Fortia-FBK at SemEval-2017 Task 5: Bullish or Bearish? Inferring Sentiment towards Brands from Financial News Headlines

Youness Mansar[⋆], **Lorenzo Gatti**[◦], **Sira Ferradans**[⋆], **Marco Guerini**[◦], **and Jacopo Staiano**[⋆]
[⋆]Fortia Financial Solutions, Paris, France
[◦]Fondazione Bruno Kessler, Povo, Italy
{youness.mansar,sira.ferradans,jacopo.staiano}@fortia.fr
{l.gatti,guerini}@fbk.eu

Abstract

In this paper, we describe a methodology to infer *Bullish* or *Bearish* sentiment towards companies/brands. More specifically, our approach leverages affective lexica and word embeddings in combination with convolutional neural networks to infer the sentiment of financial news headlines towards a target company. Such architecture was used and evaluated in the context of the SemEval 2017 challenge (task 5, subtask 2), in which it obtained the best performance.

1 Introduction

Real time information is key for decision making in highly technical domains such as finance. The explosive growth of financial technology industry (*Fintech*) continued in 2016, partially due to the current interest in the market for Artificial Intelligence-based technologies[1].

Opinion-rich texts such as micro-blogging and news can have an important impact in the financial sector (*e.g.* raise or fall in stock value) or in the overall economy (*e.g.* the Greek public debt crisis). In such a context, having granular access to the opinions of an important part of the population is of key importance to any public and private actor in the field. In order to take advantage of this raw data, it is thus needed to develop machine learning methods allowing to convert unstructured text into information that can be managed and exploited.

[1]F. Desai, "The Age of Artificial Intelligence in Fintech" https://www.forbes.com/sites/falgunidesai/2016/06/30/the-age-of-artificial-intelligence-in-fintech

S. Delventhal, "Global Fintech Investment Hits Record High in 2016" http://www.investopedia.com/articles/markets/061316/global-fintech-investment-hits-record-high-2016.asp

In this paper, we address the sentiment analysis problem applied to financial headlines, where the goal is, for a given news headline and target company, to infer its polarity score *i.e.* how positive (or negative) the sentence is with respect to the target company. Previous research (Goonatilake and Herath, 2007) has highlighted the association between news items and market fluctiations; hence, in the financial domain, sentiment analysis can be used as a proxy for *bullish* (i.e. positive, upwards trend) or *bearish* (i.e. negative, downwards trend) attitude towards a specific financial actor, allowing to identify and monitor in real-time the sentiment associated with e.g. stocks or brands.

Our contribution leverages pre-trained word embeddings (`GloVe`, trained on wikipedia+gigaword corpus), the `DepecheMood` affective lexicon, and convolutional neural networks.

2 Related Works

While image and sound come with a natural high dimensional embedding, the issue of *which is the best representation* is still an open research problem in the context of natural language and text. It is beyond the scope of this paper to do a thorough overview of word representations, for this we refer the interest reader to the excellent review provided by (Mandelbaum and Shalev, 2016). Here, we will just introduce the main representations that are related to the proposed method.

Word embeddings. In the seminal paper (Bengio et al., 2003), the authors introduce a statistical language model computed in an unsupervised training context using shallow neural networks. The goal was to predict the following word, given the previous context in the sentence, showing a major advance with respect to n-grams. Collobert *et al.* (Collobert et al., 2011) empirically proved

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 817–822,
Vancouver, Canada, August 3 - 4, 2017.

the usefulness of using unsupervised word representations for a variety of different NLP tasks and set the neural network architecture for many current approaches. Mikolov *et al.* (Mikolov et al., 2013) proposed a simplified model (`word2vec`) that allows to train on larger corpora, and showed how semantic relationships emerge from this training. Pennington *et al.* (Pennington et al., 2014), with the `GloVe` approach, maintain the semantic capacity of `word2vec` while introducing the statistical information from latent semantic analysis (LSA) showing that they can improve in semantic and syntactic tasks.

Sentiment and Affective Lexica. In recent years, several approaches have been proposed to build lexica containing prior sentiment polarities (sentiment lexica) or multi-dimensional affective scores (affective lexica). The goal of these methods is to associate such scores to raw tokens or tuples, e.g. `lemma#pos` where `lemma` is the lemma of a token, and `pos` its part of speech.

There is usually a trade-off between coverage (the amount of entries) and precision (the accuracy of the sentiment information). For instance, regarding sentiment lexica, *SentiWordNet* (Esuli and Sebastiani, 2006), (Baccianella et al., 2010), associates each entry with the numerical scores, ranging from 0 (negative) to 1 (positive); following this approach, it has been possible to automatically obtain a list of 155k words, compensating a low precision with a high coverage (Gatti et al., 2016). On the other side of the spectrum, we have methods such as (Bradley and Lang, 1999), (Taboada et al., 2011), (Warriner et al., 2013) with low coverage (from 1k to 14k words), but for which the precision is maximized. These scores were manually assigned by multiple annotators, and in some cases validated by crowdsourcing (Taboada et al., 2011).

Finally, a binary sentiment score is provided in the *General Inquirer* lexicon (Stone et al., 1966), covering 4k sentiment-bearing words, and expanded to 6k words by (Wilson et al., 2005).

Turning to affective lexica, where multiple dimensions of affect are taken into account, we mention *WordNetAffect* (Strapparava and Valitutti, 2004), which provides manual affective annotations of WordNet synsets (ANGER, JOY, FEAR, etc.): it contains 900 annotated synsets and 1.6k words in the form `lemma#PoS#sense`, which correspond to roughly 1k `lemma#PoS` entries.

AffectNet (Cambria and Hussain, 2012), contains 10k words taken from ConceptNet and aligned with WordNetAffect, and extends the latter to concepts like 'have breakfast'. *Fuzzy Affect Lexicon* (Subasic and Huettner, 2001) contains roughly 4k `lemma#PoS` manually annotated by one linguist using 80 emotion labels. *EmoLex* (Mohammad and Turney, 2013) contains almost 10k lemmas annotated with an intensity label for each emotion using Mechanical Turk. Finally, *Affect database* is an extension of *Senti-Ful* (Neviarouskaya et al., 2007) and contains 2.5k words in the form `lemma#PoS`. The latter is the only lexicon providing words annotated also with emotion scores rather than only with labels.

In this work, we exploit the `DepecheMood` affective lexicon proposed by (Staiano and Guerini, 2014): this resource has been built in a completely unsupervised fashion, from affective scores assigned by readers to news articles; notably, due to its automated crowd-sourcing-based approach, `DepecheMood` allows for both high-coverage and high-precision. `DepecheMood` provides scores for more than 37k entries, on the following affective dimensions: *Afraid*, *Happy*, *Angry*, *Sad*, *Inspired*, *Don't Care*, *Inspired*, *Amused*, *Annoyed*. We refer the reader to (Staiano and Guerini, 2014; Guerini and Staiano, 2015) for more details.

The affective dimensions encoded in `DepecheMood` are directly connected to the emotions evoked by a news article in the readers, hence it seemed a natural choice for the SemEval 2017 task at hand.

Sentence Classification. A modification of (Collobert et al., 2011) was proposed by Kim (Kim, 2014) for sentence classification, showing how a simple model together with pre-trained word representations can be highly performing. Our method builds on this conv-net method. Further, we took advantage of the rule-based sentiment analyser `VADER` (Hutto and Gilbert, 2014) (for Valence Aware Dictionary for sEntiment Reasoning), which builds upon a sentiment lexicon and a predefined set of simple rules.

3 Data

The data consists of a set of financial news headlines, crawled from several online outlets such as Yahoo Finance, where each sentence contains one or more company names/brands.

Each tuple (headline, company) is annotated with a sentiment score ranging from -1 (very negative, bearish) to 1 (very positive, bullish). The training/test sets provided contain 1142 and 491 annotated sentences, respectively.

A sample instance is reported below:

> Headline: *"Morrisons book second consecutive quarter of sales growth"*
>
> Company name: *"Morrisons"*
>
> Sentiment score: *0.43*

4 Method

In Figure 1, we can see the overall architecture of our model.

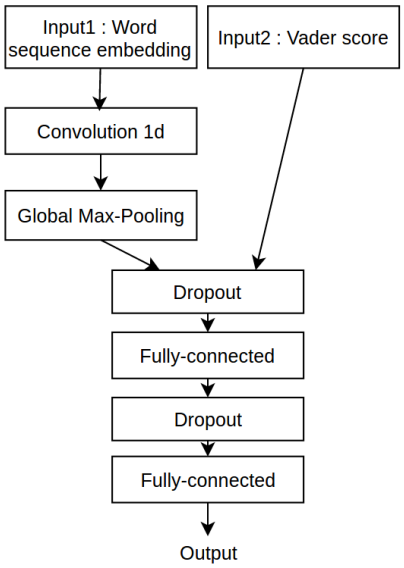

Figure 1: Network architecture

4.1 Sentence representation and preprocessing

Pre-processing. Minimal preprocessing was adopted in our approach: we replaced the target company's name with a fixed word <company> and numbers with <number>. The sentences were then tokenized using spaces as separator and keeping punctuation symbols as separate tokens.

Sentence representation. The words are represented as fixed length vectors u_i resulting from the concatenation of GloVe pre-trained embeddings and DepecheMood (Staiano and Guerini, 2014) lexicon representation. Since we cannot directly concatenate token-based embeddings (provided in GloVe) with the lemma#PoS-based representation available in DepecheMood, we proceeded to re-build the latter in token-based form, applying the exact same methodology albeit with two differences: we started from a larger dataset (51.9K news articles instead of 25.3K) and used a frequency cut-off, *i.e.* keeping only those tokens that appear at least 5 times in the corpus[2].

These word-level representation are used as the first layer of our network. During training we allow the weights of the representation to be updated. We further add the VADER score for the sentence under analysis. The complete sentence representation is presented in Algorithm 1.

Algorithm 1: Sentence representation

Input : An input sentence s, and the GloVe word embeddings W

Output: The sentence embedding x

$v = \text{VADER}(s)$

foreach w_i *in* W **do**

  u_i = [GloVe(w_i,W), DepecheMood(w_i)]

end

$x = [v, \{u_i\}_{i=1,\ldots,|W|}]$

4.2 Architectural Details

Convolutional Layer. A 1D convolutional layer with filters of multiple sizes {2, 3, 4} is applied to the sequence of word embeddings. The filters are used to learn useful translation-invariant representations of the sequential input data. A global max-pooling is then applied across the sequence for each filter output.

Concat Layer. We apply the concatenation layer to the output of the global max-pooling and the output of VADER.

Activation functions. The activation function used between layers is ReLU (Nair and Hinton, 2010) except for the out layer where tanh is used to map the output into [-1, 1] range.

[2]Our tests showed that: (i) the larger dataset allowed improving both precision on the SemEval2007 Affective Text Task (Strapparava and Mihalcea, 2007) dataset, originally used for the evaluation of DepecheMood, and coverage (from the initial 183K unique tokens we went to 292K entries) of the lexicon; (ii) we found no significant difference in performance between lemma#PoS and token versions built starting from the same dataset.

Regularization. Dropout (Srivastava et al., 2014) was used to avoid over-fitting to the training data: it prevents the co-adaptation of the neurones and it also provides an inexpensive way to average an exponential number of networks. In addition, we averaged the output of multiple networks with the same architecture but trained independently with different random seeds in order to reduce noise.

Loss function. The loss function used is the cosine distance between the predicted scores and the gold standard for each batch. Even though stochastic optimization methods like Adam (Kingma and Ba, 2014) are usually applied to loss functions that are written as a sum of per-sample loss, which is not the case for the cosine, it converges to an acceptable solution. The loss can be written as :

$$\text{Loss} = \sum_{\text{B}\in\text{Batches}} 1 - \cos(\hat{\mathbf{V}}_B, \mathbf{V}_B), \quad (1)$$

where $\hat{\mathbf{V}}_B$ and $\mathbf{V}_B$ are the predicted and true sentiment scores for batch B, respectively.

The algorithm for training/testing our model is reported in Algorithm 2.

Algorithm 2: Training/Testing algorithm. To build our model, we set N=10.

Input : A set of training instances S, with ground-truth scores y, and the set of test sentences S_o

Output : A set of trained models M, and the predictions y_o for the test set S_o

Parameters: The number N of models to train

preprocess(X) `// see sec 3.1`
foreach s_i *in* S **do**
  X_i = sentence_representation(s_i) `// see Alg. 1`
end
foreach $n \in N$ **do**
  $M_n = \min \text{Loss}(X)$ `// see Eq. 1`
end
foreach $n \in N$ **do**
  $y_n = \text{evaluate}(X_o, M_n)$
end
$y_o(u) = \frac{1}{N}\sum_n^N y_n(u)$

5 Results

In this section, we report the results obtained by our model according to challenge official evaluation metric, which is based cosine-similarity and described in (Ghosh et al., 2015). Results are reported for three diverse configurations: (i) the full system; (ii) the system without using word embeddings (*i.e.* `Glove` and `DepecheMood`); and (iii) the system without using pre-processing. In Table 1 we show model's performances on the challenge training data, in a 5-fold cross-validation setting.

Algorithm	mean±std
Full	0.701 ±0.023
No embeddings	0.586 ±0.017
No pre-processing	0.648 ±0.022

Table 1: Cross-validation results

Further, the final performances obtained with our approach on the challenge test set are reported in Table 2. Consistently with the cross-validation performances shown earlier, we observe the beneficial impact of word-representations and basic pre-processing.

Algorithm	Test scores
Full	0.745
No embeddings	0.660
No pre-processing	0.678

Table 2: Final results

6 Conclusions

In this paper, we presented the network architecture used for the Fortia-FBK submission to the Semeval-2017 Task 5 (Cortis et al., 2017), Subtask 2 challenge, with the goal of predicting positive (bullish) or negative (bearish) attitude towards a target brand from financial news headlines. The proposed system ranked 1st in such challenge.

Our approach is based on 1d convolutions and uses fine-tuning of unsupervised word representations and a rule based sentiment model in its inputs. We showed that the use of pre-computed word representations allows to reduce over-fitting and to achieve significantly better generalization, while some basic pre-processing was needed to further improve the performance.

References

S. Baccianella, A. Esuli, and F. Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC 2010*. Valletta, Malta, pages 2200–2204.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3(Feb):1137–1155.

M.M. Bradley and P.J. Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. *Technical Report C-1, University of Florida* .

Erik Cambria and Amir Hussain. 2012. *Sentic computing*. Springer.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

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.

A. Esuli and F. Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC 2006*. Genova, IT, pages 417–422.

Lorenzo Gatti, Marco Guerini, and Marco Turchi. 2016. SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing* 7(4):409–421.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. pages 470–478.

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.

Marco Guerini and Jacopo Staiano. 2015. Deep feelings: A massive cross-lingual study on the relation between emotions and virality. In *Proceedings of WWW 2015*. pages 299–305.

C.J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of ICWSM 2014*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*. pages 1746–1751.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980. http://arxiv.org/abs/1412.6980.

Amit Mandelbaum and Adi Shalev. 2016. Word embeddings and their use in sentence classification tasks. *arXiv preprint arXiv:1610.08229* .

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29(3):436–465.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2007. Textual affect sensing for sociable and expressive online communication. In *Affective Computing and Intelligent Interaction*, Springer Berlin Heidelberg, volume 4738, pages 218–229.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP 2014*. volume 14, pages 1532–43.

Nitish Srivastava, Geoffrey E 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(1):1929–1958.

Jacopo Staiano and Marco Guerini. 2014. Depeche Mood: a lexicon for emotion analysis from crowd annotated news. In *Proceedings of ACL 2014*. The Association for Computer Linguistics, volume 2, pages 427–433.

P.J. Stone, D.C. Dunphy, and M.S. Smith. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT press.

C. Strapparava and A. Valitutti. 2004. WordNet-Affect: an affective extension of WordNet. In *Proceedings of LREC 2004*. Lisbon, pages 1083 – 1086.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, pages 70–74.

Pero Subasic and Alison Huettner. 2001. Affect analysis of text using fuzzy semantic typing. *Fuzzy Systems, IEEE Transactions on* 9(4):483–496.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37(2):267–307.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 45(4):1191–1207.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on HLT/EMNLP 2005*. Vancouver, Canada.