Lancaster A at SemEval-2017 Task 5: Evaluation metrics matter: predicting sentiment from financial news headlines

Andrew Moore and Paul Rayson

February 27, 2018

School of Computing and Communications, Lancaster University.

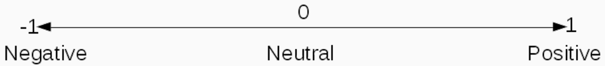
Task

The task

Example sentence

'Why AstraZeneca plc & Dixons Carphone PLC Are Red-Hot Growth Stars!'

Sentiment scale



Data

Training data: 1142 samples, 960 headlines/sentences.
Testing data: 491 samples, 461 headlines/sentences.

Approach

Models

1. Support Vector Regression (SVR) [1]
2. Bi-directional Long Short-Term Memory BLSTM [2][3]

Pre-Processing and Additional data used

Pre-Processing

1. Lower cased.
2. Tokenised.

Word2Vec model

Used 189, 206 financial articles (e.g. Financial Times) that were manually downloaded from Factiva[1] to create a Word2Vec model [5][2].

These were created using Gensim[3].

[1] https://global.factiva.com/factivalogin/login.asp?productname=global

[2] https://github.com/apmoore1/semeval/tree/master/models/word2vec_models

[3] https://radimrehurek.com/gensim/models/word2vec.html

Support Vector Regression (SVR) [1]

Features and settings that we changed

1. Tokenisation - Whitespace or Unitok[4]
2. N-grams - uni-grams, bi-grams and both.
3. SVR settings - penalty parameter C and epsilon parameter.
4. Target aspect.
5. Word Replacements.

[4] http://corpus.tools/wiki/Unitok

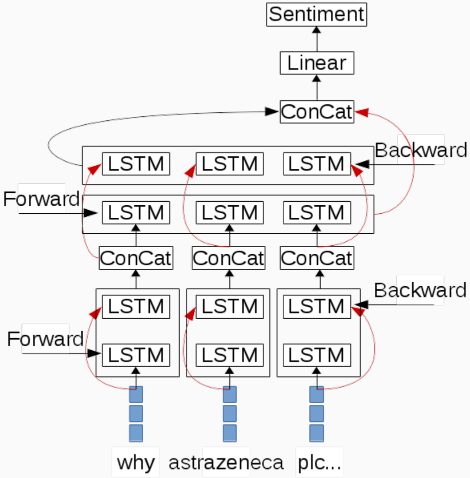
Word Replacements

Example Sentence

'AstraZeneca PLC had an improved performance where as Dixons performed poorly'

'companyname had an posword performance where as companyname performed negword'

Two BLSTM models



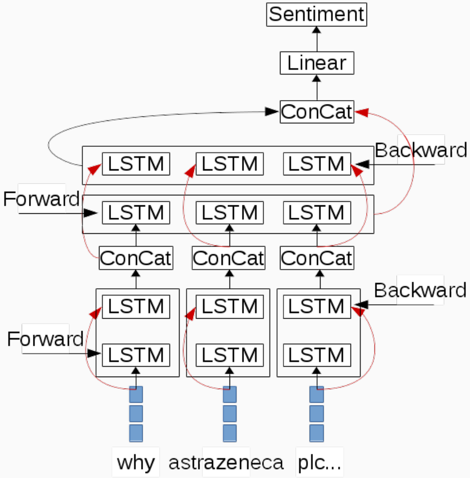
Standard Model (SLSTM)

- Drop out between layers and connections.
- 25 times trained over the data (epoch of 25).

Early stopping model (ELSTM)

- Drop out between layers only.
- Early stopping used to determine the epoch.

BLSTM loss function



Loss function

Mean Square Error (MSE)

$$\frac{1}{Y}\sum_{i=1}^{Y}(\hat{y}_i - y)^2 \qquad (1)$$

Findings and Results

SVR best features

Features

- Using uni-grams and bi-grams to be the best. 2.4% improvement over uni-grams.
- Using a tokeniser always better. Affects bi-gram results the most. 1% improvement using Unitok[5] over whitespace.
- SVR parameter settings important 8% difference between using C=0.1 and C=0.01.
- Incorporating the target aspect increased performance. 0.3% improvement.
- Using all word replacements. N=10 for POS and NEG words and N=0 for company. 0.8% improvement using company and 0.2% for POS and NEG.

[5] `http://corpus.tools/wiki/Unitok`

The three different metrics

Cosine Similarity (CS) Metric 1

$$\frac{\sum\limits_{i=1}^{K} y_i \hat{y}_i}{\sqrt{\sum\limits_{i=1}^{K} y_i^2}\sqrt{\sum\limits_{i=1}^{K} \hat{y}_i^2}} \quad (2)$$

Metric 2

$$\frac{\sum_{n=1}^{N} \mathrm{CS}(\hat{y}_n, y_n)}{N} \quad (3)$$

Metric 3

$$\frac{\sum_{n=1}^{N} \begin{cases} len(\hat{y}_n) * \mathrm{CS}(\hat{y}_n, y_n), & \text{if } len(\hat{y}_n) > 1 \\ 1 - |y - \hat{y}_n|, & \text{if } \frac{\hat{y}_n}{y} \geq 0 \end{cases}}{K} \quad (4)$$

K = Total number of samples.

N = Total number of sentences.

Results across the different metrics

Model	Metric 1	Metric 2	Metric 3
SVR	62.14	54.59	62.34
SLSTM	72.89	61.55	68.64
ELSTM	73.20	61.98	69.24
Fortia-FBK[4]	74.50	-	-

Metric 1 was the final metric used.

Error Analysis

‘uk stocks little changed as ashtead gains, housing shares drop’
Predicted: -0.43, Real: 0.23

‘standard life chief agrees 600000 bonus cut’
Predicted: -0.54, Real: 0.08

‘why i would put j sainsbury plc in my trolley before wm morrison supermarkets ...’
Predicted: 0.11, Real: 0.76

Future Work

Future Work

1. Incorporate aspects into the BLSTM's shown to be useful by Wang et al. [7].
2. Improve BLSTM's by using an attention model Wang et al. [7].
3. Add known financial sentiment lexicon into the LSTM model [6].

Summary

1. BLSTM outperform SVRs with minimal feature engineering.
2. The future is to incorporate more financial information into the LSTM's.

Questions?

a.moore@lancaster.ac.uk @apmoore94
p.rayson@lancaster.ac.uk @perayson

All the code can be found here[6]

Presentation can be found here [7]

[6] https://github.com/apmoore1/semeval

[7] https://github.com/apmoore1/semeval/blob/master/presentation/semeval.pdf

References I

H. Drucker, C. J. Burges, L. Kaufman, A. Smola, V. Vapnik, et al.
Support vector regression machines.
Advances in neural information processing systems, 9:155–161, 1997.

A. Graves and J. Schmidhuber.
Framewise phoneme classification with bidirectional lstm and other neural network architectures.
Neural Networks, 18(5):602–610, 2005.

S. Hochreiter and J. Schmidhuber.
Long short-term memory.
Neural computation, 9(8):1735–1780, 1997.

References II

Y. Mansar, L. Gatti, S. Ferradans, M. Guerini, and J. Staiano.
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. Association for Computational Linguistics, 2017.

T. Mikolov, K. Chen, G. Corrado, and J. Dean.
Efficient estimation of word representations in vector space.
arXiv preprint arXiv:1301.3781, 2013.

Q. Qian, M. Huang, J. Lei, and X. Zhu.
Linguistically regularized lstm for sentiment classification.
In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1679–1689. Association for Computational Linguistics, 2017.

References III

Y. Wang, M. Huang, x. zhu, and L. Zhao.
Attention-based lstm for aspect-level sentiment classification.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615. Association for Computational Linguistics, 2016.