COMMIT at SemEval-2017 Task 5: Ontology-based Method for Sentiment Analysis of Financial Headlines

Kim Schouten **Flavius Frasincar** **Franciska de Jong**
Erasmus University Rotterdam
P.O. Box 1738, NL-3000 DR
Rotterdam, The Netherlands
{schouten,frasincar}@ese.eur.nl
f.m.g.dejong@eshcc.eur.nl

Abstract

This paper describes our submission to Task 5 of SemEval 2017, *Fine-Grained Sentiment Analysis on Financial Microblogs and News*, where we limit ourselves to performing sentiment analysis on news headlines only (track 2). The approach presented in this paper uses a Support Vector Machine to do the required regression, and besides unigrams and a sentiment tool, we use various ontology-based features. To this end we created a domain ontology that models various concepts from the financial domain. This allows us to model the sentiment of actions depending on which entity they are affecting (e.g., *decreasing debt* is positive, but *decreasing profit* is negative). The presented approach yielded a cosine distance of 0.6810 on the official test data, resulting in the 12th position.

1 Introduction

Many companies in the financial sector are in the business of gathering and selling information, including news and sentiment analysis information, because of the profound influence this has on investor behavior (Van de Kauter et al., 2015). The relation between news and movements in the financial market is intricate, with news influencing the market (Schuster, 2003) as well as the market being a source of news itself. Price fluctuations of financial instruments can be linked to supply and demand and thus to the desirability of that financial product, which changes when new facts related to this product are published. Sentiment analysis in the context of financial news headlines aims to capture the change in desirability of a given product. Assigning a negative sentiment score to a certain news headline for a given product then represents a decrease in desirability and thus a decrease in price for that product, while assigning a positive sentiment has the opposite meaning.

In track 2 of Task 5 at SemEval 2017 (Cortis et al., 2017), each news headline contains one or more company names, and for a given company name, the sentiment, modeled as a real number between -1 and 1, needs to be determined. When multiple companies are mentioned, the same sentence can appear multiple times in the data, each time asking for the sentiment with respect to a different company. In financial headlines, there are two reasons why the expressed sentiment can differ for the various companies that are mentioned. The first is that expressed sentiment is often opposite for competitors, while the second is that news often reflects on the stock movements of the day mentioning both the biggest winners and biggest losers in the same headline.

Besides directly mentioning the stock movements of a company, news headlines often report on changes with respect to a certain aspect of a company (e.g., its profit or debt) or on actions that influence the company (e.g., opening stores or being sued). The expressed sentiment often depends on what particular aspect is in scope. A decrease in profit, for example, is considered negative, while a decrease in debt is generally considered positive.

The issue of aspect-dependent sentiment is addressed in our approach by classifying aspects and actions in such a way that an ontology reasoner, with the help of a set of class axioms, can infer which sentiment is expressed by a given pair of aspect and (increase or decrease) action. Besides aspect-dependent sentiment expressions, there are also sentiment expressions that always convey a positive (e.g., lift or good) or negative sentiment (e.g., drown or bad), and those are also stored in the ontology.

The ontology information is used a source of features for the Support Vector Regression model that is employed in our approach. Hence, we present an approach that is a hybrid between knowledge-based methods and machine learning methods (Cambria, 2016).

This paper is structured as follows. In Sect. 2, the method is presented, followed by an extensive evaluation in Sect. 3. Conclusions and suggestions for future work are given in Sect. 4.

2 Method

At the heart of the method is a Support Vector Regression (SVR) model, for which we use the Weka implementation (Frank et al., 2016; Shevade et al., 2000). To provide features that describe the news headline, all headlines are preprocessed using the Stanford CoreNLP library. This involves tokenization, Part-of-Speech tagging, lemmatization, dependency parsing, and sentiment annotation. Furthermore, after tokenization, the headlines are scanned for company names that are in the ontology, and all text is set to lowercase. The company field in the annotations is also linked to a URI in the ontology. The sentiment tool (Socher et al., 2013) that is part of the CoreNLP package assigns a sentiment score to various parts of the text (using the parse tree), but for this research we use only the sentiment assigned to the complete headline. This is a number in the range of -2 to 2, but in practice, sentence sentiment tends to be between -1 and 1. Besides the sentiment value, which is a feature for the SVR, we also use the presence or absence of unigrams as features (i.e., classical bag-of-words), denoting presence with 1 and absence with 0. This unigrams plus sentence sentiment forms our baseline method.

2.1 Ontology Design

To the baseline method, we add various ontology features. To that end, we first designed and manually populated an ontology that models expressions in the financial domain (Schouten et al., 2017). The ontology contains four main classes: `Sentiment`, modeling mentions of a certain sentiment value, `Entity`, modeling nouns that represent entities like companies or aspects of entities like profit and debt, `Property`, modeling adjectives like *lower*, *better*, etc., and `Action`, representing verbs in the text. Hence, the ontology is a model of mentions or expressions of the concepts in the financial domain rather than a model of the concepts themselves.

In accordance with the two main polar directions: up or increase, and down or decrease, all subclasses of `Entity` are split into two groups that correspond to these two directions. The first group consist of positively oriented entities for which an ‘up’ or ‘increase’ movement is positive (e.g., profit). The second group is comprised of negatively oriented entities for which a ‘down’ or ‘decrease’ movement is positive (e.g., debt).

Actions and properties are giving information about some entity and these are divided into four categories. The first two are aspect-dependent, representing an `Increase` and `Decrease` action, while the other two categories represent actions that are inherently `Positive` and `Negative`. Actions in the `Increase` or `Decrease` category can only be assigned a sentiment if they are linked with an entity from the ontology, while actions in the two sentiment classes always denote that sentiment value regardless of what entity they affect. A similar reasoning holds for properties. An overview of the main ontology classes is given in Figure 1.

2.2 Ontology Features

The presence or absence of subclasses of `Entity`, `Property`, and `Action`, which are the domain components of our ontology, are recorded as additional features for the SVR. To avoid fitting the model on certain companies that occur in predominantly positive (or negative) headlines in this particular set of news headlines, we filter out company name URLs from the set of features. Ontology concepts are linked to the text by means of lexicalizations that have been added to each non-abstract concept in the ontology. Once a concept has been found, all its superclasses are also added as features to the SVR model. Hence, if we find the action `Lift`, we also add the concepts `Action`, `Positive`, and `Sentiment`, since the concept `Lift` always denotes a positive sentiment in our domain ontology.

On top of these ontology lookup features, we define a set of class axioms that will allow the reasoner to infer the sentiment of a given combination of an `Entity` and either a `Property` or a `Action`, where the action or property on its own is not already a subclass of `Sentiment`. Using the two polar categories of entities (i.e., the positively oriented group and the negatively ori-

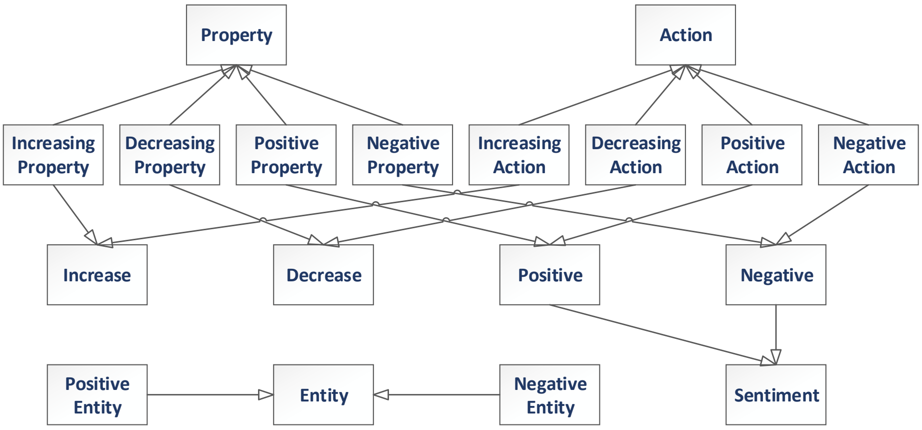

Figure 1: The main classes of the used ontology

ented group) in conjunction with the `Increase` and `Decrease` classes that contain actions and properties, we can infer the sentiment of the four different combinations that are possible. The class axioms that describe this behavior are:

1. $\texttt{Increase} \sqcap \texttt{PosEntity} \sqsubseteq \texttt{Positive}$
2. $\texttt{Increase} \sqcap \texttt{NegEntity} \sqsubseteq \texttt{Negative}$
3. $\texttt{Decrease} \sqcap \texttt{PosEntity} \sqsubseteq \texttt{Negative}$
4. $\texttt{Decrease} \sqcap \texttt{NegEntity} \sqsubseteq \texttt{Positive}$

Besides these general class axioms, we also defined a number of specific axioms that will allow the reasoner to infer the sentiment for certain particular expressions. For example, *closing a deal* is considered positive, while *closing stores* generally is not. While we could get the right behavior by classifying `Store` as a positively oriented entity, and `Deal` as a negatively oriented entity, this did not match with our intuition that a deal is something positive and more deals is not necessarily bad, which is a conclusion that would follow from classifying `Deal` as a negatively oriented entity. Hence we have specific axioms that deal with this scenario and the related `Open` action:

1. $\texttt{Close} \sqcap \texttt{Deal} \sqsubseteq \texttt{Positive}$
2. $\texttt{Close} \sqcap \texttt{CompanyPart} \sqsubseteq \texttt{Negative}$
3. $\texttt{Open} \sqcap \texttt{CompanyPart} \sqsubseteq \texttt{Positive}$

In the above axioms, a `CompanyPart` is the class that models all parts of a certain company, including things like headquarters, stores, webshops, departments, etc. An example of the reasoner in action is visualized in Figure 2.

2.3 Company-specific Sentiment

The above model, with all the described ontology features, would still result in a sentence-level sentiment algorithm that would not be able to give different sentiment scores for different companies in the same sentence. Since this problem does appear, we add a company-specific sentiment feature to the feature set. This feature denotes a positive (1), neutral (0), or negative (-1) sentiment score for the company that is mentioned in the company field of the annotation. Since we already annotated this field with a URL, we can locate the company within the headline. After that, we use the grammatical dependencies to find all words directly connected to the company. If these words are either a property or an action, we can use those to compute the company-specific sentiment as we can safely assume that directly connected words in the dependency graph pertain to that company.

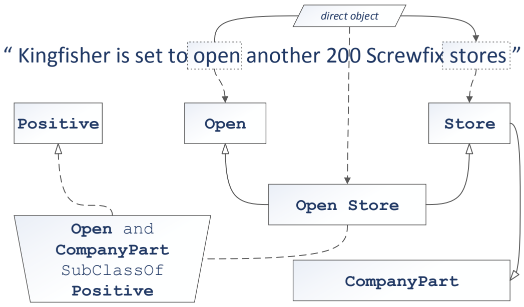
direct object
“ Kingfisher is set to open another 200 Screwfix stores ”
Positive
Open
Store
Open Store
Open and CompanyPart SubClassOf Positive
CompanyPart

Figure 2: A schematic overview of the given reasoning example.

Since a company is positively oriented entity, all actions and properties with superclass `Decrease` or `Negative` convey a negative sentiment towards the company, while `Increase` or `Positive` communicate a positive sentiment. A positive sentiment is quantified as +1, and a negative sentiment is represented by -1. Then, the company-specific sentiment feature is computed as the sum of all sentiment conveying words directly related in the dependency graph to the company mention in the headline.

3 Evaluation

In this section we evaluate our submission on the training data and report the results obtained on the official test data. The data consist of news headlines in the financial domain, and each headline is annotated with the name of the target company. For the training data the target sentiment score is also provided. Note that the same headline can appear multiple times in the data, each time with a different target company. On the official test data, a cosine distance of 0.6810 is achieved, resulting in the 12th position. The feature set experiments have been performed on the training data by running 5 times a 10-fold cross-validation setup, each time with different random folds, to ensure robust results.

To measure the effect of the various employed ontology features, the method is run with different subsets of all features. The results of this experiment are reported in Table 1. In this way, we can compare the benefit of adding entities, properties, and actions from the ontology, separately. From the reported results we can see that entities and properties are not particularly useful for sentiment analysis. For entities, this makes sense, as these convey no sentiment information. For properties, it is less intuitive, as adjectives, the word types that usually correspond to a subclass of `Property` from the ontology, are often strong indicators of sentiment (e.g., *good*, *bad*, etc.).

On the other hand, matching verbs in the text to subclasses of `Action` shows a large improvement to sentiment analysis. We hypothesize that the division into four categories (i.e., `Positive`, `Negative`, `Increase`, and `Decrease`) is a meaningful categorization in the domain of financial news. We observe that verbs are often the central word in conveying information to the reader, and hence, a lot of sentiment information is communicated using this type of concept.

Adding the class axioms to determine the sentiment of combinations with subclasses of `Increase` and `Decrease` is also useful, with a 2% increase compared to not using it. Adding the company specific sentiment, however, does not seem to help much.

Table 1: The change in performance when using different feature sets, reporting the average performance on the training data, using 5 runs with 10-fold cross-validation. Feature sets that are statistically indistinguishable from each other in terms of performance are grouped together

	avg. cosine dist.	st.dev.
base (B)	0.6311	0.0482
B + entities (E)	0.6361	0.0455
B + properties (P)	0.6300	0.0478
B + actions (A)	0.6815	0.0498
B + E + P + A	0.6883	0.0502
B + E + P + A + class axioms	0.7041	0.0450
B + E + P + A + class axioms + company-specific sentiment	0.7050	0.0441

4 Conclusions

In this work we presented our submission to Task 5 of SemEval 2017: fine-grained sentiment analysis on financial news headlines (track 2). We showed that by categorizing entities (nouns), properties (adjectives), and actions (verbs), and linking them to concepts in an ontology, we can harness the power of the ontology reasoner to infer the sentiment of expressions that indicate a typical up/increase or down/decrease movement. This is achieved by defining class axioms within the ontology. In terms of contribution to performance, we can state that the categorization of actions into `Positive`, `Negative`, `Increase`, and `Decrease` gave the highest increase in performance, followed by adding class axioms for sentiment inference.

For future work, we want to invest more in the company-specific sentiment so we can assign different sentiment values to different companies in the news headline. Given the fact that headlines often contain companies with opposite sentiment, this is a highly desirable feature to have. By using a form of spreading activation, we could compute the sentiment for the whole dependency graph, not with respect to the root which would result in the sentence sentiment, but with respect to the node in the graph representing the company. Negators and other valence shifters can be used to properly spread the sentiment from one node to the next.

References

Erik Cambria. 2016. Affective Computing and Sentiment Analysis. *IEEE Intelligent Systems* 31(2):102–107.

Keith Cortis, Andre Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, and Brian Davis. 2017. SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Association for Computational Linguistics. http://alt.qcri.org/semeval2017/task5/.

Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Fourth Edition*. Morgan Kaufmann.

Kim Schouten, Flavius Frasincar, and Franciska de Jong. 2017. COMMIT at SemEval-2017 Task 5: Ontology-based Method for Sentiment Analysis of Financial Headlines - Ontology. http://www.kimschouten.com/publications/#semeval2017.

Thomas Schuster. 2003. Meta-Communication and Market Dynamics. Reflexive Interactions of Financial Markets and the Mass Media. Technical report, EconWPA. http://EconPapers.repec.org/RePEc:wpa:wuwpfi:0307014.

S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, and K.R.K. Murthy. 2000. Improvements to the SMO Algorithm for SVM Regression. *IEEE Transactions on Neural Networks* 11(5):188 – 1993.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods on Natural Language Processing (EMNLP 2013)*. Association for Computational Linguistics, pages 1631–1642.

Marjan Van de Kauter, Diane Breesch, and Véronique Hoste. 2015. Fine-grained Analysis of Explicit and Implicit Sentiment in Financial News Articles. *Expert Systems with Applications* 42(11):4999–5010.