiment analysis to predict news impact on the stock price, by leveraging specific resources such as the Harvard psychological dictionary and Loughran–McDonald financial sentiment dictionary [17] to construct a sentiment space. They have also conducted experiments on news article summarization, showing that prediction based on summarizations can effectively outperform prediction based on full-length articles on both validation and independent testing sets [18]. Extracting topics in the financial domain has been as well studied by Feuerriegel et al. [19] to detect how stock prices are affected by financial news using Latent Dirichlet Allocation. As far as it has been reported, many of the current research works in Sentiment Analysis within the financial domain are still too much focused on word occurrence methods and they rarely even use WordNet [1], ignoring consequently advancements of techniques in semantics. However, semantics is crucial to the text classification problem. Following this trend, authors in [8], recently have proposed a novel approach to predict intraday directional-movements of currency-pair in the foreign exchange market based on the text of breaking financial news-headlines using a semantic abstraction layer that addresses the problem of co-reference in text mining. Their work produces selection which creates a way to recognize words with the same parent-word to be regarded as one entity.

The work we present in this paper lies within this context of semantics investigation for Sentiment Analysis within the financial domain. But, going beyond them and in addition to co-reference resolution, we aim at using wide coverage linguistic resources such as FrameNet [20], WordNet [7], BabelNet [21], and others to leverage semantics to more accurate Sentiment Analysis. This leads to the investigation of the so-called *sentic computing* which helps to better exploitation of both computer and human sciences to better interpret and process user-generated data within the financial domain.

3 Material

In this section, we introduce the resources we need for supporting our approach of fine-grained sentiment analysis in the financial domain using *frame semantics* and *BabelNet synsets*, and dealing with the user-generated data in the financial domain as big data.

Framester[2] [9] is a frame-based ontological resource, which acts as a hub between several linguistic resources such as *FrameNet*[3] [20], *WordNet*[4] [7], *VerbNet*[5] [22], *BabelNet*[6] [21], *DBpedia*[7] [23], *Yago*[8] [24], DOLCE-Zero[9] [25], as well as other resources to provide a wide coverage and formal linkage of lexical and factual resources. Framester plays an important role in word sense disambiguation through its frame-detection based application called *Word Frame Disambiguation* (*WFD*) that represents a new detour approach to frame detection and aims at complete coverage of the frames evoked in a sentence.

In this work, Framester is used to extract semantic features such as semantic frames and BabelNet (BN) synsets with a guarantee of resolving the problem of word sense disambiguation that existing work suffers from.

Stanford CoreNLP toolkit[10] [26] is an extensible pipeline that provides core natural language analysis consisting in converting the raw input text in an annotated and structured representation.

In this paper, we use two modules of the CoreNLP library adopted within our approach of Sentiment Analysis: *tokenization (TokenizerAnnotator)* and *lemmatization (MorphaAnnotator)*.

Apache Spark[11] [27] is an open-source cluster computing framework that provides high-level APIs in *Java*, *Scala*, *Python* and *R*, and an optimized engine that supports general execution graphs. Apache Spark has been chosen mainly for two reasons: (1) *its speed of computation*, and (2) *its scalable library MLlib*[12] for machine learning as we run experiments with five different machine learning algorithms (*logistic regression*, *linear regression*, *lasso (Least Absolute Shrinkage and Selection Operator)) regression* and *decision trees*) using seven different features set applied for two different datasets.

Weka[13] [28] is a collection of machine learning algorithms for data mining tasks. It contains tools for data preprocessing, classification, regression, clustering, association rules and visualization.

2 http://framester.com/.

3 https://framenet.icsi.berkeley.edu/.

4 https://wordnet.princeton.edu/.

5 https://verbs.colorado.edu/~mpalmer/projects/verbnet.html.

6 http://babelnet.org/.

7 http://wiki.dbpedia.org/.

8 http://www.yago-knowledge.org/.

9 http://www.loa.istc.cnr.it/old/DOLCE.html.

10 http://stanfordnlp.github.io/CoreNLP/.

11 https://spark.apache.org/.

12 https://spark.apache.org/mllib/.

13 http://www.cs.waikato.ac.nz/ml/weka/.

Table 1 Statistics of the two financial datasets

Dataset	#Messages	#Positive	#Negative	#Neutral
Microblog messages	1694	1086	581	27
News headlines messages	1142	653	451	38

We use Weka for *support vector regression* (*SVR*) as it is not directly embedded within the MLlib APIs of Apache Spark APIs and therefore we had to call it as an external program from the developed Apache Spark software.

4 Financial data description

In this section we describe the two financial datasets (see Table 1) we used to perform our Sentiment Analysis task, which identifies bullish (optimistic, believing that the stock price will increase) and bearish (pessimistic, believing that the stock price will decline) sentiments associated with companies and stocks in a fine-grained manner. The two financial datasets, (1) *microblog messages* which consist of *Stocktwits* and *Twitter* messages, and (2) *news statements and headlines* which consist of sentences taken from news headlines as well as news text, are taken from SemEval-2017 Task 5. Organizers of the task 5 of SemEval-2017 also provided the annotations using the approach explained in the following subsections.

4.1 Microblog messages dataset

The microblog messages dataset consists of a collection of financially relevant microblog messages from Twitter and StockTwits which have been annotated for fine-grained sentiment analysis. StockTwits messages focus on stock market events and typically contain references to company stock symbols, which are called cashtags. The corpus has been created by performing an initial random sampling on a pool of StockTwits messages and tweets, to construct an unbiased set of statements. which has been manually annotated to construct the gold standard. Each message in the dataset is annotated with the following information:

- **source** either Twitter or Stocktwits;
- **id** identifies the unique Twitter or StockTwits ID of the message;
- **cashtag** identifies the stock ticker symbol that the sentiment and span relate to;
- **spans** a list of strings from the message which express sentiment;
- **sentiment** a floating-point value between −1 (very bearish/negative) and 1 (very bullish/positive).

Messages containing information that reveals a positive trend for a company or stock are annotated with positive values, while messages implying a negative trend are annotated with negative sentiment scores. Clearly, if the sentiment score is 0, then the message is classified as neutral. The total number of microblog messages is 1694, with 1086 positives, 581 negatives and 27 neutral.

4.2 News headlines dataset

The news headlines dataset consists of a collection of sentences taken from news headlines as well as news text. The textual content is crawled from different sources on the Internet, such as AP News, Reuters, Forbes and Handelsblatt. This collection of financially relevant news headlines has been manually annotated by three domain experts for fine-grained sentiment analysis, using the same process employed for the microblog messages dataset. Each message in the dataset is annotated with the following information:

- **id** unique ID of the instance in our data;
- **text** text content of the headline;
- **company** company that the sentiment relates to;
- **sentiment** a floating-point value between −1 (very bearish/negative) and 1 (very bullish/positive) denoting the sentiment expressed towards the company. 0 denotes neutral sentiment.

The total number of messages in the dataset is 1142, with 653 positives, 451 negatives and 38 neutral.

5 Fine-grained, supervised sentiment analysis

The aim of the approach proposed in this work is to take microblog messages or news headlines as input and predict the sentiment score of each of the companies or stocks mentioned in the text instance. Sentiment values need to be floating point values in the range of − 1 (very negative/bearish) to 1 (very positive/bullish), with 0 designating neutral sentiment. This prediction is realized by making a decision on assigning a real-valued score to the overall sentiment in order to provide precise, fine-grained assessments of sentiment in the financial text. In other words, the role of machine learning techniques in our approach is predicting the sentiment score given to the mentioned companies or stock for each message. These methods are supervised and, therefore, require a training dataset for their learning stage that needs features selection task to create the vectorial space.

In this section, we introduce a feature engineering approach for term selection including semantic enrichment. Then, we present the fine-grained Sentiment Analysis within

Table 2 Different features used in our experiments

Lexical features	*n*-grams *(unigrams + bigrams + 3-grams)*	f_1
	BN synsets	f_2
Semantic features	Semantic frames	f_3
	BN synsets + Semantic frames	f_4
	n-grams + BN synsets	f_5
Lexical + semantic features	*n*-grams + semantic frames	f_6
	n-grams + BN synsets + semantic frames	f_7

financial domain as a sentiment regression problem that summarizes the overall sentiment of a message or a news headline with a real-valued score. We use *Linear Regression*, *LassoWithSGD*, *Ridge Regression*, *Support Vector Regression* and *Random Forest* as five supervised machine learning algorithms for this analysis. Our pipeline consists of the following modules:

Preprocessing Tokenization, lemmatization and stop-word removal are applied to each message in both datasets. The output of this phase represents the input for the feature extraction phase. The evaluation follows the machine learning phase and its goal is the assessment of the training model.

Feature selection For each microblog message or news headline, a feature vector is prepared. Our features can be divided into three main categories: (1) *lexical features* (*n-grams*), (@) *semantic features* (*BN synsets* and *semantic frames*) and (3) a combination of the *lexical* and *semantic features*. Therefore, we extract seven kinds of features, which are shown in Table 2. In addition to the three kinds of extracted features, we select an additional set of features extracted from SentiWordNet.[14] The reason was to perform a comparison between our supervised approach and unsupervised approaches. In the following subsections we will detail each of the three kinds of features.

1. *Lexical features* are extracted in the form of *n-grams*. The process of *n*-grams extraction is preceded by text tokenization and stop-word removal. At first, the text of the grouped microblog messages and news headlines is tokenized and lemmatized using Stanford CoreNLP. Then, the stop-words are removed using Stanford CoreNLP stop-word list.[15] From this standard stop-word list, we removed the two words "*up*" and "*down*" since they are important keywords in the financial domain that represent sentiment towards stocks and companies. For instance, our dataset contains a lot of messages like "*up almost*" 11% "*now*". It is clear here that the word "*up*" is the keyword that gives important information about the sentiment of this sentence. After tokenization and stop-word removal, we create the lexical feature-vector for each text instance in our dataset. The vector contains (1) *unigrams* that are resulted after the lemmatization step realized by Stanford CoreNLP and (2) *bigrams and 3-grams* obtained by using Apache Spark APIs, in particular the class *org.apache.spark.ml.feature.NGram*.[16]
2. *Semantic features* correspond to the *semantic frames* and the *BabelNet synsets* returned by Framester for each microblog message or news headline. Semantic frames and BabelNet synsets have been extracted using the profile *b* of the Framester APIs.[17] We use *semantic replacement* and *semantic augmentation* methods as detailed in [29].
3. *SentiWordNet features* are extracted from WordNet synsets, which are automatically annotated according to their degrees of positivity, negativity and neutrality, with a score ranging from − 1 to +1. SentiWordNet allows us to construct four attributes for the textual representation of each message or news headline: (1) *sum of positive scores*, (2) *sum of negative scores* and (3) *average polarity score*.

Sentiment score granularity We propose to use *SVM regression* to conduct the quantitative sentiment score by performing sentiment analysis on a real-valued scale. To do so, at first it is crucial to realize that the extracted features above have different levels of impact in terms of the sentiment that they entail. Therefore, we propose to represent features in a scaled manner by *TF.IDF*. Then, it is important to determine the positively and negatively correlated words because the used algorithms learn to predict the score of a text instance from microblogs or news headlines based solely on presence/absence of words in the text instance.

In order to determine the positively and negatively correlated words, we use the *word-score correlation metric* presented in [30]. We note that a word could be unigram, bi-gram or 3-gram. This metric reveals how much a word's presence/absence tends to cause a message's score to deviate from the mean on average. Table 3 shows the top ten positively and negatively correlated words over the entire set of microblog messages and news headlines separately.

[14] http://sentiwordnet.isti.cnr.it/.

[15] https://github.com/stanfordnlp/CoreNLP/blob/master/data/edu/stanford/nlp/patterns/surface/stopwords.txt.

[16] https://spark.apache.org/docs/1.5.1/api/java/org/apache/spark/ml/feature/NGram.html.

[17] https://github.com/framester/Framester/wiki/Framester-Documentation.

Table 3 The top ten positively and negatively correlated words, according to the word-score correlation metric

Microblog messages				News headlines			
Positive		Negative		Positive		Negative	
Long	0.0251	Short	0.0421	Rise	0.0264	Oil	0.0240
Stock	0.0019	Sell	0.0303	Buy	0.0260	Barclays	0.0022
Up	0.0018	Down	0.0242	up	0.0215	Cut	0.0199
Buy	0.0171	$Spy	0.0191	Group	0.0164	Fall	0.0199
Call	0.0168	Downgrade	0.0132	Astrazeneca	0.0150	Loss	0.0185
Add	0.0129	Overbought	0.0012	Ftse	0.0138	Fine	0.0016
$Amzn	0.0124	Put	0.0105	Billion	0.0129	Hit	0.0150
Rally	0.0109	Market	0.0105	Boost	0.0123	Bp	0.0147
Bullish	0.0106	Will	0.0105	Drug	0.0121	Job	0.0118
$Wmt	0.0100	End	0.0098	Jump	0.0119	Rbs	0.0109

Table 4 10-fold-cross validation results of fine-grained classification on microblog messages dataset using the whole message text

	f_1	f_2	f_3	f_4	f_5	f_6	f_7
Random Forest	**0.641**	0.533	0.444	0.539	0.632	0.634	0.635
Linear regression with SGD	0.632	0.615	0.355	0.609	0.653	0.625	**0.647**
Lasso with SGD	0.541	0.451	0.313	0.450	0.562	0.552	**0.566**
Ridge regression with SGD	0.633	0.615	0.355	0.608	0.654	0.626	**0.648**
Support vector regression	0.633	0.600	0.393	0.603	**0.677**	0.665	0.676

Bold expressed the best value along all the features set f1, f2, ..., f7

Table 5 10-fold-cross validation results of fine-grained classification on microblog messages dataset using only spans

	f_1	f_2	f_3	f_4	f_5	f_6	f_7
Random Forest	**0.680**	0.570	0.444	0.572	0.679	0.674	0.675
Linear regression with SGD	0.718	0.663	0.383	0.660	**0.725**	0.717	0.722
Lasso with SGD	0.582	0.467	0.314	0.476	0.590	0.583	**0.592**
Ridge regression with SGD	0.718	0.662	0.383	0.659	**0.725**	0.717	**0.722**
Support vector regression	0.712	0.654	0.383	0.661	0.724	0.715	**0.726**

Bold expressed the best value along all the features set f1, f2, ..., f7

6 Experiments

We have evaluated our proposed fine-grained sentiment analysis approach using the two datasets described in Sect. 4: the *microblog messages* dataset and the *news headlines* dataset. For the first dataset, we have carried out the experiments twice: the first time using the whole text of the message and the second time using only the *spans*, which have been defined in Sect. 4.1.

We have considered the different feature representations for each microblog message or news headline outlined in Sect. 5. We have considered the first feature representation f_1 which represents the lexical features as a baseline and we have compared the results obtained by constructing a classifier trained on each feature set.

We have compared five classifiers: a *decision tree classifier* (*Random Forest*), a *linear regression model* (*LinearRegressionWithSGD*), a *Lasso regression model* (*LassoWithSGD*), a *Ridge regression model* (*RidgeRegressionWithSGD*) and a *Support Vector Regression (SVR) model*. The Apache Spark machine learning library *MLlib* implementation was used for the first 4 classifiers, while we used Weka machine learning library for the last classifier.

Ten-fold cross validation was used for each of the segmentation experiments, with the results averaged over the ten folds. We use *cosine similarity* as the performance metric.

6.1 Results

Fine-grained classification and regression results using lexical-based, semantic-based and a combination of lexical and semantic-based features for the two datasets are shown in Tables 4, 5 and 6.

Table 6 10-fold-cross validation results of fine-grained classification on news headlines dataset

	f_1	f_2	f_3	f_4	f_5	f_6	f_7
Random Forest	**0.563**	0.456	0.337	0.465	0.556	0.559	0.544
Linear regression with SGD	**0.633**	0.550	0.329	0.540	0.626	0.624	0.619
Lasso with SGD	**0.516**	0.384	0.294	0.389	0.510	0.506	0.502
Ridge regression with SGD	**0.634**	0.551	0.329	0.541	0.627	0.624	0.620
Support vector regression	0.647	0.530	0.328	0.543	0.649	0.625	**0.655**

Bold expressed the best value along all the features set f1, f2, ..., f7

Tables 4 and 5 show cosine similarity scores related to the microblog messages dataset. Table 4 is related to the whole message text while Table 5 shows the results computed using only spans. Table 6 lists results of the news headlines dataset.

The overall results shown in the three tables reveal that the best results are given when experimenting with microblog messages dataset. This could be justified by the fact that news headlines are more objective. For instance, the best result given overall by the five algorithms in news headlines is 0.655 while for the microblog messages it reaches 0.726 with SVR for spans text, with a gain of more than 10% in cosine similarity.

For the microblog messages dataset, we have experimented first with the whole message text (Table 4) and then with only spans (Table 5). The obtained results demonstrate the effectiveness of the spans comparing to the whole message text as the granularity of the sentiment is more accurate with this list of strings that capture sentiments in microblog message. The spans effectiveness is shown by comparing the results in Tables 4 and 5. This substantial improvement from the text-level classification to the sentence-level (spans) classification underlines the importance of text extraction techniques in fine-grained sentiment analysis Interestingly, the results indicate that it is possible to achieve large improvements over message-based sentiment classification using quite simple text-extraction approaches to extract the most relevant segments of the messages.

For the semantic incorporation, our experimental results on the microblog message dataset show that the integration of semantic features performs better than simply using lexical features (*n*-grams) for the four regression algorithms, except for Random Forest. Even if the baseline (*n*-grams) gives the best results with Random Forest (0.680 in microblog spans), the highest accuracy is given when semantic features are introduced (0.726 in microblog spans with *n*-grams + BN synsets + semantic frames).

For news headlines dataset, results of four out of five algorithms show that the baseline is performing better except SVR. This returns to the objectivity of news headlines while semantic features are resulted of the sentic computing principle which gives better results with opinionated text.

For overall experimental results on both datasets, semantic integration with enrichment either by BN synsets (f_5, *n*-grams + BN synsets) or by both BN synsets and semantic frames (f_7, *n*-grams + BN synsets + semantic frames) gives better results. For instance, in the microblog text dataset (Table 4) the best cosine similarity is reached when *n*-grams were enriched by BN synsets, passing from 0.663 with *n*-grams to 0.677. In the microblog spans dataset (Table 5) the best accuracy is given by f_7 when *n*-grams were enriched by both BN synsets and semantic frames, passing from 0.712 to 0.726. On the other hand, for the news headlines dataset, the only gain in cosine similarity is given by the semantic enrichment of *n*-grams with BN synsets and semantic frames with a slightly increase of about 1.2%.

6.2 Unsupervised approaches

In the following, we present an analysis of the performances of our approach by means of a comparative study against existing unsupervised approaches.

The comparative evaluation task was designed to evaluate two learners: *K*-means (L_1) and LDA (L_2) after removing all messages within a sentiment score in the $[-0.25, 0.25]$, in order to keep only highly polarized messages. For the first learner *K*-means, we have tested the system using several combinations of features extracted from SentiWordNet. The combination that yields the best results was given by the sum of positive scores, the sum of negative scores and the average polarity score. *K*-means has been then used to divide the messages into two clusters: *Cluster 0* for positive classes and *Cluster 1* for negative classes. For the microblog dataset, *Cluster 0* contains 700 positive messages and 393 negative messages, while *Custer 1* contains 176 positive

Table 7 Cosine similarity results of the comparative study between our approach and semi-supervised and unsupervised approaches

	Unsupervised Approach		Supervised Approach
	L_1	L_2	
Microblogs	0.19	0.41	0.68
Headlines	0.26	0.40	0.66

messages and 62 negative messages. However, for news headlines dataset, *Cluster 0* contains 376 positive headlines and 280 negative headlines, and *Cluster 1* contains 21 positive headlines and 35 negative headlines. It is clearly shown that both clusters contain mostly positive messages. Hence, this unsupervised learner was not able to make distinction between positive and negative messages, resulting in low cosine similarity scores, as shown in Table 7.

Similarly, for the second learner LDA, two topics (*Topic 0* and *Topic 1*) have been revealed using the same SentiWordNet features as the first learner. For the microblog dataset, *Topic 0* contains 445 positive messages and 231 negative messages, while *Topic 1* contains 431 positive messages and 224 negative messages. On the other hand, for the news headlines dataset, *Topic 0* contains 279 positive messages and 207 negative messages, while *Topic 1* contains 118 positive messages and 108 negative messages. We notice that both topics are mixed, which shows the limitation of LDA to predict sentiment. This is justified by the fact that LDA does not work well for short documents. The first problem comes from the shortness of text. For instance, LDA models a document as a mixture of topics, and each word is drawn from one of its topics. For short documents, only few words are available. Consequently, the observations are obviously too few to infer the parameters. In other words, when working on smaller documents, the extra topic layer does not add anything to the classification. In case of short documents like tweets or news headlines, it is really hard to break documents into topics.

Overall, the comparative study of our supervised approach against existing semi-supervised and unsupervised approaches, summarized in Table 7, shows that the performance of our proposed approach based on boosting textual representation by semantic features extracted from the knowledge graph of Framester is significantly better in the financial domain. This clearly proves that grasping common-sense knowledge bases and semantics adds a deep understanding of opinions expressed in natural language.

7 Conclusion and future work

Sentiment Analysis in the financial domain using user-generated data is challenging. This work addressed this challenge in an accurate way by bringing together natural language processing and Semantic Web as well as fine-grained Sentiment Analysis to propose an approach that predicts a real-valued sentiment score of each of the companies or stocks mentioned in the text instance of microblog messages or time-stamped news headlines in our two datasets.

We have considered three main categories of features: *lexical features*, *semantic features* and a combination of the lexical and semantic features. Then, using these features, we have compared the performance of five learning methods: one classification-based and four regression-based algorithms. The approach succeeded in performing the accuracy level of more than 72% in some cases when the training model was boosted by semantics through replacement and augmentation.

For the first dataset based on the user-generated content, we have performed two types of experiments: one using the whole text of microblog messages and the other one using only spans. Interestingly, our results indicated that spans perform significantly better than the whole text. On the other hand, the news headlines dataset did not perform well in term of accuracy compared to the microblog messages dataset. This could be justified by the fact that news headlines are more objective than microblog messages.

As future work, we plan to study implicit sentiment which represents a substantial amount of the polar expressions encountered in most datasets.

Acknowledgements This work has been supported by Sardinia Regional Government (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2014-2020 - Axis IV Human Resources, Objective l.3, Line of Activity l.3.1.) The authors gratefully acknowledge Sardinia Regional Government for the financial support (Convenzione triennale tra la Fondazione di Sardegna e gli Atenei Sardi Regione Sardegna L.R. 7/2007 annualità 2016 DGR 28/21 del 17.05.2016, CUP: F72F16003030002). Moreover, the research leading to these results has received funding from the European Union Horizon 2020 the Framework Programme for Research and Innovation (2014-2020) under Grant Agreement 643808 Project MARIO Managing active and healthy aging with use of caring service robots.

References

1. Mostafa MM (2013) More than words: social networks' text mining for consumer brand sentiments. Expert Syst. Appl. 40(10):4241–4251
2. O'Hare N, Davy M, Bermingham A, Ferguson P, Sheridan P, Gurrin C, Smeaton AF (2009) Topic-dependent sentiment analysis of financial blogs. In: Proceedings of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion, TSA '09, ACM, New York, pp 9–16
3. Ghiassi M, Skinner J, Zimbra D (2013) Twitter brand sentiment analysis: a hybrid system using N-gram analysis and dynamic artificial neural network. Expert Syst Appl 40(16):6266–6282
4. Paul F, Neil O, Michael D, Adam B, Scott T, Paraic S, Cathal G, Alan FS (2009) Exploring the use of paragraph-level annotations for sentiment analysis of financial blogs. In: Proceedings of the 1st workshop on opinion mining and sentiment analysis, WOMSA 2009, pp 42–52
5. Van de Kauter M, Breesch D, Hoste V (2015) Fine-grained analysis of explicit and implicit sentiment in financial news articles. Expert Syst. Appl 42(11):4999–5010
6. Raina P (2013) Sentiment analysis in news articles using sentic computing. In: Proceedings of the 2013 IEEE 13th international conference on data mining workshops, ICDMW '13, IEEE Computer Society, Washington, DC, pp 959–962

7. Fellbaum C (ed) (1998) WordNet: an electronic lexical database. MIT Press, Cambridge
8. Khadjeh Nassirtoussi A, Aghabozorgi S, Ying Wah T, Ngo DCL (2015) Text mining of news-headlines for FOREX market prediction. Expert Syst Appl 42(1):306–324
9. Gangemi A, Alam M, Asprino L, Presutti V, Recupero DR (2016) Framester: a wide coverage linguistic linked data hub. In: EKAW 2016, Bologna, 19–23 Nov 2016, Proceedings, pp 239–254
10. Baccianella S, Esuli A, Sebastiani F (2010) Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of LREC'10. European Language Resources Association (ELRA), Valletta, pp 2200–2204
11. Blei DM (2012) Probabilistic topic models. Commun ACM 55(4):77–84
12. Khadjeh Nassirtoussi A, Aghabozorgi S, Ying Wah T, Ngo DCL (2014) Review: text mining for market prediction: a systematic review. Expert Syst Appl 41(16):7653–7670
13. Sprenger TO, Tumasjan A, Sandner PG, Welpe IM (2014) Tweets and trades: the information content of stock microblogs. Eur Financ Manag 20(5):926–957
14. Du J, Xu H, Huang X (2014) Box office prediction based on microblog. Expert Syst Appl 41(4):1680–1689
15. Schulz A, Thanh TD, Paulheim H, Schweizer I (2013) A fine-grained sentiment analysis approach for detecting crisis related microposts. In: 10th proceedings of the international conference on information systems for crisis response and management, Baden-Baden, 12–15 May 2013, pp 846–851
16. Li X, Xie H, Chen L, Wang J, Deng X (2014) News impact on stock price return via sentiment analysis. Knowl Based Syst 69(Supplement C):14–23
17. Loughran T, McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J Finance 66(1):35–65
18. Li X, Xie H, Song Y, Zhu S, Li Q, Wang FL (2015) Does summarization help stock prediction? A news impact analysis. IEEE Intell Syst 30(3):26–34
19. Feuerriegel S, Ratku A, Neumann D (2016) Analysis of how underlying topics in financial news affect stock prices using latent Dirichlet allocation. In: Proceedings of HICSS, HICSS '16. IEEE Computer Society, Washington, DC, pp 1072–1081
20. Baker CF, Fillmore CJ, Lowe JB (1998) The Berkeley FrameNet project. In: Proceedings of the 17th international conference on computational linguistics—volume 1, COLING '98. Association for Computational Linguistics, Stroudsburg, pp 86–90
21. Navigli R, Ponzetto SP (2012) BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif Intell 193:217–250
22. Kipper K, Dang HT, Palmer M (2000) Class-based construction of a verb lexicon. In: Proceedings of the seventeenth national conference on artificial intelligence and twelfth conference on innovative applications of artificial intelligence. AAAI Press/The MIT Press, Cambridge, pp 691–696
23. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2007) DBpedia: a nucleus for a web of open data. In: The semantic web: 6th ISWC 2007 + ASWC 2007, Busan, 11–15 Nov 2007, pp 722–735
24. Suchanek FM, Kasneci G, Weikum G (2007) Yago: a core of semantic knowledge. In: Proceedings of the 16th international conference on world wide web, WWW '07, ACM, New York, pp 697–706
25. Lando P, Lapujade A, Kassel G, Furst F (2007) Towards a general ontology of computer programs. In: Filipe J, Shishkov B, Helfert M (eds) ICSOFT (PL/DPS/KE/MUSE), INSTICC Press, Funchal, pp 163–170
26. Manning CD, Surdeanu M, Bauer J, Finkel JR, Bethard S, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: Proceedings of the 52nd ACL, Baltimore, 22–27 June 2014. System demonstrations, pp 55–60
27. Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I (2010) Spark: cluster computing with working sets. In: Proceedings of the 2Nd USENIX conference on hot topics in cloud computing, HotCloud'10. USENIX Association, Berkeley, p 10
28. Smith TC, Frank E (2016) Statistical genomics: methods and protocols, chap. introducing machine learning concepts with WEKA. Springer, New York, pp 353–378
29. Dridi A, Reforgiato Recupero D (2017) Leveraging semantics for sentiment polarity detection in social media. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-017-0727-z
30. Drake A, Ringger EK, Ventura D (2008) Sentiment regression: using real-valued scores to summarize overall document sentiment. In: Proceedings of ICSC 2008, 4–7 Aug 2008, Santa Clara, pp 152–157