Proceedings of the 4th Financial Narrative Processing Workshop @ LREC 2022, pages 17–26
Marseille, 24 June 2022

Sentiment Classification by Incorporating Background Knowledge from Financial Ontologies

Timen Stepišnik-Perdih[1,3], Andraž Pelicon[1,2], Blaž Škrlj[1,2], Martin Žnidaršič[1], Igor Lončarski[4], Senja Pollak[1]
[1]Jožef Stefan Institute, Ljubljana, Slovenia
[2]Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[3]Faculty of Computer and Information Science, Ljubljana, Slovenia
[4]School of Economics and Business, University of Ljubljana, Slovenia

tstepisnikp@gmail.com
igor.loncarski@ef.uni-lj.si
{andraz.pelicon, blaz.skrlj, martin.znidarsic, senja.pollak}@ijs.si

Abstract

Ontologies are increasingly used for machine reasoning over the last few years. They can provide explanations of concepts or be used for concept classification if there exists a mapping from the desired labels to the relevant ontology. This paper presents a practical use of an ontology for the purpose of data set generalization in an oversampling setting, with the aim of improving classification models. We demonstrate our solution on a novel financial sentiment data set using the Financial Industry Business Ontology (FIBO). The results show that generalization-based data enrichment benefits simpler models in a general setting and more complex models such as BERT in low-data setting.

Keywords: Sentiment classification, Financial ontology, Generalization

1. Introduction

From the perspective of financial economics, capturing and understanding the impact of non-financial information, such as sentiment or subjectivity conveyed in textual (language) form, has become increasingly more important. Identification and estimation of the sentiment are important in order to better understand and be able to predict investor behavior and the impact on supply and demand for financial assets and, in turn, the effect on asset prices, mainly from the perspective of disentangling fundamental drivers of asset prices from those based on perceptions/sentiment.

There has been a range of natural language processing approaches to automatically assess financial sentiment from texts (Man et al., 2019; Xing et al., 2020), which can be categorized into unsupervised, semi-supervised and supervised approaches. Financial text sentiment analysis is most frequently used for predictive analytics on financial markets (e.g., (Jin et al., 2019; Xing et al., 2018; Day and Lee, 2016; Smailović et al., 2014)), while on the other hand, a growing body of literature (e.g. (Smailović et al., 2017b)) is dedicated to the analysis of relations between financial and non-financial information in financial reports, which is motivated by the fact that the issue of the quality of financial reporting has become one of the central issues during the recent financial crisis and has received considerable attention from the society at large ever since.

Domain ontologies are becoming increasingly available. Containing background knowledge in a computer-readable form inspires the creation of new systems that try to solve problems in a way, more similar to domain experts. Provision of semantic information allows the learner to use features on a higher semantic level, possibly enabling better data generalizations. The methods, leveraging background knowledge (from domain or general resources), have been proposed in various fields (e.g. biology (Kim et al., 2018; Chang et al., 2015), sociology (Freeman, 2017)), short text classification (e.g. (Škrlj et al., 2020), fake news detection (Koloski et al., 2021)). Developing and using domain ontologies in financial economics could thus facilitate more accurate identification and classification of sentiment.

This paper discusses the use of background knowledge in the form of financial ontologies, more specifically the FIBO ontology, for improving classification models by text generalizations. While FIBO ontology has been previously used in automated approaches to classify the financial concepts (Stepišnik Perdih et al., 2021b), the potential of domain ontologies has not yet been sufficiently exploited for financial sentiment analysis. We use FIBO for text generalization, and more specifically assess it as a method for oversampling, where one transforms the original data set so that the new one is potentially more suitable for learning with the aim of improving a model's performance. The main contributions of this paper are as follows:

- we propose new text generalization methods using FIBO financial ontology;

- we assess their potential to be used in financial sentiment classification tasks using simple symbolic as well as neural transformer models in high- and low-data settings;
- we evaluate the method on a novel sentiment-annotated data set of random sentences from a selection of annual reports of companies listed on US or UK stock exchanges.

Our paper is structured as follows. In Section 2 we discuss work related to this paper: approaches modeling sentiment in financial texts and data upsampling approaches. Section 3 presents the financial sentiment data we use in our study and the Financial Industry Business Ontology (FIBO) which we use as background knowledge. In Section 4 we discuss the methodology of term generalization and our approaches of enriching data sets with generalized terms. In Section 5 we lay out the experimental evaluation of our methods and present the results. We draw conclusions and discuss further work in Section 6 and in Section 7 we discuss the reproducibility of our experiments.

2. Related work

There has been a range of natural language processing approaches developed to automatically assess financial sentiment from texts (Man et al., 2019; Xing et al., 2020). Financial sentiment analysis can be performed on various data sources including microblog posts (Cortis et al., 2017a), news (Cortis et al., 2017a) or corporate disclosures. El-Haj et al. (2016) gathered a dataset, similar to the one presented in this work, and annotated it for tone expressed in the text. In contrast to our dataset, which was gathered from financial reports, their dataset was gathered from earning announcements of UK comapnies.

In terms of annual reports, which are also the source of our data, several approaches have been proposed for prediction of financial phenomena such as: next year performance through indicators such as return on equity (Qiu et al., 2006; Butler and Kešelj, 2009; Li, 2010; Balakrishnan et al., 2010), contemporaneous returns around filing dates (Feldman et al., 2008; Amel-Zadeh and Faasse, 2016), stock return volatility (Kogan et al., 2009; Loughran and McDonald, 2011a), earnings forecast dispersion (Kothari et al., 2009; Loughran and McDonald, 2011a), costs of capital (Kothari et al., 2009), financial distress (Hájek and Olej, 2013; Hajek et al., 2014) and bank failure (Gupta et al., 2016). Another line of research (e.g. (Smailovic et al., 2017a), is dedicated to the analysis of relations between financial and non-financial information in financial reports, which is motivated by the fact that the issue of the quality of financial reporting has become one of the central issues during the recent financial crisis and has received considerable attention from the society at large ever since.

In terms of methods, we can distinguish between dictionary-based, supervised and hybrid methods. In the first category, the collection of dictionaries by (Loughran and McDonald, 2011b) is the most widely-used resource. In addition, general lexica like Opinion Lexicon (Hu and Liu, 2004) and MPQA Subjectivity Lexicon (Wilson et al., 2009) are being used by various researchers (e.g. (Chen et al., 2013; Goel and Uzuner, 2016) including by high-ranked teams in SemEval 2017 competition(Cortis et al., 2017a).

On the other side, supervised approaches are being developed. In older research, a lot of attention has been put on feature engineering, and several algorithms have been used. In the context of analyses of the financial reports it has been employed to categorize tone and content of forward-looking statements in 10-K fillings (Li, 2010) and to detect financial constraints based on word stem frequencies (Buehlmaier and Whited, 2015). Decision trees are not common in financial sentiment analysis, but were used among several other approaches in the study by (Hajek et al., 2014) on relations of report text sentiments and financial performance indicators, Random Forest approach has been used to predict short-term stock price changes on the basis of sentiment in 8-K reports (Lee et al., 2014), other non-neural approaches use logistic regression (e.g. (Hajek et al., 2014)), while the most frequently used algorithm is Support Vector Machine (SVM), e.g. in fraud detection models (Goel and Uzuner, 2016), classification of companies as out-performing or under-performing on the basis of narrative of disclosures (Balakrishnan et al., 2010), discriminating between failed and non-failed banks based on the sentiment of their reports (Gupta et al., 2016), picking out financially distressed companies on the basis of sentiment in reports (Hájek and Olej, 2013; Hajek et al., 2014), predicting financial risk from text features of reports (Kogan et al., 2009), predicting risk through stock return volatility on the basis of sentiment (Wang et al., 2013) or ranking companies as to their risk level on the basis of textual information on their reports (Tsai and Wang, 2012).

Several recent works tackled the problem of modeling sentiment in financial texts using deep learning methods. (Zhang et al., 2018) developed a neural architecture based on gated recurrent units which embedded textual and user information into a shared embedding space for mining financial opinions (e.g., bullish or bearish) from Twitter data. (Dong and Liu, 2021) note that quality annotated data for financial sentiment classification is scarce. They try to mitigate this limitation by training their convolutional neural network model on cross-domain data with the addition of an adversarial domain-adaptation module. (Araci, 2019) performed additional pretraining of the original BERT language model on texts from the financial domain. The updated finBERT model has shown improvement on two financial sentiment analysis data sets over the baselines. (Lee, 2021) use the adapted finBERT model

in their work to train a financial sentiment classifier on social media posts and include its predictions as features for predicting stock returns. Additionally, through investigation of feature importance, they are able to quantify the impact these features have on stock return predictions.

The branch of research of high relevance to the presented publication considers *data upsampling*. This process, given the input data set, outputs a *transformed* data set which is potentially more suitable for learning. Upsampling regimes can be based solely on the input data (Halterman and Radford, 2021), however, upsampling based on external knowledge has also been of increasing interest in the last decades (Schneider et al., 2016; Lu et al., 2006). Incorporation of taxonomy-like background knowledge, however, was recently also shown to have performance-beneficial effects when considering texts(Škrlj et al., 2020). Semantic enrichment has shown promising results also when annotating scientific literature (Bertin and Atanassova, 2012). Another line of research related to data upsampling is *data augmentation*. With this process, given the original dataset, we obtain an upsampled dataset by adding slightly modified instances of the original instances or newly created synthetic instances. Several data augmentation approaches were developed explicitly for textual data: using WordNet as a dictionary to randomly replace words/phrases with their synonyms in an instance (Zhang et al., 2015), replacing words using the nearest neighbour of the word from a given word embedding (Wang and Yang, 2015), or replacing random words in a sentence based on the predictions of those words from a BERT model conditioned on the label for a particular instance. (Wu et al., 2019)

3. Data

In this section, we introduce the corpus of annotated sentences that we have used (Section 3.1), as well as the financial ontology used in our generalization method (Section 3.2).

3.1. Corpus

For our experiments, we have created a new data set of sentences from annual reports of companies that are listed on US or UK stock exchanges and cover the period between the years 2017 and 2019. Reports in PDF format were transformed into raw texts with the pdfminer[1] library, as well as some post-processing editing steps. Annual reports were first split into sentences and for each annotator, we created a data set containing 480 randomly sampled sentences. In order to include proper and relevant sentences from the reports, we have included only sentences from the first part of the report that begins with a capital letter and end with a full stop, contain at least 20 words or numbers and where at most 15% of characters are numeric. We additionally randomly sampled 20 sentences from the reports for annotation by all the annotators, for a total of 500 sentences per annotator. For annotating the data set, we engaged thirteen annotators. Annotators were the second-year graduate students of MSc in Quantitative Finance and Actuarial Sciences at the School of Economics and Business, University of Ljubljana. Given their field and length of studies, we believe they were very much suitable for the task of annotating financial texts from the perspective of domain experts in the area of financial sentiment. Annotators were then asked to annotate each of the sentences according to several criteria. First, whether the sentence is relevant from the perspective of corporate business. Second, whether the sentence conveys positive/negative/neutral financial sentiment. Third, whether the sentence expresses an opinion (subjectivity) or states the facts (objectivity). Four, whether it is forward-looking or not. Finally, whether it relates to sustainability issues or not. In this work, we are using the labels with regards to the sentiment of the sentences to train a financial sentiment text classifier. The financial sentiment classification is posed as a three-class classification problem where each sentence can be classified as either positive, negative or neutral.

Given that our data set was annotated by several annotators, we estimate the agreement in annotations using labels for the 20 common sentences which were labeled by all the annotators. The inter-annotator agreement was estimated using Krippendorff's Alpha-reliability (Krippendorff, 2018), an established measure for estimating agreement between human annotators. While the measured alpha reliability was relatively low (α=0.3937) we note that it is comparable to other studies in the domain of sentiment analysis, especially when it comes to annotating sentiment in short texts (Pelicon et al., 2020; Bobicev and Sokolova, 2017; Santos et al., 2021). The low scores can be attributed to the fact that sentiment classification is a hard and rather subjective task.

This final financial sentiment data set was additionally preprocessed before conducting the experiments. We included the 20 common sentences, originally used to calculate the inter-annotator agreement, in the data only once and averaged the labels from several annotators into the final gold standard label. Next, we removed all the instances that were not labeled by the annotators. The final data set used for experiments contained 5994 labeled instances. The class distribution of this data set is presented in Table 1.

positive	neutral	negative	total
2194	3033	767	5994

Table 1: Class distribution of the financial sentiment classification data set.

We opted for the development of this specific new data set as most of the available financial texts with sentiment annotations originate from social media or news

[1] https://github.com/euske/pdfminer

(e.g., (Cortis et al., 2017b)) and we are not aware of any suitable sentiment annotations of texts from annual reports.

3.2. The Financial Industry Business Ontology

Ontologies are studies of all that exists in a given domain. In information science, an ontology is a model representing knowledge as a set of concepts connected with different relations. They are often represented by a directed graph where nodes represent concepts of the domain and edges represent relations connecting concepts.

The Financial Industry Business Ontology (FIBO), that we use in our generalization approach presented in Section 4, defines the sets of things that are of interest in financial business applications and the ways that those things can relate to one another. In this way, FIBO can give meaning to any data (e.g., spreadsheets, relational databases, XML documents) that describe the business of finance (fib, 2021a). Visualization of a part of FIBO is presented in Figure 1.

In this work we considered the following FIBO relations: "subClassOf", "isProvidedBy", "type", "isUsedBy", "isMemberOf", "hasJurisdiction" and "isPartOf", which connect 36,344 FIBO concepts. These represent the subset of FIBO that we use.

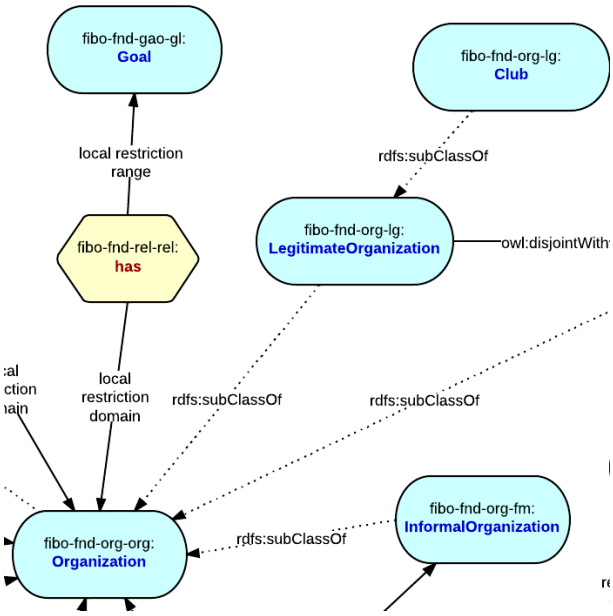

Figure 1: Visualization of a part of FIBO (fib, 2021b). The graph shows domain entities like "Organization", "LegitimateOrganization" and "Club" connected with relations like "subClassOf".

4. Methodology of term generalization for data set enrichment

In this section, we discuss generalizing terms using a domain ontology. In section 4.1, we first explain how we generalize financial terms using the Financial Industry Business Ontology (which was introduced in Section 3.2), and next, Section 4.2 proposes two generalization-based data enrichment methods.

4.1. Ontology-based generalization

Semantic reasoning from model-agnostic explanations (Stepišnik Perdih et al., 2021a) introduces a way of generalizing sets of terms using a domain ontology represented as a directed graph. It uses relations within the ontology (edges in the graph) that connect terms to more general ones. Each term is generalized relation by relation (in steps) until found generalization(s) are too connected to terms of other sets we are generalizing. This way the resulting sets contain more general terms but remain specific because we control the allowed intersection between terms of different sets during the process of generalization.

We have modified the mentioned approach so that, instead of generalizing sets of terms that represent model explanations, it generalizes individual terms found in the financial sentiment classification data set. Each resulting set contains all found generalizations of a single term.

The search for generalizations can be **constrained** or a **full** search of the ontology. The constrained setting only considers found generalizations that also satisfy the condition of being specific for the given term, meaning that generalizations common to multiple terms (to more than 1% of terms) are not considered, while the full ontology search generalizes each term to its top-level generalizations.

4.2. Data set enrichment

In this section, we describe ways in which we enrich the data set using acquired generalizations with the aim of improving the performance of prediction models. We try swapping terms we have successfully generalized with their generalizations as described in 4.2.1 and concatenating found generalizations of terms present in a sentence at the end of the sentence as described in 4.2.2. Both methods support the constrained and the full ontology-based generalization search.

Before any of the two approaches is employed we lemmatize the sentences with Lemmagen3 (pyp, 2021) so that we can recognize the terms in sentences for which generalizations have been found.

4.2.1. Term swapping

With term swapping, we augment the train subset with new samples. We acquire the new samples by swapping terms in sentences with their generalizations. This is done by iterating over terms that have been generalized and creating t new samples from each sentence that contains at least one occurrence of the term, where t is the number of found generalizations for that term. These new samples are immediately added to the subset so that other terms can be generalized in the next

iteration. We also keep the original sample in the subset. In each iteration, we get $n \cdot t$ new samples, where n is the number of samples in the subset containing the term we are swapping.
Because in this way the number of samples increases very quickly we introduce the parameter k which serves as a target factor of upscaling the number of samples in the train subset. If after any iteration, the number of samples in the subset is larger than $k \cdot N$, where N is the number of original samples in the subset before the swapping, the swapping stops.
Let us look at an example of a financial sentiment classification text after lemmatization:

- through real-time information and visualisation, *USA* help reduce business waste

and two of the new training instances acquired using term swapping with the constrained search of FIBO generalizations:

- through real-time information and visualisation, *united states of america* help reduce business waste
- through real-time information and visualisation, *geographic region identifier* help reduce business waste

4.2.2. Generalization concatenation

Generalization concatenation keeps the number of instances but appends all possible generalizations for FIBO terms found in a sentence to the end of the sentence of a training set.
An example of this approach is:

- through real-time information and visualisation, USA help reduce business waste, *code element, united states of america, geographic region identifier*

5. Experiments

In this section, we evaluate our method of enriching the data sets used for model training. First, we describe the train and test split of the evaluation data sets (Section 5.1). Next we present the models used in our experiments (section 5.2), followed by presenting the experimental setting (Section 5.3), and finally describing the results (Section 5.4).

5.1. Evaluation data sets

We split the data set of 5994 into train and test subsets with a random 10% of samples being included in the test subset. The proposed methods were benchmarked in both high- and low-data settings. In the high-data setting, the methods were benchmarked using the whole training data set (train_{ALL}). In the low-data setting, we have reduced the training data set to the 10% of the size of the original training data (train_{LOW}) while keeping the class distribution intact. The test set was kept the same in both experimental settings. Class distribution of the two subsets is presented in Table 2.

subset	positive	neutral	negative	total
train_{ALL}	1991	2731	672	5394
train_{LOW}	199	273	67	539
test	203	302	95	600

Table 2: Class distribution of train and test subsets. The number of training instances differs in the high- and low-data setting experiments, while the test set is the same.

A total of **818** different terms were generalized and the train subset contains an average of **11.04** occurrences of these terms per sentence. Table 3 shows examples of frequently generalized terms and their generalizations.

term	generalization
group	collection
report	document
executive	agent in role
customer	agent in role
shareholder	agent in role
future	agreement

Table 3: Examples of some of the most frequently generalized terms and one of their possible generalizations. Generalizations were found using the constrained search of the ontology.

5.2. Models

For the evaluation of our method we use the following models: logistic regression with doc2vec (*lr-doc2vec*) (Le and Mikolov, 2014), linear regression using character features (*lr-char*), linear regression using word features (*lr-word*), Support Vector Machine classifier using "all-mpnet-base-v2" representations from Simple Transformers (*svm-mpnet*), Support Vector Machine classifier using character features (*svm-char*), Support Vector Machine classifier using word features (*svm-word*) and TPOT (*tpot*) - an AutoML tool based on genetic programming which learns to normalize and model the data based on an internal validation procedure (Le et al., 2020). Because this approach is not able to preprocess raw text data, word features are extracted from the input documents using TF-IDF.
Additionally, we test our oversampling method in combination with the fine-tuning technique for transformer-based language models. For this purpose, we use two monolingual language models based on the BERT architecture, namely the original base version of the BERT language model (Devlin et al., 2019) and the finBERT language model (Araci, 2019) which was additionally trained on a corpus of unlabeled financial texts. For fine-tuning a classifier based on the language model, we added a linear layer with a softmax activation function at the output to serve as the classification layer. As input to the classifier, we took the

representation of the special (CLS) token from the last layer of the language model. The whole model was then jointly trained on the downstream task of financial sentiment classification. During training, we split the original training set into training and validation subsets in 90%-10% ratio. We used the Adam optimizer with the learning rate of $2e-5$ and learning rate warmup over the first 10% of the training instances. We used a weight decay set to 0.01 for regularization. All models were trained for maximum of 3 epochs with batch size 32. We performed the training of the models using the HuggingFace Transformers library (Wolf et al., 2019). We tokenized the textual input for the neural models with the respective language model's tokenizer. For performing matrix operations efficiently, all inputs were adjusted to the same length, which is a standard procedure. After tokenizing all inputs, their maximum length was set to 256 tokens. Longer sequences were truncated, while shorter sequences were zero-padded.

5.3. Evaluation setting

Models described in 5.2 were trained on the unchanged training set that we use as a baseline (*baseline)* and subsets enriched with term swapping *(swp)* introduced in Section 4.2.1 or generalization concatenation (*cnct*) introduced in Section 4.2.2. When using term swapping for data enrichment we used different values of the k parameter: 2 and 10. Every model in all of the settings was evaluated on the original test set without any modifications.

As we consider term swapping as a data oversampling approach, we additionally compared our methods with two simple and widely used oversampling techniques in natural language processing. The first method is *random* oversampling where the original training set was oversampled by duplicating random instances in the training set so that the original class distribution remained the same. The second, more widely used technique, was the minority class oversampling. In this method, at each iteration, an instance from the current *minority* class is chosen at random and duplicated. This way the final oversampled training set has an approximately balanced class distribution. To control for the effect of the size of the data set on model performance, each baseline oversampling technique oversampled the original data set to the sizes of the proposed term swapping method.

Methods that employ ontology-based generalization search on the low-data setting are tested using both the *constrained* and *full* generalization search (see Section 4.1). While testing on the high-data setting we only explored the constrained generalization search due to longer model training times.

We measure the performance of the models using macro-F1 score, which is defined as the harmonic mean between recall and precision scores averaged across all classes. Formally, it is defined as follows:

$$F1 = \frac{1}{N}\sum_{i=1}^{N}\frac{2 * P_i * R_i}{P_i + R_i}$$

where i represents the class label, N represents the number of classes, P_i represents the i-th class precision score and R_i represents the i-th class recall score. We use this standard metric because it is shown to be more robust for problems with a highly unbalanced distribution of classes.

5.4. Evaluation results

5.4.1. High-data setting results

Table 4 shows F1-scores of the trained models described in Section 5.2 on the full data set.

All simpler models using logistic regression or SVM show increased performance by at least one of the data enrichment or oversampling methods, but the differences are rather small. The main difference (5 percentage points) can be observed when using term swapping generalization (parameter $k = 2$) with doc2vec.

For language models fine-tuned end-to-end (BERT and finBERT), the oversampling methods (swp, random, minority) seem to generally degrade the performance of the final model when the model is trained on the full data set. The results for the original BERT model show worse performance when oversampling methods are utilized, while for finBERT model only results with minority and random oversampling stay comparable with the baseline. This result indicates that oversampling in general is not a viable method for improving performance of language model-based classifiers when ample training data is already available. The language models also do not seem to gain enough additional information by introducing background knowledge from financial ontologies through concatenation of generalized terms at the end of the sentences (cnct). The performance of the BERT-based language model stays the same as without the introduction of background knowledge while a slight drop in performance is observed with the finBERT-based model. This effect might be explained by the fact that language models trained with self-attention and relatively long context windows weight every part of the input in relation to one another to construct the final representations. For this reason, these models are robust to minor perturbances in textual input, especially when the textual input is shorter than the input window.

5.4.2. Low-data setting results

The results of the trained models in terms of F1 scores on the downsized training set (where only 10% of the original training data is used) are presented in Table 5. In contrast to the high-data setting (see Section 5.4.1, the results on the downsized training data show that generalization and oversampling techniques help in improving the performance of all of the classifiers. In terms of our proposed methods, we see that the most

		lr-doc2vec	lr-char	lr-word	svm-mpnet	svm-char	svm-word	tpot	BERT	finBERT
baseline		0.45	0.46	0.47	0.58	0.47	0.46	**0.54**	**0.60**	**0.57**
cnct (constr.)		0.44	**0.49**	**0.49**	**0.59**	0.45	0.46	0.47	**0.60**	0.56
swp (constr.)	k2	**0.50**	0.44	0.48	0.57	0.42	**0.47**	0.42	0.59	0.55
	k10	0.49	0.44	0.48	0.57	0.42	**0.47**	0.41	0.53	0.51
minority	k2	0.38	**0.49**	0.43	0.57	0.48	**0.47**	0.46	0.57	**0.57**
	k10	0.32	**0.49**	0.47	0.56	**0.50**	**0.47**	0.26	0.54	0.55
random	k2	0.40	0.44	0.45	0.54	0.43	0.45	0.45	0.59	**0.57**
	k10	0.40	0.37	0.45	0.54	0.40	0.45	0.36	0.58	0.56

Table 4: Test set results of all models trained in **high data setting,** on the baseline data set (no generalizations or oversampling), the enriched training subsets with concatenation (*cnct*) and term swapping oversampling (*swp*), as well as the oversampled data using *minority* and *random* oversampling methods. Generalizations are obtained with the constrained ontology-based search. The models are evaluated with the F1-score. Bold results represent the best result for individual models.

		lr-doc2vec	lr-char	lr-word	svm-mpnet	svm-char	svm-word	tpot	BERT	finBERT
baseline		0.33	0.22	0.39	0.53	0.34	0.39	0.37	0.26	0.22
cnct (constr.)		0.27	0.22	0.38	0.53	0.36	0.38	0.38	0.22	0.30
cnct (full)		0.31	0.22	0.38	**0.56**	0.36	0.38	0.38	0.22	0.30
swp (constr.)	k2	0.33	0.22	0.39	0.37	0.34	0.39	0.37	0.35	0.31
	k10	0.35	0.33	0.34	0.51	0.34	0.30	0.34	0.38	0.36
swp (full)	k2	0.40	0.38	0.36	0.52	0.37	0.38	**0.41**	0.43	0.41
	k10	**0.44**	**0.42**	0.39	0.51	0.41	0.39	0.36	0.50	0.46
minority	k2	0.41	0.41	**0.44**	0.55	**0.42**	**0.44**	0.23	**0.52**	0.43
	k10	0.36	**0.42**	0.42	0.54	**0.42**	0.42	0.23	0.49	**0.49**
random	k2	0.35	0.37	0.38	0.53	0.37	0.38	0.31	0.36	0.27
	k10	0.36	0.35	0.38	0.53	0.36	0.38	0.31	0.24	0.29

Table 5: Test set results of all models trained in **low data setting**, on the baseline data set (no generalizations or oversampling), the enriched training subsets with concatenation (*cnct*) and term swapping oversampling (*swp*), as well as the oversampled data using *minority* and *random* oversampling methods. Generalizations are obtained with both constrained and full ontology-based searches. The models are evaluated with the F1-score. Bold results represent the best result for individual models.

consistent improvements are obtained using *full* search of the ontology. Overall, the best results are obtained using *svm-mpnet* model with our proposed background knowledge-enriched method *cnct(full)*; in this case the performance in low-data setting nearly reaches the performance of the finBERT model in high-data setting (see Table 4). We also see that term swapping (*swp*) leads to several large improvements, although *minority* oversampling is a very competitive approach (most frequently improving the individual classifier's performance).

In contrast to the high-data setting, for language model-based classifiers the performance can be increased using oversampling. Using our proposed term-swapping approach, we generally observe an increase in the final model performance as the size of the data set increases (k=10 vs. k=2), even though the highest improvements for BERT-based models are obtained with minority oversampling approach. The fine-tuned language models trained in low-data regimes generally lag behind the same models trained in high-data regimes, they however surpass other machine learning models in high-data settings.

6. Conclusion and future work

In our paper, we propose two generalization methods using the FIBO ontology as background knowledge. In the first one generalized terms are concatenated to the original training set instances, while in the second one, generalized terms are used in the oversampling setting, creating new generalized instances of the training set. We evaluate the potential of these methods in high- and low-data settings, and also more generally assess the potential of oversampling for financial sentiment analysis.

The results show that while in high-data setting best results are obtained using fine-tuned BERT-based models, where generalizations and oversampling do not lead to any improvement, simpler models using logistic regression or SVMs can be improved when integrating background knowledge (however improvements are rather small). More interestingly, we show that in low-data scenarios, large improvements can be obtained by our generalizations, as well as with simpler oversampling methods, leading to performances similar to those when 90% more data is available.

In future work, we aim to proceed in the following way. First, we will apply our method to other classification problems using annotations of our data set (rel-

evance for corporate business, subjectivity, sustainability issues relevance). Next, we aim to test our methods on other financial sentiment data sets (e.g. SemEval 2017 data for fine-grained sentiment analysis of financial microblog posts and news headlines (Cortis et al., 2017a)). Next, as in our paper (Stepišnik Perdih et al., 2021a), we have already shown that generalizations can be used for model explainability, we will continue this line of research, which would lead to improved interpretability of financial sentiment classification models for financial domain experts. Last but not least, we plan to use our sentiment classifiers to annotate a larger corpus of annual reports, where correlation analysis of financial indicators and text sentiment will be assessed, continuing and improving over our work in (Smailović et al., 2017b).

7. Reproducibility and reusability

The code of all our experiments is publicly available at the following GitLab repository: `https://gitlab.com/Andrazp/sentiment_classification_with_financial_ontologies.git`. The data identifying the sentences of annual reports that we used in our experiments and their sentiment annotations is available at `http://kt.ijs.si/data/sentences_financial_sentiment.zip`. The FIBO ontology is public and accessible at (fib, 2021a).

Acknowledgements

This work was supported by the Slovenian Research Agency (ARRS) grants for the core programme Knowledge technologies (P2-0103) and the project quantitative and qualitative analysis of the unregulated corporate financial reporting (J5-2554). B.S. was funded through the ARRS junior researcher grant. We also want to thank Matthew Purver and Aljosa Valentincic for their help in data set preparation, as well as students of the School of Business and Economics for their effort in data annotation.

8. References

Amel-Zadeh, A. and Faasse, J. (2016). The information content of 10-K narratives: Comparing MD&A and footnotes disclosures. 01.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Balakrishnan, R., Qiu, X. Y., and Srinivasan, P. (2010). On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research*, 202(3):789–801.

Bertin, M. and Atanassova, I. (2012). Semantic enrichment of scientific publications and metadata. *D-lib Magazine*, 18(7/8).

Bobicev, V. and Sokolova, M. (2017). Inter-annotator agreement in sentiment analysis: Machine learning perspective.

Buehlmaier, M. and Whited, T. M. (2015). Looking for risk in words: A narrative approach to measuring the pricing implications of financial constraints. In *Annual Conference of the Western Finance Association (WFA)*.

Butler, M. and Kešelj, V. (2009). Financial forecasting using character n-gram analysis and readability scores of annual reports. In *Canadian Conference on Artificial Intelligence*, pages 39–51. Springer.

Chang, S., Han, W., Tang, J., Qi, G.-J., Aggarwal, C. C., and Huang, T. S. (2015). Heterogeneous network embedding via deep architectures. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 119–128, New York, NY, USA. ACM.

Chen, C., Liu, C., Chang, Y., and Tsai, H. (2013). Opinion mining for relating subjective expressions and annual earnings in US financial statements. *Journal of Information Science and Engineering*, 29(4):743–764.

Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., and Davis, B. (2017a). 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)*, pages 519–535, Vancouver, Canada, August. Association for Computational Linguistics.

Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., and Davis, B. (2017b). 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)*, pages 519–535, Vancouver, Canada, August. Association for Computational Linguistics.

Day, M.-Y. and Lee, C.-C. (2016). Deep learning for financial sentiment analysis on finance news providers. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1127–1134. IEEE.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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. Association for Computational Linguistics.

Dong, S. and Liu, C. (2021). Sentiment classification for financial texts based on deep learning. *Computational Intelligence and Neuroscience*, 2021.

El-Haj, M., Rayson, P. E., Young, S. E., Walker, M., Moore, A., Athanasakou, V., and Schleicher, T. (2016). Learning tone and attribution for financial text mining.

Feldman, R., Govindaraj, S., Livnat, J., and Segal, B. (2008). The incremental information content of tone change in management discussion and analysis. *Available at SSRN 1126962*.

(2021a). The Financial Industry Business Ontology. https://spec.edmcouncil.org/fibo/, December.

(2021b). Visualising FIBO. https://datalanguage.com/blog/visualising-fibo, December.

Freeman, L. C. (2017). *Research Methods in Social Network Analysis*. Routledge.

Goel, S. and Uzuner, O. (2016). Do sentiments matter in fraud detection? estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management*, 23(3):215–239.

Gupta, A., Simaan, M., and Zaki, M. J. (2016). When positive sentiment is not so positive: Textual analytics and bank failures. *Available at SSRN 2773939*.

Hájek, P. and Olej, V. (2013). Evaluating sentiment in annual reports for financial distress prediction using neural networks and support vector machines. In *International Conference on Engineering Applications of Neural Networks*, pages 1–10. Springer.

Hajek, P., Olej, V., and Myskova, R. (2014). Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. *Technological and Economic Development of Economy*, 20(4):721–738.

Halterman, A. and Radford, B. J. (2021). Few-shot upsampling for protest size detection. *arXiv preprint arXiv:2105.11260*.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Jin, Z., Yang, Y., and Liu, Y. (2019). Stock closing price prediction based on sentiment analysis and lstm. *Neural Computing and Applications*, pages 1–17.

Kim, C., Yin, P., Soto, C. X., Blaby, I. K., and Yoo, S. (2018). Multimodal biological analysis using NLP and expression profile. In *2018 New York Scientific Data Summit (NYSDS)*, pages 1–4, Aug.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (2009). Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280. Association for Computational Linguistics.

Koloski, B., Perdih, T., Pollak, S., and Škrlj, B. (2021). Identification of covid-19 related fake news via neural stacking. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers*, page 177. Springer Nature.

Kothari, S., Li, X., and Short, J. E. (2009). The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review*, 84(5):1639–1670.

Krippendorff, K. (2018). *Content Analysis, An Introduction to its methodology*. Sage Publications, Thousand Oaks, CA, USA, 4th edition.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Le, T. T., Fu, W., and Moore, J. H. (2020). Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics*, 36(1):250–256.

Lee, H., Surdeanu, M., MacCartney, B., and Jurafsky, D. (2014). On the importance of text analysis for stock price prediction. In *LREC*, pages 1170–1175.

Lee, S. S. (2021). Feature investigation for stock returns prediction using xgboost and deep learning sentiment classification.

Li, F. (2010). The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102.

Loughran, T. and McDonald, B. (2011a). Barron's red flags: Do they actually work? *Journal of Behavioral Finance*, 12(2):90–97.

Loughran, T. and McDonald, B. (2011b). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Lu, X., Zheng, B., Velivelli, A., and Zhai, C. (2006). Enhancing text categorization with semantic-enriched representation and training data augmentation. *Journal of the American Medical Informatics Association*, 13(5):526–535.

Man, X., Luo, T., and Lin, J. (2019). Financial sentiment analysis (fsa): A survey. In *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)*, pages 617–622. IEEE.

Pelicon, A., Pranjić, M., Miljković, D., Škrlj, B., and Pollak, S. (2020). Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, 10(17):5993.

(2021). Lemmagen3. https://pypi.org/project/lemmagen3/#description, December.

Qiu, X. Y., Srinivasan, P., and Street, N. (2006). Exploring the forecasting potential of company annual reports. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–15.

Santos, J. S., Bernardini, F., and Paes, A. (2021). Measuring the degree of divergence when labeling tweets in the electoral scenario. In *Anais do X Brazilian*

Workshop on Social Network Analysis and Mining, pages 127–138. SBC.

Schneider, N., Schneider, L., Pinggera, P., Franke, U., Pollefeys, M., and Stiller, C. (2016). Semantically guided depth upsampling. In *German conference on pattern recognition*, pages 37–48. Springer.

Smailović, J., Grčar, M., Lavrač, N., and Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, 285:181–203.

Smailovic, J., Znidarsic, M., Valentincic, A., Loncarski, I., Pahor, M., Martins, P., and Pollak, S. (2017a). Automatic analysis of annual financial reports: A case study. *Computación y Sistemas*, 21(4).

Smailović, J., Žnidaršič, M., Valentinčič, A., Lončarski, I., Pahor, M., Martins, P. T., and Pollak, S. (2017b). Automatic analysis of annual financial reports: A case study. *Computación y Sistemas*, 21(4):809–818.

Stepišnik Perdih, T., Lavrač, N., and Škrlj, B. (2021a). Semantic reasoning from model-agnostic explanations. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI) [Elektronski vir]: on-line conference*, 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI) [Elektronski vir]: online conference, page 105–110. Nasl. z nasl. zaslona Opis vira z dne 23. 3. 2021 Bibliografija: str. 10.

Stepišnik Perdih, T., Pollak, S., and Škrlj, B. (2021b). Jsi at the finsim-2 task: ontology-augmented financial concept classification. In Jurij Leskovec, et al., editors, *The Web Conference [Elektronski vir]: companion of the World Wide Web conference (WWW 2021): [30th edition, Ljubljana, 19th - 23rd April, 2021]*, The Web Conference [Elektronski vir]: companion of the World Wide Web conference (WWW 2021): [30th edition, Ljubljana, 19th - 23rd April, 2021], page 298–301. Association for Computing Machinery. Soavtorji: Vlado Dimovski, Judita Peterlin, Maja Meško, Vasja Roblek Opis vira z dne 14. 6. 2021 Nasl. z nasl. zaslona Bibliografija: str. 299-301 Abstract.

Tsai, M.-F. and Wang, C.-J. (2012). Visualization on financial terms via risk ranking from financial reports. In *COLING (Demos)*, pages 447–452.

Wang, W. Y. and Yang, D. (2015). That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2557–2563.

Wang, C.-J., Tsai, M.-F., Liu, T., and Chang, C.-T. (2013). Financial sentiment analysis for risk prediction. In *IJCNLP*, pages 802–808.

Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Articles: Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433, September.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Wu, X., Lv, S., Zang, L., Han, J., and Hu, S. (2019). Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95. Springer.

Xing, F. Z., Cambria, E., and Welsch, R. E. (2018). Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73.

Xing, F., Malandri, L., Zhang, Y., and Cambria, E. (2020). Financial sentiment analysis: an investigation into common mistakes and silver bullets. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 978–987.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Zhang, L., Xiao, K., Zhu, H., Liu, C., Yang, J., and Jin, B. (2018). Caden: A context-aware deep embedding network for financial opinions mining. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 757–766. IEEE.

Škrlj, B., Martinc, M., Kralj, J., Lavrač, N., and Pollak, S. (2020). tax2vec: Constructing interpretable features from taxonomies for short text classification. *Computer Speech Language*, page 101104.