# IITPB at SemEval-2017 Task 5: Sentiment Prediction in Financial Text

**Abhishek Kumar[a], Abhishek Sethi[b], Md Shad Akhtar[a], Asif Ekbal[a]**
**Chris Biemann[c], Pushpak Bhattacharyya[a,b]**
[a]Indian Institute of Technology Patna, India
[b]Indian Institute of Technology Bombay, India
[c]Universität Hamburg, Germany
{abhishek.ee14,shad.pcs15,asif,pb}@iitp.ac.in
{abhisethi,pb}@cse.iitb.ac.in
biemann@informatik.uni-hamburg.de

## Abstract

This paper reports team IITPB's participation in the SemEval 2017 Task 5 on 'Fine-grained sentiment analysis on financial microblogs and news'. We developed 2 systems for the two tracks. One system is based on an ensemble of Support Vector Classifier and Logistic Regression. This system relis on Distributional Thesaurus (DT), word embeddings and lexicon features to predict a continuous sentiment value between -1 and +1. The other system is based on Support Vector Regression using word embeddings, lexicon features, and PMI scores as features. Our systems are ranked $5^{th}$ in track 1 and $8^{th}$ in track 2.

## 1 Introduction

We are living in a world where stock market directly affects the economic system of a country. Therefore, a reliable and prompt delivery of information plays an important role in the financial market. Up until the last decade printed/television news were the major source of stock market-related information. However, with the introduction of micro-blogging websites (e.g. Twitter etc.) the trend has been shifted. The rise of Twitter and StockTwits has given the people and organizations an opportunity to vent out their feelings and views. This information can be used by an individual or an organization to make an informed prediction related to any company or stock (Si et al., 2013). This opens a new avenue for sentiment analysis in the financial domain of microblogs and news.

News headlines are a short piece of text describing the nature of an article. Due to space constraints, headlines normally follow a compact writing style, known as *headlinese*, which limits the usage of articles, the verb form of to be, conjunctions etc.

Similarly, social media platforms text is prone to noise. There is a very high possibility of the data lacking a proper structure, grammar and appropriate punctuations. These inconsistencies make it challenging to solve any NLP problems including sentiment analysis (Khanarian and Alwarez-Melis, 2012). Moreover, each tweet can have reference to multiple company names (or stock symbols) and the expressed sentiment can be different towards different companies. Hence, there is a need to perform fine-grained sentiment analysis wherein, generally, a context is used to decide the relevant portion of a tweet for a particular company. Another inherent challenge with the microblog and news data is the use of short languages, hashtag, emoticons and embedded URL. Special attention should be given to these as they can provide some important hidden information (Mohammad et al., 2013). Example - *#bullishMarket* and *#increasingProfit* can reflect *positive* sentiment. These are some of the major challenges associated with fine-grained sentiment analysis of microblogging and news data.

The SemEval-2017 task 5 (Fine-Grained Sentiment Analysis on Financial Microblogs and News) has two tracks (Cortis et al., 2017). For both the tracks, the overall aim was to assign a sentiment score to a cashtag/company over a continuous range of -1 (very negative/bearish) to 1 (very positive/bullish).

First track involves finding a sentiment score towards a given 'cashtag' (stock symbol preceded by a $, e.g. $AAPL for Apple Inc.) in microblog messages while the second track involves finding a sentiment score towards a given company name in the news headlines.Instances in track 1 datasets also contain 'span'. It is the section of a tweet from where sentiment score should be derived.

| Track 1 | Microblogs |
|---|---|
| Message: | Putting on a little $F short, prevailing wisdom notwithstanding. |
| Score: | -0.454 |
| Span: | Putting on a little $F short |
| Cashtag: | $F |

| Track 2 | News headlines |
|---|---|
| Message: | RBS and Barclays shares temporarily suspended amid heavy losses. |
| Score: | -0.941 |
| Company: | Royal Bank of Scotland Group |

Table 1: Instances of of microblog and news headline dataset.

We participated and submitted our system for both the tracks. A total of 27 and 29 teams participated in track 1 and track 2 respectively. Our system ranked $5^{th}$ in the first track with a cosine similarity of 0.725. In the second track, our system scored cosine similarity of 0.695 and ranked $8^{th}$ overall.

The rest of the paper is organized as follows: Section 2 briefly describes the proposed systems. Description of the feature set is given in Section 3. Section 4 is devoted to experimental result and error analysis. Lastly, we conclude in Section 5.

## 2 System Overview

In this section, we present a brief description of the proposed systems. We adopted a supervised approach for solving the problem of both the tasks. We employed Logistics Regression, Support Vector Machine (SVM) and Support Vector Regression (SVR) as the base classifier for the prediction. We tried various combinations of the feature set for training the model. Following this approach, we select a feature set that best suited for the problem at hand. To further improve the efficacy of the system we ensemble the outputs of various classifiers at the end. For ensemble, the final sentiment value was calculated by taking the harmonic mean of both the system's prediction and then, linearly scaling it in between -1 and +1.

### 2.1 Distributional Thesaurus

Missing words in word2vec or Glove vector representation makes it non-trivial to learn from the data. We employ Distributional Thesaurus (DT) (Biemann and Riedl, 2013) expansion strategy for those words whose representation was missing in word2vec or GloVe model. Distributional Thesaurus is an automatically computed word list which ranks words according to their semantic similarity. It finds words that tend to occur in similar contexts as the target word. We use a pre-trained DT model to expand a source word. If the representation of a word is not present in word2vec or GloVe model, then its corresponding most similar expanded word is used to replace it. If the replaced word does not have its corresponding representation also we select next similar word and so on. For a source word, we took top 5 similar words in the expanded list as targets. An example is listed in Table 2. For the source word 'drinks', its DT expanded word list contains 'beer', 'wine', 'coffee', 'liquids' and 'beverages'.

| Word | DT expanded list |
|---|---|
| drinks | beer, wines, coffee, liquids, beverages |
| price | prices, pricing, cash, cost, pennies |
| laptop | pc, computer, notebook, tablet, imac |

Table 2: Example of DT expansion

## 3 Feature set

We use following set of features for training the model.

### 3.1 Track 1 - Microblogs messages

- **Word Embedding:** Word embeddings are known to capture the syntactic and semantic similarity in a better and representative way. We used 200 dimensional twitter based pretrained GloVe vectors[1] for word representation. Averaging of words representation was done for calculating sentence embeddings.

- **Tf-Idf Score:** We use Tf-Idf score as a feature value in the work. The score reflects how important a word is to a document in a corpus.

- **Sentiment Lexicon:** We compiled a list of positive and negative words using NRC Hashtag Sentiment Lexicon (Kiritchenko et al., 2014), MPQA Subjectivity Lexicon (Wilson et al., 2009) and Bing Liu Opinion Lexicon (Hu and Liu, 2004). Using these we created hand-engineered features. $M_{pos}$ and

[1] http://nlp.stanford.edu/projects/glove/

$M_{neg}$ are the number of positive and negative words in span and text.

- **Agreement Score:** It is the agreement value of the positive and negative words in the data instance. This was calculated both for span or text. If we have all positive or all negative words then A = 1. We have modified the proposal in (Rao and Srivastava, 2012) to make the feature more effective.

$$A = 1 - \sqrt{1 - \left| \frac{M_{pos} - M_{neg}}{M_{pos} + M_{neg}} \right|}$$

- **Polar word occurrence:** We count the number of occurrences of all positive and negative words in the text and assign values +1, -1 and 0 if the difference between $M_{pos}$ & $M_{neg}$ are positive, negative and zero respectively.

### 3.2 Track 2 - News headlines

- **Word Ngrams:** We extracted and used unigrams and bigrams as features for this task.

- **Sentiment Lexicon:** Sentiment lexicons have been known to be a decisive feature in sentiment analysis tasks. We use the following four sentiment lexicons to get lexicon based features:

  - Bing Liu's Sentiment Lexicon (Hu and Liu, 2004)
  - Harvard General Inquirer (Stone et al., 1966)
  - SentiWordNet (Baccianella et al., 2010)
  - Loughran and McDocnald's Finance Lexicon (Loughran and McDonald, 2011)

  For each instance, we extract 3 features: positive score, negative score, and cumulative score. Each token is assigned a score of +1 or -1 if it belongs to positive or negative list respectively. We followed stated approach for all lexicons except SentiWordNet. In the case of SentiWordNet lexicon, we use the positive and negative score as given in the lexicon rather than +1 or -1.

- **Semantic Orientation (SO):** Semantic orientation (Hatzivassiloglou and McKeown, 1997) finds the association of a token with respect to its positivity and negativity. We calculate a score for each term in our training corpus to get the association value.

$$score(w) = PMI(w, pos) - PMI(w, neg)$$

  where PMI is point-wise mutual information and calculated as follows:

$$PMI(w, pos) = \log_2 \frac{freq(w, pos) * N}{freq(w) * freq(pos)}$$

  In the above equation *pos* is the collection of positive reviews and N is the total number of tokens in the corpus.

- **Word Embeddings:** We use the 300-dimensional pre-trained word2vec (Mikolov et al., 2013) vectors trained on part of Google News dataset (about 100 billion words). The sentence embedding is obtained by averaging the embedding vectors of all words in the sentence.

## 4 Experiments and Results

### 4.1 Dataset

The training datasets contains 1700 and 1142 instances of microblog messages and news headlines respectively. Test data comprises of 800 and 491 resp. of such instances for the two tracks. We use 20% of the training dataset as validation set.

### 4.2 Preprocessing

We used CMU ARK toolkit[2] for tokenization of microblog tweets. For preprocessing the text, each url, username and number was replaced by <*url*>, <*user*> and <*number*> respectively. Example - '*www.twitter.com*' by <*url*>, '*@johnSnow*' by <*user*> and '*9.7*' by <*number*>. Since the data was collected from the web all HTML entities were converted to their corresponding unicode characters e.g. '*&amp;*' to '*and*'. Datasets analysis suggests that few hashtags convey explicit sentiment in the text. Therefore, we replace hashtags by '*#*' followed by the associated word with the hashtag. For example - '*#happy*' by '*# happy*'. Lastly, all the characters are converted to lower case and for the news headline we use NLTK[3] for the tokenization.

[2] http://www.cs.cmu.edu/~ark/

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

### 4.3 Experiments

We used python based machine learning package scikit-learn[4] for the implementation. As classification algorithm, we used Logistic Regression (LR), Support Vector Machine (SVM) and Support Vector Regression (SVR). As discussed earlier, each instance of the dataset need a score over a continuous range of -1 to +1. Since SVM predicts discrete class labels, as post-processing we use the probability of predicted class as the score. During validation phase we observed that models trained on SVM work better than that of SVR for the microblog datasets. In contrast, SVR works better than SVM in news headline datasets. The hyperparameters of the SVM were $C = 30$ and $\gamma = 0.01$, for SVR we used $C = 10$ and $\gamma = 0.01$ and for LR we set $C = 6$. Cosine similarity of various combinations of the feature set is listed in Table 3 and 4 for microblogs and news headlines validation set respectively. For fine tuning of hyper-parameters, we did an exhaustive grid search evaluated through ten-fold cross-validation on the training set.

| Model | Cosine Similarity | | |
|---|---|---|---|
| | LR | SVM | SVR |
| W.E | 0.649 | 0.654 | 0.691 |
| Tf-Idf | 0.727 | 0.729 | 0.736 |
| W.E + Lexicon | 0.656 | 0.678 | 0.684 |
| W.E + Tf-Idf | 0.745 | 0.762 | 0.726 |
| Tf-Idf + Lexicon | 0.749 | 0.752 | 0.759 |
| **W.E + Tf-Idf + Lexicon** | 0.760 | **0.775** | 0.717 |

Table 3: Microblog messages: Cosine similarity on validation set.

| Model | Cosine similarity | | |
|---|---|---|---|
| | LR | SVM | SVR |
| Unigrams | 0.507 | 0.58 | 0.566 |
| Unigrams + Lexicon | 0.598 | 0.609 | 0.640 |
| (Uni+Bi)grams + Lexicon | 0.603 | 0.609 | 0.648 |
| (Uni+Bi)grams + Lexicon + SO | 0.738 | 0.713 | 0.794 |
| Unigrams + Lexicon + SO | 0.736 | 0.713 | 0.789 |
| W.E | 0.619 | 0.584 | 0.673 |
| W.E + Lexicon | 0.613 | 0.580 | 0.639 |
| **W.E + Lexicon + PMI** | 0.746 | 0.708 | **0.80** |

Table 4: News headline: Cosine similarity on validation set.

As a result, we observed that the word embedding along with lexicon based features produce the best cosine similarity for both the datasets. Further, we observed the output of different classifier are contrasting in nature, therefore we merge the outputs of different classifiers using averaging and harmonic mean. We found that harmonic mean of LR and SVM produces better cosine similarity score than other combinations for microblogs messages. However, for news headline performance did not improve on the ensemble, so we choose the best feature combination to train an SVR. Table 5 shows the results for harmonic mean of SVM and LR cosine similarities in microblogs datasets.

| Model | Cosine similarity |
|---|---|
| W.E | 0.687 |
| Tf-Idf | 0.733 |
| W.E + Lexicon | 0.697 |
| W.E + Tf-Idf | 0.768 |
| Tf-Idf + Lexicon | 0.755 |
| **W.E + Tf-Idf + Lexicon** | **0.778** |

Table 5: Microblog messages: Ensemble of SVM & LR on validation set.

After finalizing the proposed approach on validation set, we evaluated it on the test datasets. For microblogs messages we got the cosine similarity of 0.725. In news headline, our system produces cosine similarity of 0.695. Table 6 depicts evaluation results on test datasets.

| Datasets | Cosine similarity |
|---|---|
| Track 1: Microblogs | 0.725 |
| Track 2: News headlines | 0.695 |

Table 6: Cosine similarity on test dataset.

## 5 Conclusion

In this paper we proposed a supervised sentiment analyzer for financial texts as part of our participation in SemEval 2017 shared task. As base classification algorithm we used Logistic Regression (LR), Support Vector Machine (SVM) and Support Vector Regression (SVR) for predicting the sentiment score. In second stage we combine the predictions of two best performing models using harmonic mean. Evaluation shows encouraging results on the shared task dataset. In future we would like to explore other relevant features to improve the performance of the system.

[4] http://scikit-learn.org

## References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*. volume 10, pages 2200–2204.

Chris Biemann and Martin Riedl. 2013. Text: now in 2D! A framework for lexical expansion with contextual similarity. *J. Language Modelling* 1(1):55–95. https://doi.org/10.15398/jlm.v1i1.60.

Keith Cortis, Andre Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. *Proceedings of SemEval* .

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Madrid, Spain, pages 174–181. https://doi.org/10.3115/976909.979640.

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.

Michael Khanarian and David Alwarez-Melis. 2012. Sentiment classification in twitter: A comparison between domain adaptation and distant supervision. Technical report, CSAIL, MIT. Statistical NLP Final Project. http://people.csail.mit.edu/davidam/docs/ SentClassifTwitter.pdf.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *J. Artif. Intell. Res. (JAIR)* 50:723–762. https://doi.org/10.1613/jair.4272.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. Lake Tahoe, NV, USA, pages 3111–3119.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA, pages 321–327. http://www.aclweb.org/anthology/S13-2053.

Tushar Rao and Saket Srivastava. 2012. Analyzing stock market movements using twitter sentiment analysis. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, pages 119–123.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. 2013. Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, pages 24–29. http://www.aclweb.org/anthology/P13-2005.

Philip J. Stone, Dexter C. Dunphy, and Marshall S. Smith. 1966. The general inquirer: A computer approach to content analysis .

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics* 35(3):399–433. https://doi.org/10.1162/coli.08-012-R1-06-90.