

# DICoE@FinSim-3: Financial Hypernym Detection using Augmented Terms and Distance-based Features

Lefteris Loukas<sup>1,2</sup>, Konstantinos Bougiatiotis<sup>1</sup>, Manos Fergadiotis<sup>1,2</sup>, Dimitris Mavroeidis<sup>1</sup>, Elias Zavitsanos<sup>1</sup>

<sup>1</sup>Institute of Informatics and Telecommunications,

National Centre for Scientific Research “Demokritos”, Greece

<sup>2</sup> Department of Informatics, Athens University of Economics and Business, Greece  
{eloukas, bogas.ko, mfergadiotis, dmavroeidis, izavits}@iit.demokritos.gr

## Abstract

We present the submission of team DICoE for FINSIM-3, the 3rd Shared Task on Learning Semantic Similarities for the Financial Domain. The task provides a set of terms in the financial domain and requires to classify them into the most relevant hypernym from a financial ontology. After augmenting the terms with their Investopedia definitions, our system employs a Logistic Regression classifier over financial word embeddings and a mix of hand-crafted and distance-based features. Also, for the first time in this task, we employ different replacement methods for out-of-vocabulary terms, leading to improved performance. Finally, we have also experimented with word representations generated from various financial corpora. Our best-performing submission ranked 4th on the task’s leaderboard.

## 1 Introduction

Taxonomies constitute the backbone of many knowledge representation schemas, especially in the context of the Semantic web and ontologies. Such hierarchies model among others the *hyponymy* relation, a significant semantic relation between concepts. It is an asymmetric relation between two concepts, a hyponym (subordinate) and a hypernym (superordinate), as in “car-vehicle”, where the hyponym necessarily implies the hypernym, but not vice versa.

In the context of hypernym detection, the shared task on Learning Semantic Similarities for the Financial Domain (FINSIM-3) focuses on the evaluation of semantic representations by assessing the classification of a given list of terms from the financial domain against a domain ontology. A list of carefully selected terms from the financial domain is provided, such as “European depositary receipt”, “Interest rate swaps”, and others, and the task is to design a system that can classify them into the most relevant hypernym (or top-level) concept in an external ontology. The referenced ontology is the Financial Industry Business Ontology (FIBO).<sup>1</sup> For

<sup>1</sup> Visit <https://spec.edmcouncil.org/fibo/ontology> for more information.

instance, given the set of concepts “Bonds”, “Unclassified”, “Share”, “Loan”, the most relevant hypernym of “European depositary receipt” is “Share”. Figure 1 illustrates hypernym examples based on small parts of the FIBO ontology.

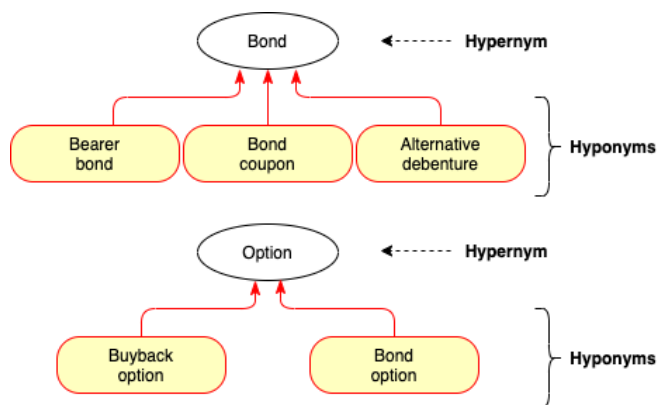


Figure 1: Examples of hypernym relations from the FIBO ontology. Interestingly, “Bond coupon” is a kind of “Bond”, but “Bond option” is a kind of “Option”.

In this paper, we present our solution to the FINSIM-3 task. Our system starts by augmenting all given terms with their Investopedia definitions. Then, we employ a Logistic Regression classifier over financial word embeddings and a mix of hand-crafted and distance-based features. Moreover, for the first time in this task, we explore various replacement methods for out-of-vocabulary (OOV) terms, leading to improved performance in our experiments. Our best-performing submission ranked 4th on the task’s leaderboard.

In what follows, Section 2 gives a brief overview of the related work in the context of the previous FINSIM tasks. Section 3 presents the data that are provided in this task. Then, Section 4 presents our solution to the task, while Section 5 provides empirical evaluation results. Finally, Section 6 summarizes the paper and presents future directions.

## 2 Related work

**Hypernym modeling:** Unsupervised hypernym modeling and classification mainly rely on measures assuming a dis-

tributional inclusion. This means that if a term  $c$  is semantically narrower than term  $p$ , then a number of distributional features of  $c$  should also be included in the feature vector of  $p$  [Lenci and Benotto, 2012]. Similarly, the work in [San-tus *et al.*, 2014] is based on the distributional informativeness hypothesis, which assumes that hypernyms tend to be less informative than hyponyms. Such distributional approaches rely on vector semantics and represent words as vectors.

Supervised methods are mainly based on word embeddings to represent words as low dimensional vectors in a latent space. Hypernym/hyponym pairs are encoded as combinations of two-word vectors, and hypernym relation classification is usually performed by training a classifier given the latter combinations of vectors as input [Baroni *et al.*, 2012; Roller *et al.*, 2014; Weeds *et al.*, 2014]. Other approaches rely on pre-extracted taxonomic relation data to create word embeddings that are later used as input to a Support Vector Machine (SVM) to learn the hypernym relation [Tuan *et al.*, 2016; Yu *et al.*, 2015].

In the context of FINSIM-3, we approach the task as a classification problem, aiming to classify input terms to their correct hypernym. We follow a supervised distributional approach without explicitly modeling hypernym relations or other ontological relations.

**FINSIM:** The first task to propose hypernym categorization in the financial domain was FINSIM-1 [Maarouf *et al.*, 2020], having a total tagset of 8 FIBO classes/hypernyms. The 1st-year winner system [Keswani *et al.*, 2020] combined rules and a Naive Bayes classifier over word2vec embeddings [Mikolov *et al.*, 2013], overperforming BERT [Devlin *et al.*, 2019] embeddings. The runner-up system [Saini, 2020] augmented all terms with their Investopedia definition and used a linear SVM over some hand-crafted and bi-gram TF-IDF features.

One year later, FINSIM-2 [Mansar *et al.*, 2021] held place, expanding the tagset to 10 financial hypernyms. The FINSIM-2 winners [Chersoni and Huang, 2021] used a Logistic Regression classifier over word embeddings, semantic and string similarities, along with BERT-derived masking probabilities to classify each term to a hypernym. The second-place system [Pei and Zhang, 2021] also used a Logistic Regression classifier over fine-tuned word embeddings derived from various financial text corpora.

Our system augments all terms with their Investopedia definitions, following [Saini, 2020], and incorporates hand-crafted and distance-based features. However, in contrast to previous works, we also experiment with different OOV replacement methods for unknown terms to deal with the many words in the dataset that are not contained in the vocabulary of the pre-trained word embeddings.

### 3 Data

In this section, we briefly present the data provided by the task organizers, which include the training dataset, the FIBO ontology, the prospectus corpus, and the word embeddings.

**Dataset:** The dataset consists of one-word or multi-word concepts from the financial domain and their labels. It is

Table 1: Class distribution in the training set.

Class	Count
Equity Index	286
Regulatory Agency	205
Credit Index	129
Central Securities Depository	107
Debt pricing and yields	58
Bonds	55
Swap	36
Stock Corporation	25
Option	24
Funds	22
Future	19
Credit Events	18
Stocks	17
MMIs	17
Parametric schedules	15
Forward	9
Securities restrictions	8

separated into training and test sets, with the former being released a month before the test set.

The training set comprises 1050 examples with their corresponding labels. The unique labels are 17 in total, and their frequencies in the training data are provided in Table 1 in descending order. The most frequent label is *Equity Index*, which appears in 27% of the training examples, while the rarest labels are *Forward* and *Securities restrictions*. Both of them appear less than 10 times in the training data.

**FIBO Ontology:** The Financial Industry Business Ontology (FIBO) is a pioneering effort to formalize the semantics of the financial domain using a large number of ontologies. At the time of this writing, FIBO is still a work in progress. However, it already defines large sets of concepts that are of interest in financial business applications and how these concepts relate to one another.

For each concept, FIBO provides additional textual information, including a definition, a generated description, labels, titles, and in many cases, a small abstract. We combine the FIBO textual information with the prospectus corpus (see below) for training custom embeddings. In particular, we traverse the ontologies and concepts related to the labels in the training set, and we fetch the corresponding textual information. This way, we augment the prospectus corpus with additional concept-specific documents from the ontology that include small snippets, descriptions, and definitions.

**Prospectus corpus:** A corpus of documents is also provided in the English language for training word embeddings. The corpus has been compiled from various websites, comprising financial prospectuses, and it consists of approximately 14M tokens. Those files are given in pdf format. We have used this corpus and text parts of the FIBO ontology to train custom word embeddings.

**Word embeddings:** Finally, two sets of pre-trained word embeddings are also provided in the context of the task. Both

sets are based on the Word2Vec model and trained in an internal financial corpus by the organizers, which comprises key investor information documents and financial prospectuses. The difference between the two word embeddings is the number of dimensions of the vectors and the vocabulary size. The first comprises 100-dimensional vectors and 17328 words, while the second has 300 dimensions and 34437 words.

## 4 System description

Our system extends baseline 2, which is a Logistic Regression classifier over the word2vec embeddings from the FINSIM-3 organizers. Figure 2 illustrates the pipeline of our system. After augmenting all terms with their Investopedia definitions, we concatenate and scale the OOV-aware embeddings with the hand-crafted and distance features. We then use a Logistic Regression classifier over these features.

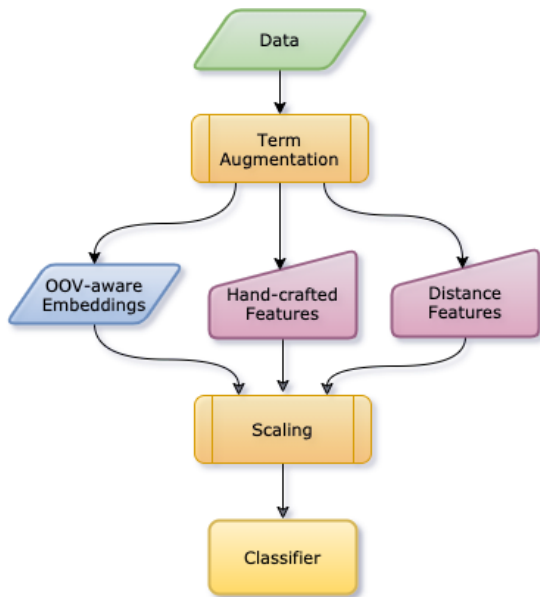


Figure 2: The pipeline of our best system.

In the following subsections, we present how we worked towards building our pipeline.

### 4.1 Feature Engineering

First, we created a set of hand-crafted features that are indicative of specific classes. In order to gain insights on terms that could be indicative of each class, we did a preliminary error analysis using the provided baseline 2 system as a predictor. We devised 7 simple boolean hand-crafted features that denoted the existence of a specific string in the term. These strings were common in miss-classified terms of the baseline model while also being highly indicative of the true class of the term. For example, the occurrence of the string “Inc.” was very common in errors of the baseline, while at the same time, almost all terms having this string belonged to the *Stock Corporation* class.

Moreover, specific classes such as *Credit Index* have a lot of upper-case letters or unusual patterns. To capture these

intricacies, we have also added 3 features that correspond to the number of characters in the term, the number of upper-case letters, and the ratio of upper-case to lower-case letters. We ended up with 10 such hand-crafted features.

Furthermore, in order to represent the latent space distance between the terms and the classes, we calculated the cosine distance between the term’s embedding and each class’ embedding, adding 17 features in total.<sup>2</sup> Last but not least, we calculated the Levenshtein distances between the term and each class label, adding another 17 features. After concatenating all of the features, we scale them to a uniform range of [-1, 1].

### 4.2 Out-of-Vocabulary Words

Inside the 1050 training terms, we found at least 214 words that were out of vocabulary, using the 300-dimensional organizers’ embeddings. Such words were mainly instances of credit indexes and regulatory agencies. Since terms contain a single word or a small number of words, we identified the need to deal with OOV occurrences instead of using zero embeddings. For example, the word “asiacorporate” and “t-bill” are OOV and they are represented by zero embeddings. Ideally, we would like to match them correctly to the most similar in-vocabulary words “corporate” and “treasury-bill” in order to retrieve the best vector representations available to help the classification process.

First, we tried replacing each OOV word with its closest in-vocabulary word in terms of Levenshtein distance, which helped performance. Then, we replaced this relatively simple mechanism by utilizing the Magnitude toolkit [Patel *et al.*, 2018]. Magnitude works as a wrapper for already-trained word2vec models. It allows advanced OOV lookup as it combines different character n-grams, string similarities, and morphology-aware matching.

### 4.3 In-domain representations

We use the 300-dimensional embeddings provided by the organizers for the word vector representations since they perform better in the development set than the 100-dimensional ones.

We also experimented with a wide variety of other domain-specific word representations. First, we trained our financial word2vec embeddings (d=200) on the prospectus corpus and the FIBO ontology provided. We extracted the text from the prospectuses PDF files using the provided pdf-to-text toolkit<sup>3</sup>. However, they proved to be less beneficial than the provided embeddings.

Furthermore, we extracted the embedding layer of FINBERT [Yang *et al.*, 2020], a financial BERT, and converted it to word2vec format. To our surprise, the provided embeddings outperformed the FINBERT embeddings. The provided embeddings also outperformed our custom word2vec and BERT embeddings, trained on financial documents from the Securities & Exchange Commission.

<sup>2</sup>The vector representation for multi-token terms/classes is the sum of each token’s vector representation. We also tried the centroids of the embeddings which reduced the performance.

<sup>3</sup>Consult <https://poppler.freedesktop.org/> for more information.

#### 4.4 Augmentation with Investopedia Terms

During the later stages of our experimentation, we noticed that many of the misclassifications were attributed to acronyms (e.g. "FRN") or common words being present ("Long call/put"). In order to alleviate this problem and provide more context for all terms, we utilized Investopedia<sup>4</sup> as a dictionary of definitions for each term. To do this, we built a scraper that pinpoints the closest match of a given term that has a definition in the terms dictionary of the website. The scraper first tries to find exact matches of the query term in the dictionary of the website. If this fails, we utilize the search functionality of the site to identify the closest matching term. Having found a corresponding match (exact or approximate), we fetch the definition of the matched term and keep only the first sentence of the definition, as it usually is in the format: "[Term] is a ..", where [Term] denotes the matched term for the given query term.

This process is followed both for augmenting the initial training data and the final test terms, as it can be easily incorporated at inference time. Following the above process, approximately 70% of both the train and test terms were augmented. In case the augmentation process did not retrieve any definition for a given term (this was common for *Credit Indexes* for example), the term was left as is.

#### 4.5 Classifier tuning

Apart from Logistic Regression, we also evaluated a battery of different classifiers implemented in the scikit-learn library, like the Naive Bayes Classifier, Decision Trees, linear SVMs, Multi-Layer Perceptron, XGBoost, and RUSBoost [Seiffert *et al.*, 2008], without any improvement in performance. Indicatively, classifiers based on trees (Decision Trees, XGBoost, RUSBoost) scored the lowest, possibly due to the extensive feature space.

Thus, we chose to continue with the simple -yet powerful- Logistic Regression classifier, where we tuned the regularization strength hyperparameter C. We defined a search space of {0.001, 0.01, 0.1, 1.0, 10.0, 100.0}. We found that C=0.1 is the best option in terms of mean rank and accuracy based on a stratified 5-fold cross-validation setting.

## 5 Experimentation

### 5.1 Metrics

We evaluate our performance using accuracy, mean rank and macro-average F1 score, as shown in equations 1, 2 and 3.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$MeanRank = \frac{1}{n} * \sum_{i=1}^n rank_i \quad (2)$$

$$MacroF1 = \frac{1}{n} * \sum_{i=0}^n F1_i \quad (3)$$

In the context of the shared task, apart from accuracy, we also had to generate all labels in ranked order and measure

<sup>4</sup><https://www.investopedia.com/>

Table 2: Experimental results based on stratified 5-fold cross validation. Results are shown using a tuned Logistic Regression Classifier (C=0.1).

Model	Mean Rank	Accuracy	Macro F1
BL	1.196	87.6	80.0
BL.HF	1.166	90.0	82.0
BL.HF.OOV <sub>1</sub>	1.156	90.5	83.0
BL.HF.OOV <sub>1</sub> .D	1.148	90.7	84.0
BL.HF.OOV <sub>1</sub> .D <sup>2</sup>	1.147	90.8	83.8
BL.HF.OOV <sub>m</sub> .D <sup>2</sup>	1.144	91.2	84.1
<b>BL.HF.OOV<sub>m</sub>.D<sup>2</sup>.+</b>	<b>1.132</b>	<b>91.5</b>	<b>85.0</b>

the mean rank. For each term  $x_i$  with a label  $y_i$  from the  $n$  samples in the test set, the expected prediction is a top-3 list of labels ranked from most to least likely to be equal to the ground truth. In equation 2,  $rank_i$  is 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.

### 5.2 Results

Table 2 presents the experimental results of our system's variations. We used the Logistic Regression classifier in a stratified 5-fold cross-validation setting. Since the training set is small and imbalanced we selected that setting in order to ensure that all classes will be represented in each fold. We provide empirical results using the best hyperparameters after tuning (see Subsection 4.5). In particular, the following variations were evaluated:

- **BL**: The baseline model. Logistic Regression classifier with the given input embeddings as features.
- **BL.HF**: BL with additional hand-crafted features (Subsection 4.1).
- **BL.HF.OOV<sub>1</sub>**: BL.HF plus the Levenshtein-based OOV words handling (Subsection 4.2).
- **BL.HF.OOV<sub>1</sub>.D**: BL.HF.OOV<sub>1</sub> plus additional features based on the cosine distance between term and class embeddings (Subsection 4.1).
- **BL.HF.OOV<sub>1</sub>.D<sup>2</sup>**: BL.HF.OOV<sub>1</sub> plus the character distance between terms and classes (Subsection 4.1).
- **BL.HF.OOV<sub>m</sub>.D<sup>2</sup>**: BL.HF plus the Magnitude-based OOV words handling (Subsection 4.2). This variation constitutes the first submission of our system (DICOE 1) to the shared task.
- **BL.HF.OOV<sub>m</sub>.D<sup>2</sup>.+**: This is BL.HF.OOV<sub>m</sub>.D<sup>2</sup> using Investopedia-based augmented terms (Subsection 4.4). This variation constitutes the second submission of our system (DICOE 2) to the shared task.

Table 2 shows that incorporating the hand-crafted features improves the baseline by 2.4% in terms of accuracy. The mean rank also improves from 1.196 to 1.166, suggesting that simple substring binary features may indicate the specific class of each term. Then, we first leverage the Levenshtein distance to replace each OOV word found in the terms with its closest in-vocabulary word. This boosts accuracy by 0.5%,

while the mean rank is reduced to 1.156. Thus, handling OOV words with replacements allows us to retrieve better vector representations than zero embeddings. Our next improvement combines the Levenshtein character distance between the term’s words and the class labels, as well as the cosine distance between their representations in the latent space, improving accuracy to 90.8% and mean rank to 1.147.

Next, we combine all of the above and replace the simple Levenshtein OOV mechanism with the Magnitude toolkit [Patel *et al.*, 2018]. This advanced OOV lookup method scored 91.2% in terms of accuracy and 1.144 in terms of mean rank and represents our first submission to the FINSIM-3 shared task (see BL.HF.OOV<sub>1</sub>.D<sup>2</sup> in Table 2).

Our final system extends the first submission by augmenting the financial terms with their Investopedia definitions (Section 4.4) in order to provide more context for the classification. This was our best system and scored a 1.132 mean rank, 91.5% accuracy, and 85.0% Macro F1 Score in the 5-fold cross validation (see BL.HF.OOV<sub>m</sub>.D<sup>2</sup>+ in Table 2).

## 6 Conclusions and Future Work

We presented DICOE team’s submissions to FINSIM-3. Our Investopedia-augmented system ranked 4th on the leaderboard. We leveraged hand-crafted and distance-based features which led to significant improvements over the baseline. To our surprise, external and modern financial word representations, such as FINBERT, did not contribute positively to the results. Moreover, for the first time in this shared task, we introduced the application of OOV word replacement methods. Using OOV replacements, we can successfully retrieve correct vector representations for unknown tokens that share the same morphology with in-vocabulary words.

In future work, we plan to investigate other ways of augmenting terms with their definitions and broad context, as well as creating new external financial resources for experimentation. An additional future direction is the direct modeling of the hypernym relation using pairs of tokens and labels to explicitly learn that type of relation.

## References

- [Baroni *et al.*, 2012] Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung chieh Shan. Entailment above the word level in distributional semantics. In *13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 23–32, Avignon, France, 2012.
- [Chersoni and Huang, 2021] Emmanuele Chersoni and Churen Huang. *PolyU-CBS at the FinSim-2 Task: Combining Distributional, String-Based and Transformers-Based Features for Hypernymy Detection in the Financial Domain*, page 316–319. New York, NY, USA, 2021.
- [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.
- [Keswani *et al.*, 2020] Vishal Keswani, Sakshi Singh, and Ashutosh Modi. IITK at the FinSim task: Hypernym detection in financial domain via context-free and contextualized word embeddings. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 87–92, Kyoto, Japan, 5 January 2020.
- [Lenci and Benotto, 2012] Alessandro Lenci and Giulia Benotto. Identifying hypernyms in distributional semantic spaces. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics, Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval)*, pages 75–79, Montreal, Canada, 2012.
- [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. In Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia, editors, *Companion of The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 288–292. ACM / IW3C2, 2021.
- [Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.
- [Patel *et al.*, 2018] Ajay Patel, Alexander Sands, Chris Callison-Burch, and Marianna Apidianaki. Magnitude: A fast, efficient universal vector embedding utility package. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 120–126, Brussels, Belgium, November 2018.
- [Pei and Zhang, 2021] Yulong Pei and Qian Zhang. *GOAT at the FinSim-2 Task: Learning Word Representations of Financial Data with Customized Corpus*, page 307–310. New York, NY, USA, 2021.
- [Roller *et al.*, 2014] Stephen Roller, Katrin Erk, and Gemma Boleda. Inclusive yet selective: Supervised distributional hypernymy detection. In *25th International Conference on Computational Linguistics (COLING)*, pages 1025–1036, Dublin, Ireland, 2014.
- [Saini, 2020] Anuj Saini. Anuj@FINSIM-Learning Semantic Representation of Financial Domain with Investopedia. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 93–97, Kyoto, Japan, 5 2020.

- [Santus *et al.*, 2014] Enrico Santus, Alessandro Lenci, Qin Lu, , and Sabine Schulte Im Walde. Chasing hypernyms in vector spaces with entropy. In *14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 38–42, Gothenburg, Sweden, 2014.
- [Seiffert *et al.*, 2008] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Rusboost: Improving classification performance when training data is skewed. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, 2008.
- [Tuan *et al.*, 2016] Luu Anh Tuan, Yi Tay, Siu Cheung Hui, and See Kiong Ng. Learning term embeddings for taxonomic relation identification using dynamic weighting neural network. In *In Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 403–413, Austin, Texas, 2016.
- [Weeds *et al.*, 2014] Julie Weeds, Daoud Clarke, Jeremy Reffin, David J. Weir, and Bill Keller. Learning to distinguish hypernyms and co-hyponyms. In *25th International Conference on Computational Linguistics (COLING)*, pages 2249–2259, Dublin, Ireland, 2014.
- [Yang *et al.*, 2020] Yi Yang, Mark Christopher Siy UY, and Allen Huang. Finbert: A pretrained language model for financial communications. *arXiv*, abs/2006.08097, 2020.
- [Yu *et al.*, 2015] Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. Learning term embeddings for hypernymy identification. In *24th International Conference on Artificial Intelligence (IJCAI)*, pages 1390–1397, Buenos Aires, Argentina, 2015.