

# IBA-Sys at SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News

**Zarmeen Nasim**

Institute of Business Administration (IBA)

Karachi, Pakistan

znasim@khi.iba.edu.pk

## Abstract

This paper presents the details of our system IBA-Sys that participated in SemEval Task: Fine-grained sentiment analysis on Financial Microblogs and News. Our system participated in both tracks. For microblogs track, a supervised learning approach was adopted and the regressor was trained using XgBoost regression algorithm on lexicon features. For news headlines track, an ensemble of regressors was used to predict sentiment score. One regressor was trained using TF-IDF features and another was trained using the n-gram features. The source code is available at Github <sup>1</sup>

## 1 Introduction

Sentiment Analysis has become a very active area of research during the last decade. The reason behind this rising popularity is twofold. First, sentiment analysis has a great number of applications varying from academia to commercial domains such as customer support, brand management, social media marketing e.t.c. Second, sentiment analysis involves a number of challenges such as handling unstructured and noisy text, anaphora resolution, context understanding and many others.

Sentiment Analysis now becomes an interesting area of research in the Financial domain also. Researchers have shown that the consumer opinions and sentiments have a profound impact on market dynamics [(Goonatilake and Herath, 2007),(Van de Kauter et al., 2015)]. This further leads to the research interest in predicting stock market from social media discussions and

<sup>1</sup><https://github.com/zarmeen92/IBA-Sys-SemEval-2017>

news text (Bollen et al., 2011). Earlier attempts of sentiment analysis in Financial domain includes the work of McDonald and Loughran (Loughran and McDonald, 2011) in 2011. They developed the list of words with associated sentiment polarities for classifying sentiment in financial text. The SemEval 2017 Fine-grained sentiment analysis on financial microblogs and news task (Cortis et al., 2017) aims at identifying bullish (optimistic) and bearish (pessimistic) sentiment associated with companies and stocks. This task involved two tracks. Track 1 included microblog messages and track 2 included the dataset of news statements and headlines. In both tracks, the task was to predict the sentiment score for a stock (in track 1) and for a company (in track 2) in a given instance of text. The challenging part was that the sentiment values are on a continuous scale between -1 (very negative) to +1 (very positive) rather than discrete labels.

IBA-Sys participated in both subtasks. For subtask 1, our system was trained to predict the sentiment score on the given microblog with relevance to the given cashtag. For subtask 2, our system was trained to predict the sentiment score on the given piece of headline with relevance to the given company name. Our system IBA-Sys participated in both tracks. In track 1, we were among top 5 teams whereas, in track 2, our system secured 14th position.

The remainder of this paper is organized as follows. Section 2 describes the datasets in detail. Section 3 presents the preprocessing steps applied to clean the dataset. Section 4 and Section 5 discusses the features and methodology used to build our system. Section 6 discusses experimental results and official submission. Finally, Section 7 concludes this paper.

Table 1: Statistic of Dataset

Subtask	Training Set	Test Set
Microblogs	1694	799
News Headlines	1142	491

## 2 Datasets

This section presents the details of the datasets provided by SemEval organizers.

### 2.1 Subtask 1 - Microblogs

The dataset provided by SemEval for fine-grained sentiment analysis on Microblogs comprises of microblog messages related to the Financial domain. Each message is annotated with the following information. Table 1 presents statistics of dataset provided by SemEval organizers.

1. Source: Source identifies the name of the platform where the message was posted. This contains either "Twitter" or "Stocktwits".
2. Id: Id provides a unique identifier of the message.
3. Cashtag: Cashtag provides the stock ticker symbol to which the span and sentiment are related.
4. Sentiment: Sentiment is a floating point value between -1 and 1 (very negative to very positive).
5. Spans: Spans contains piece of message expressing sentiment.

The data set contains 1694 microblog messages for training and 799 microblog messages for evaluation purpose.

### 2.2 Subtask 2 - News Headlines

The dataset provided for this subtask consisted of news headlines. Each message is annotated with the following information. Table 1 presents statistics of dataset provided by SemEval organizers.

1. Id: Id provides a unique identifier of the message.
2. title: Title contains the textual content of headline.
3. Sentiment: Sentiment is a floating point value between -1 and 1 (very negative to very positive).

4. Company: Company contains the name of a company to which the sentiment is related to.

The data set contains 1142 headlines for training and 491 headlines for evaluation purpose.

## 3 Preprocessing

Preprocessing is an important step in any natural language processing task. This section describes the preprocessing steps applied on the datasets.

### 3.1 Subtask 1 - Microblogs

Microblogs often contain noisy text such as special characters, URLs, punctuations e.t.c. Preprocessing is an important step applied in machine learning before proceeding to train phase. For preprocessing the actual microblog message, following tasks were performed.

1. Removal of special characters, punctuations and numbers.
2. Removal of URLs, user names mentioned in a tweet message.
3. Removal of words with length less than three in order to reduce the dimensionality of feature space.
4. Conversion of tweet text into lower case.
5. Concatenation of *spans* to form a unified string. For the empty *spans* field, we considered the whole preprocessed message text for feature extraction.

### 3.2 Subtask 2 - News Headlines

The textual content of news headlines contains the name of organizations. In the train and test datasets, the organization for which the sentiment needs to be extracted was given. However, it was found that often more than one organization name was mentioned in the headline content. Therefore, we applied named entity recognition to extract names of organizations that were included in the given headline. To extract the names of organizations we used NLTK Named Entity Recognition (NER) Tagger (Bird et al., 2009). After applying NER tagging, following steps were performed.

1. Removal of special characters and punctuations and numbers.

2. Removal of words with length less than three in order to reduce the dimensionality of feature space.
3. News text often contains important words beginning with capital letter. After applying NER tagging, words with Named Entity tags other than  $\{Person, Organization\}$  were converted to lower case.

## 4 Features

This section describes features used in training our system for predicting sentiment score on the given microblog message or news headline.

### 4.1 Subtask 1 - Microblogs

Following features were used in system training for subtask 1.

#### 4.1.1 Lexicon Features

We used sentiment lexicons constructed for the Financial domain to compute sentiment polarity score of the microblog message under consideration. Lexicons have been widely used for sentiment analysis. The use of domain-specific lexicon can greatly improve the performance of the system. We used following lexicons to compute lexicon based features.

1. **Loughran and McDonald Sentiment Word Lists**  
(Loughran and McDonald, 2011) identified that the sentiment lexicons constructed for other domains often misclassify words commonly used in financial blogs. They developed a list of positive and negative words used in the financial text. For modeling our system to predict sentiment score of microblog text, we used the word list constructed by (Loughran and McDonald, 2011). For each message, we compute a positive word count and negative word count. Positive word count refers to the number of positive words occurred in the message and negative word count refers to the number of negative words occurred in the message.
2. **Stock Market Lexicon**  
(Oliveira et al., 2016) created a lexicon using a large set of labeled messages from Stock-Twits. For each word with a Part of Speech (POS) tag in a lexicon, sentiment score in range  $-\infty$  to  $+\infty$  is determined in positive

context and negated context. In order to compute sentiment score of a microblog message using Stock Market Lexicon, a message is tagged with POS tags using NLTK POS tagger (Bird et al., 2009). Then for each word in a message, a positive and negative sentiment score was determined using the lexicon. The total positivity of a message was determined by the sum of positive scores of each word in a message and the total negativity of a message was determined using the sum of negative scores of each word in a message.

#### 4.1.2 Term Frequency - Inverse Document Frequency (TF-IDF)

TF-IDF feature determines the importance of a word to a document in a collection or corpus. TF-IDF assigns higher weights to words occurring less frequently in a corpus. This helps in reducing the importance of commonly used words. A matrix of TF-IDF features was computed using sklearn library (Pedregosa et al., 2011).

### 4.2 Subtask 2 - News Headlines

Following features were used in system training for subtask 2.

#### 4.2.1 Lexicon Features

For subtask 2, we used following lexicons to compute sentiment polarity scores.

1. **Loughran and McDonald Sentiment Word Lists**  
Lexicon scores using Loughran and McDonald Sentiment Word Lists were computed in a similar way as done for subtask 1.
2. **Harvard Inquirer Sentiment Lexicon**  
Harvard IV sentiment lexicon was used to determine the sentiment polarity of a given headline.
3. **NRC Hashtag Sentiment Lexicon** (Mohammad et al., 2013) constructed a list of words associated with a positive and negative sentiment score. Sentiment score is a real number, where values greater than zero indicates positive sentiment and values less than zero indicates negative sentiment. For each headline text, the polarity score was computed by summing the sentiment score of each word in the text.

### 4.2.2 Term Frequency - Inverse Document Frequency (TF-IDF)

TF-IDF feature determines the importance of a word to a document in a collection or corpus. A matrix of TF-IDF features was computed using sklearn library (Pedregosa et al., 2011).

### 4.2.3 N-gram Features

N-gram refers to the sequence of  $N$  words in the given text. In this paper, we used unigrams to learn a vocabulary from the given training set and then constructed a square matrix of size equal to the size of vocabulary. Each entry in a matrix represents the occurrence of a corresponding word in a given text.

## 5 Modeling

This section describes our approach for training system.

### 5.1 Subtask 1 - Microblogs

We trained our system on the provided training data using features described in Section 4. Since the task was determined the sentiment score as a real number ranging from -1 to +1, we trained our model using XGBoost Regression algorithm<sup>2</sup>. 3-fold cross-validation was also performed to tune XGBoost regression parameters.

### 5.2 Subtask 2 - News Headlines

For subtask 2, we used ensembling of two regressors trained on the different set of features. For model 1, we trained XgBoost Regressor on feature set including McDonald and Loughran Positive Word Count, McDonald and Loughran Negative Word Count, Sentiment score computed using NRC Hashtag sentiment lexicon, sentiment polarity score computed using Harvard IV sentiment lexicon and TF-IDF features.

Our second model was trained using XgBoost Regressor on features including same lexicon features as used in model 1 training and n-gram features. For predicting sentiment score on test data, we computed the average of the sentiment scores predicted by each of our models.

## 6 Results and Discussion

This section presents evaluation results of our system on subtask 1 and subtask 2. Evaluation of the

<sup>2</sup><https://github.com/dmlc/xgboost>

participating systems was based on cosine similarity metric. Cosine similarity was computed as follows,

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

where,  $G$  represents the vector of true sentiment polarity values and  $P$  represents the vector of predicted sentiment polarity values by the system.

Table 2 presents evaluation results of our official submission. Our system secured **4th** position in subtask 1 and **14th** position in subtask 2.

On subtask 2 which was related to News headlines dataset, our system did not perform well. The subtask2 was more challenging as compared to subtask 1. In subtask 1, participants are also given with the spans related to cashtag towards which the sentiment is expressed. Whereas, in subtask 2, spans were not given. It was quite challenging to identify the orientation of sentiment towards a company under consideration, in cases where more than one company is mentioned in the headline. We did not consider this issue while modeling our system and considered the whole text for extracting features.

Table 2: Official Results of System Evaluation

Subtask	Test Set	Cosine Similarity Score
Microblogs	799	0.655
News Headlines	491	0.547

## 7 Conclusion

This paper presented our approach to fine grained sentiment analysis on financial microblogs and news headlines SemEval Task 5. The task includes two subtasks including Sentiment analysis on Financial Microblogs and sentiment analysis on News Headlines. Our system was among top scorers for subtask 1. However, we did not performed well in subtask 2. In future, we can improve the system by further integrating dependency parsing to extract phrases from sentences. This will help in identifying different sentiments oriented towards specific companies with in the same text.

## References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing*

ing text with the natural language toolkit. ” O’Reilly Media, Inc.”.

- Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2(1):1–8.
- Keith Cortis, André Freitas, Tobias Dauert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 517–533. <http://www.aclweb.org/anthology/S17-2089>.
- 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.
- 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.
- Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*. Atlanta, Georgia, USA.
- Nuno Oliveira, Paulo Cortez, and Nelson Areal. 2016. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems* 85:62–73.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.
- Marjan Van de Kauter, Diane Breesch, and Véronique Hoste. 2015. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with applications* 42(11):4999–5010.