FinSim-3: The 3rd Shared Task on Learning Semantic Similarities for the Financial Domain

Juyeon Kang[*] , **Sandra Bellato** , **Mei Gan** and **Ismail El Maarouf**
Fortia Financial Solutions
Paris, France
{juyeon.kang,sandra.bellato, mei.gan, ismail.elmaarouf }@fortia.fr

Abstract

The FinSim-3 is the third edition of FinSim shared task on Learning Semantic Similarities for the Financial Domain, held in conjunction with IJCAI 2021 @Online as part of the FinNLP-2021 workshop. FinSim Shared Task proposes the challenge to automatically learn effective and precise semantic models for the financial domain. The third edition of the FinSim offered an extended dataset with more diversified financial concepts in order to increase its coverage in the semantic models, which were limited to those of instruments and market indices in the previous editions, and interested in systems which make creative use of relevant resources such as ontologies and lexica, as well as systems which make use of contextual word embeddings such as BERT [Devlin *et al.*, 2019].

This year, 11 system runs were submitted by 5 teams and systems were ranked according to two metrics, Accuracy and Mean rank. All the systems largely beat our baselines 1 and 2, and the best system of this edition beats the performance of the previous editions [Mansar *et al.*, 2021] in both of the metrics.

1 Introduction

The FinSim-3, organized by Fortia Financial Solutions[1], a French AI startup with expertise in Financial Natural Language Processing (NLP), is held as part of the third workshop on Financial Technology and Natural Language Processing (FinNLP)[2] at IJCAI 2021[3]. It focused on automatically learning effective and precise semantic models adapted to the financial domain. More specifically, it addressed an automatic classification of the financial terms of the large scope from instruments, market indices, securities, market data, regulatory agencies, financial services entities, corporate bodies to credit events. This is typically addressed using either unsupervised corpus-derived representations like word embeddings, which are typically opaque to human understanding but very useful in NLP applications or manually created resources such as taxonomies and ontologies, which typically have low coverage and contain inconsistencies, but provide a deeper understanding of the target domain.

The range of those terms is vast and the categories encompass all sorts of tradable contracts. Financial instruments can be challenging as, while traditional instruments such as Bonds or Stocks are straightforward, other instruments such as Futures pose a number of difficulties as they may apply to various underlying instrument types (e.g. bond futures, equities futures). Market indices also bring a complexity to the classification task for a given term-label pair, as for example, a "credit index" like *Standard Latin America Corporate Bond* or *Standard US Municipal General Fund* can be confused with an instrument label. Thus, the challenge of automatic classification of financial terms, is coupled with a challenge of semantic analysis.

Also, those concepts do not fairly appear in the real world financial documents, consequently in the dataset too, and it results a data imbalance issue as several participants of the FinSim are already mentioned in their papers and made effort to tackle this issue.

The FinSim-3 focuses on the evaluation of semantic representations by assessing the quality of the automatic classification of a given list of carefully selected term-label pairs from the Financial domain against a domain ontology. Participants will be given a list of carefully selected terms from the Financial domain such as *Alphabet Inc. US-CA*, *CDX Tranche XO* and will be asked to design a system which can automatically classify them into the most relevant hypernym (or top-level) concept in an external ontology. For example, given the set of concepts "Fund", "Unclassified", "Credit index", "Stock Corporation", and "Equity index", the most relevant hypernym of *Alphabet Inc. US-CA* will be "Stock Corporation". The FinSim-3 task also provided a raw corpus of financial prospectuses, from which to derive automatic representations, a train set of financial terms corresponding to 17 concepts, as well as mappings to an ontology of the financial domain, namely FIBO (Financial Industry Business Ontology)[4].

The paper is structured as follows: the section 2 gives a description on the related shared tasks, the section 3 explains

[*]Contact Author
[1]https://www.fortia.fr/
[2]https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp2021
[3]https://ijcai-21.org/
[4]https://spec.edmcouncil.org/fibo/

Proceedings of the Third Workshop on Financial Technology and Natural Language Processing
(FinNLP@IJCAI 2021), pages 31-35, Online, August 19, 2021.

the FinSim-3 task in detail. The participating systems are introduced along with the results in the section 4. We conclude the paper with some open problems and perspectives for the next edition of the FinSim.

2 Related Shared Tasks

The FinSim-3 proposes a task of predicting a hypernym or "is a" relation for financial term's categorization: given a training set of terms and a fixed set of labels, participants are asked to propose systems allowing to categorize new terms to their most likely hypernym. Two terms are considered as in a hypernymy-hyponymy relation if one of them can be conceived as a more generic term(e.g. "vehicle" - "car").

A taxonomy represents a semantic relation of term pairs, which is "isa" pairs, largely used in NLP and IE tasks. The taxonomy extraction task from a domain specific corpora as a competition is first proposed by the shared task TExEval [Bordea *et al.*, 2015] and TExEval-2 [Bordea *et al.*, 2016] as part of SemEval-2015 and 2016. Several works introduced methods to learn hypernymy from the corpora and showed how to induce taxonomies from "isa" pairs. While those systems largely exploit semantic lexical resources like WordNet, BabelNet, YAGO, Wiki, DBPedia, etc., distributional approach was not meaningfully adopted.
More recently, as part of SemEval 2020, the shared task on Predicting Multilingual and Cross-Lingual (Graded) Lexical Entailment [Glavaš *et al.*, 2020] proposed a challenge for detecting semantic hierarchical relation, hypernym-hyponym, from multilingual and cross-lingual datasets. The Distributional track was newly added in order to evaluate only distributional systems. The participating systems make use of rule-based approach by exploiting Wiktionary definitions of concepts [Kovács *et al.*, 2020] or distributional approach combining distributional word vectors, multilingual lexical resources and translated parallel corpora to obtain cross lingual synonyms, then to extract a set of terms which are semantically most similar to a seed term [Hauer *et al.*, 2020][Wang *et al.*, 2020].

3 Task Description

3.1 Financial Domain Ontology: FIBO

FIBO is an interesting, pioneering and ongoing work to formalize the semantics of the financial domain in the form of a large number of ontologies. More detail can be found on their website. Participants were encouraged to use this resource (as well as others) in designing their system and this is why we provided a number of scripts to facilitate its processing. We also provided a mapping from each of the categories used in FinSim-3 to a concept in the FIBO ontology (in the file data/outputs/fibo mapping.json).

3.2 Dataset

Prospectus Corpus

This year, we provided 210 English financial prospectuses in PDF format to be used for training embeddings and/or other experimental investigation. Financial prospectuses provide key information for investors and detail investment rules related to a fund. The corpus size is estimated to about 15 million tokens and each document is composed of a dozen pages to several hundreds.

Labels

We expanded the scope of the label's type, which was limited to the **financial instrument** category with `Bonds, Forwards, Funds, Future, MMIs, Option, Stocks` and `Swap` labels and the **market indices** category with `Credit index` and `Equity index` labels, by adding five new categories **securities**, **market data**, **FBC functional entities**, **legal entities**, **FBC debt and equities**, with `Securities restrictions`, `Parametric schedules`, `Debt pricing and yields`, `Credit Events`, `Stock Corporation`, `Central Securities Depository` and `Regulatory Agency` labels. As an experienced RegTech with the expertise in financial domain, Fortia elaborated a FundTree[5], an ontological representation of financial domain's concepts. The FundTree is partially presented in the FinSim Asset Tree, as described in [Maarouf *et al.*, 2020]. The new labels are carefully selected based on our FundTree and the FIBO ontology. Our investigation on the financial documents confirm that the proposed 17 labels are more and less frequently used in the documents like prospectus, KIID (Key Investor Information Document), reporting, etc. The concepts of each label are defined in the FIBO ontology[6].

Terms

The data creation is manually processed by the way that, for a given label, the annotators first identified the corresponding terms from the FundTree, then if more samples needed, the external resources like FIBO is considered. The final dataset is then validated by our financial domain expert.

Labels	Training	Test
Instrument related	217	114
Equity Index	286	80
Credit Index	129	18
Regulatory Agency	205	51
Central Securities Depository	107	27
Stock Corporation	25	6
Credit Events	18	6
Debt pricing and yields	58	14
Parametric schedules	15	6
Securities restrictions	8	4

Table 1: Data set of terms for FinSim-3

3.3 Metrics

We use the same metrics as the previous editions of FinSim, Accuracy and Mean Rank. For each term x_i with a label y_i, the expected prediction is a top 3 list of labels ranked from

[5]It is comparable and can be aligned to Openfunds, for more details on Openfunds, refer to https://openfunds.org/.

[6]https://spec.edmcouncil.org/fibo/ontology/

most to least likely to be equal to the ground truth by the predictive system $\hat{y}_i^l$. We note by $rank_i$ the rank of the correct label in the top-3 prediction list, if the ground truth does not appear in the top-3 then $rank_i$ is equal to 4. Given those notation, the accuracy can be expressed as:

$$Accuracy = \frac{1}{n} * \sum_{i=1}^{n} I(y_i = \hat{y}_i^l[0])$$

And the Mean Rank as:

$$Mean\_Rank = \frac{1}{n} * \sum_{i=1}^{n} rank_i$$

3.4 Baseline systems

We prepared two simple baselines in order to help the participants get started. Both baselines are based on a custom Word2vec model that was trained on a financial corpus. The vector representation for each term is computed as the average of the word embeddings of their tokens. For each test sample, the first baseline ranks all the possible hypernyms using the hyponym-hypernym similarity in the embedding space. The second baseline trains a logistic regression model that classifies each test sample into ten different classes where each class represents one possible hypernym.

4 Participants and Submitted Systems

We received 11 system runs from 5 private and public institutions. The participating institutions, organized into 5 teams, are listed in Table 2.

Team	Affiliation
DICoE	NCSR "Demokritos"/AUEB
MiniTrue	University of Zurich/Harbin Institute of Technology
MXX	MXX
Lipi	Fidelity Investments
Yseop	Yseop

Table 2: List of the 5 participating teams in the FinSim-3.

Participating teams investigated and implemented a wide variety of techniques and features from sentence-level embeddings to knowledge graph embeddings and the data augmentation methods are also largely explored. We present a summary of the methods proposed by each participating team in this section. More detailed methods and results of each team can be found in the Proceedings of FinNLP-2021 published at ACL anthology.

DICoE

DICoE team submitted two systems by exploring a data augmentation method based on the term's definitions from Investopedia[7] in order to enlarge the initial training data, and training a Logistic Regression classifier over the hand-crafted and distance based features. They also handle the out of vocabulary (OOV) words by replacing them with the semantically closest in-vocabulary words using Levenshtein distance. They show that the Logistic Regression classifier, trained on the augmented training data based on the Investopedia's terms/definitions using the combined features between Word2Vec [Mikolov *et al.*, 2013] embeddings (provided by FinSIm), hand-crafted features and the Levenshtein-based OOV words handling, achieves the best result in their experiments.

MiniTrue

MiniTrue team proposes an approach combining contextuel embeddings BERT [Devlin *et al.*, 2019] with knowledge graph embeddings using the RotatE algorithm which is able to model and infer various relation patterns including: symmetry/antisymmetry, inversion, and composition [Sun *et al.*, 2019]. In addition to these features, their voting mechanism allows to joint the inference values predicted by the basic models and predict the best final output. Their experiments show that a simple classifier network based on the voting function using the financial text pretrained language model FinBERT [Araci, 2019] combined with Knowledge graph embedding as features, leads to an improved classification results and also the voting function brings an extra benefit to the final output.

MXX

MXX team explores a recurrent neural networks based classifier using a custom word embeddings trained on the NLP preprocessed set of prospectuses and the flat FIBO ontology definitions provided by FinSim. They build two-stage framework to first train the custom word embeddings, then a Bi-directional Long Short Term Memory (Bi-LSTM) networks [Greff *et al.*, 2017] to map a sequence of word embeddings to its potential hypernym. They also propose a data augmentation method to expand the volume of the instances in training set based on the semantic similarity between the words, which results a total of 8000 terms from 1050 terms provided in the training set by FinSim. Their framework applied to the augmented data outperforms all other participating systems including two baseline methods.

Lipi

Lipi team submitted three systems, the first two systems tackle the shared task as a classification problem and the last as a phrase similarity problem, fully exploring the external resources like DBPedia[8], Investopedia and FIBO for a data augmentation and using pre-trained FinBERT embeddings. The FinBERT model is also fine-tuned using Sentence BERT [Reimers and Gurevych, 2019] over an extended labelled set based on FIBO ontology allowing to achieve the best performance in their experiments comparing to other pre-trained models provided by the Huggingface's transformers repository[9].

Yseop

Yseop team proposes methods of combining sentence-level embeddings SRoBERTa [Reimers and Gurevych, 2019] with word embeddings to improve the classification performance and experiments Logistic Regression and Random Forest classifier using these features. For the best performance, they

[7] https://www.investopedia.com/

[8] https://lookup.dbpedia.org/api/search

[9] https://huggingface.co/transformers/

customize the data by augmenting it with an extracted corpus from FIBO and trained from scratch a sentence transformer model, and also two custom word embeddings models based on Word2Vec and FastText are trained and used in their comparative experiments. The dual word-sentence embeddings model used in a Logistic Regression classifier shows the best result in their experiments comparing to that of the sentence-level embeddings used in a Logistic Regression or Random Forest as classifier.

4.1 Results and Discussion

The Tables 3 and 4 show the results of the 11 system runs in terms of accuracy and mean rank along with the overall results obtained by combining those of mean Rank and accuracy. Mxx team won first place, yseop_2 came second and lipi_3 third for both metrics.

Rank	Team
1	mxx
2	yseop_2
3	lipi_3
4	dicoe_2
5	dicoe_1
6	lipi_2
7	yseop_1
8	lipi_1
9	MiniTrue_2
10	MiniTrue_1
11	MiniTrue_3
12	Baseline_2
13	Baseline_1

Table 3: Overall Results

Team	Mean Rank	Accuracy
Baseline_1	1.941	0.564
Baseline_2	1.75	0.669
dicoe_1	1.180	0.889
dicoe_2	1.162	0.904
lipi_1	1.257	0.886
lipi_2	1.22	0.895
MiniTrue_1	1.346	0.855
MiniTrue_2	1.315	0.865
MiniTrue_3	1.337	0.825
mxx	1.113	0.941
yseop_1	1.236	0.883
yseop_2	1.141	0.917

Table 4: Mean Rank and Accuracy (listed alphabetically)

The teams presented the experiments based on a wide range of approaches in order to train a predictive model for the hypernym-hyponym relation in the financial domain covering 17 financial concepts. All the teams explored distributional semantic models like contextual word embeddings, BERT [Devlin *et al.*, 2019] and FinBERT [Araci, 2019], context-free embeddings, Word2Vec [Mikolov *et al.*, 2013], sentence-level embeddings, SRoBERTa [Reimers and Gurevych, 2019], knowledge graph embeddings, RotatE [Sun *et al.*, 2019]. Also, most of the teams investigated on the data augmentation methods by increasing significantly the initial volume of the training set. In this purpose, they used, either external sources of data, DBPedia, Investopedia along with the FIBO definitions, or similarity distance based approaches to cover the OOV words or to expand the size of the training data.

The data imbalance issue is addressed by the participants as in the previous editions, which we could not totally improve as some concepts are more frequently used in the real world financial documents than others beyond their importance.

5 Conclusion and Perspectives

The third edition of FimSim task attracted 5 teams and all of the system runs showed very performing results using a wide variety of NLP and ML/DL techniques and features like sentence-level embeddings, word embeddings, knowledge-graph embeddings, distance-based and hand-crafted features. The systems with the best performance achieved very high accuracy of 90.4%$\sim$94.1% comparing to the results of the previous editions of FinSim. All the participating systems largely explored not only semantic distributional methods for the similarity measures and classification but also the data augmentation methods, and the results showed that using distributed and contextual features for training a classifier on an augmented training set improve the performance of their systems. The impact of AI technologies grows more and more in financial domain like banking, investment fund management, stock market prediction, legal documents understanding, ESG (Environment, Social, Governance) scoring, etc. and this requires an investigation on how to exploit a large scale of financial domain's concepts and build a knowledge representation of those concepts. As the results of the participating systems show an improved coverage of concepts, the future task can propose a more challenging training set to consider all the main financial concepts defined in the FIBO ontology. Also, the current task is focused on a monolingual data processing. Knowing that the financial documents exist in different language and the majority of the financial concepts are language-independent, it will be interesting to extend the task to a multilingual data processing. Or, it will be also possible to improve the FinSim task by tackling new concepts related to ESG financial products.

6 Acknowledgments

This shared task is fully supported by Fortia Financial Solutions.

References

[Araci, 2019] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models, 2019.

[Bordea *et al.*, 2015] Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. SemEval-2015 task 17:

Taxonomy extraction evaluation (TExEval). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 902–910, Denver, Colorado, June 2015. Association for Computational Linguistics.

[Bordea *et al.*, 2016] Georgeta Bordea, Els Lefever, and Paul Buitelaar. SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1081–1091, San Diego, California, June 2016. Association for Computational Linguistics.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Glavaš *et al.*, 2020] Goran Glavaš, Ivan Vulić, Anna Korhonen, and Simone Paolo Ponzetto. SemEval-2020 task 2: Predicting multilingual and cross-lingual (graded) lexical entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 24–35, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[Greff *et al.*, 2017] Klaus Greff, Rupesh K. Srivastava, Jan Koutnik, Bas R. Steunebrink, and Jurgen Schmidhuber. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232, Oct 2017.

[Hauer *et al.*, 2020] Bradley Hauer, Amir Ahmad Habibi, Yixing Luan, Arnob Mallik, and Grzegorz Kondrak. UAlberta at SemEval-2020 task 2: Using translations to predict cross-lingual entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 263–269, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[Kovács *et al.*, 2020] Ádám Kovács, Kinga Gémes, Andras Kornai, and Gábor Recski. BMEAUT at SemEval-2020 task 2: Lexical entailment with semantic graphs. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 135–141, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[Maarouf *et al.*, 2020] Ismail El Maarouf, Youness Mansar, Virginie Mouilleron, and Dialekti Valsamou-Stanislawski. The FinSim 2020 shared task: Learning semantic representations for the financial domain. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 81–86, Kyoto, Japan, 5 January 2020. -.

[Mansar *et al.*, 2021] Youness Mansar, Juyeon Kang, and Ismail El Maarouf. *The FinSim-2 2021 Shared Task: Learning Semantic Similarities for the Financial Domain*, page 288–292. Association for Computing Machinery, New York, NY, USA, 2021.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Sun *et al.*, 2019] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space, 2019.

[Wang *et al.*, 2020] Shike Wang, Yuchen Fan, Xiangying Luo, and Dong Yu. SHIKEBLCU at SemEval-2020 task 2: An external knowledge-enhanced matrix for multilingual and cross-lingual lexical entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 255–262, Barcelona (online), December 2020. International Committee for Computational Linguistics.