Bearish-Bullish Sentiment Analysis on Financial Microblogs

Amna Dridi[1], Mattia Atzeni[1], and Diego Reforgiato Recupero[1]

University of Cagliari, Mathematics and Computer Science Department, Via Ospedale 72, 09124, Cagliari, Italy
{amna, diego.reforgiato}@unica.it

Abstract. User-generated data in blogs and social networks has recently become a valuable resource for sentiment analysis in the financial domain since it has been shown to be extremely significant to marketing research companies and public opinion organizations.
In this paper a fine-grained approach is proposed to predict a real-valued sentiment score. We use several feature sets consisting of lexical features, semantic features and combination of lexical and semantic features. To evaluate our approach a microblog messages dataset is used. Since our dataset includes confidence scores of real numbers within the [0-1] range, we compare the performance of two learning methods: Random Forest and SVR. We test the results of the training model boosted by semantics against classification results obtained by n-grams. Our results indicate that our approach succeeds in performing the accuracy level of more than 72% in some cases.

1 Introduction

Sentiment analysis in financial domain is becoming more and more a big concern for businesses, organizations and marketing researchers, mainly due to their high subjectivity as users express freely their opinions through opinionated sentences, contrary to news articles which are known by their objectivity and implicit opinions [1].

Both lexicon-based [2, 3] and machine learning methods [4, 1] have been used for mining user's opinion in the financial domain. Most of lexicon-based methods have focused on the coarse-grained analysis of sentiment expressed in 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 we are interested in [5]. To tackle this problem, machine learning techniques have been recently proposed [1, 5, 6] that mainly investigated fine-grained schema to allow pinpointing the particular phrases in a text express sentiment and analyzing 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 to text classification problem. 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 studies have focused on investigating Semantic Web in sentiment analysis in the financial domain. In this research work, we aim to fill this gap. We believe that by grasping common-sense 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], as 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 semantics for sentiment analysis in the financial domain. Additionally, by focusing solely on user-generated texts, rather than on traditional texts such as news papers, and testing our fine-grained sentiment approach on a collection of financially relevant microblog messages from Twitter[1] and Stocktwits[2], this work enriches the knowledge base of financial user-generated data. Finally, by training two machine learning classifiers, boosting the training model by semantics through *replacement* and *augmentation*, and using Apache Spark to deal with user-generated big data, this research shows that the accuracy of fine-grained polarity detection in financial domain when using semantic features is slightly better in term of cosine similarity score comparing to the baseline in the microblogs dataset. To the best of our knowledge the proposed approach represents the first attempt towards harnessing Semantic Web in sentiment analysis in the financial domain.

2 Related Work

Sentiment analysis in the financial domain has been applied for a wide range of economic and financial fields [10], such as market prediction [8, 10, 11], box office prediction for movies [12], analyzing consumer's attitudes towards certain brands [3, 2], determining the financial blogger's sentiment towards companies and their stock [1].

Both lexicon-based [3, 2] and machine learning methods [1, 4] have been used. Mostafa [3], for instance, has used an *expert-predefined lexicon* including around 6800 seed adjectives with known orientation to conduct the analysis of consumer brand sentiments. He has shown that his study added breadth and depth to the debate over attitudes towards cosmopolitan brands. In the same context, Ghiassi et al. (2013) [2] 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.

[1] https://twitter.com/
[2] stocktwits.com/

On the other hand, Ferguson et al. (2009) [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. (2009) [1] have proposed 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).

As far as it has been reported, many of the current works of research in sentiment analysis in financial domain are still too much focused on word occurrence methods and they rarely even use WordNet [3], ignoring consequently advancements of techniques in semantics. However, semantics is crucial to the text classification problem. Following this trend, Khadjeh Nassirtoussi et al. (2015) [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 in financial domain. But, going beyond them and in addition to co-reference resolution, we aim at using a wide coverage linguistic resources such as FrameNet [13], WordNet [7], BabelNet [14], and others to leverage semantics to more accurate sentiment analysis following the novel sentic computing system presented in [15, 16] that combines natural language processing techniques with knowledge representation. This leads to better exploitation of both computer and human sciences to better interpret and process user-generated data in financial domain.

3 Fine-grained sentiment analysis

The aim of our approach is to take microblog messages 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 score given by the annotator. These methods are supervised and, therefore, require a training dataset of their learning stage. For the learning stage, a feature selection task is required.

3.1 Feature selection

For each microblog message, a feature-vector is prepared. Our features can be divided into three main categories which are *lexical features* (*n-grams*), *semantic*

features (*BN synsets* and *semantic frames*) and a combination of the *lexical* and *semantic features*.

Lexical features. In this work, we use *word n-grams* as lexical features. The process of n-grams extraction is preceded by a step of text tokenization and stop-word removal. At first, the text of the grouped microblog messages is tokenized and lemmatized using Stanford coreNLP. Then, the stop-words are removed using Stanford coreNLP stop-word list[3]. 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 *(i) unigrams* that are resulted after the lemmatization step realized by Stanford coreNLP, *(ii) bigrams and 3-grams* that are given by Apache Spark APIs, in particular the class *org.apache.spark.ml.feature.NGram*[4].

Semantic features. The semantic features correspond to the *semantic frames* and the *BabelNet synsets* returned by Framester for each microblog message. Semantic frames and BabelNet synsets have been extracted using the profile *b* of the Framester APIs.

- *BabelNet synsets* are sets of synonyms in different languages grouped by *BabelNet* which is an encyclopedic dictionary that provides concepts and named entities lexicalized in many languages and connected with large amounts of semantic relations, automatically created by linking Wikipedia[5] to WordNet [14].
- *Semantic Frames* are a collection of facts that specify *"characteristic features, attributes, and functions of a denotatum, and its characteristic interactions with things necessarily or typically associated with it"* [17].

We use a **semantic replacement** method to incorporate semantic features into the classifier. *Semantic replacement* means replacing lexical features (n-grams) by semantic features (BN synsets and/or semantic frames). In other words, instead of using the textual representation of a microblog message, we substitute it by BN synsets, semantic frames or both of them (BN synsets+semantic frames).

[3] https://github.com/stanfordnlp/CoreNLP/blob/master/data/edu/stanford/nlp/patterns/surface/stopwords.txt
[4] https://spark.apache.org/docs/1.5.1/api/java/org/apache/spark/ml/feature/NGram.html
[5] http://www.wikipedia.org/

Combination of lexical and semantic features. This consists in augmenting the original n-grams feature space (lexical features) with the semantic features (BN synsets and semantic frames) as additional features for the classifier training in three different ways: (i) augment the original lexical features (n-grams) with semantic frames, (ii) augment lexical features with BabelNet synsets, and (iii) augment lexical features with both semantic frames and BabelNet synsets.

The size of the vocabulary in this case is enlarged by the introduced semantic features. In other words, we use a **semantic augmentation** method to incorporate semantic features into the classifier. This means instead of using only the textual representation of the microblog message, we augment it by BN synsets and/or semantic frames.

3.2 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 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 algorithms we will use learn to predict the score of a text instance from microblogs based solely on presence/absence of words in the text instance.

Word-score correlation metric. In order to determine the positively and negatively correlated words, we use the *word-score correlation metric* presented in [18]. We note that a word could be unigram, bi-gram or 3-gram.

The correlation of a word w with the scores of a set of financial messages M, denoted $c(w, M)$, is defined by the following:

$$c(w, M) = \frac{1}{|M|} \sum_{m \in M} \left\{ I(w, m) \cdot \left((S(m) - \frac{1}{|M|} \sum_{m' \in M} S(m') \right) \right\} \qquad (1)$$

where $S(m)$ is the real-valued score associated with message m and $I(w, m)$ is a function that outputs 1 if m contains word w and outputs -1 otherwise. Note that the $\frac{1}{|M|} \sum_{m' \in M} S(m')$ term is the average message score. Intuitively, if a word is positively correlated with message scores then it would tend to appear in documents with above average scores and be absent from messages with below average scores. Similarly, if a word is negatively correlated with message scores then it would tend to appear in documents with below average scores and be absent from messages with above average scores.

To see how this applies in the correlation metric defined above, notice that if a word w appears in a message m and ms score is above average, then both $I(w, m)$ and $\left((S(m) - \frac{1}{|M|} \sum_{m' \in M} S(m') \right)$ are positive, and the correlation goes up. If w does not appear in message m and ms score is below average, then both terms are negative, and again the correlation goes up. Meanwhile, in the other

two cases (when w is not in m and ms score is above average and when w is in m and ms score is below average), the terms have different signs and the correlation drops.

This metric reveals how much a words presence/absence tends to cause a messages score to deviate from the mean on average. A large positive value indicates that the word tends to occur in reviews with above average scores and be absent from messages with below average scores, while a large negative value indicates the opposite. A value near 0 indicates that the words presence does not tend to influence the score significantly in either a positive or negative direction. This metric implicitly tends to remove words that occur too rarely or too frequently to be useful for learning.

4 Evaluation

4.1 Financial data description

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. The dataset identifies bullish (optimistic; believing that the stock price will increase) and bearish (pessimistic; believing that the stock price will decline) sentiment associated with companies and stocks in a fine-grained manner. The total number of microblog messages is 1694, with 1086 positives, 581 negatives and 27 neutral.

Each message in the dataset is annotated with the following information: **source** which presents the platform where the message was posted, either Twitter or Stocktwits, **id** which identifies the unique Twitter or StockTwits ID of the message, **cashtag** which identifies the stock ticker symbol that the sentiment and span relate to. For example, *"$amzn"* is a cashtag related to *Amazon*, **sentiment** which is a floating point value between -1 (very bearish/negative) and 1 (very bullish/positive) denoting the sentiment expressed towards cashtag. 0 denotes neutral sentiment, and **spans** which is a list of strings from the message which express sentiment.

4.2 Evaluation methodology

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* related to the message which are defined as a list of strings from the message that express sentiment.

We have considered the different feature representations for each microblog message outlined in Section 3.1. We have considered the first feature representation which represents the lexical features (n-grams) as a baseline and we have compared the accuracy obtained by constructing a classifier trained on each feature.

We have compared two classifiers: a *decision tree classifier* (*Random Forest*) and a *Support Vector Regression (SVR)* classifier. The Apache Spark machine

learning library *MLlib* implementation was used for the first classifier, while we used Weka machine learning library for the second 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.

As the sentiment score predicted by the learned classifiers lie on a continuous scale between -1 and 1, cosine distance enables comparing the degree of agreement between gold standard and predicted results. At the same time, while not requiring exact correspondence between the gold and predicted score, a given instance does not need to be identical in order to achieve a good evaluation result. The scores are conceptualized as vectors, where each dimension represents a stock symbol or company within a given microblog message or headline . Note that the both vectors have the same number of dimensions as the stock symbols and companies for which sentiment needs to be assigned was given in the input data [19].

Cosine similarity is calculated according the following equation, where G is the vector of gold standard scores and P is the vector of scores predicted by the classifier:

$$cosine(G, P) = \frac{\sum_{i=1}^{n} G_i \times P_i}{\sqrt{\sum_{i=1}^{n} G_i^2} \times \sqrt{\sum_{i=1}^{n} P_i^2}} \tag{2}$$

In order to reward classifiers which attempt to answer all problems in the gold standard, the final score is obtained by weighting the cosine from Equation 2 with the ratio of answered problems (scored instances), given below (as given in [19]).

$$cosine\_weight = \frac{|P|}{|G|} \tag{3}$$

The equation for the final score is the product of the cosine and the weight, given below:

$$final\_cosine\_score = cosine\_weight \times cosine(G, P) \tag{4}$$

4.3 Results

Fine-grained classification (*Random Forest*) and regression (*SVR*) results using lexical-based, semantic-based and combination of lexical and semantic-based features for our dataset are shown in Table 1.

Table 1 shows cosine similarity scores related to the microblog messages related to the whole message text as well as those related to spans. 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 of *Microblogs-Text* and *Microblogs-Spans* rows where the accuracy of spans outperforms the accuracy of the messages text in each row; for the two algorithms and with the 7 features, notably by more than 12% with SVR algorithm using n-grams. This substantial improvement from the

text-level classification to the sentence-level (spans) classification underlines the importance of the text extraction techniques in fine-grained sentiment analysis. Interestingly, the results indicate that 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 se-

	Microblogs - Text		Microblogs- Spans	
	Random Forest	SVR	Random Forest	SVR
n-grams	**0.641**	0.633	**0.680**	0.712
BN synsets	0.533	0.600	0.570	0.654
Semantic frames	0.444	0.393	0.444	0.383
BN synsets+semantic frames	0.539	0.603	0.572	0.661
n-grams+BN synsets	0.632	**0.677**	0.679	0.724
n-grams+semantic frames	0.634	0.665	0.674	0.715
n-grams+BN synsets+semantic frames	0.635	0.676	0.675	**0.726**

Table 1. 10-fold-cross validation results of Random Forest and SVR algorithms on microblog messages dataset

mantic incorporation, our experimental results show that the integration of semantic features performs better than simply using lexical features (n-grams) for SVR, but not for *Random Forest*. Our baseline (n-grams) keeps the best performance. This could be justified by the principle of decision tree algorithms where the rules are composed of words, and words have meaning, then the rules themselves can be insightful. More than just attempting to assign a label, a set of decision rules may suggest a pattern of words found in newswire prior to the rise of a stock price. The downside of rules is that they can be less predictive if the underlying concept is complex [10]. Even the baseline (n-grams) gives the best results with Random Forest (0.680 in microblog spans), the high accuracy is given when semantic features are introduced (0.726 in microblog spans with n-grams+BN synsets+semantic frames).

For overall experimental results,semantic integration with enrichment either by BN synsets (n-grams+ BN synsets) or by both BN synsets and semantic frames(n-grams+BN synsets+semantic frames) gives better results. For instance, for *Microblogs-Text* the best cosine similarity is reached when n-grams were enriched by BN synsets, passing from 0.663 with n-grams to 0.677 when BN synsets are incorporated, giving a gain in accuracy of more than 2%. In the microblog spans dataset the best accuracy is given when n-grams were enriched by both BN synsets and semantic frames, passing from 0.712 to 0.726 giving again a gain in accuracy of approximately 2%.

Noteworthy is the fact that the SVR algorithm is the top performer in all experiments. This shows that regression approach for fine-grained sentiment analysis will likely be best.

5 Conclusion

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-value sentiment score of each of the companies or stocks mentioned in the text instance of microblog messages.

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 two learning methods: one classification-based and one 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 our dataset, 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 performs significantly better than the whole text. This indicates that 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. In our dataset, these segments are already given in form of list of strings expressing sentiments and called spans. However, the approach is interesting and could be investigated in future work by developing techniques to extract most relevant segments for sentiment classification over different text levels.

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.).

References

1. O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., Smeaton, A.F.: Topic-dependent sentiment analysis of financial blogs. In: Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion. TSA '09, New York, NY, USA, ACM (2009) 9–16
2. Ghiassi, M., Skinner, J., Zimbra, D.: Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Syst. Appl. **40**(16) (November 2013) 6266–6282
3. Mostafa, M.M.: More than words: Social networks' text mining for consumer brand sentiments. Expert Syst. Appl. **40**(10) (August 2013) 4241–4251

4. Paul, F., Neil, O., Michael, D., Adam, B., Scott, T., Paraic, S., Cathal, G., Alan, F.S.: 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). WOMSA 2009 (2009) 42–52
5. Van de Kauter, M., Breesch, D., Hoste, V.: Fine-grained analysis of explicit and implicit sentiment in financial news articles. Expert Syst. Appl. **42**(11) (July 2015) 4999–5010
6. Raina, P.: Sentiment analysis in news articles using sentic computing. In: Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops. ICDMW '13, Washington, DC, USA, IEEE Computer Society (2013) 959–962
7. Fellbaum, C., ed.: WordNet: an electronic lexical database. MIT Press (1998)
8. Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., Ngo, D.C.L.: Text mining of news-headlines for forex market prediction. Expert Syst. Appl. **42**(1) (January 2015) 306–324
9. Gangemi, A., Alam, M., Asprino, L., Presutti, V., Recupero, D.R.: Framester: A wide coverage linguistic linked data hub. In: Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings. (2016) 239–254
10. Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., Ngo, D.C.L.: Review: Text mining for market prediction: A systematic review. Expert Syst. Appl. **41**(16) (November 2014) 7653–7670
11. Sprenger, T.O., Tumasjan, A., Sandner, P.G., Welpe, I.M.: Tweets and trades: the information content of stock microblogs. European Financial Management **20**(5) (2014) 926–957
12. Du, J., Xu, H., Huang, X.: Box office prediction based on microblog. Expert Syst. Appl. **41**(4) (March 2014) 1680–1689
13. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: Proceedings of the 17th International Conference on Computational Linguistics - Volume 1. COLING '98, Stroudsburg, PA, USA, Association for Computational Linguistics (1998) 86–90
14. Navigli, R., Ponzetto, S.P.: Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193** (December 2012) 217–250
15. Recupero, D.R., Presutti, V., Consoli, S., Gangemi, A., Nuzzolese, A.G.: Sentilo: Frame-based sentiment analysis. Cognitive Computation **7**(2) (2015) 211–225
16. Gangemi, A., Presutti, V., Recupero, D.R.: Frame-based detection of opinion holders and topics: A model and a tool. IEEE Comp. Int. Mag. **9**(1) (2014) 20–30
17. Allan, K.: Natural Language Semantics. Wiley (2001)
18. Drake, A., Ringger, E.K., Ventura, D.: Sentiment regression: Using real-valued scores to summarize overall document sentiment. In: Proceedings of the 2th IEEE International Conference on Semantic Computing (ICSC 2008), August 4-7, 2008, Santa Clara, California, USA. (2008) 152–157
19. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., Reyes, A.: Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, Association for Computational Linguistics (June 2015) 470–478