Numeral Understanding in Financial Social Media Data

Tom Neuhäuser
Computer Science / 2021
tomneu@stud.ntnu.no

Abstract

This report presents my approach to tackle the first subtask of the NTCIR-14 FinNum Shared Task on Fine-Grained Numeral Understanding in Financial Social Media Data. The first subtask asked the participants to classify given target numerals in financial tweets into 7 categories. My approach may be summarized as follows. I view the task as a sequence labeling problem. I enrich word level representations of tweets by linguistic features (Part-Of-Speech labels and Named Entity categories) as well as non-linguistic features (positions of target numerals) . I feed the enriched representations into a Convolutional Neural Network in order to extract local contexts. I then use the contextual information to predict the classes of target numerals.

1 Introduction

In recent years, social media's influence on all our lifes and on every part of life grew substantially. This trend motivated the idea to automatically retrieve informations from texts published on various social media platforms using techniques originating from the area of computer science and using background knowledge provided by the area of linguistics. The shared research field, right between the areas of linguistics and computer science, is commonly known under the name of Natual Language Processing. Even though social media posts are harder to analyse than news articles or other kinds of official documents, due to the informal writing style, the value of infomation social media posts provide justify the Natural Language Processing community's interest in them.
One particular influential social media platform is the micro-blogging service Twitter. In particular, Twitter was able to attract the attention of researchers and practioners developing financial technologies, as many individuals use the service to share their opinions on many finance related topics. One example for the successful application of Natural Language Processing in order to retrieve moods of investors from financial tweets is that of Sentiment Analysis. Another example is that of Stock Market Movement Prediction.
The analysis of a financial instruments, which is required in order to predict the market, can be approached in two different ways – fundamental analysis and technical analysis. In the case of fundamental analysis, one attempts to measure the intrinsic value of a financial instrument. On the other hand, for technical analysis, one studies historical market data. For both kinds of analyses, and thus for Stock Market Movement Prediction, numerals play an important role. For example, when performing technical analysis, one may consider technical indicators like the moving average.
The remainder of this report is structured as follows. In section 2, the data which we will work with is examined. Section 3 introduces the key concepts which we will make use of in the further discussion. In section 4, related work is discussed. In particular, other peoples submissions to the shared task are summarized. Section 5 describes my approach to tackle the first subtask. In section 6 presents the results of my efforts. Section 7 concludes this report by giving suggestions on possible future work.

2 Data

Before we proceed, it is crucial to gain a proper understanding of the dataset we will be working with. The dataset the NTCIR-14 FinNum Shared Task was based on is called FinNum 2.0 [Chen et al. (2018)].

The data was collected from StockTwits[1], a social trading platform similar to Twitter, where individuals can share their opinions on finance related topics. An exemplaric tweet, published on StockTwits, is shown below.

$WRN My fav $WRN pattern on my watchlist for 11/09/17.
Very nice support at 0.96 so well see if we can get an entry aro

Note that at least one cashtag, such as $WRN, and at least one numeral, such as 0.96, can be found in each tweet. A cashtag stands for a stock. For example, the cashtag $WRN stands for the stock of the Western Copper & Gold Corporation.
Two experts were involved in the process of annotating the tweets. They located and classified numerals into 7 categories: Monetary, Temporal, Percentage, Quantity, Indicator, Option and Product Number. The annotations are available under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license.

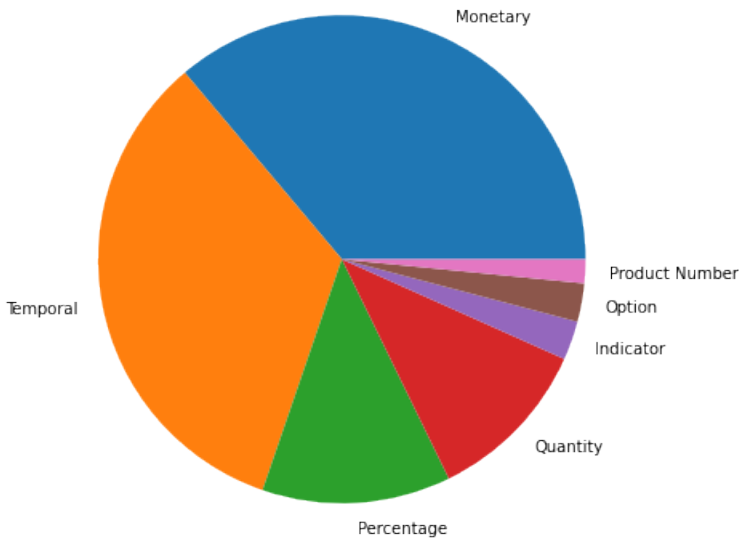

Figure 1: Category Distribution

I was able to rebuild 4039 of the 7613 tweets the FinNum 2.0 dataset originally consisted of. I split the 4039 samples into 3231 training samples, 407 samples to use for validation purposes and 401 samples to use for testing, resulting in an approximate 80%, 10%, 10% split. The category distribution of the 5196 target numerals found in the 3231 training samples is visualized in figure 1. By having a look at the figure, it becomes obvious, that we are dealing with an imbalanced classification problem.

3 Background

To be able to follow the remainder of this report, it is important to first give a brief introduction to the key concepts which we will make use of in the further discussion.

3.1 Sequence Labeling

The task of sequence labeling is a pattern recognition task which asks for the assignment of a label to each member of a sequence. A popular example for a sequence labeling problem is Part-of-Speech Tagging. In our case, where we want to classify given target numerals in financial tweets into 7 categories, we want to label each target numeral with its respective class and every other member of a tweet with the placeholder **O**. An exemplaric tweet and the labels of each of its members are shown

[1] https://stocktwits.com/

below.

$WRN	My	fav	$WRN	pattern	on	my	watchlist	for	11	/	09	/	17	.
O	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**Temporal**	**O**	**Temporal**	**O**	**Temporal**	**O**
Very	nice	support	at	0.96	so	well	see	if	we	can	get	an	entry	aro
O	**O**	**O**	**O**	**Monetary**	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**O**

3.2 GloVe Vectors

GloVe (Global Vectors) is a learning algorithm to obtain vector representations of words. The training is performed on aggregated global word-word co-occurrence statistics from a given corpus [Pennington et al. (2014)]. Pre-trained word vectors are available under the Open Data Commons Public Domain Dedication and License (PDDL).

3.3 Part-of-Speech Tagging

Part-of-Speech Tagging is an example for a sequence labeling problem. When performing Part-of-Speech Tagging, one is asked to assign Part-of-Speech tags to each word in a given corpus. Part-of-Speech Tagging can be approached in two different ways – rule-based and stochastic. In both cases, to asign a Part-of-Speech tag to a word, the word itself as well as its context is considered. An exemplaric tweet and the Part-of-Speech tags of each of its members are shown below.

$WRN	My	fav	$WRN	pattern	on	my	watchlist	for	11	/	09	/	17	.
NN	**PRP$**	**NN**	**NN**	**NN**	**IN**	**PRP$**	**NN**	**IN**	**CD**	**SYM**	**CD**	**SYM**	**CD**	**.**
Very	nice	support	at	0.96	so	well	see	if	we	can	get	an	entry	aro
RB	**JJ**	**NN**	**IN**	**CD**	**RB**	**RB**	**VB**	**IN**	**PRP**	**MD**	**VB**	**DT**	**NN**	**NN**

3.4 Named Entity Recognition

Named-Entity Recognition is an information extraction problem. When performing Named-Entity Recognition, one is asked to locate and classify named entities in a given text into categories such as persons, organizations or locations. In our case, we want to label each recognized Named-Entity with its respective class and every other member of a tweet with the placeholder **O**. An exemplaric tweet and the recognized Named-Entities of each of its members are shown below.

$WRN	My	fav	$WRN	pattern	on	my	watchlist	for	11	/	09	/	17	.
O	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**DATE**	**DATE**	**DATE**	**DATE**	**DATE**	**O**
Very	nice	support	at	0.96	so	well	see	if	we	can	get	an	entry	aro
O	**O**	**O**	**O**	**NUMBER**	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**O**

3.5 Convolutional Neural Networks

Convolutional Neural Networks build a class of neural networks, most well known for their application in the field of computer vision. A typical Convolutional Neural Network to solve the task of image classification is shown in figure 2. The Convolutional Neural Network consists of two parts, the feature extractor and the actual classifier. The feature extractor applies so called convolutional filters to an input image, in order to detect features such as the edges of an object. Then, the classifier computes, based on the features detected by the feature extractor, the likelihoods of an input image belonging to the predefined classes.

In recent years, Convolutional Neural Networks have been successfully applied to solve many Natural Language Processing tasks, as they are good at recognizing local patterns and are computationally cheaper than Recurrent Neural Networks [e.g. Kim (2014)].

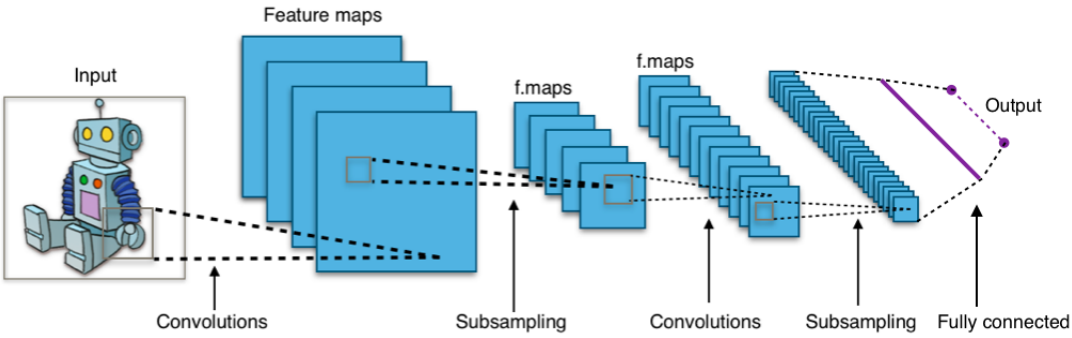

Figure 2: A Convolutional Neural Network [Aphex34, CC BY-SA 4.0]

3.6 F_1-Score

The F_1-Score is a measure of accuracy. It is the harmonic mean of the precision and the recall, where the precision is the number of true positives divided by the number of all positives (true and false) and the recall is the number of true positives divided by the number of all true positives and false negatives. The F_1 score is the harmonic mean of the precision and recall. The F_1-Score is calculated by the equation shown below.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

4 Related Work

In *Numeral Understanding in Financial Tweets for Fine-Grained Crowd-Based Forecasting*, Chen et al. made a first attempt to understand numerals in financial social media data [Chen et al. (2018)]. They base their research on the FinNum 1.0 dataset and compare the performance of Support Vector Machines, Convolutional Neural Networks and Recurrent Neural Networks at the task of classifying given target numerals in financial tweets into 7 categories and 14 subcategories.
In 2018, the 14th NTCIR Conference on Evaluation of Information Access Technologies took place. In the context of the conference, the NTCIR-14 FinNum Shared Task, based on the FinNum 2.0 dataset, was published. Again, participants were asked to classify given target numerals in financial tweets into 7 categories and 14 subcategories. In *Overview of the NTCIR-14 FinNum Task: Fine-Grained Numeral Understanding in Financial Social Media Data*, Chen et al. give an overview of the task and the submissions they received [Chen et al. (2019)]. There were six submissions from six different groups of researchers which got accepted. Two groups viewed the task as a sequence labeling problem [Azzi and Bouamor (2019), Liang and Su (2019)] while the other four groups viewed the task as a classification problem [Spark (2019), Wang et al. (2019), Tian and Peng (2019), Wu et al. (2019)].
Azzi and Bouamor (2019) proposed a Convolutional Neural Network working on enriched word level representations of tweets and Liang and Su (2019) developed a model combining Convolutional Filters with Recurrent Neural Networks. The approaches followed by the groups, which viewed the task as a classification problem, range from the usage of Support Vector Machines and BERT to the development of Convolutional Neural Networks and Recurrent Neural Networks.

5 Model

First, I preprocess the tweets. I enrich word level representations of tweets by linguistic features (Part-Of-Speech labels and Named Entity categories) as well as non-linguistic features (positions of target numerals) . Then, I feed the enriched representations into a Convolutional Neural Network in order to extract local contexts. Finally, I use the contextual information to predict the classes of target numerals.

5.1 Preprocessing

I remove hashtags, mentions, URLs, emojis and smileys using preprocessor[2]. I replace cashtags by a special token. I remove all special characters except '$' and '%' as well as special characters occuring between digits (e.g. '.' or '/'). I uncouple numbers from other characters (e.g. '15th' becomes '15 th'). I transform all characters to lowercase. I tokenize tweets and lemmatize tokens using CoreNLP [Manning et al. (2014)]. I remove stop words using NLTK[3].

5.2 Representation

I use pre-trained GloVe word vectors to represent a word as a vector [Pennington et al. (2014)]. More precisely, I use 50-dimensional word vectors which were pre-trained on a joined corpus consisting of the Wikipedia 2014 and the Gigaword 5 corpuses, comprising 6 billion tokens, resulting in a vocabulary (uncased) of size 400000.
I use CoreNLP to perform Part-of-Speech Tagging and Named-Entity Recognition [Manning et al. (2014)]. I use One-Hot-Encoding to represent Part-of-Speech labels and Named-Entity categories as vectors. There are 36 Part-of-Speech labels and 13 Named-Entity categories, resulting in 36-dimensional vectors and 13-dimensional vectors respectively.
I represent whether a word is a target numeral or not using a 1-dimensional vector, which shall be equal to 1 if the token is a target numeral and equal to 0 if not.
I concatenate the 50-dimensional vector representing a word, the 36-dimensional vector representing a Part-of-Speech label, the 13-dimensional vector representing a Named-Entity category and the 1-dimensional vector representing whether a word is a target numeral or not, to obtain an enriched 100-dimensional vector representation.

5.3 Architecture

To classify given target numerals in financial tweets, I use a Convolutional Neural Network. The Convolutional Neural Network is shown in figure 3. It consists of one convolutional layer and one linear layer.

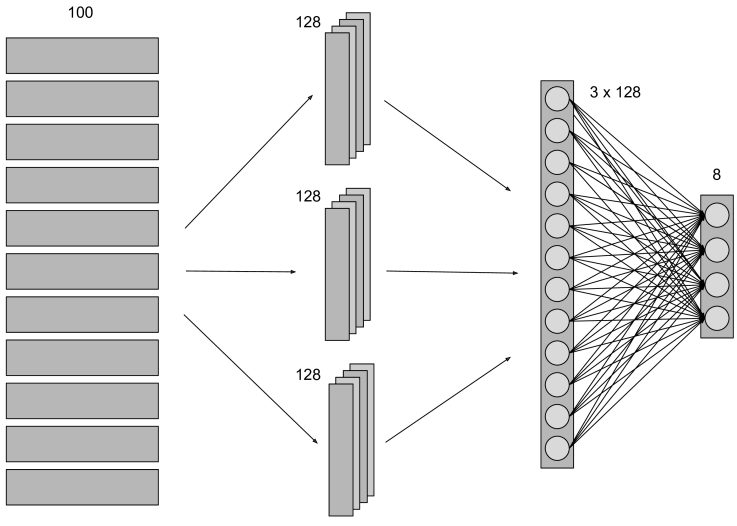

Figure 3: Architecture

The convolutional layer takes the enriched representation as input and applies three convolutions in parallel. Each convolution uses a kernel of different size. The first convolution uses a kernel of size 3, the second convolution uses a kernel of size 5 and the third convolution uses a kernel of size 7. The number

[2]https://github.com/s/preprocessor
[3]https://www.nltk.org/

of output filters of each convolution is 128. After applications of the ReLU activation function, the three outputs of the three convolutions get concatenated and serve as the input to the linear layer, after a dropout with dropout rate 0.5 is applied. The linear layer outputs the likelihoods of a word to belong to one of the 7 categories, after the application of the Softmax activation function.

6 Experiments and Results

I performed an experiment to draw conclusions about the effectiveness of the enrichment of the word level reprepresentations of tweets. In one run, I enriched the word level representations, and in another, I did not enrich the word level representations. In both runs, I used the Adam optimizer and set the learning rate to 0.001. I used a batch size of 32 in both cases. In table 1, the F_1-Scores of both runs can be found.

enriched?	Micro F_1-Score	Macro F_1-Score
no	0.845	0.717
yes	0.865	0.698

Table 1: F_1-Scores

7 Conclusion and Future Work

To conclude, my model performed well and would have been able to compete with the models of the participants of the NTCIR-14 FinNum Shared Task. Having experimented with enrichment of the word level representations of tweets was a good first step. A next step, which I strongly believe could be worthwhile to take, is to experiment with the combination of Convolutional Neural Networks and Recurrent Neural Networks, given the time and computing resources.

References

Abderrahim Ait Azzi and Houda Bouamor. Fortia1 at the NTCIR-14 FinNum Task: Enriched sequence labeling for numeral classi
cation. In *In: Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, 2019.

Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. Numeral understanding in financial tweets for fine-grained crowd-based forecasting. In *Proceedings of the 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2018)*, pages 136–143, Santiago, Chile, 2018. doi: 10.1109/WI.2018.00-97.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the NTCIR-14 FinNum Task: Fine-grained numeral understanding in financial social media data, 2019.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL `https://www.aclweb.org/anthology/D14-1181`.

Chao-Chun Liang and Keh-Yih Su. ASNLU at the NTCIR-14 FinNum Task: Incorporating knowledge into dnn for financial numeral classification. In *In: Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, 2019.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`.

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

Alan Spark. BRNIR at the NTCIR-14 Fi
nNum Task: Scalable feature extraction technique for number classi
cation. In *In: Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, 2019.

Ke Tian and Zi Jun Peng. aiai at the NTCIR-14 FinNum Task: Financial numeral tweets fine-grained classification using deep word and character embedding-based attention model. In *In: Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, 2019.

Wei Wang, Maofu Liu, and Zhenlian Zhang. WUST at the NTCIR-14 Fi
nNum Task. In *In: Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, 2019.

Qianhui Wu, Guoxin Wang, Yuying Zhu, Haoyan Liu, and Börje F. Karlsson. DeepMRT at the NTCIR-14 FinNum Task: A hybrid neural model for numeral type classi
cation in financial tweets. In *In: Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, 2019.