Knowledge-enriched Two-layered Attention Network for Sentiment Analysis

Abhishek Kumar[a], **Daisuke Kawahara**[b], **Sadao Kurohashi**[b]
[a]Indian Institute of Technology Patna, India
[b]Kyoto University, Japan
{abhishek.ee14}@iitp.ac.in
{dk,kuro}@i.kyoto-u.ac.jp

Abstract

We propose a novel two-layered attention network based on Bidirectional Long Short-Term Memory for sentiment analysis. The novel two-layered attention network takes advantage of the external knowledge bases to improve the sentiment prediction. It uses the Knowledge Graph Embedding generated using the WordNet. We build our model by combining the two-layered attention network with the supervised model based on Support Vector Regression using a Multilayer Perceptron network for sentiment analysis. We evaluate our model on the benchmark dataset of SemEval 2017 Task 5. Experimental results show that the proposed model surpasses the top system of SemEval 2017 Task 5. The model performs significantly better by improving the state-of-the-art system at SemEval 2017 Task 5 by 1.7 and 3.7 points for sub-tracks 1 and 2 respectively.

1 Introduction

With the rise of microblogging websites, people have access and option to reach to the large crowd using as few words as possible. Microblog and news headlines are one of the common ways to dispense information online. The dynamic nature of these texts can be effectively used in the financial domain to track and predict the stock prices (Goonatilake and Herath, 2007). These 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 gives rise to an interesting problem of sentiment analysis in financial domain. 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). An efficient system for sentiment analysis is a core component of a company involved in financial stock market price prediction.

Social media texts are prone to word shortening, exaggeration, lack of grammar and appropriate punctuations. Moreover, the word limit constraint forces a user to limit their content and squeeze in their opinion about companies. These inconsistencies make it challenging to solve any natural language processing tasks including sentiment analysis (Khanarian and Alwarez-Melis, 2012).

Bag-of-words and named entities were used by Schumaker and Chen (2009) for predicting stock market. For predicting the explicit and implicit sentiment in the financial text, de Kauter et al. (2015) used a fine-grained sentiment annotation scheme. Kumar et al. (2017) used a classical supervised approach based on Support Vector Regression for sentiment analysis in financial domain. Oliveira et al. (2013) relied on multiple regression models. Akhtar et al. (2017) used an ensemble of four different systems for predicting the sentiment. It used a combination of Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), Gated Recurrent Unit (GRU) (Cho et al., 2014), Convolutional Neural Network (CNN) (Kim, 2014) and Support Vector Regression (SVR) (Smola and Schölkopf, 2004). Yang et al. (2016) used a hierarchical attention network to build the document representation incrementally for document classification.

Our model focuses on interpretability and usage of knowledge bases. Knowledge bases have been recognized important for natural language understanding tasks (Minsky, 1986). Our main contribution is a two-layered attention network which utilizes background knowledge bases to build good word level representation at the primary level. The secondary attention mechanism works on top of the primary layer to build meaningful sentence representations. This provides a good intuitive working insight of the attention network.

Proceedings of NAACL-HLT 2018, pages 253–258
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

2 Proposed Methodology

We propose a two-layered attention network which leverages external knowledge for sentiment analysis. It consists of a bidirectional Long Short-Term Memory (BiLSTM) (Graves et al., 2005) based word encoder, word level attention mechanism for capturing the background knowledge and a sentence level attention mechanism aimed at grasping the context and the important words. The output of the two-layered attention network is then ensembled with the output of the feature based SVR using the Multilayer perceptron based approach described in Akhtar et al. (2017). The overall ensembled system is shown in Figure 2. Each of the components is explained in the following subsections and an overview of the two-layered attention network is depicted in Figure 1.

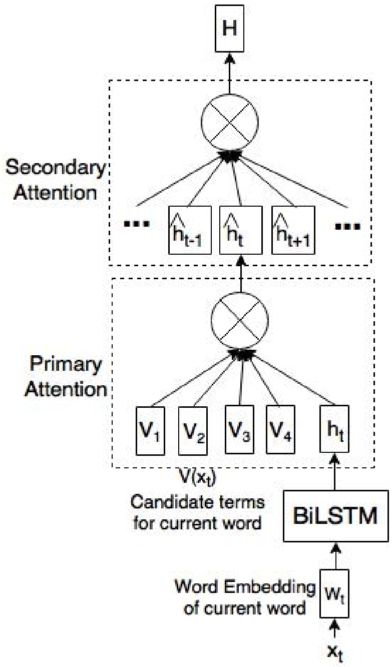

Figure 1: Two-layered attention network

2.1 Two Layered Attention Model

2.1.1 BiLSTM based word encoder

A Long-Short Term Memory (LSTM) is a special kind of Recurrent Neural Network. It handles the long-term dependencies where the current output is dependent on many prior inputs. BiLSTM, in essence, is a combination of two different LSTM - one working in forward and the other working in the backward direction. The contextual information about both past and future helps in determining the current output.

The two hidden states $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ for forward and backward LSTM are the information about past and future respectively at any time step t. Their concatenation h_t = $[\overrightarrow{h_t}, \overleftarrow{h_t}]$ provides complete information. Each word of the sentence is fed to the network in form of word embeddings which are encoded using the BiLSTM.

2.1.2 Word Level Attention

External knowledge in form of Knowledge Graph Embedding (Yang et al., 2015) or top-k similar words are captured by using the word level attention mechanism. This serves the purpose of primary attention which leverages the external knowledge to get the best representation for each word. At each time step we get V(x_t) relevant terms of each input x_t with v_i being the embedding for each term. (Relevant terms and embeddings are described in next section). The primary attention mechanism assigns an attention coefficient to each of relevant term having index $i \in$ V(x_t):

$$\alpha_{ti} \propto h_t^T W_v v_i \tag{1}$$

where W_v is a parameter matrix to be learned.

$$m_t = \sum_{i \in V(x_t)} \alpha_{ti} v_i \tag{2}$$

$$\widehat{h_t} = m_t + h_t \tag{3}$$

The knowledge aware vector (m_t) is calculated as Equation 2, which is concatenated with the hidden state vector to get the final vector representation for each word.

2.1.3 Sentence Level Attention

The secondary attention mechanism captures important words in a sentence with the help of context vectors. Each final vector representing the words is assigned a weight indicating its relative importance with respect to other words. The attention coefficient α_t for each final vector representation is calculated as:

$$\alpha_t \propto \widehat{h_t}^T W_s u_s \tag{4}$$

$$H = \sum_t \alpha_t \widehat{h_t} \tag{5}$$

where W_s is a parameter matrix and u_s is the context vector to be learned. H is finally fed to a one layer feed forward neural network.

2.2 Relevant Terms and Embeddings

External knowledge can provide explicit information for the model which the training data lacks. This helps the model to make better predictions. We relied on Knowledge Graph Embeddings based on WordNet and Distributional Thesaurus to get relevant terms and their corresponding embeddings for each word in the text.

2.2.1 Knowledge Graph Embedding

WordNet[1] is a lexical database which contains triplets in the form of (subject, relation, object). Both subject and object are synsets in WordNet. Each word in the text serves as the subject of the triplet. The relevant terms for the current word are the triplets having the current word as the subject. We then employ Knowledge Graph Embeddings to learn the representation of the triplet. A 100-dimensional dense vector representation for each subject, relation and object were learned using the DistMult approach (Yang et al., 2015) and concatenated. These served as the relevant embeddings. An example of triplet in WordNet is (*bronze_age*, *part_of*, *prehistory*).

2.2.2 Distributional Thesaurus

Distributional Thesaurus (DT) (Biemann and Riedl, 2013) is an automatically computed word list which ranks words according to their semantic similarity. We use a pre-trained DT to expand a current word. For each current word, top-4 target words are found which are the relevant terms. The relevant embeddings are obtained by using a 300-dimensional pre-trained Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) model. An example of the DT expansion of the word 'touchpad' is *mouse*, *trackball*, *joystick* and *trackpad*.

2.3 Feature Based Model - SVR

The following hand-crafted features are extracted and used to train a Support Vector Regression (SVR).

- Tf-Idf: Term frequency-inverse document frequency (Tf-Idf) reflects the importance of each word in a document. We use Tf-Idf score as a feature value for each word.

- Lexicon Features: Sentiment lexicons are an important resource for sentiment analysis. We employ the following lexicons: Bing Liu opinion lexicon (Ding et al., 2008) and MPQA subjectivity lexicon (Wilson et al., 2005), SentiWordNet (Baccianella et al., 2010) and Vader sentiment (Gilbert, 2014). From the above lexicons we extracted the agreement score (Rao and Srivastava, 2012) and the count of the number of occurrences of all positive and negative words in the text.

- Word embedding: We use the 300-dimensional pre-trained Word2Vec and GloVe embedding. The sentence embedding was obtained by concatenating the embedding for all words in the sentence.

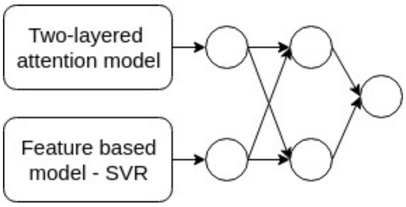

Figure 2: Multilayer perceptron based ensemble

3 Experiments

3.1 Dataset

We evaluate our proposed approach for sentiment analysis on the benchmark datasets of SemEval-2017 shared task 5. The task 'Fine-Grained Sentiment Analysis on Financial Microblogs and News' (Keith Cortis and Davis, 2017) had two sub-tracks. Track 1 - 'Microblog Messages' had 1,700 and 800 train and test instances respectively. Track 2 - 'News Statements & Headlines' had 1,142 and 491 train and test instances respectively. The task was to predict a regression score in between -1 and 1 indicating the sentiment with -1 being negative and +1 being positive.

3.2 Implementation Details

We implement our model using Tensorflow and Scikit-learn on a single GPU. We use a single layer BiLSTM with the two-layered attention mechanism followed by a one layer feed forward neural network. The number of units in each LSTM cell of the BiLSTM was 150. The batch size was 64 and the dropout was 0.3 (Srivastava et al., 2014) with the Adam (Kingma and Ba, 2014) optimizer. The length of context vector in the secondary attention network was 300. For each experiment, we report the average of five random runs. Cosine similarity is a measure of similarity. It represents the degree of agreement between the predicted and gold values. Cosine similarity was used for evaluation as per the guideline.

[1] https://wordnet.princeton.edu

3.3 Results

We compare our system with the state-of-the-art systems of SemEval 2017 Task 5 and the system proposed by Akhtar et al. (2017). Table 1 shows evaluation of our various models. Team ECNU (Lan et al., 2017) and Fortia-FBK (Mansar et al., 2017) were the top systems for sub-tracks 1 and 2 respectively. Team ECNU and Fortia-FBK reported a cosine similarity of 0.777 and 0.745 for sub-tracks 1 and 2 respectively. Team ECNU employed a number of systems - Support Vector Regression, XGBoost Regressor, AdaBoost Regressor and Bagging Regressor ensembled together. Team Fortia-FBK used a Convolutional Neural Network for this task. The system proposed by Akhtar et al. utilizes an ensemble of LSTM, GRU, CNN and a SVR and reported a cosine similarity of 0.797 and 0.786 for the two sub-tracks.

Our proposed system has a cosine similarity of 0.794 and 0.782 for sub-tracks 1 and 2 respectively. The proposed system performs significantly better than top systems of SemEval 2017 Task 5 for both the tasks. Moreover, the system performs at par with the system proposed by Akhtar et al. with half the number of subsystems involved in the ensemble. This shows that our proposed system is not only robust since it performs for both the task equally well but also powerful as it involves fewer subcomponents while having the same expressive power.

The two-layered attention network alone performs better than the best system of SemEval 2017 Task for both the sub-track. It manages to achieve much higher score than any of the deep learning component utilized by the system proposed by Akhtar et al. (2017) as shown in Table 2. This shows that the two-layered attention network helps to reduce overall model complexity without compromising the performance.

	Models	Microblog	News
Layered Attention Network			
L1	Knowledge Graph Embedding	0.758	0.727
L2	Distributional Thesaurus + GloVe	0.764	0.749
L3	Distributional Thesaurus + Word2Vec	0.779	0.763
Support Vector Regression			
S1	Tf-Idf + Lexicon	0.735	0.720
S2	Tf-Idf + Lexicon + GloVe	0.755	0.753
S3	Tf-Idf + Lexicon + Word2Vec	0.743	0.740
Ensemble			
E1	L3 + S2	0.794	0.782

Table 1: Cosine similarity score of various models on test dataset.

Models	Microblog	News
Single systems		
Mansar et al. (Team Fortia-FBK)	-	0.745
Akhtar et al. - LSTM	0.727	0.720
Akhtar et al. - GRU	0.721	0.721
Akhtar et al. - CNN	0.724	0.722
L3 (proposed)	0.779	0.763
Ensembled systems		
Lan et al. (Team ECNU)	0.777	0.710
Akhtar et al.	0.797	0.786
E1 (proposed)	0.794	0.782

Table 2: Comparison with the state-of-the-art systems.

3.4 Error Analysis

We performed error analysis and observed that the proposed system faces difficulty at times. Following are the situations when the system failed and incorrectly predicted values of the opposite polarity:

• Sometimes the system fails to identify an intensifier. In the example below, 'pure' is used as an intensifier.
Text : Pure garbage stock
Actual: -0.946 **Predicted:** 0.042

• The system fails when it does not have enough real-world information. In the example below, a low share price is a good opportunity to buy for an individual but from a company's point of view, a low share price does not indicate a prosperous situation.
Text : Good opportunity to buy
Actual: -0.771 **Predicted:** 0.260

4 Conclusion

In this paper, we proposed an ensemble of a novel two-layered attention network and a classical supervised Support Vector Regression for sentiment analysis. The two-layered attention network has an intuitive working. It builds the representation hierarchically from word to sentence level utilizing the knowledge bases. The proposed system performed remarkably well on the benchmark datasets of SemEval 2017 Task 5. It outperformed the existing top systems for both the sub-tracks comfortably. Experimental results demonstrate that the system improves the state-of-the-art system of SemEval 2017 Task 5 by 1.7 and 3.7 points for sub-tracks 1 and 2 respectively. This robust system can be effectively used as a submodule in an end-to-end stock market price prediction system.

5 Acknowledgements

This work was supported by the Kyoto University Cooperative Intelligence Short-term Internship Program.

References

Md Shad Akhtar, Abhishek Kumar, Deepanway Ghosal, Asif Ekbal, and Pushpak Bhattacharyya. 2017. A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 551–557.

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.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *CoRR*, abs/1409.1259.

Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A Holistic Lexicon-Based Approach to Opinion Mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240. ACM.

CJ Hutto Eric Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf.*

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.

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional lstm networks for improved phoneme classification and recognition. In *Proceedings of the 15th International Conference on Artificial Neural Networks: Formal Models and Their Applications - Volume Part II*, ICANN'05, pages 799–804, Berlin, Heidelberg. Springer-Verlag.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural computation*, 9(8):1735–1780.

Marjan Van de Kauter, Diane Breesch, and Vronique Hoste. 2015. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications*, 42(11):4999 – 5010.

Tobias Daudert Manuela Huerlimann Manel Zarrouk Keith Cortis, Andre Freitas 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 Evaluations (SemEval-2017)*, pages 519–535, Vancouver, Canada. ACL.

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.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

Abhishek Kumar, Abhishek Sethi, Md Shad Akhtar, Asif Ekbal, Chris Biemann, and Pushpak Bhattacharyya. 2017. Iitpb at semeval-2017 task 5: Sentiment prediction in financial text. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 894–898.

Man Lan, Mengxiao Jiang, and Yuanbin Wu. 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)*, pages 888–893, Vancouver, Canada. ACL.

Youness Mansar, Lorenzo Gatti, Sira Ferradans, Marco Guerini, and Jacopo Staiano. 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)*, pages 817–822, Vancouver, Canada. ACL.

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, Lake Tahoe, NV, USA.

Marvin Minsky. 1986. *The Society of Mind*. Simon & Schuster, Inc., New York, NY, USA.

Nuno Oliveira, Paulo Cortez, and Nelson Areal. 2013. On the Predictability of Stock Market Behavior Using StockTwits Sentiment and Posting Volume. In *EPIA*, volume 8154 of *Lecture Notes in Computer Science*, pages 355–365. Springer.

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.

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)*, pages 119–123. IEEE Computer Society.

Robert P. Schumaker and Hsinchun Chen. 2009. Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System. *ACM Transactions on Information Systems*, 27(2).

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)*, pages 24–29, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Alex J. Smola and Bernhard Schölkopf. 2004. A Tutorial on Support Vector Regression. *Statistics and Computing*, 14(3):199–222.

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.

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*, pages 347–354. Association for Computational Linguistics.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *International Conference on Learning Representations (ICLR)*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.