FinSentiA: Sentiment Analysis in English Financial Microblogs

Thomas Gaillat (1), Annanda Sousa (1), Manel Zarrouk (1), Brian Davis (2)
(1) Insight Centre for Data Analytics, IDA Business Park, Ireland
(2) Maynooth University, Ireland
firstname.surname@insight-centre.org, brian.davis@mu.ie

RÉSUMÉ

FinSentiA: Analyse de Sentiments dans les Microblogs Financiers en Anglais

L'objectif de cet article est de présenter la construction d'un système d'analyse de sentiments dans le domaine des microblogs financiers en anglais. Le but de notre travail est de construire un classifieur pour la prédiction de sentiments chez les investisseurs financiers sur les plateformes de microblogs telles que StockTwits et Twitter. Notre contribution montre qu'il est possible de mener une analyse fine des sentiments. Après extraction des entités financières et leurs contextes, le système attribue des scores en valeurs continues. Il repose sur une approche par réseaux profonds pour la méthode de classification. Les résultats montrent un F1-score de 0.85 (2 classes) et une valeur de similarité cosinus de 0.62.

ABSTRACT

FinSentiA: Sentiment Analysis in English Financial Microblogs

The objective of this paper is to report on the building of a Sentiment Analysis (SA) system dedicated to financial microblogs in English. The purpose of our work is to build a financial classifier that predicts the sentiment of stock investors in microblog platforms such as StockTwits and Twitter. Our contribution shows that it is possible to conduct such tasks in order to provide fine-grained SA of financial microblogs. We extracted financial entities with relevant contexts and assigned scores on a continuous scale by adopting a deep learning method for the classification. Results show a 0.85 F1-Score on a two-class basis and a 0.62 cosine similarity score.

MOTS-CLÉS : Analyse de Sentiments, Entités Financières, Valeurs Continues, Fouille de Données d'Opinions, Granularité Entité

KEYWORDS: Sentiment Analysis, Financial Entities, Continuous Scale, Opinion Mining, Entity level

1 Introduction

Stock investment platforms like StockTwits[1] provide their users with many indicators of live events occurring on markets. To anticipate volatility, seen as a risk, investors use indexes such as the VIX[2] (Volatility IndeX) which is interpreted as a measure of investor confidence with respect to S&P 500 stock index option prices (Blitzer, 2017). The VIX is calculated from numerical values which module investor sentiment but there is an increasing interest in capturing opinions on finance from social media as an alternative source of sentiment. In a highly competitive and volatile domain such as investing, 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 messages posted on social media such as Twitter or StockTwits are central to determining these economic signals.

Sentiment Analysis (SA) plays a central role for this purpose. Intersecting between the fields of Computational Linguistics and Natural Language Processing, the field has a long tradition in the use of tools to automatically determine sentiments in documents (Liu, 2012; Pang & Lee, 2008). It is admitted that SA tends to perform best when restricted to a specific domain and much work, so far, has focused on domains such as hotel and product reviews. In the financial domain some studies have used press articles with economic and financial focuses (Malo et al., 2013a; Malo et al., 2013b) but few studies were conducted on financial microblog messages (Bollen et al., 2011). And yet, due to the specific style of this type of messages, financial microblog SA requires adapted NLP methods. In addition, despite the need of financial experts to have a more revealing fine-grained analysis, most studies so far have concentrated on classifying sentiments in categories at sentence or document level. Recently, there have been some studies focusing on entities on a continuous scale (Cortis et al., 2017), but they were based on already-identified text spans which simplified the sentiment assignment task. Finally, many studies have used supervised learning approaches but, as far as we know, few have applied deep learning methods.

The purpose of our research is to develop a fine-grained financial SA classifier for the SSIX[3] project. In this paper, we present a supervised learning approach[4] which relies on a financial microblog Gold Standard for the training of a neural network. The SA is carried out in a fine-grained fashion: i) by extracting entities, i.e. stocks, along with relevant contexts ii) by assigning sentiment scores on a continuous polarity scale ranging from -1 to 1. The remaining of this paper is divided into four sections. In Section 2, we provide a short review of related work in the domain of financial SA. Section 3 covers the method used to build the financial classifier. In Section 4, we present the results followed by a short discussion. We conclude in Section 5.

---

1 See https://stocktwits.com

2 Cboe created the Volatility Index® (VIX® Index) which is a benchmark index to measure the market's expectation of future volatility.

3 Social Sentiment IndeX is a sentiment analysis platform dedicated to financial microblogs

4 This financial classifier is operational and available for users as it is implemented in the SSIX online platform. See https://ssix-project.eu/knowledgebase_category/demos/

2 Related work in financial sentiment analysis

Given the broad diversity of studies conducted in the domain of SA to date (see Liu, 2012; Pang & Lee, 2008 for a detailed overview of the domain) and due to the scope of the SSIX project, we choose to focus this review of related work to the financial domain only. There have been two types of approaches: rule-based and supervised learning approaches. Rule-based approaches consist in dictionary lookups for the assignment of polarities. Typically, the first step of the task consists in building a lexicon of terms from a domain-specific corpus together with their respective polarities (Loughran & McDonald, 2011; Moreno-Ortiz & Fernández-Cruz, 2015; Tetlock, 2007). The second step of the rule-based approach is to create rules for polarity lookups. (see (Loughran & McDonald, 2011; Malo et al., 2013a; Malo et al., 2013b; Tetlock, 2007; Wiebe et al. 2005) for variations in the approach).

The second type of approach in SA relies on supervised learning methods. When applied to financial texts, these Machine Learning (ML) methods make use of annotated documents that are used to "learn" specific financial features in order to subsequently classify. Some studies use common feature engineering strategies relying on internal text features such as bag-of-words, Part-of Speech (POS) tags and Named Entities (Antweiler & Frank, 2004; O'Hare et al., 2009; Schumaker et al., 2012; Sprenger et al. 2014). Other studies add polarities from lexicons as features of vectors (Bollen et al., 2011; Malo, et al., 2013b) and assign sentiment classes at sentence level. In terms of results, (Bollen et al., 2011) reported best results with a 0.869 three-class accuracy.

More recently, as part of the SSIX project (Davis et al., 2016), (Cortis et al., 2017) gave the opportunity for SemEval-2017 candidates to conduct experiments based on the data set used in our experiment. The best scores were obtained by combining features such as POS, Word embeddings (Mikolov et al., 2013) and financial lexicons in ML approaches including neural network architectures (Ghoshal et al., 2017; Jiang et al., 2017).

The presented approaches focus on SA at sentence/document level while our finer-grained approach targets entities on a continuous scale. We use a different implementation of cosine similarity that targets entities, and we do not use manually selected text spans for sentiment assignment.

3 Method

This section covers the method used to build the financial classifier. We describe the evaluation corpus before detailing the experimental setup.

3.1 Evaluation corpus

The corpus[5] is a data set that was made available for SemEval-2017 Task 5 (Cortis et al., 2017). It was specifically created for the purpose of financial SA in microblogs. It is a collection of 2,002 microblog messages in English from the Twitter and StockTwits platform. The messages were manually annotated to assign one sentiment score per financial entity (i.e. stocks) of each message.

[5] This data set was published as a Gold Standard and is publicly available from https://ssix-project.eu. For legal reasons, it only includes 1,336 StockTwits messages.

Table 1 shows the distribution of sentiments. To rate the sentiments on stocks—i.e, *cashtags* such as $AAPL for Apple—the scores were given on a continuous scale of [-1; 1] from *bearish*[6] to *bullish*[7]. To measure reliability, inter-rater agreement was calculated at entity level using Fleiss's Kappa and Krippendorff's alpha (0.69 and 0.61 respectively) prior to consolidating raters' scores. For more on this Gold Standard (GS) see (Gaillat, Zarrouk, Freitas, & Davis, 2018). For the experiment (Section 3.3), the corpus is divided into two subsets on a 80-20% basis. The training set accounts for 1,761 messages.

	Positive Sentiments	Negative Sentiments
Training set	1,931	934
Test set	393	197

TABLE 1: Distribution of positive and negative sentiments for entities in the training and test sets

3.2 Evaluation metrics

We use two evaluation metrics for the purpose of our study. The first and main evaluation metric, implemented in Semeval-2017 task 5 (Cortis et al., 2017), is the cosine similarity function which uses two vectors of values, i.e. the Gold Standard (GS) scores per entity and the predicted scores by the classifier. A score of 1 indicating perfect similarity. The second metric (not required in SemEval-2017) is provided as an additional measure. It is based on recoding continuous values into two positive and negative categories ([-1, 0[and [0, 1] intervals) the two possible polarities. We then conduct precision and recall as well as overall accuracy measurements per class. This evaluation method is implemented to provide a basis for comparison with other studies even though they may use a different number of classes. Calculating the cosine of the angle between the two vectors provides a measurement of the overall similarity between the two sets of values (Ghosh et al., 2015; Jurafsky & Martin, 2009). The metric gives results on a scale of [-1;1], with 1 indicating perfect match between the two lists of values (See Formula (1)). Let G be the vector of values from the evaluation corpus, and let P be the vector of predicted values.

$$(1) \quad cosine(G,P) = \frac{\sum_{i=0}^{n} (Gi.Pi)}{\sqrt{\sum_{i=0}^{n} Gi^2}.\sqrt{\sum_{i=0}^{n} Pi^2}}$$

3.3 Experimental set up

This section details the experimental set up involving linguistic preprocessing (message splitting and cleaning), feature engineering and deep learning classification.

6 *Bearish* indicates the belief that the price of an asset will fall.

7 *Bullish* is the opposite of *bearish*

3.3.1 Linguistic preprocessing

(Cortis et al., 2017) relied on the message spans that matched entities and scores, thus avoiding noisy data. However, in our case, we dealt with entire messages to mirror the reality of incoming data into the system. Consequently, it was necessary to apply a method in order to split sentences into segments that contain entities with relevant contexts. Example (1) shows that $IBM and $WYNN are target entities that are not adjacent and are separated contextually with different sentiments.

(1) "$IBM was THE play today *but* this $WYNN ain't bad either"

The splitting strategy is twofold. Firstly, an algorithm determines whether entities should be treated as one single target or independently on the basis of adjacency. Adjacent entities are treated as one and non-adjacent entities are treated separately. Secondly, another algorithm determines whether messages should be split and how. This algorithm uses the targets determined by the first algorithm. It essentially splits sentences if targets are separated by more than one word. As a result, the sentence in Example 1 is split in two segments at the *but* coordination conjunction. Sentences also undergo stemming along with stop-word and punctuation elimination in order to strip messages of noisy data. To create an abstract representation of each segment, feature engineering is applied next. The influence of the performance of this pre-processing was evaluated by conducting classification with and without it. Results with version 2.2 of the model show an improvement in cosine similarity scores from 0.57 to 0.62.

3.3.2 Feature engineering

At this point, the messages are converted into a machine-readable vector representation. Each split message goes through a vectorisation process in which words are turned into feature values. To build the vector space, we use different four types of features. Compared with (Jiang et al., 2017), their work presented a bigger selection of features (12 types), several of them are the same as those we use, e.g. Bag of Words and sentiment lexicons such as Afinn[8] and SentiWordNet[9]. We experimented with some features implemented in Jiang's model such as Word Embedding (Google W2V[10]), and more sentiment lexicons (Bing Liu opinion lexicon[11], General Inquirer lexicon[12], MPQA[13]). However, in our proposed model the best result of features combination is shown in Table 1. As regard (Ghosal et al., 2017), they only used features from Word Embedding models.

[8] Available at http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010
[9] Available at http://sentiwordnet.isti.cnr.it/
[10] Google word to vector at https://code.google.com/archive/p/word2vec/
[11] Available at https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
[12] Available at http://www.wjh.harvard.edu/~inquirer/homecat.htm
[13] Available at http://mpqa.cs.pitt.edu/

Features	Description
Bag of words	After stemming, the vocabulary size is 1,006 dimensions. Feature binary values are embedded within a token representation of messages that keeps track of word order.
Sentiword Net lexicon	Feature values are extracted from SentiWordNet (Baccianella et al., 2010) in which words are assigned positive and negative polarity in a [0;1] interval. They are embedded in a token representation of each sentence to keep track of the order of sentiments in a sentence.
AFINN[14]	Feature values are extracted from a list of 2,477 words and phrases with polarity information ranging from -5 to 5 (Nielsen, 2011)
Vader predicted sentiment	Feature value provided by Vader (Hutto & Gilbert, 2014), a rule-based heuristic relying on lexicon and aggregation for scoring.

TABLE 2: Lexicon and linguistic features used for classification

3.3.3 Deep learning classification

The classification process is run through a neural network using a deep learning model with recurrent LSTM and dense layers. These layers have 5 and 7 hidden layers respectively. This number of layers is the result of a layer incremental process to maximise results. (Figure 1 shows the neural network structure.

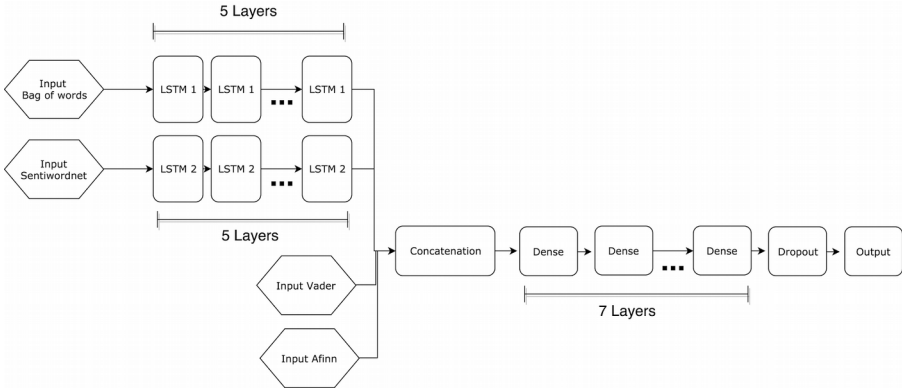

FIGURE 1: The structure of the neural network used in the financial classifier

The process is divided into two stages. Firstly, bag of words and Sentiwordnet (Baccianella et al., 2010) feature vectors are used as input for two parallel sets of five LSTM layers. These reduce the initial feature representation to optimize the number of dimensions by keeping the most significant features. The LSTM layers result in two vectors which are subsequently concatenated with the other feature vectors built with Vader (Hutto and Gilbert 2014) and Afinn (Nielsen, 2011). Secondly, the resulting concatenation is used as input for seven dense layers. A 25% dropout layer (Srivastava et al., 2014) is added because it is an efficient approach to prevent dataset overfitting. The output is a sentiment value included in the [-1;1] interval. Our neural network structure is similar to (Ghosal et

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

al., 2017) structure on the matter of having a set of LSTM with Dense layers. But their model differs from ours regarding the number (smaller) of layers and nodes from both the LSTM and Dense layers. In addition, they created an ensemble of neural networks to produce their final classifier model.

4 Results and discussion

Figure 2 shows the distribution of the predicted (right) and the test set (left) of sentiment scores at entity level. We can see clearly that the scores predicted by the classifier have the same distribution pattern as the ones from the test set from the Gold Standard. We notice, though, that there is a lack of predicted scores in the intervals [0.5, 1] and [-0.5, -1]. This can be explained by the inability of our model to learn from the training set due to the very small number of scores in these intervals in the GS. This can be seen in the scatter plot of the training set occurrences (Figure 3).

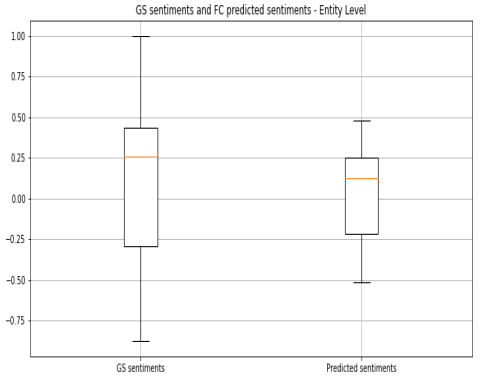

FIGURE 2: Distribution of the predicted and GS sets of sentiment scores at entity level

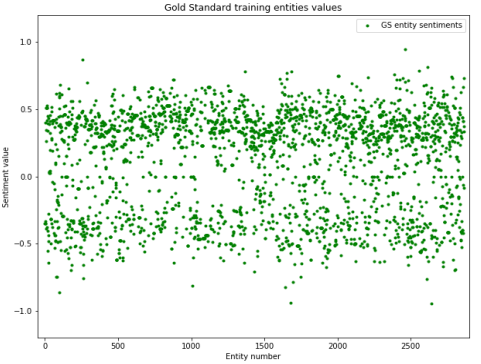

FIGURE 3: Dispersion of sentiment scores in the training set

As the closest existing work to ours is (Cortis et al. 2017), we computed our system performance using the exact same evaluation method they used. We decided not to use the predefined manually selected spans they use. Instead, we kept the real world configuration, which means dealing with entire messages, as they are, even if that implies adding a preprocessing step and dealing with the noise. Our system had 0.713 versus 0.733 for (Jiang et al. 2017).

As their evaluation method calculate the average of the cosine similarities at message level and not at entity level (both for the predicted and the test set), we decided to modify the method (see section 3.2) for it to take into consideration the entity scores distinctively, which keeps the fine-grained characteristic.

As we cannot compare totally our work to theirs we decided to have an internal baseline. We built a baseline model with Word2vec and SentiWordNet as features run through a 1-hidden-layer neural network. The cosine similarity function gives results taking into consideration the continuous values of the scores (see Table 3 for comparisons with the baseline model).

One point of discussion is about the cosine similarity measure which is the most used in the sentiment analysis domain. A more suitable evaluation metric that takes into consideration not only

the semantic similarity but also the sentiment similarity (Mohtarami et al. 2012) as not only the semantic space is important but also the emotional space.

To add another perspective to our results we computed as well the Accuracy and the F1-Scores of our model on a negative and positive classes basis ( ]-1,0] and]0,1]). Our main evaluation metric still the one described above (section 3.2) as it maintains the fine grained aspect of our work.

Entity Level	Financial Classifier version 1.0 (Baseline)	Financial Classifier version 2.2
Cosine Similarity	0.3352	**0.6269**
Accuracy	0.7801	0.8028
F1-Scores	0.8098	0.8522

TABLE 3: Accuracy and cosine similarity between predicted scores and Gold Standard results for the classification of financial entities in the Gold Standard

One point of discussion is about the cosine similarity measure. It may be relevant for semantic similarity but, in the case of sentiment similarity, it is not only the proximity of the vectors that is important but also the polarities of the values. Two vectors may be close to each other and yet have opposite values. This distinction is important in the measurement of sentiments.

5 Conclusion

Building a financial classifier for SA in the domain of financial microblogs faces many challenges involving linguistic engineering and machine learning tasks. Our contribution shows that it is possible to conduct such tasks in order to provide fine-grained SA of financial microblogs. We extracted financial entities with relevant contexts and assigned scores on a continuous scale by adopting a deep learning method for the classification. Further steps involve exploring a more suitable evaluation metric to our case, broadening the classifier features, increasing the training sets and applying some sensitivity analysis.

References

ANTWEILER, W., & FRANK, M. Z. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance*, *59*(3), 1259–1294.

BACCIANELLA, S., ESULI, A., & SEBASTIANI, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Presented at the Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta: European Language Resources Association (ELRA).

BLITZER, D. M. (2017). *S&P 500® - S&P Dow Jones Indices - Index Methodology*. New York, USA: S&P Dow Jones Indices.

BOLLEN, J., MAO, H., & ZENG, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

CORTIS, K., FREITAS, A., DAUDERT, T., HUERLIMANN, M., ZARROUK, M., HANDSCHUH, S., & DAVIS, B. (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)* (pp. 519–535). Vancouver, Canada: ACL.

DAVIS, B., CORTIS, K., VASILIU, L., KOUMPIS, A., MCDERMOTT, R., & HANDSCHUH, S. (2016). Social Sentiment Indices Powered by X-Scores. In *ALLDATA 2016 , The Second International Conference on Big Data, Small Data, Linked Data and Open Data*. Lisbon, Portugal: International Academy, Research, and Industry Association (IARIA).

GAILLAT, T., ZARROUK, M., FREITAS, A., & DAVIS, B. (2018). The SSIX Corpus: A Trilingual Gold Standard Corpus for Sentiment Analysis in Financial Microblogs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: ELRA.

GHOSH, A., LI, G., VEALE, T., ROSSO, P., SHUTOVA, E., REYES, A., & BARNDEN, J. (2015). SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*. Denver, USA: Association for Computational Linguistics.

GHOSHAL, D., BHATNAGAR, S., EKBAL, A., SHAD AKHTAR, M., & BHATTACHARYYA, P. (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)*. Vancouver, Canada: Association for Computational Linguistics.

HUTTO, C. J., & GILBERT, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Eighth International AAAI Conference on Weblogs and Social Media*. Ann Arbor, Michigan, USA: University of Michigan.

JIANG, M., LAN, M., & WU, Y. (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)* (pp. 888–893). Vancouver, Canada: Association for Computational Linguistics.

JURAFSKY, D., & MARTIN, J. H. (2009). *Speech and Language Processing (2nd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

LIU, B. (2012). *Sentiment Analysis and Opinion Mining*. San Rafael, Calif.: Morgan & Claypool Publishers.

LOUGHRAN, T., & MCDONALD, B. (2011). When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, *1*(66), 35–66.

MALO, P., SINHA, A., TAKALA, P., AHLGREN, O., & LAPPALAINEN, I. (2013). Learning the Roles of Directional Expressions and Domain Concepts in Financial News Analysis. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops* (pp. 945–954). Washington, DC, USA: IEEE Computer Society.

MALO, P., SINHA, A., TAKALA, P., KORHONEN, P., & WALLENIUS, J. (2013). Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. *Journal of the Association for Information Science and Technology*, *65*(4), 782–796.

MOHTARAMI, MITRA, HADI AMIRI, MAN LAN, THANH PHU TRAN, AND CHEW LIM TAN. (2012). Sense Sentiment Similarity: An Analysis. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 1706–1712. AAAI'12. Toronto, Ontario, Canada: AAAI Press.

MORENO-ORTIZ, A., & FERNÁNDEZ-CRUZ, J. (2015). Identifying Polarity in Financial Texts for Sentiment Analysis: A Corpus-based Approach. *Procedia - Social and Behavioral Sciences*, *198*, 330–338.

NIELSEN, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages* (Vol. 718, pp. 93–98). Heraklion, Greece: CEUR Workshop Proceedings.

O'HARE, N., DAVY, M., BERMINGHAM, A., FERGUSON, P., SHERIDAN, P., GURRIN, C., & SMEATON, A. F. (2009). Topic-dependent Sentiment Analysis of Financial Blogs. In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion* (pp. 9–16). New York, USA: ACM.

PANG, B., & LEE, L. (2008). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, *2*(1–2), 1–135.

SCHUMAKER, R. P., ZHANG, Y., HUANG, C.-N., & CHEN, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, *53*(3), 458–464.

SPRENGER, T. O., TUMASJAN, A., SANDNER, P. G., & WELPE, I. M. (2014). Tweets and Trades: the Information Content of Stock Microblogs. *European Financial Management*, *20*(5), 926–957.

SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I., & SALAKHUTDINOV, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

TETLOCK, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, *62*(3), 1139–1168.

WIEBE, J., WILSON, T., & CARDIE, C. (2005). Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, *39*(2–3), 165–210.