Implicit and Explicit Aspect Extraction in Financial Microblogs
Thomas Gaillat, Bernardo Stearns, Gopal Sridhar, Ross Mcdermott, Manel Zarrouk, Brian Davis

To cite this version:

HAL Id: hal-02280371
https://hal.univ-rennes2.fr/hal-02280371
Submitted on 10 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Implicit and Explicit Aspect Extraction in Financial Microblogs

Thomas Gaillat, Bernardo Stearns, Gopal Sridhar, Ross McDermott, Manel Zarrouk
Insight Centre for Data Analytics NUI Galway
firstname.surname@insight-centre.org

Abstract

This paper focuses on aspect extraction which is a sub-task of Aspect-based Sentiment Analysis. The goal is to report an extraction method of financial aspects in microblog messages. Our approach uses a stock-investment taxonomy for the identification of explicit and implicit aspects. We compare supervised and unsupervised methods to assign predefined categories at message level. Results on 7 aspect classes show 0.71 accuracy, while the 32 class classification gives 0.82 accuracy for messages containing explicit aspects and 0.35 for implicit aspects.

1 Introduction

Sentiment Analysis (SA) in the financial domain has shown a growing interest in recent years. Acquiring an insight into the public opinion of relevant and valuable economic signals can give a competitive edge and allow more informed investment decisions to be executed. Microblog platforms such as Twitter and StockTwits, are central to determining these economic signals (Bollen et al., 2011; Zhang et al., 2011). Investors share their opinions about stocks, companies and products, and these contents are valuable for whomever is interested in predicting market trends. Research in the area of SA tries to shed some light on this problem. Its purpose is to identify opinions and sentiments that are directed towards entities such as stocks and companies or towards the attributes, or aspects, of these entities.

The authors are involved in SSIX1 (Davis et al., 2016), a project focused on SA in financial markets. It currently offers sentiment scores for stocks and companies and intends to provide finer-grained SA by including aspects. In order to conduct Aspect-Based SA in this project, the first step is to identify aspects in microblog messages, which is the focus of this paper.

As stated in SemEval-2015, the problem in Aspect-based SA can be divided into three sub-tasks, i.e. aspect category identification, Opinion Target Expression (OTE) extraction and sentiment polarity assignment (Pontiki et al., 2015). In this paper, we focus on the first sub-task of aspect category assignment. There have been two types of approaches to conduct this subtask. In the first type, aspect words are extracted and clustered (Qiu et al., 2011; Chen and Liu, 2014; Shu et al., 2016; Poria et al., 2016). In the second type, predefined aspects categories are assigned to entity-attribute pairs at sentence level (Pontiki et al., 2015). The first type of approaches targets explicit aspects while the second one also includes implicit aspects, i.e. aspects that are not explicitly mentioned in the text strings (Liu, 2012, p. 77). Using predefined aspects corresponds to the project requirements but most approaches deal with hotel, restaurant and product-related data. To the best of our knowledge none of them use a corpus of annotated aspects in the financial domain.

We present a method that focuses on the aspect category identification of implicit and explicit aspects. The originality of our work is to evaluate different aspect category identification approaches based on a predefined taxonomy of stock-investment aspects. Work is carried out on a limited data set with a view to expanding it should results be satisfactory. Our approach relies on using a corpus of annotated messages to build several types of models based on distributional semantics and supervised learning methods. Also original is that our work focuses on the stock-investment domain as it is to be added to the SSIX
platform. The remainder of this paper is divided as follows. Section 2 covers related work. Section 3 gives details about the corpus that was used. In Section 4 the different models are described. Results are presented in Section 5, followed by the conclusion in Section 6.

2 Related Work

Available methods in aspect category identification can be divided into supervised and unsupervised approaches. Unsupervised approaches include a number of lexicon-based strategies relying on i) frequency measures used with association measures such as Point-wise Mutual Information (PMI) to link words with lexicon entries (Popescu and Etzioni, 2005; Long et al., 2010), ii) syntactic relations to relate core sentiment words, expressed by adjectives, to target aspect words expressed by nouns (Liu et al., 2016; Fang and Huang, 2012; Jo and Oh, 2011; Brody and Elhadad, 2010; Chen and Liu, 2014), and iii) on word association measures for topic extractions and clustering methods (Fang and Huang, 2012; Jo and Oh, 2011; Brody and Elhadad, 2010; Chen and Liu, 2014). All these methods rely on lexicons to search for explicit words linked to aspects.

Supervised approaches rely on Machine Learning (ML) algorithms that are trained on classified instances of aspects prior to performing classification of new instances. Many studies have proposed different types of Conditional Random Fields (CRF) models (Jakob and Gurevych, 2010; Mitchell et al., 2013; Shu et al., 2016; Cruz et al., 2014; Poria et al., 2016) that distinguish aspects from non-aspects in text sequences. In parallel, other methods apply aspect category identification on the basis of predefined aspects linked to Entity (E) and Attribute (A) pairs (Pontiki et al., 2015, 2014). The current SemEval framework requires the extraction of explicit mentions of E and of all mentions of A (implicit and explicit) (Pontiki et al., 2015).

With respect to the implicit / explicit distinction, traditional approaches have focused on explicit aspects (Liu et al., 2016; Schouten et al., 2018), hence relying on word occurrences to determine aspects. Other, more novel, methods have focused on identifying implicitly-referred-to aspects (Pontiki et al., 2015). (Dosoula et al., 2016) developed an implicit feature algorithm that uses co-occurrences to assign implicit aspects at sentence level in online restaurant reviews.

Our framework is similar to SemEval-2015 Task 12 (Pontiki et al., 2015) insofar as we used predefined categories of aspects (A) for stocks considered as entities (E). Likewise, our approach includes the extraction of aspects that are not necessarily mentioned in messages. The difference is that we use a two-level aspect taxonomy for coarse and fine-grained characterization, which gives 32 fine-grained classes as opposed to the 9 classes of the laptop data set of SE-2015 task 12 for instance. We also conduct category identification at message level without creating E/A pairs. For the requirements of the project, we use a specific financial aspect taxonomy. Albeit applied to a different domain, results show higher or equivalent F1-Scores depending on the granularity.

3 Corpus

The approach relies on a corpus of messages specialised in stock trading \(^2\). Microblog messages were posted by stock traders who share investment ideas and intelligence. The data set is described in Table 1.

<table>
<thead>
<tr>
<th>Aspect type</th>
<th>Number of messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>All types</td>
<td>368</td>
</tr>
<tr>
<td>Implicit aspects</td>
<td>218</td>
</tr>
<tr>
<td>Explicit aspects</td>
<td>150</td>
</tr>
</tbody>
</table>

Table 1: Number of implicit and explicit messages in the data set

3.1 Taxonomy of Stock-Investment Aspects

As a preliminary step to aspect identification, a financial expert defined a taxonomy of trading aspects (See Appendix). They were grouped on the basis of hypo/hypernym relations following a general to more specific hierarchy. The final taxonomy consists in an aspect class dominating an aspect sub-class. No related terms, nor synonyms, were added to these subclasses. There are 7 aspect classes, e.g. User Action, Asset Direction and 32 aspect subclasses, e.g. User Action > Buying Intention. Aspect classes do not include the same number of subclasses. For instance, the User Action class includes 5 aspect subclasses while the User Outlook class includes 2 aspect subclasses. The

\(^2\)The dataset is available at https://bitbucket.org/ssix-project/stock-investment-aspect-extraction
taxonomy is used i) to compute the semantic relatedness between taxonomy labels and textual candidates (DSM approach. See Section 4.1) and ii) to relate message features with taxonomy classes (Supervised-learning approach. See Section 4.2.

3.2 Annotation Scheme

The messages were manually classified by one financial expert according to the afore-mentioned taxonomy by matching aspect classes and subclasses with messages. Annotation includes the message ID and the OTE that substantiates the selected class. The following example is a JSON-type extract of the first message classified as *User Outlook > Negative Outlook*.

```json
{
  "ID": 1,
  "AspectClass": "User Outlook",
  "Aspect": "Negative Outlook",
  "OTE": "Could easily see $AMZN drop 200 points after hours tomorrow",
  "Message": "Could easily see $AMZN drop 200 points after hours tomorrow after earnings"
}
```

4 Building a Classification Model

This section focuses on the method used to build different models for the aspect extraction task. The task of the classifier is to assign (i) aspect classes and (ii) subclasses to messages. In this section, we present the two approaches. The first one applies a distributional semantics model, while the second one is based on several Machine Learning algorithms.

4.1 Distributional Semantics Model (DSM)

This approach relies on word embeddings for the computation of semantic relatedness with Word2vec (Mikolov et al., 2013). Word embeddings fall in the category of distributional semantics methods in which the meaning of a word is related to the distribution of words around it (Jurafsky and Martin, 2009, p.659-665).

Word2vec, in its skip-gram architecture, is such a model and was trained on the Google news corpus. The vector values are the weights computed by the hidden layer of a Neural Network trained on a corpus. The Word2vec skip-gram model allows to find words that appear frequently together, and infrequently in other contexts (Mikolov et al., 2013).

The task of identifying aspects can be formulated as mapping textual elements of messages to their most related aspect class label in the taxonomy. There are two steps: extracting candidates and computing relatedness with the classes.

4.1.1 Extracting Candidates

After preprocessing (tokenisation and Part-of-Speech (POS) tagging) The extraction of candidates relies on rule-based heuristics using morphosyntactic patterns to select relevant Noun Phrases and Verb Phrases including modifiers such as adverbs, adjectives and present participles. The purpose is to capture fine-grained senses of these phrases. Example (1) illustrates the extraction of the item *declining revenue*.

1) $MCD with *declining revenue* for a good while

In example (1) only *declining revenue* is extracted. This segment is semantically relevant for the classification as *Revenue Down*, while the remainder of the NP does not procure any information regarding the type of aspect.

4.1.2 Computing Semantic Relatedness

Computing semantic relatedness consists of comparing vectors of candidates with vectors of aspect subclasses. First, multi-word candidates or labels are combined into single vectors to obtain pairs of candidate-aspect vectors. The method is the sum of the vectors of multi-word expressions. To compute relatedness between vectors, we use the Indra implementation (Freitas et al., 2016) of the cosine similarity metric. The system computes cosine similarity for all possible pairwise combinations of tokens in each message. We retain the pair with the highest score.

4.2 Supervised Learning Models

This approach relies on training several machine-learning models. Building the classifier consists in a multi-class supervised classification task.

4.2.1 Feature Engineering

After preprocessing (tokenisation, accent removal, lower-casing and POS tagging), messages were converted into vectors including the following features:

- **Bag of Words (BoW)** - They are used to create a numerical representation of the vocabulary of messages. We use three types of statis-
tics (binary count, frequency count and tf-idf) applied on n-gram clusters.

- **Part of Speech** - PoS are used to create a numerical representation of the POS present in each message. This representation is based on the Penn Treebank POS tagset (Marcus et al., 1993).

- **Numericals** - These are used to create a representation of financial values mentioned in the messages such as percentages, ratios, stock prices and amounts (e.g. $55).

- **Predicted sentiment of entity** - The sentiment predicted on the financial entities included in the messages that may contain aspects. It is a continuous value on a [-1;1] range.

### 4.2.2 Machine-Learning Algorithms and Optimization

A number of Machine Learning Python-based models were tested. Two methods are based on decision trees with XGboost (Chen and Guestrin, 2016) and Random Forests (Breiman, 2001). We also used Support Vector Machines (Vapnik, 2000) and Conditional Random Fields (Lafferty et al., 2001). Each of these methods use the same vector representation created in the feature engineering phase.

In order to find the best hyper-parameters for the tested models, we used the Particle Swarm Optimization (PSO) method. This method was appropriate due to the fact that hyper-parameters are numbers, mostly in a continuous space. PSO (Kennedy and Eberhart, 1995) was applied using 100 particles (specific hyper-parameter configurations) during 100 iterations, using same weights for velocity, particle best and global best. For each particle position, the average accuracy in 10-fold cross validation was calculated.

### 4.3 Model Selection, Validation and Evaluation

Choosing the best classifier is done in two stages. Firstly, a model selection procedure helps select the best model among the DSM and ML models. All models were tested with 10-fold cross-validation whereby the dataset is divided in ten parts. Each part is used as a test set once in the ten iterations of the process. Secondly, the selected model is validated by using the leave-one-out option, meaning that the training is conducted on all instances except one. The process is repeated until all instances have been used as a test instance.

In the model selection stage we computed global accuracy for 32 classes. In the validation stage, we used F1-Score for 7 and 32 classes to measure the effects of the coarse and fine-grained annotation levels. The annotated corpus described in Section 3 was used for training and testing. In the DSM approach, 172 initially annotated messages were used as test set.

### 5 Results and Discussion

In the model selection stage all of the approaches show different results as shown in Table 2.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSM (baseline) 0.425</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>ML Methods</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xgboost 0.5689</td>
<td>0.046</td>
<td></td>
</tr>
<tr>
<td>Random Forest 0.5435</td>
<td>0.038</td>
<td></td>
</tr>
<tr>
<td>SVC 0.449</td>
<td>0.027</td>
<td></td>
</tr>
<tr>
<td>CRF 0.431</td>
<td>0.052</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Model selection stage: Accuracy for each model for the 32 aspect classification task

Xgboost was selected and validation showed results (see Table 3) in line with the best scores obtained in SemEval-2015 Task 12.

Table 4 shows the accuracy for message classification according to the implicit or explicit nature of the 32 aspects. The distinction between implicit and explicit aspect messages shows that explicit aspects are better classified while implicit aspects are only correctly handled in about 35% of cases. This suggests that the classifier lacks significant features to identify implicit aspects. The size of the data set appears to be a limitation but the size of sentences may also impair the classifier by adding noise to the data. Using aspect-relevant OTEs as a BoW feature could help address this point.

### 6 Conclusion and Future Work

In this paper, we have reported on a series of experiments in the domain of Aspect Extraction. The experiments focused on the sub-task of aspect cat-
### Table 3: Model validation stage: Accuracy, F1-Score, Precision (P) and Recall (R) for the 32 and 7 aspect classification task

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc</th>
<th>F1-Score</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xgboost</td>
<td>0.565</td>
<td>0.49</td>
<td>0.52</td>
<td>0.49</td>
</tr>
<tr>
<td>(32 classes)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xgboost</td>
<td>0.712</td>
<td>0.71</td>
<td>0.70</td>
<td>0.71</td>
</tr>
<tr>
<td>(7 classes)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 4: Accuracy according to messages including 32 implicit and explicit aspects

<table>
<thead>
<tr>
<th>Aspects</th>
<th>Acc</th>
<th>F1-Score</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implicit</td>
<td>0.351</td>
<td>0.32</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>Explicit</td>
<td>0.826</td>
<td>0.8</td>
<td>0.84</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Results show that explicit aspect identification performs well, but implicit aspect identification remains an issue that can be tackled with larger data set and improved feature engineering. Despite the size of the training data set, results suggest that more efforts can be invested in the development of a larger data set.

## 7 Appendix

Taxonomy of stock-investment aspects

- **User Action**
  - Buying Intention
  - Selling Intention
  - Bought
  - Sold
  - Shorting

- **User Outlook**
  - Positive Outlook
  - Negative Outlook

- **Insider Activity**
  - Insider Selling
  - Insider Buying

- **Asset Direction**
  - Moving Higher
  - Moving Lower
  - Breakout
  - New High
  - Trending Higher
  - Trending Lower
  - Trending Sideways

- **Asset Behaviour**
  - Oversold
  - Overbought
  - Overvalued
  - Undervalued
  - Short Squeeze
  - Selling Pressure

- **Financial Results**
  - Earnings Beat
  - Earnings Miss
  - Revenue Up
  - Revenue Down
  - Profit Warning

- **Analyst Ratings**
  - Buy Recommendation
  - Sell Recommendation
  - Rating Upgrade
  - Rating Downgrade

## References


Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androustopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4:


