IITP at SemEval-2017 Task 5: An Ensemble of Deep Learning and Feature Based Models for Financial Sentiment Analysis

Deepanway Ghosal, Shobhit Bhatnagar, Md Shad Akhtar,
Asif Ekbal, Pushpak Bhattacharyya
Indian Institute of Technology Patna, India
{deepanway.me14,shobhit.ee14,shad.pcs15,asif,pb}@iitp.ac.in

Abstract

In this paper we propose an ensemble based model which combines state of the art deep learning sentiment analysis algorithms like Convolution Neural Network (CNN) and Long Short Term Memory (LSTM) along with feature based models to identify optimistic or pessimistic sentiments associated with companies and stocks in financial texts. We build our system to participate in a competition organized by Semantic Evaluation 2017 International Workshop. We combined predictions from various models using an artificial neural network to determine the opinion towards an entity in (a) Microblog Messages and (b) News Headlines data. Our models achieved a cosine similarity score of 0.751 and 0.697 for the above two tracks giving us the rank of 2nd and 7th best team respectively.

1 Introduction

Sentiment analysis of financial text is an important area of research. It has been shown that sentiments and opinions can affect market dynamics (Goonatilake and Herath, 2007). Social media has created a new world of venting customer voice. People tend to express their personal sentiment about the stock market through tweets. On the other hand, news presents the macroeconomic factors, company-specific or political information. Positive news tend to bring optimism and lift the market where as negative news effect the market in opposite direction (Van de Kauter et al., 2015). Sentiment analysis gives organizations the ability to observe the various social media sites in real time and then act accordingly. Twitter is considered to be an ocean of sentiment data.

A study indicates that sentiment analysis of public mood derived from Twitter feeds can be used to eventually forecast movements of individual stock prices (Smailović et al., 2014). All these evidences show us that financial sentiment analysis has a lot of untapped power and extensive research in the field can help us gain great insight about the financial market. The fundamental problem with classifying financial tweets is the presence of noise. The natural use of short, informal languages, emoticons, hashtag and sarcasm in tweets makes the sentiment analysis problem especially challenging.

News headlines usually use limited number of words to summarize the article. Moreover, aspects like language patterns, writing style, irony usage differs notably among different news categories and articles. Use of articles, verb form of 'to be', conjunction are very rare in practice.

In this paper we describe our proposed system as part of the 'SemEval-2017 Task 5 on Fine-Grained Sentiment Analysis for Financial Microblogs and News' (Cortis et al., 2017). We propose a multilayer perceptron (MLP) based ensemble method that leverages the combination of deep learning and feature based models for the prediction. Our system produces 4th and 8th best cosine similarity score for microblogs messages and news headline respectively. A total of 25 teams participated for the microblogs messages task while 29 teams submitted their systems for the news headline track.

The task defines sentiment score prediction in two separate tracks i.e. microblogs and news headlines. The objective of the task is to predict a sentiment score associated with a company/cashtag in the text. The sentiment score lies in a continuous range of -1(very bearish) to +1(very bullish). Cashtag refers to a stock symbol that uniquely identifies a company. For e.g. $AAPL represents

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 899–903,
Vancouver, Canada, August 3 - 4, 2017.

stock symbol for the company Apple Inc. Every instance of microblogs messages also include a span which indicates a part of text from where prediction should be derived.

This rest of the paper is organized as follows: Section 2 illustrates our system architecture in detail. We present our experimental results in Section 3. Finally, Section 4 presents our conclusions.

2 System Description

In this section we discuss our proposed system for the task. We developed a multi-layer perceptron (MLP) based ensemble approach which learns on top of a convolution neural network (CNN), a long short term memory network (LSTM), a vector averaging MLP and a feature driven MLP model. We separately train and tune all the models and then feed the prediction scores of each model as input to an MLP for ensembling. Training and tuning of this system is performed separately. The resultant pipeline is used to predict the final sentiment score.

2.1 Word Embeddings

Word embeddings are generally helpful in many natural processing tasks due to it's excellence in capturing hidden semantic structures. For word embeddings we used two pre-trained embedding models: GloVe[1] and Word2Vec[2]. For microblogs messages we used GloVe (Pennington et al., 2014) and Word2Vec (Godin et al., 2015) twitter model trained on 2 billion and 400 million tweets respectively. For news headline we used GloVe common crawl model trained on 802 billion words and Word2Vec Google News model (Mikolov et al., 2013). We experimented with 200, 300 and 400 dimension vectors and observed that 200 & 300 dimension vectors are the near-optimal case for microblogs messages and news headlines respectively. We have used concatenation of word embeddings to form sentence embeddings.

2.2 Convolutional Neural Network (CNN)

Convolutional neural network consists of one or more convolution and pooling layers followed by one or more dense layers. Our system uses 2 convolution layers followed by a max pool layer, 2 dense layers and an output layer. Size of convolution filters dictates the hidden features to be extracted. We employ 50 such filters while sliding over 1, 2, 3 and 4 word(s) at a time.

2.3 Long Short Term Memory (LSTM)

LSTMs are special kind of recurrent neural network which can efficiently learn long-term dependencies. We use two layers of LSTM on top of each other followed by 2 dense layers and a output layer. We fix number of neurons on each LSTM layers as 100. For the dense layer we use 50 and 10 neurons in the hidden layers.

2.4 Multilayer Perceptron (MLP) - Vector Averaging Model

Concatenation of word vectors for generating sentence embeddings often face the curse of high-dimensionality. In an attempt to get a constant low-dimensional feature vector we employ vector averaging technique for producing sentence vector. We perform an element wise averaging of the word vectors in a tokenized tweet/headline. We then use the sentence embeddings to train a 3-layered neural network for the prediction.

2.5 Multilayer Perceptron (MLP) - Feature Driven Model

This model is based on various lexical and semantic features. We trained a multilayer perceptron on top of the following features.

- **Character ngrams:** tf-idf weighted counts of continuous sequences of 2, 3, and 4 characters;
- **Word ngrams:** tf-idf weighted counts of continuous sequences of 1, 2, 3, and 4 words;
- **POS-tag**: parts of speech tags of each token in the text;
- **Lexicons:**
  - Following set of features are used for each of the four lexicons: Opinion Lexicon (Liu et al., 2005), Loughran and McDonald Sentiment Word Lists (Loughran and McDonald, 2011), MPQA Lexicon [+1.0 for strong positive, +0.5 for weak positive, similarly for negative] (Wilson et al., 2005) and Harvard's General Inquirer (Stone et al., 1962):
    - **positive count:** number of positive tokens in a tweet/title.

[1] http://nlp.stanford.edu/projects/glove/
[2] https://code.google.com/archive/p/word2vec/

    - ∗ **negative count:** number of negative tokens in a tweet/title.
    - ∗ **net count:** positive count - negative count in tweet/title.
  - In addition we use four NRC Lexicons: Hashtag Context, Hashtag Sentiment, Sentiment140, Sentiment140 Context (Svetlana Kiritchenko and Mohammad, 2014; Mohammad et al., 2013) for the microblogs messages. Following set of features are extracted for each of them:
    - ∗ positive count, negative count and net count.
    - ∗ sum of positive scores, negative scores and all scores.
    - ∗ maximum of positive and negative scores.

- **Pointwise Mutual Information (PMI):** We calculate a sentiment score for each term in our training corpus to get the association of each term with positive as well as negative sentiment.

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

  PMI is calculated as follows:-

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

  In the above equation $freq(w, pos)$ is the frequency of word w in positive text, $freq(pos)$ is the number of words in positive headlines and N is the total number of tokens in the corpus.

- **Microblog Specific Features:** We use following features only for microblogs messages track:
  - the number of words with all characters in upper case.
  - the number of favorite and retweet counts of a message (tweet).
  - the number of hashtags in the message.

The multilayer perceptron network has three hidden layers and one output layer consisting of 500, 50, 10 and 1 neurons respectively.

2.6 Ensemble Model

Ensemble of various systems is an effective technique to improve the overall performance by assisting each other. Ensembling usually reduces the generalization error, which in turn reduces overfitting. Here we discuss second stage of the our proposed system. We merge predicted sentiment scores of all four models (CNN, LSTM, Vector Averaging, Feature Driven) to create a new feature vector, and then fed it into a multilayer perceptron (MLP) network for training. Figure 1 shows, an overall schema of the proposed approach.

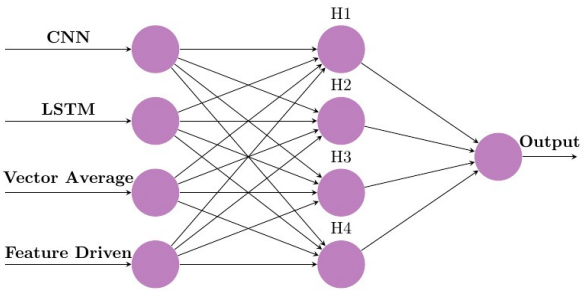

Figure 1: Ensembling Network Structure

3 Dataset, Experiments and Results

3.1 Datasets

The training datasets comprises of 1700 and 1142 instances of microblogs messages and news headlines respectively. We used the span in microblogs message track and the title in news headlines track as the textual feature for all our experiments described in this paper. For validation we did a 80:20, train:development split of the full datasets. The split was done such that the relative percentage of sources (twitter and stocktwits), mean and standard deviation of sentiment scores were same in the training and development data. We trained our model on the train data and selected models for ensembling, based on results on development data. Figure 2 and 3 shows the distribution of sentiment scores for the two datasets.

3.2 Experiments

We used python based neural network package Keras[3] for the implementation. We use ReLU activations for the intermediate layers and tanh activation for the final layer. Dropout (Srivastava et al., 2014) is a very effective regularization technique to prevent over-fitting of a network. It restrict convergence of weights to identical positions by randomly turning off the neurons during forward propagation. We use 15% dropout and '*Adam*' optimizer (Kingma and Ba, 2014) for regularization and optimization.

[3] www.keras.io

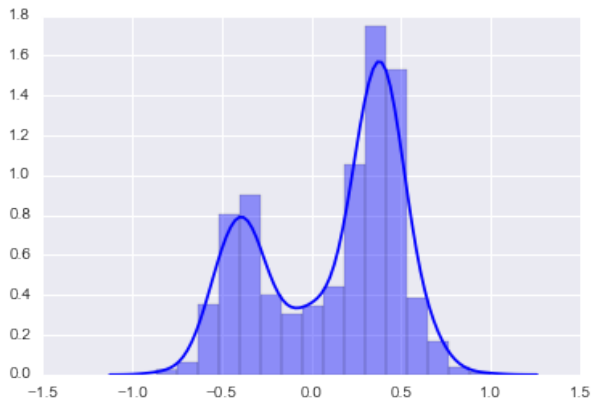

Figure 2: Histogram plot of sentiment scores in microblogs messages

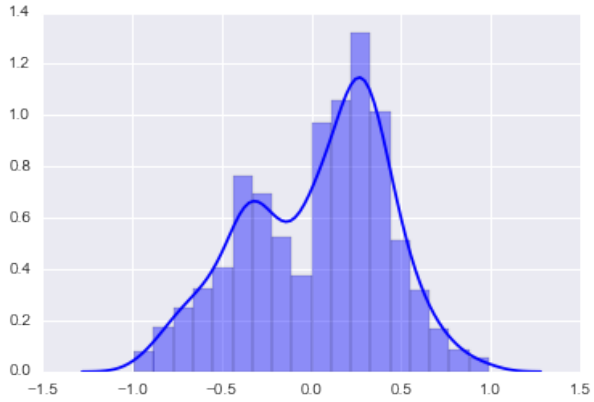

Figure 3: Histogram plot of sentiment scores in news headlines

We train and validate each model on 80% & 20% of the full data respectively. Table 1 shows our results of deep learning models (D), feature based model (F) and vector averaging models (V) on the validation set. It also depicts the results of our ensemble model (E) on the development set. It should be observed that use of ensemble improves the performance by a margin of 2-3%.

We submitted the E1 and E6 systems for the final evaluation and got a test cosine similarity score of 0.751 and 0.697 for microblogs messages and news headlines tracks respectively. Table 2 reports cosine similarity of our system.

4 Conclusion

In this paper we presented an MLP based ensemble technique for predicting the sentiment score. The proposed approach is a robust regression algorithm which predicts optimistic or pessimistic sentiments of associated stocks and companies

SNo	Models	Cosine Similarity	
		Microblogs	Headlines
D1	W2V CNN	0.752	0.670
D2	W2V LSTM	0.725	0.652
D3	GloVe CNN	0.768	0.649
D4	GloVe LSTM	0.765	0.644
F1	Feature Based	0.792	0.784
V1	W2V Average	0.804	0.663
V2	GloVe Average	0.781	0.643
E1	D3 + D4 + F1 + V1	**0.834**	-
E2	D3 + F1 + V1	0.826	-
E3	D4 + F1 + V1	0.821	-
E4	D3 + D4 + V1	0.812	-
E5	D3 + D4 + F1	0.799	-
E6	D1 + D2 + F1 + V1	-	**0.802**
E7	D1 + F1 + V1	-	0.795
E8	D2 + F1 + V1	-	0.788
E9	D1 + D2 + V1	-	0.683
E10	D1 + D2 + F1	-	0.791

Table 1: Cosine similarity score on validation set.

Tracks	Cos Sim	Rank
Track-1: Microblogs Messages	0.751	2nd
Track-2: News Headlines	0.697	7th

Table 2: Cosine similarity score on test dataset.

in financial text. We implemented a variety of semantic and linguistic features for our analysis of the noisy text such as tweets and news headlines. We combined predictions of four models (i.e. CNN, LSTM, Vector Averaging MLP and Feature Driven MLP) for calculation of final prediction. Our submission stood 2nd and 7th in two tracks that involves microblogs messages and news headlines respectively in SemEval 2017 shared task on 'Fine-Grained Sentiment Analysis of Financial Microblogs and News'.

References

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* .

Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia lab@ acl w-nut ner shared task: named entity recognition for twitter microposts using distributed word representations. *ACL-IJCNLP* 2015:146–153.

Rohitha Goonatilake and Susantha Herath. 2007. The volatility of the stock market and news. *Interna-*

tional Research Journal of Finance and Economics 3(11):53–65.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980. http://dblp.uni-trier.de/db/journals/corr/corr1412.html.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*. ACM, pages 342–351.

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*. 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 *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*. Atlanta, Georgia, USA.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. 2014. Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences* 285:181–203.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15:1929–1958. http://jmlr.org/papers/v15/srivastava14a.html.

Philip J Stone, Robert F Bales, J Zvi Namenwirth, and Daniel M Ogilvie. 1962. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science* 7(4):484–498.

Xiaodan Zhu Svetlana Kiritchenko and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts 50:723–762.

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.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, pages 347–354.