SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News

Keith Cortis[*], André Freitas[*], Tobias Daudert[**], Manuela Hürlimann[**],
Manel Zarrouk[**], Siegfried Handschuh[*], and Brian Davis[**]

[*]Department of Computer Science and Mathematics, University of Passau, Germany
[**]Insight Centre for Data Analytics, National University of Ireland, Galway

Abstract

This paper discusses the "Fine-Grained Sentiment Analysis on Financial Microblogs and News" task as part of SemEval-2017, specifically under the "Detecting sentiment, humour, and truth" theme. This task contains two tracks, where the first one concerns Microblog messages and the second one covers News Statements and Headlines. The main goal behind both tracks was to predict the sentiment score for each of the mentioned companies/stocks. The sentiment scores for each text instance adopted floating point values in the range of -1 (very negative/bearish) to 1 (very positive/bullish), with 0 designating neutral sentiment. This task attracted a total of 32 participants, with 25 participating in Track 1 and 29 in Track 2.

1 Overview

Our task is focused on Sentiment Analysis in the domain of financial microblogs and news. Domain-specific Sentiment Analysis has received much attention within the NLP community, motivated by the highly domain-dependent language used to express sentiment (Liu, 2012). Domain-specificity impacts all levels of analysis. On the lexical level, which is crucial in sentiment analysis, Liu (2012) notes that positive words in one domain can be negative in another, and vice versa. For instance, Loughran and McDonald (2011a) show that many words which are considered negative in general-purpose polarity lexicon have a neutral meaning in the financial domain (e.g. "liability"). This makes it difficult to transport sentiment classifiers across domains and underlines the need for domain-specific tools.

The financial domain is a high-impact use case for Sentiment Analysis because it has been shown that sentiments and opinions can affect market dynamics (Goonatilake and Herath, 2007; de Kauter et al., 2015). Sentiments are in some cases derived from news which discuss macroeconomic factors, company-specific, or political information as all of these can be market-relevant (Sinha, 2014). Good news tends to lift markets and increase optimism (de Kauter et al., 2015; Schuster, 2003). Evidence has been found that both quantitative measures (e.g. the quantity of news, market fluctuation) and qualitative indicators, (e.g. linguistic style and tone) affect investors' behaviour (Tetlock et al., 2008; Loughran and McDonald, 2011a; Takala et al., 2014). (Bollen et al., 2011) showed that changes in public mood reflect value shifts in the Dow Jones Industrial Index three to four days later.

Given the link between sentiment and market dynamics, the analysis of public sentiment becomes a powerful method to predict the market reaction. However, the accuracy of machine learning-based sentiment analysis approaches rarely exceeds seventy percent (Takala et al., 2014; Eagle Alpha, 2016). Research effort is required to overcome and address complex linguistic issues, such as sarcasm, irony and poorly-structured and/or colloquial language (Eagle Alpha, 2016). In addition, text that is short in length (such as microblog messages) can be quite opinionated, dense in information, dependent on the modelling of economic context and challenging to parse, due to the different vocabularies used (Sinha, 2014). Our task is motivated by the interest of this field and the great potential for improvement. It aims at assessing the overall market sentiment as well as sentiment about specific stocks and thus to make use of their predictive power.

More specifically, the aim of organising this task and creating this test collection was to achieve the following goals:

1. Developing state-of-the-art on classification

methods for sentiment analysis in the domain of financial short texts.

2. Incentivising the creation of new lexical resources for the financial domain.

3. Understanding how state-of-the-art sentiment analysis performs on a domain-specific / highly technical corpus.

4. Improving the understanding of linguistic phenomena and the creation of semantic models for the financial domain.

The domain of finance has unique linguistic and semantic features, whose interpretation depends on the formulation of semantic models which reflect the economic and mathematical tools used by the experts to assess financial information. Moreover, the accurate interpretation of financial text requires the orchestration of large volumes of common sense and domain-specific financial/economic knowledge. Additionally, as much of the financial discourse is mediated by terms which demand precise definitions, many times associated with the quantification of economic phenomena, the semantic interpretation processes in the financial domain require fine-grained semantic interpretation approaches.

From a linguistic standpoint, topics of interest in this task include (but are not limited to):

- Low-level linguistic analysis tools for the financial domain (e.g. tokenization, part-of-speech tagging, parsing)
- Sentiment classification on financial texts;
- Understanding of linguistic phenomena associated with financial tweets;
- New semantic models for finance;
- Construction and application of distributional semantic models on finance;
- Sentiment compositionality;
- Machine learning approaches for sentiment classification;
- Lexical resources for the financial domain;

2 Data

2.1 Tracks

The test collection consists of two tracks:

1. **Microblog Messages** derived from two sources:
   (a) *StockTwits Messages*: Consists of microblog messages focusing on stock market events and assessments from investors and traders, exchanged via the StockTwits microblogging platform[1]. Typical stocktwits consist of references to company stock symbols (so-called cashtags - a stock symbol preceded by "$", e.g. "$AAPL" for the company Apple Inc.), a short supporting text or references to a link or pictures (typically containing charts showing stock values analysis).
   (b) *Twitter Messages*: Some stock market discussion also takes place on the Twitter platform[2]. In order to extend and diversify our data sources, we extract Twitter posts containing company stock symbols (cashtags).
2. **News Statements & Headlines** Sentences have been taken from news headlines as well as news text. The textual content was crawled from different sources on the Internet, such as Yahoo Finance[3]. The Enterprise identification for this track was based on company names and abbreviations, as cashtags are not typically used in news statements and headlines.

2.2 Corpus Creation

The corpus of statements was created by conducting random sampling and an initial filtering process over a pool of StockTwits messages, tweets and RSS News feeds.

While the random sampling ensured an unbiased set of statements, the filtering mechanism aimed at removing messages from the set microblog messages which are spam. The filtering mechanism was based on a manual curation of the set of microblog users which are classified as spammers. The goal of data sampling is to come up with a most representative and manageable amount of data for manual annotation. The first step in our case is to ap-

[1] http://stocktwits.com/
[2] https://twitter.com
[3] http://finance.yahoo.com/

ply a stratified random sampling by objects δ per the smallest time unit level θ we determine (in our case it is stock's messages per day) to ensure that all different objects are adequately represented in the sample with respect to their distribution in the population. Then, the random samples of a time-unit level θ_i are pooled into a time-unit level θ_{i+1} and randomly sampled.

The purpose of re-sampling at different time-unit levels is to make the resulted random sample more random, more balanced and more representative of the entire time-span of our data. A general negative sentiment in a certain sub-sample will be counter-balanced by the other sub-samples.

StockTwits data have been provided by StockTwits in a batch export and refer to the period from October 2011 to June 2015. The original pool before sampling contains 27 million StockTwits, from which 1847 messages were sampled. Twitter data was collected between March 11th and 18th 2016 using the official Streaming APIs. Sampling was also applied to this data and resulted in a sample of 1591 messages.

The News Statements and Headlines have been collected from a pool of 20.000 RSS feeds in the period between August and November 2015 (e.g. AP News, Reuters, Handelsblatt, Bloomberg and Forbes). A final set of about 1780 News Statements and Headlines has been produced.

2.3 Annotation

To create the Gold Standard, the final sample has been annotated by 3 independent financial expert annotators using a Web platform developed for that purpose and according to the annotations guidelines we defined. A fourth domain expert consolidated the ratings to create the final data set. The total time the experts spend on annotating and consolidating the data set is 120 hours (30 hours per expert). The costs of annotation and consolidation have been covered by ICT-15-2014 Grant: 645425 (SSIX project).

Each statement (instance) is annotated with the following information:

- **Cashtag (subtask1) / Company (subtask2)**: A stock company symbol (for microblogs) or reference to a company (for news/headlines) to which a sentiment score is assigned.

- **Sentiment Score**: A sentiment between -1 (very negative/bearish) and 1 (very positive/bullish), with 0 representing neutral/no sentiment is assigned to each cashtag or company. The sentiment is assigned from the point of view of an investor and the sentiment annotation is carried out by domain experts. Textual data containing information implying a positive prospective trend for a company or stock, the markets, or the economy, in general, constitutes a positive sentiment, whereas information revealing negative trends constitutes a negative sentiment since it may impact companies, markets or the economy negatively.

- **Span (subtask 1)**: extract of a text string in which sentiment is expressed.

- **Message (subtask 1) / Title (subtask2)**: Text string in which sentiment is expressed.

- **Source (subtask 1)**: Either "twitter" or "stocktwits" dependent on the origin of the text message.

Examples of annotated microblog messages and news headlines are provided in Section 2.6 below.

The quality of the annotations was assessed following a similar methodology as proposed in Takala et al. (2014), where inter-annotator agreements measures for continuous data is calculated for the sentiment classifications. Spearman's Rank Correlation on sentiment scores was calculated for each pair of annotators, then averaged across annotator pairs. This yielded the following results: 0.54 for news headlines (three annotators, three pairs) and 0.69 for microblogs (four annotators, six pairs).

2.4 Gold Standard

After annotating and consolidating the data, the gold standard for subtask 1 consists of 2510 Twitter and StockTwit messages. The gold standard for subtask 2 contains 1647 Headlines and News Statements.

2.5 Task Formulation

Participating systems needed to fulfil the following task: given a text instance (microblog message in Track 1, news statement or headline in Track 2; cp. Section 2.1), predicting the sentiment score for each of the companies/stocks mentioned. Sentiment values needed to be floating point values within the range of -1 (very negative/bearish) to 1 (very positive/bullish), with 0 designating neutral sentiment.

2.6 Examples

Below we present annotated example statements, two for microblogs and one for news. Please note that sentiment score agreement as per Section 2.3 is not given as annotations for these examples were provided by a single expert. Also, the string covered by the 'span' is given for ease of reading.

Microblogs

> Este Lauder beats on Revenues and EPS and boosts dividend 25% - global growth in the Middle Class trend continues. $EL $NKE $SBUX $AAPL

- **Sentiment Score**:
  - $EL: 0.95
  - $NKE: 0.5
  - $SBUX: 0.5
  - $AAPL : 0.5

- **Cashtag**
  - $EL
  - $NKE
  - $SBUX
  - $AAPL

- **Span**
  - $EL:
    - (13, 38) - "beats on Revenues and EPS"
    - (43, 62) - "boosts dividend 25%"
    - (65, 144) - "global growth in the Middle Class trend continues"
  - $NKE, $SBUX, $AAPL:
    - (65, 144) - "global growth in the Middle Class trend continues"

> Awaiting These Sell Signals on the $SPY & $QQQ - https://t.co/GF9PRk5OUF $TQQQ $SQQQ https://t.co/W97yN4Zb4N

- **Sentiment Score**:
  - $SPY: -0.25
  - $QQQ: -0.15
  - $TQQQ: -0.15
  - $SQQQ : 0.10

- **Cashtag**
  - $SPY
  - $QQQ
  - $TQQQ
  - $SQQQ

- **Span**
  - $SPY:
    - (0, 41) - "Awaiting These Sell Signals on the $SPY"
    - (From the blog post) - "this bearish rising wedge for the next sell signal in the SPY"
    - (From the blog post) - Chart shows a bearish rising wedge
  - $QQQ, $TQQQ:
    - The message and blog make reference to shorting the SPY, but as but indexes are strongly correlated so some of the sentiment for SPY could be transferred to these ETFs.
  - $SQQQ:
    - The message and blog make reference to shorting the SPY, but as indexes are strongly correlated so some of the sentiment for SPY could be transferred to this ETF but inverted.

News Statements & Headlines

> First Solar, Vivint Solar Lead Short Interest Trend

- **Sentiment Score**:
  - First Solar: -0.7
  - Vivint Solar: -0.7

- **Company**
  - First Solar
  - Vivint Solar

2.7 Assessment Infrastructure & Baselines

Two classification baselines were provided:

- **Random selection:** Consists of a random number generated within the sentiment range.
- **SentiWordNet-based average and maximum functions:** Consist of the maximum and averaging of all the sentiment words using a simple SentiWordNet-based lookup.

For the Microblogs test set, SentiWordNet lexicon-based look-up (average) achieved an average score of 0.3021, while the same look-up method using the max/min score achieved 0.2428. The random baseline achieved 0.0148.

For the Financial Headlines test set, a SentiWordNet lexicon-based look-up classifier, which averages all the sentiment scores of individual lemmatised words, achieved a score of 0.290, while the same look-up method using a max/min score achieved 0.2184. The random baseline achieved 0.1064.

3 Pilot Task

A pilot dataset consisting of financial social data was collected from two on-line social networking services, specifically Twitter and StockTwits, as part of a pilot study carried out within the **SSIX: Social Sentiment analysis financial IndeXes**[4] project as part of the European Horizon 2020 Research and Innovation programme (Davis et al., 2016). A domain expert experienced in trading annotated 100 tweets and 100 StockTwits messages selected randomly. He annotated the messages for sentiment following the guideline of assuming the point of view of an investor in the given stock(s) (see Section 2.3 above).

The results from the pilot study provided valuable insights with regards to the distribution of sentiment and the need for improved filtering (Figures 3 and 2). These insights proved to be valuable when building the data set for this task, enabling us to provide a higher-quality data collection.

The results (Figure 1) showed a relatively even distribution of positive and negative sentiment, with slight differences between the StockTwits and Twitter sources as regards the intensity of the sentiment .

[4] http://ssix-project.eu/

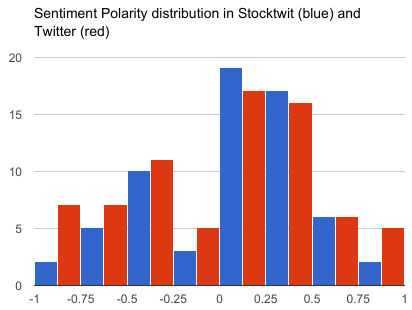

Figure 1: Pilot results on sentiment distribution for StockTwits and Twitter

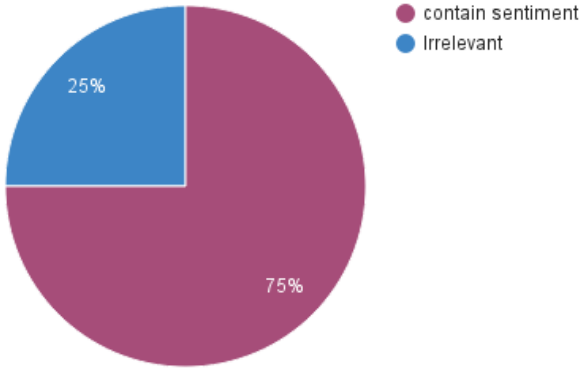

Figure 2
Pilot results on percentage of sentiment-containing and irrelevant messages on Twitter.

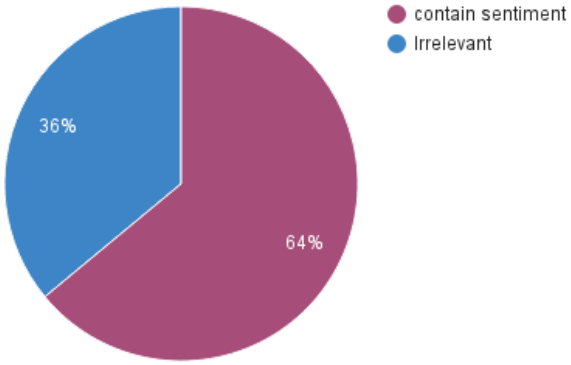

Figure 3
Pilot results on percentage of sentiment-containing and irrelevant messages on StockTwits.

The pilot study also pointed to the need to improve the filtering phase as 25% - 36% of Twitter and StockTwits messages, respectively, have been deemed irrelevant (i.e. spam and/or not providing any relevant financial sentiment) by the annotator (figure 3 and 2). As a consequence, filtering rules have been added to the filtering phase and the data for the gold standard proposed in this task under-

went additionally manual post-filtering by a domain expert prior to sentiment annotation. This is ensuring that only relevant messages are included in the data set.

4 Evaluation

The Evaluation of the participating systems was based on cosine similarity, in a spirit similar to Ghosh et al. (2015). As the sentiment scores to be predicted by systems lie on a continuous scale between -1 and 1 (cp. Section 2.5), cosine enables us to compare the proximity between gold standard and predicted results (conceptualized as vectors), while not requiring exact correspondence between the gold and predicted score for a given instance. An instance is a message or headline which can include several entities (companies or cashtags). Cosine similarity is calculated according to equation (1), where `G` is the vector of gold standard scores and `P` is the vector of corresponding scores predicted by the system:

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

In order to reward systems which attempt to answer all problems in the gold standard, the final score is obtained by weighting the cosine similarity from (1) with the ratio of answered problems (scored instances), given in (2) in line with Ghosh et al. (2015).

$$cos\_weight = \frac{|P|}{|G|} \quad (2)$$

The equation for the final score is the product of the cosine similarity (1) and the weight (2), given in (3).

$$final\_score = cos\_weight \times cosine(G,P) \quad (3)$$

5 Results and Participants

Task 5 attracted a total of 32 participants: 25 teams participated in Track 1 and 29 in Track 2, of which 22 teams addressed both tracks. The analysis and results for each track are discussed in more detail in the sub-sections below. Given that 19 out of the 32 participants submitted a paper with their approach and findings, we opted to include the system analysis and ranking of results of only the submitted participants.

Analysis of the systems consisted of the following criteria: pre-processing methods, techniques used, external sources, data sets and/or lexica used, tools utilised, why the adopted approach was chosen and if it is (i) multilingual/cross-lingual and/or (ii) domain dependent/independent, any issues encountered and how they were tackled and potentially solved.

5.1 Track 1 - Microblog Messages

Figure 4 shows the results of the 25 participants in Track 1. Results of all participants were ranked (first column) according to the evaluation metric (last column) described in the Section 4 - Evaluation. The second column specifies the team's/participant's name. Please note that the analysis of systems discussed in this sub-section includes only the participants highlighted in yellow since only these submitted a paper with their approach and findings.

Track 1 - Microblog Messages			
Rank	**Team Name**		**Score**
1	ECNU	Jiang et al. (2017)	0.778
	CodersGoneCrazy		0.760
	zhiqiang		0.759
2	IITP	Ghosal et al. (2017)	0.751
3	SSN_MLRG1	Deborah et al. (2017)	0.735
4	HHU	Cabanski et al. (2017)	0.730
5	IITPB	Kumar et al. (2017)	0.726
6	RiTUAL-UH	Kar et al. (2017)	0.723
7	IBA-Sys	Nasim (2017)	0.720
	cbaziotis		0.714
8	SentiHeros	Seyeditabari et al. (2017)	0.707
	mattia-atzeni		0.703
9	FEUP	Saleiro et al. (2017)	0.693
10	funSentiment	Li et al. (2017)	0.677
	hittle2008		0.655
	xiwu		0.636
11	INF-UFRGS	Zini et al. (2017)	0.614
	lhurtado		0.614
12	HCS	Pivovarova et al. (2017)	0.607
	amouma		0.507
13	NLG301	Chen et al. (2017)	0.383
	vpekar		0.337
	four_u		0.314
	gloncakv		0.186
14	DUTH	Symeonidis et al. (2017)	0.003

Figure 4: Track 1 Results

5.1.1 Pre-processing

In terms of pre-processing, all 14 participants adopted some methods in order to clean the microblog messages before further processing. The most common methods used were: removal of special characters and/or punctuations, removal of URLs and user mentions ('@username') and/or substitution of certain expressions by specific words (e.g., replace 'full urls' with 'url' and 'company names' with 'company'), stop word removal, tokenisation, lemmatisation and lowercase conversion. Some participants also performed Named Entity Recognition (NER), emoticon removal, Part-of-Speech (POS) tagging, stemming and URL resolution, besides other specific tasks, such as concatenation of spans to form a unified string (Nasim, 2017). NLTK[5] was the tool mostly used (Seyeditabari et al., 2017; Deborah et al., 2017; Kumar et al., 2017; Symeonidis et al., 2017; Jiang et al., 2017) for pre-processing tasks, such as lemmatisation, stemming and lowercase conversion.

5.1.2 Techniques

All the techniques used by each system of the 14 participants are shown in Figure 5. Each system was analysed and in-turn categorised under one of the following techniques: Hybrid, Machine Learning (ML), Deep Learning (DL) and Lexicon-based (Lex).

It is clear that most techniques were of a Hybrid nature with the Machine Learning and Lexicon-based approach being the most popular choice, followed by Machine Learning-based approaches. Authors of some systems experimented with multiple approaches to find the one that fared best in the competition. In fact, Cabanski et al. (2017) implemented two-hybrid techniques (as noted above), where the Hybrid (DL, Lex) approach produced their best result for this track. On the other hand, Kumar et al. (2017) implemented two Hybrid (ML, Lex) systems, one adopting Support Vector Machine and Logistic Regression and the other adopting SVR.

The Hybrid (ML, Lex) technique by Jiang et al. (2017) ranked first for this track, whereas the Hybrid (DL, Lex) technique by Ghosal et al. (2017) ranked second. The system placing third (Deborah et al., 2017) adopted a ML technique.

The Machine Learning-based techniques made use of the following algorithms:

- Artificial Neural Network (ANN) - adopted by Li (2017); Symeonidis et al. (2017); Saleiro et al. (2017)
- Random Forests - adopted by Seyeditabari et al. (2017); Symeonidis et al. (2017); Jiang et al. (2017); Saleiro et al. (2017)
- Support Vector Machine (SVM) - adopted by Seyeditabari et al. (2017); Cabanski et al. (2017); Kumar et al. (2017); Saleiro et al. (2017)
- Support Vector Regression (SVR) - adopted by Zini et al. (2017); Kumar et al. (2017); Chen et al. (2017); Jiang et al. (2017)
- Linear Regression (LiR) - adopted by Symeonidis et al. (2017)
- Logistic Regression (LoR) - adopted by Seyeditabari et al. (2017); Kumar et al. (2017)
- Naive Bayes (NB) - adopted by Seyeditabari et al. (2017)
- Multi-Kernel Gaussian Process (MKGP) - adopted by Deborah et al. (2017)
- XGBoost Regressor (XGB) - adopted by Nasim (2017); Jiang et al. (2017)
- Boosted Decision Tree Regression (BDTR) - adopted by Symeonidis et al. (2017)
- AdaBoost Regressor (ABR) - adopted by Jiang et al. (2017)
- Bagging Regressor (BR) - adopted by Jiang et al. (2017)
- Gradient Boosting Regressor (GBR) - adopted by Jiang et al. (2017)
- Least Absolute Shrinkage and Selection Operator (LASSO) - adopted by Jiang et al. (2017)

The most common ML techniques used overall –by 4 participants– were RF, SVM and SVR. The SVR was part of the ensemble regression model used by the system that ranked first for this track (Jiang et al., 2017). The RF classifier was ultimately used by Seyeditabari et al. (2017), since it is the best performer in terms of tweets classification. Regarding the ANN computational approach, both Li (2017) and Symeonidis et al. (2017) use a regression method, whereas Saleiro et al. (2017) use a Multilayer Perceptron (MLP).

The Deep Learning-based techniques made use of the following algorithms:

[5] http://www.nltk.org/

Technique	System
Hybrid (ML, Lex)	Nasim (2017), Seyeditabari et al. (2017), Cabanski et al. (2017), Kumar et al. (2017), Chen et al. (2017), Jiang et al. (2017), Saleiro et al. (2017)
Hybrid (DL, Lex)	Ghosal et al. (2017), Cabanski et al. (2017), Kar et al. (2017)
ML	Li (2017), Zini et al. (2017), Symeonidis et al. (2017), Deborah et al. (2017)
DL	Pivovarova et al. (2017)

Figure 5: Techniques used by systems in Track 1

- Convolution Neural Network (CNN) - adopted by Pivovarova et al. (2017); Kar et al. (2017); Ghosal et al. (2017)
- Recurrent Neural Network (RNN) : Long Short-Term Memory (LSTM) - adopted by Cabanski et al. (2017); Ghosal et al. (2017)
- Bidirectional Gated Recurrent Unit (Bi-GRU) - adopted by Kar et al. (2017)

The MLP based ensemble model in Ghosal et al. (2017) that combines the CNN and LSTM Deep Learning algorithms ranked second in this track. In Cabanski et al. (2017), their best submission for this track was provided by the RNN (as opposed to SVR).

Lexicon-based methods made use of the following known sentiment lexica:

- Loughran and McDonald Sentiment Word Lists (Loughran and McDonald, 2011b) - adopted by Nasim (2017); Seyeditabari et al. (2017); Saleiro et al. (2017); Ghosal et al. (2017)
- Stock Market Lexicon[6] - adopted by Nasim (2017)
- SentiWordNet[7] - adopted by Cabanski et al. (2017); Chen et al. (2017); Jiang et al. (2017)
- SenticNet 4[8] - adopted by Chen et al. (2017); Kar et al. (2017)
- VADER (Hutto and Gilbert, 2014) - adopted by Cabanski et al. (2017)
- Opinion Lexicon(Hu and Liu, 2004) - adopted by Cabanski et al. (2017); Kumar et al. (2017); Jiang et al. (2017); Ghosal et al. (2017)
- MPQA Subjectivity Lexicon (Wilson et al., 2009) - adopted by Kumar et al. (2017); Jiang et al. (2017); Saleiro et al. (2017); Ghosal et al. (2017)
- NRC Hashtag Sentiment Lexicon (Kiritchenko et al., 2014) - adopted by Cabanski et al. (2017); Kumar et al. (2017); Jiang et al. (2017); Ghosal et al. (2017)
- NRC Hashtag Emotion Lexicon (Kiritchenko et al., 2014) - adopted by Chen et al. (2017)
- NRC Hashtag Affirmative Context Sentiment Lexicon (Kiritchenko et al., 2014) - adopted by Chen et al. (2017); Ghosal et al. (2017)
- NRC Hashtag Negated Context Sentiment Lexicon (Kiritchenko et al., 2014) - adopted by Chen et al. (2017)
- NRC Word-Emotion Association Lexicon / NRC Emotion Lexicon (Kiritchenko et al., 2014) - adopted by (Chen et al., 2017)
- Emoticon Lexicon / Sentiment140 Lexicon[9] - adopted by Chen et al. (2017); Jiang et al. (2017); Ghosal et al. (2017)
- Sentiment140 Affirmative Context Lexicon (Kiritchenko et al., 2014) - adopted by Ghosal et al. (2017); Chen et al. (2017)
- Yelp Restaurant Sentiment Lexicon[10] - adopted by Chen et al. (2017)
- Amazon Laptop Sentiment Lexicon[11] - adopted by Chen et al. (2017)
- Macquarie Semantic Orientation Lexicon[12] - adopted by Chen et al. (2017)

[6] https://github.com/nunomroliveira/stock_market_lexicon
[7] http://sentiwordnet.isti.cnr.it/
[8] http://sentic.net/senticnet-4.pdf
[9] http://saifmohammad.com/Lexicons/Sentiment140-Lexicon-v0.1.zip
[10] http://saifmohammad.com/Lexicons/Yelp-restaurant-reviews.zip
[11] http://saifmohammad.com/Lexicons/Amazon-laptop-electronics-reviews.zip
[12] http://saifmohammad.com/Lexicons/MSOL-June15-09.txt.zip

- Harvard's General Inquirer Lexicon[13] - adopted by Jiang et al. (2017); Ghosal et al. (2017)
- IMDB (Zhu et al., 2013) - adopted by Jiang et al. (2017)
- AFINN[14] - adopted by Jiang et al. (2017)
- Corpus of Business News (Pivovarova et al., 2013) - adopted by Pivovarova et al. (2017)

The following three lexica listed are the ones mostly used overall: (i) the Loughran and McDonald Sentiment Word (rank 2), (ii) Opinion Lexicon (rank 1, 2) and (iii) MPQA Subjectivity Lexicon (rank 1, 2). An interesting observation is that the systems that ranked first (Jiang et al., 2017) and second (Ghosal et al., 2017) in this track utilised all three lexicons (ranked system using the particular lexicon represented next to each name), whereby lexica (ii) and (iii) were used by both.

Seyeditabari et al. (2017) extended Loughran and McDonald's word list of positive and negative words with 10,000 financial reports containing a summary of the company's performances in order to add features to the training dataset In Cabanski et al. (2017), the authors, besides using the sentiment lexica identified above, also built and used a custom Financial Sentiment Lexicon.

5.2 Track 2 - News Statements and Headlines

Figure 6 shows the results of the 29 participants in Track 2. Results of all participants were ranked (first column) according to the evaluation metric (last column) described in Section 4 - Evaluation. The second column specifies the team/participant name. Please note that the analysis of systems discussed in this sub-section includes only the participants highlighted in yellow, which are the participants who submitted a paper with their approach and findings.

5.2.1 Pre-processing

In terms of pre-processing – same as for Track 1 – all 17 participants adopted some methods in order to clean the news statements and headlines before further processing. The most common methods used were: removal of numbers, special characters and/or punctuations, removal of URLs and user mentions and/or substitution of certain expressions with tags (e.g. replace 'company name' with '⟨company⟩' and 'numbers' with '⟨number⟩'), stop word removal, tokenisation, lemmatisation, lowercase conversion and NER on certain entities (e.g., Organisation and Person). Some participants also performed dependency parsing, POS tagging, stemming and URL resolution, besides other specific tasks, such as filtering out all named entities and keeping only "general" tokens given that they are generally the ones carrying the sentiment (Rotim et al., 2017). Same as track 1, NLTK was the tool mostly used (Ghosal et al., 2017; Deborah et al., 2017; Kumar et al., 2017; Symeonidis et al., 2017; Jiang et al., 2017) for pre-processing, whereas Stanford CoreNLP[15] was used for performing NER, sentence breaking and parsing. (Nasim, 2017; Rotim et al., 2017; Schouten et al., 2017; Chen et al., 2017; Jiang et al., 2017)

Track 2 - News Statements & Headlines			
Rank	Team Name	Reference	Score
1	Fortia-FBK	Mansar et al. (2017)	0.745
2	RiTUAL-UH	Kar et al. (2017)	0.744
3	TakeLab	Rotim et al. (2017)	0.733
4	Lancaster A	Moore and Rayson (2017)	0.732
	CodersGoneCrazy		0.726
5	ECNU	Jiang et al. (2017)	0.710
6	HHU	Cabanski et al. (2017)	0.702
7	IITP	Ghosal et al. (2017)	0.697
8	IITPB	Kumar et al. (2017)	0.696
	cbaziotis		0.686
	zhiqiang		0.681
9	COMMIT	Schouten et al. (2017)	0.681
10	HCS	Pivovarova et al. (2017)	0.680
11	FEUP	Saleiro et al. (2017)	0.670
12	SSN_MLRG1	Deborah et al. (2017)	0.666
13	IBA-Sys	Nasim (2017)	0.656
14	UW-FinSent	John and Vechtomova (2017)	0.645
	MarinaChern		0.627
	bonson		0.615
	mattia-atzeni		0.613
15	INF-UFRGS	Zini et al. (2017)	0.608
	lhurtado		0.607
	xiwu		0.603
16	DUTH	Symeonidis et al. (2017)	0.588
	amouma		0.431
17	NLG301	Chen et al. (2017)	0.415
	vpekar		0.352
	hittle2008		0.251
	four_u		0.016

Figure 6: Track 2 Results

[13] http://www.wjh.harvard.edu/~inquirer/

[14] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

[15] http://stanfordnlp.github.io/CoreNLP/

5.2.2 Techniques

Figure 7 shows all the techniques used by each system of the 17 participants. Each system has been analysed and categorised under one of the following techniques: Hybrid, Machine Learning (ML), Deep Learning (DL), Lexicon (Lex) and Ontology (Ont).

Similar to track 1, the Machine Learning and a Machine Learning/Lexicon-based Hybrid approach were the ones mostly used (six participants). However, the techniques were more balanced in this track, with six participants adopting a Machine Learning-based approach. It is worth noting that one of the systems used a Machine Learning and Ontology-based Hybrid approach, which technique is unique in both tracks. In this system,Schouten et al. (2017) used the SVR algorithm with ontology features (including features derived from ontology reasoning), which ontology was self-designed by the authors. Multiple techniques were used by some authors in order to find the best one to use in this competition within their system. Moore and Rayson (2017) experimented with an ML and DL algorithm respectively, with the latter performing better. On the other hand, Cabanski et al. (2017) implemented two-hybrid techniques, where the Hybrid (DL, Lex) approach produced their best result for this track , same as for track 1.

The systems that ranked first (Mansar et al., 2017) and second (Kar et al., 2017) both adopted a Hybrid (DL, Lex) technique, whereas an ML technique was used by the system in rank three.

The Machine Learning-based techniques made use of the following algorithms:

- Artificial Neural Network (ANN) - adopted by Symeonidis et al. (2017)
- Random Forests - adopted by Symeonidis et al. (2017); Jiang et al. (2017); Saleiro et al. (2017)
- Support Vector Machine (SVM) - adopted by Kumar et al. (2017); Saleiro et al. (2017)
- Support Vector Regression (SVR) - adopted by Rotim et al. (2017); Schouten et al. (2017); Moore and Rayson (2017); John and Vechtomova (2017); Zini et al. (2017); Cabanski et al. (2017); Kumar et al. (2017); Chen et al. (2017); Jiang et al. (2017)
- Linear Regression (LiR) - adopted by John and Vechtomova (2017); Symeonidis et al. (2017)
- Logistic Regression (LoR) - adopted by Kumar et al. (2017)
- Multi-Kernel Gaussian Process (MKGP) - adopted by Deborah et al. (2017)
- XGBoost Regressor (XGB) - adopted by Nasim (2017); John and Vechtomova (2017); Jiang et al. (2017)
- Boosted Decision Tree Regression (BDTR) - adopted by Symeonidis et al. (2017)
- AdaBoost Regressor (ABR) - adopted by Jiang et al. (2017)
- Bagging Regressor (BR) - adopted by Jiang et al. (2017)
- Gradient Boosting Regressor (GBR) - adopted by Jiang et al. (2017)
- Least Absolute Shrinkage and Selection Operator (LASSO) - adopted by Jiang et al. (2017)

As can be seen above, the most common ML technique used within the systems was SVR by 9 participants. This was used by the system that ranked third for this track (Rotim et al., 2017).

The Deep Learning-based techniques made use of the following algorithms:

- Convolution Neural Network (CNN) - adopted by Mansar et al. (2017); Pivovarova et al. (2017); Ghosal et al. (2017); Kar et al. (2017)
- Recurrent Neural Network (RNN) : Long Short-Term Memory (LSTM) - adopted by Ghosal et al. (2017); Cabanski et al. (2017)
- RNN : Bidirectional Long Short-Term Memory (BLSTM) - adopted by Moore and Rayson (2017)
- Bidirectional Gated Recurrent Unit (Bi-GRU) - adopted by Kar et al. (2017)

The CNN algorithm was the most popular amongst all Deep Learning-based techniques, with both systems ranking first (Mansar et al., 2017) and second (Kar et al., 2017) using it.

Lexicon-based methods made use of the following known sentiment lexica:

- Loughran and McDonald Sentiment Word Lists - adopted by Nasim (2017); Ghosal et al. (2017); Kumar et al. (2017); Saleiro et al. (2017)

Technique	System
Hybrid (ML, Lex)	Nasim (2017), Cabanski et al. (2017), Kumar et al. (2017), Chen et al. (2017), Jiang et al. (2017), Saleiro et al. (2017)
Hybrid (DL, Lex)	Mansar et al. (2017), Ghosal et al. (2017), Cabanski et al. (2017), Kar et al. (2017)
Hybrid (DL, Ont)	Schouten et al. (2017)
ML	Rotim et al. (2017), Moore and Rayson (2017), John and Vechtomova (2017), Deborah et al. (2017), Zini et al. (2017), Symeonidis et al. (2017)
DL	Moore and Rayson (2017), Pivovarova et al. (2017)

Figure 7: Techniques used by systems in Track 2

- SentiWordNet - adopted by Cabanski et al. (2017); Kumar et al. (2017); Chen et al. (2017); Jiang et al. (2017)
- SenticNet 4 - adopted by Chen et al. (2017); Kar et al. (2017)
- VADER - adopted by Mansar et al. (2017); Cabanski et al. (2017)
- Opinion Lexicon - adopted by Ghosal et al. (2017); Cabanski et al. (2017); Kumar et al. (2017); Jiang et al. (2017)
- MPQA Subjectivity Lexicon - adopted by Ghosal et al. (2017)
- NRC Hashtag Sentiment Lexicon - adopted by Cabanski et al. (2017); Nasim (2017); Ghosal et al. (2017); Jiang et al. (2017)
- NRC Hashtag Emotion Lexicon - adopted by Chen et al. (2017)
- NRC Hashtag Affirmative Context Sentiment Lexicon - adopted by Ghosal et al. (2017); Chen et al. (2017)
- NRC Hashtag Negated Context Sentiment Lexicon - adopted by Chen et al. (2017)
- NRC Word-Emotion Association Lexicon / NRC Emotion Lexicon - adopted by Chen et al. (2017)
- Emoticon Lexicon / Sentiment140 Lexicon - adopted by Ghosal et al. (2017); Jiang et al. (2017); Chen et al. (2017)
- Sentiment140 Affirmative Context Lexicon - adopted by Ghosal et al. (2017); Chen et al. (2017)
- Yelp Restaurant Sentiment Lexicon - adopted by Chen et al. (2017)
- Amazon Laptop Sentiment Lexicon - adopted by Chen et al. (2017)
- Macquarie Semantic Orientation Lexicon - adopted by Chen et al. (2017)
- Harvard's General Inquirer Lexicon - adopted by Nasim (2017); Ghosal et al. (2017); Kumar et al. (2017); Jiang et al. (2017)
- IMDB - adopted by Jiang et al. (2017)
- AFINN - adopted by Jiang et al. (2017)
- DepecheMood Affective Lexicon (Staiano and Guerini, 2014) - adopted by Mansar et al. (2017)
- Amazon Product Reviews[16] - adopted by John and Vechtomova (2017)
- Financial Phrasebank (Malo et al., 2014a) - adopted by John and Vechtomova (2017)
- Corpus of Business News - adopted by Pivovarova et al. (2017)

In total, four lexica listed above are the ones mostly used (all by 4 participants each): (i) the Loughran and McDonald Sentiment Word, (ii) SentiWordNet, (iii) Opinion Lexicon and (iv) Harvard's General Inquirer Lexicon. Unlike the case in track 1, none of the participants ranked first till third used one of these four lexica.

Some authors constructed their own lexica from external sources, such as Moore and Rayson (2017) (rank four) who manually downloaded 189,206 financial articles which contain 161,877,425 tokens from Factiva[17] (articles come from sources such as Financial Times that relate to United States companies only).

[16] http://jmcauley.ucsd.edu/data/amazon/
[17] https://global.factiva.com

5.2.3 Tools used in both tracks

Several tools were used within the participants' systems, with the following (Figure 8) being the most popular:

Scikit-learn is a Machine Learning kit (in Python) that offers simple efficient tools (e.g., classification and regression algorithms) for data mining and data analysis. This is the tool mostly used by the participants of our task (42% in total) to compute their results. Similarly, Weka –a collection of machine learning algorithms for data mining tasks– was used by 2 participants. The Keras Deep Learning library was used by 2 participants, whereas TensorFlow – an open source software library for numerical computation using data flow graph– was also used by 3 participants (work in Pivovarova et al. (2017) built their implementation on top of it).

GloVe, an unsupervised learning algorithm for obtaining vector representations of words, was used by 6 participants for word embeddings. Word2vec –an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words– was also used for the same purpose by 4 participants.

5.3 General assessment of the task

The approaches proposed by the participating systems explored a combination of machine learning methods using lexical features, sentiment lexical resources (both generic and specific to finance) and pre-trained word embedding models. Novel features specific to the task included the creation of a domain-specific ontology (Schouten et al., 2017), a stocktwits-based embedding model and distance supervision model (Li, 2017) and domain-specific lexica (Moore and Rayson, 2017). Moreover, due to the emphasis of the task on the sentiment classification on a continuous scale, many approaches targeted regression-based models.

With the exception of Cabanski et al. (2017), few approaches explicitly tackled the problem of compositionality (Sales et al., 2016), valency shifting (Malo et al., 2014a), and clausal disembedding (Niklaus et al., 2016), a fact that is reflected by the lack of submissions which explored syntactic features.

With regard to language transportability, most approaches have a medium level of transportability, being dependent on the translation of the sentiment lexica, but not depending on syntactic parser.

Important specific aspects proposed by the task remained unexplored or poorly explored, including: (i) the use of quantitative background knowledge (e.g. stock price, financial report data), (ii) the use of the annotated text spans.

6 Alternative Evaluation Metric

Based on the evaluation metric as stated in Section 4, another evaluation metric has been developed during the competition. The intention to propose a modified way of evaluation was based on the fact that the cosine similarity (1) is treating all predicted scores with the same weight. This approach is not exploiting all information given in the data set, in specific it is not taking the link between entities and instances into consideration.

6.1 First Modification

Therefore, we proposed an approach which is using multiple vectors (one per instance) instead of only two. These instance vectors are containing one score per corresponding entity. Cosine similarity scores are calculated for each instance and added up to then divide the sum of all similarity scores by the number of submitted instance predictions in order to retrieve an average cosine similarity score.

While considering this modified evaluation metric a drawback, dividing the predictions on one hand into a regression problem but on the other hand into a classification problem, was noticed. The cosine similarity (1) for vectors with a length of 1 is resulting in either +1 or -1. However, the cosine similarity for vectors with a length greater than 1 is resulting in a floating point value. Thus, another modification of the initial evaluation formula has been conducted.

6.2 Second Modification

The second modification of the initial evaluation metric is also using one vector per instance containing one score per corresponding entity. Those instance vectors are populated into either a gold standard (GS) or predicted system (PS) vector. As both vectors are populated according to matching instances and entities, both vectors should be the same length. In contrast to the first modification (6.1), the second modification is using two different methods of evaluating the given scores dependent on the length of a vector. For each instance vector in GS/PS which has a length of 1, the absolute distance between both scores (4) is added to the total similarity score (6).

$$s\_similarity(G, P) = 1 - |G_0 - P_0| \quad (4)$$

Tool	System
scikit-learn[18]	Nasim (2017), Moore and Rayson (2017), John and Vechtomova (2017), Cabanski et al. (2017), Kumar et al. (2017), Symeonidis et al. (2017), Kar et al. (2017), Jiang et al. (2017)
word2vec[19]	Li (2017),Ghosal et al. (2017), Kumar et al. (2017), Saleiro et al. (2017)
Weka[20]	Seyeditabari et al. (2017), Zini et al. (2017)
GloVe[21]	Seyeditabari et al. (2017), Mansar et al. (2017), Rotim et al. (2017), Pivovarova et al. (2017), Ghosal et al. (2017), Kumar et al. (2017)
LIBSVM[22]	Rotim et al. (2017)
LIBLINEAR[23]	Rotim et al. (2017), Jiang et al. (2017)
Keras[24]	Moore and Rayson (2017), Ghosal et al. (2017)
XGBoost[25]	John and Vechtomova (2017), Jiang et al. (2017)
gensim[26]	John and Vechtomova (2017), Cabanski et al. (2017)
TensorFlow[27]	John and Vechtomova (2017), Pivovarova et al. (2017), Cabanski et al. (2017)

Figure 8: Tools used by systems in both tracks

For each instance vector with a length greater than 1, the cosine similarity is "length times" added to the total similarity score (5).

$$m\_similarity(G, P) = |P| \times cosine(G, P) \quad (5)$$

$$total\_similarity(GS, PS) = \sum_{i=1}^{|PS|} \begin{cases} |PS_i| = 1 & s\_similarity(GS_i, PS_i) \\ |PS_i| > 1 & m\_similarity(GS_i, PS_i) \end{cases} \quad (6)$$

Once the similarity scores are calculated for each instance vector and added to the total_similarity, the final score is calculated by dividing the total similarity score by the number of predicted entities to then multiply the quotient with a weight which consists of the quotient of all predicted entities divided by all possible entity predictions (7). In contrast to the cosine weight as stated in (2), this weight is calculated on an entity level.

$$final\_score(GS, PS) = \frac{\sum_{i=1}^{|PS|} |PS_i|}{\sum_{i=1}^{|GS|} |GS_i|} \times \frac{total\_similarity(GS, PS)}{\sum_{i=1}^{|PS|} |PS_i|} \quad (7)$$

Similarity scores produced using this alternative evaluation metric can be found in the appendix A.

6.3 Pros and Cons

On one hand, the evaluation metric as stated in 6.2 differentiating between vectors according to their lengths avoids the regression/classification problem as described in 6.1. In addition, it is considering the link between instances and entities in the final score.

On the other hand, one disadvantage of this is approach is the linearity / non-linearity of the two sub-methods used ((4), (5)). One could argue that both sub-methods are not equally impacting the total score. Balancing would be one approach to reducing discrepancy but also be subjectively influenced.

7 Related Initiatives

A number of projects have addressed questions pertaining to Sentiment Analysis and Finance. The FIRST (2010-2013) FP7 European project [28] provides sentiment extraction and analysis of market participants from social media networks in near real-time, for detecting and predicting financial market events, such as insights about financial market movements and financial market abuse. The developed tool consists of a decision support model based on Web sentiment as found within textual data extracted from Twitter or blogs, for the financial domain.

The TrendMiner (2011-2014) FP7 European project [29], presents an innovative and portable open-source real-time method for cross-lingual mining and summarisation of large-scale social media streams, such as weblogs, Twitter, Facebook, etc. One high profile case study was a financial decision support (with analysts, traders, regulators and economists).

[28] http://project-first.eu/
[29] http://www.trendminer-project.eu/

StockWatcher (Micu et al., 2008) provides a customised, aggregated view of news categorised by different topics, where it performs sentiment analysis - positive, negative or neutral effect - on particular news messages about a particular company. This tool enables the extraction of relevant news items from RSS feeds concerning the NASDAQ-100 listed companies. The sentiment of the news messages directly affects a company's respective stock price.

Mirowski et al. (Mirowski et al., 2010) present an algorithm for topic modelling, text classification and retrieval from time-stamped documents. This algorithm has been applied to predict the stock market volatility using financial news from Bloomberg. The volatility considered is estimated from daily stock prices of a particular company.

Several data sets have been created which are relevant in the context of our current endeavour. (Sanders, 2011) provide the Sanders Twitter Sentiment corpus, consisting of 5513 tweets about four topics/companies (Apple, Google, Microsoft, Twitter). One annotator manually assigned a positive, negative, neutral or irrelevant annotation to each tweet, depending on the sentiment expressed towards the given topic (company). This can refer to any aspect of the company, e.g. the service at the Apple store or the features of the iPhone in the case of Apple Inc. The current proposal will instead focus on a much larger range of companies and evaluate them specifically with respect to their stock market value. Furthermore, sentiment scoring will be more fine-grained as it will consist of floating-point numbers in the range of -1 (very negative/bearish) and 1 (very positive/bullish), with 0 representing neutral sentiment.

(Malo et al., 2014b) present the Financial Phrase Bank, a resource containing around 5000 sentences from English-language news about companies listed on the Helsinki stock exchange. Annotations at the level of syntactic phrases assigned one of three sentiment classes (positive, negative, neutral), based on the expected influence on the stock price. Each phrase was scored by between five and eight annotators. In our proposed task, the sentiment was also assigned with a view to the stock price or market development. However, our annotation is more fine-grained, ranging on a scale from -1 to 1. Furthermore, we annotate at the target (stock or company entity) rather than the phrase-level.

Over the years, many shared tasks in SemEval have focused on Sentiment Analysis, exploring various angles within the field. A series of tasks have concentrated on Sentiment Analysis in Twitter (Nakov et al., 2013; Rosenthal et al., 2014; Rosenthal and Stoyanov, 2015). They have covered tasks such as Polarity Disambiguation, document- and topic-level Polarity Classification, and topic-based Sentiment Aggregation. These tasks targeted open domains, with topics being determined using Named Entity Recognition (e.g. celebrities, places, sports clubs). The sentiment was assigned on two-point (positive, negative), three-point (positive, negative, neutral) or five-point (strongly positive, weakly positive, neutral, weakly negative, strongly negative) scales. In contrast, our proposed task aims to detect fine-grained sentiment, a scoring company- and stock-level sentiment on a floating point scale between -1 and 1. Furthermore, the data in our proposed task focuses only on the financial domain and its particular semantic challenges.

Aspect-based Sentiment Analysis has also emerged in recent editions of SemEval (Pontiki et al., 2014, 2015). Depending on the subtask, entities and their aspects are provided to the participants or need to be identified. Sentiment for entity-aspect pairs is scored according to four categories: positive, negative, neutral and conflict. In terms of data, while the 2014 task focused on isolated sentences from customer reviews, the 2015 edition dealt with full reviews. Again, our proposed task differs in the assignment of fine-grained sentiment, in the short nature of the text instances and in terms of the domain.

8 Conclusions and Future Work

We presented a new task on fine-grained sentiment analysis for the financial domain, where a sentiment in range (-1, 1) is assigned to entities. In our two subtasks, we focussed on two distinct data sources: financial microblogs (Twitter and StockTwits), where the target entities are company stock symbols ("cashtags"), and financial news headlines, where sentiment needs to be assigned to companies.

Deep Learning (word embeddings) and more traditional Machine Learning techniques account for the majority of contributions. Many participants made use of sentiment lexica, both finance-specific (e.g. the word lists from (Loughran and McDonald, 2011b)) and general domain (e.g. (Hu and Liu, 2004; Wilson et al., 2009)), as well as custom lexica created in the context of this task. A review of the results obtained by participants shows that three of the systems that performed best (top three in each

track) adopted a Hybrid (Deep Learning, Lexicon) technique, while the other three used a Machine Learning-based approach.

For a future edition of this task, we will focus on enhancing the evaluation metric in the light of the discussion in Section 6. It would be interesting to add subtasks with different sources, perhaps broadening the scope to include longer texts, such as full news articles from financial newspapers, or Facebook posts.

Acknowledgements

Horizon 2020 ICT Program Project SSIX: Social Sentiment analysis financial IndeXes, has received funding from the European Union's Horizon 2020 Research and Innovation Program ICT 2014 - Information and Communications Technologies under grant agreement No. 645425.

References

Johan Bollen, et al. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2(1):1–8.

Tobias Cabanski, et al. 2017. Hhu at semeval-2017 task 5: Fine-grained sentiment analysis on financial data using machine learning methods. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Chung-Chi Chen, et al. 2017. Nlg301 at 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, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Brian Davis, et al. 2016. Social sentiment indices powered by x-scores. In *2nd International Conference on Big Data, Small Data, Linked Data and Open Data, ALLDATA 2016*.

Marjan Van de Kauter, et al. 2015. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications* 42:4999–5010.

Angel Deborah, et al. 2017. Ssn_mlrg1 at semeval-2017 task 5: Fine-grained sentiment analysis using multiple kernel gaussian process regression model. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Eagle Alpha. 2016. Sentiment analysis in the financial domain. does it work? https://medium.com/eagle-alpha/sentiment-analysis-in-the-financial-domain-does-it-work-36fa974ea3cb#.o793xdr9h. Accessed 23-March-2016.

Deepanway Ghosal, et al. 2017. Iitp at semeval-2017 task 5: An ensemble of deep learning and feature based models for financial sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Aniruddha Ghosh, et al. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. pages 470–478.

Rohitha Goonatilake and Susantha Herath. 2007. The volatility of the stock market and news. *International Research Journal of Finance and Economics* 3(11):53–65.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 168–177.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Mengxiao Jiang, et al. 2017. Ecnu at semeval-2017 task 5: An ensemble of regression algorithms with effective features for fine-grained sentiment analysis in financial domain. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Vineet John and Olga Vechtomova. 2017. Uw-finsent at 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, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Sudipta Kar, et al. 2017. Ritual-uh at semeval-2017 task 5: Sentiment analysis on financial data using neural networks. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Svetlana Kiritchenko, et al. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research* 50:723–762.

Abhishek Kumar, et al. 2017. Iitpb at semeval-2017 task 5: Sentiment prediction in financial text. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Quanzhi Li. 2017. funsentiment at semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs using different word embeddings and target contexts. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5(1):1–167.

Tim Loughran and Bill McDonald. 2011a. "when is a liability not a liability? textual analysis, dictionaries, and 10-ks". *The Journal of Finance* 66(1):35–65.

Tim Loughran and Bill McDonald. 2011b. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance* 66(1):35–65.

Pekka Malo, et al. 2014a. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology* 65(4):782–796.

Pekka Malo, et al. 2014b. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology* 65(4):782–796.

Youness Mansar, et al. 2017. Fortia-fbk at semeval-2017 task 5:bullish or bearish? inferring sentiment towards brands from financial news headlines. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Alex Micu, et al. 2008. Financial news analysis using a semantic web approach.

Piotr Mirowski, et al. 2010. Dynamic auto-encoders for semantic indexing. In *NIPS 2010 Workshop on Deep Learning*.

Andrew Moore and Paul Rayson. 2017. Lancaster a at semeval-2017 task 5: Evaluation metrics matter: predicting sentiment from financial news headlines. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Preslav Nakov, et al. 2013. Semeval-2013 task 2: Sentiment analysis in twitter .

Zarmeen Nasim. 2017. Iba-sys at 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, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Christina Niklaus, et al. 2016. A sentence simplification system for improving relation extraction. In *26th International Conference on Computational Linguistics*.

Lidia Pivovarova, et al. 2017. Hcs at semeval-2017 task 5: Polarity detection in business news using convolutional neural networks. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Lidia Pivovarova, et al. 2013. Event representation across genres. In *NAACL HLT*. volume 2013, page 29.

Maria Pontiki, et al. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado*. pages 486–495.

Maria Pontiki, et al. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*. pages 27–35.

Sara Rosenthal, et al. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. pages 73–80.

Sara Rosenthal and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. *Proceedings of SemEval-2015* .

Leon Rotim, et al. 2017. Takelab at semeval-2017 task 5: Linear aggregation of word embeddings for fine-grained sentiment analysis of financial news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Pedro Saleiro, et al. 2017. Feup at semeval-2017 task 5: Predicting sentiment polarity and intensity with financial word embeddings. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Juan Efson Sales, et al. 2016. A compositional-distributional semantic model for searching complex entity categories. In *5th Joint Conference on Lexical and Computational Semantics (*SEM)*.

Niek J. Sanders. 2011. Sanders-twitter sentiment corpus. `http://www.sananalytics.com/lab/twitter-sentiment`. Accessed 30-March-2016.

Kim Schouten, et al. 2017. Commit at semeval-2017 task 5: Ontology-based method for sentiment analysis of financial headlines. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Thomas Schuster. 2003. Meta-communication and market dynamics. reflexive interactions of financial markets and the mass media .

Armin Seyeditabari, et al. 2017. Sentiheros at semeval-2017 task 5: An application of sentiment analysis on financial tweets. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Nitish Sinha. 2014. Using big data in finance: Example of sentiment-extraction from news articles. http://www.federalreserve.gov/econresdata/notes/feds-notes/2014/using-big-data-in-finance-example-of-sentiment-extraction-from-news-articles-20140326.html. Accessed 29-March-2016.

Jacopo Staiano and Marco Guerini. 2014. Depechemood: A lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605* .

Symeon Symeonidis, et al. 2017. Duth at semeval-2017 task 5: Sentiment predictability in financial microblogging and news articles. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

Pyry Takala, et al. 2014. Gold-standard for topic-specific sentiment analysis of economic texts. In *LREC*. Citeseer, volume 2014, pages 2152–2157.

Paul C Tetlock, et al. 2008. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance* 63(3):1437–1467.

Theresa Wilson, et al. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics* 35(3):399–433.

Tian Tian Zhu, et al. 2013. Ecnucs: A surface information based system description of sentiment analysis in twitter in the semeval-2013 (task 2). *Atlanta, Georgia, USA* page 408.

Tiago Zini, et al. 2017. Inf-ufrgs at semeval-2017 task 5: A supervised identification of sentiment score in tweets and headlines. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. http://alt.qcri.org/semeval2017/.

A Supplemental Material

Similarity scores calculated using the evaluation metric as proposed in Section 6.2:

http://alt.qcri.org/semeval2017/task5/data/uploads/results/subtask1-microblogs-siscosim.pdf

http://alt.qcri.org/semeval2017/task5/data/uploads/results/subtask2-headlines-siscosim.pdf