Preliminary experimentation with combinations and extensions of forward-looking sentence detection wordlists

Jan Štihec
University of Ljubljana
Ljubljana, Slovenia
stihec.jan@gmail.com

Senja Pollak
Jožef Stefan Institute
Ljubljana, Slovenia
senja.pollak@ijs.si

Martin Žnidaršič
Jožef Stefan Institute
Ljubljana, Slovenia
martin.znidarsic@ijs.si

Abstract

Forward-looking sentences are often a subject of studies of financial texts. Detection of such sentences is usually performed with wordlists of inclusive and exclusive keywords that are used as indicators of the forward-looking nature of the sentences at hand. In this paper we describe our assessment of potential improvements of forward-looking sentence detection wordlists by combining them together and by extending them with neighboring words in word-vector representations. Our current results indicate that simple combinations and straightforward extensions of wordlists with vector-space representation neighbors might not be suitable for FLS detection without further methodological improvements.

1 Introduction

Many studies of financial texts focus specifically on the contents of the forward-looking sentences (FLS). Detection of such sentences is then either a part of the methodological approach of a study or even one of the main aims of research.

Approaches to detection of these sentences usually employ lists of keywords, which are used as indicators whether a given sentence tends to be forward-looking or not. Keyword lists usually consist of: words that imply the future (e.g. "future"), future years numbers, conjugations of verbs that imply the future (e.g. "we intend") and combinations of certain adjectives and time indicators (e.g. "next year"). Some approaches also use lists of exclusive words, which are used to exclude a sentence that contains them from the forward-looking sentences identification process. Exclusive keywords are not always correlated with a nature of the sentence not being forward-looking, but might only indicate that a sentence containing them should not be analyzed in a specific study. It might for example contain keywords that are indicative for the parts of text, which aren't relevant for the study at hand.

The aim of our work is to study various wordlists for forward-looking sentence detection that appear in the literature, assess their combined use and experiment with wordlist extensions that are based on vector representation distances (similar to related work in terminology extraction, see e.g. (Pollak et al., 2019; Vintar et al., 2020)).

In this paper we report preliminary results of four wordlists, their combination and one wordlist's vector-space based extension on two manually labeled datasets. The current results indicate that the addition of exclusive wordlists might not always improve the results, which stands also for merging of the wordlists. The extension of the wordlists with word vector neighbors increases the amount of detected forward-looking-sentences, but it also increases the amount of sentences that are wrongly classified as FLS. With the current approach, this issue could not be alleviated with a similar extension of the corresponding exclusive wordlist.

2 Related work

Future-oriented information is recognized as very relevant to investors and is the subject of varius studies (Mio et al., 2020). Some studies rely on manual collection and analysis of FLS, while others employ automatic procedures that are mostly based on a number of widely used FLS wordlists. Each of these two approaches has its benefits and drawbacks, but we are interested in the latter one and the impact of the wordlists that are used for such purposes.

We identified four wordlists which are proposed in works that are commonly cited with regards to FLS identicifation and provide complete wordlists. Chronologically ordered the first one is the work by Li (2010), which is focused on information content and tone of FLS sentences. Next is the one by Athanasakou and Hussainey (2014) which is aimed at the assessment of the frequency of such statements and its relations with financial indicators.

The work of Muslu et al. (2015) studies the relation among FLS quantity and the firms' information environments. It suggests also use of word combination patterns, so the FLS wordlist that corresponds to this approach is relatively extensive. Tao et al. (2018) study the relationships among FLS features and IPO valuation. They use a wordlist based FLS detection approach (similar to the one by Muslu et al. (2015), but with additional consideration of sentence structure) in the stage of data preparation for machine-learning of a neural network based FLS classifier. All the listed studies provide a list of FLS inclusive keywords and all with the exception of Athanasakou and Hussainey (2014) also provide a list of FLS exclusive keywords.

3 Methodology

The approach that we used for the study described in this paper consists of: (I) selection of relevant wordlist-based approaches to FLS detection, (II) preparation of the data for testing and learning, and (III) design and running of the experiments.

We selected wordlists from four works (Li, 2010; Athanasakou and Hussainey, 2014; Muslu et al., 2015; Tao et al., 2018), which are often cited with regards to FLS detection and also provide the wordlists and detailed explanation of their FLS detection processes. We denote these wordlists as wl-Li, wl-At, wl-Mu and wl-Ta respectively. The data that was used for assessments and for learning the vector representations (also referred to as embeddings) in our experiments is described in detail in Section 3.1. For efficient experimentation with the selected wordlists we implemented a general wordlist-based labeling tool in python. Section 3.2 is dedicated to description of the methodological details of experiments.

3.1 Data

For the assessments of FLS detection approaches we used the sentences that were selected at random from recent (since 2017) annual reports of randomly selected FTSE 350 index constituents and were annotated as forward-looking/non-forward-looking by two human annotators. As the data was annotated by two annotators who worked on separate (not overlapping) groups of sentences, we treat this data as two datasets of 467 and 459 annotated sentences respectively and we denote them as D_1 and D_2. There are 260 FLS and 207 non-FLS sentences in D_1, while D_2 contains 122 FLS and 337 non-FLS sentences.

Table 1: Size of the used wordlists in terms of the amount of keywords.

	inclusive	exclusive
wl-Li	17	31
wl-At	45	/
wl-Mu	332	6
wl-Ta	373	6

Data was necessary also in the approach for extending wordlists, where it was used for learning vector space representations of words. Annotations are not needed for this purpose, but the data should be from the same domain as the task in which the vector representations are to be employed. We used a corpus of 604 periodic (10-Q and 10-K) reports. Specifically, it consisted of the 2018 Q4 reports from the Stage One 10-X Parse Data collection (from file `10-X_C_2016-2018.zip`) of the well known Notre Dame Software Repository for Accounting and Finance that was established by Loughran and McDonald (2016).

3.2 Experimental setup

In our experiments we used each individual selected wordlist and a merged wordlist that is denoted as wl-all and contains a set of all the words appearing in any of the wordlists. The wordlists were used for labelling the sentences as FLS or non-FLS. The results were calcualted separately for each of the two datasets.

With the exception of the approach by Athanasakou and Hussainey (2014), all the selected approaches provide an inclusive and an exclusive wordlist. First, we used only inclusive wordlists with a straighforward classification approach: the sentences that contained any word from a given inclusive list were classified as FLS. In the next series of experiments we used also all the corresponding exclusive wordlists in the sense that any sentence classified as FLS was re-classified into non-FLS, if it contained any word from the given list of exclusive words.

Note that our use of the wordlists is not completely comparable with most of the related works, from which the wordlists originate, as they were focused on specific sections of financial reports and some of the FLS detection approaches additionally considered numeric indications of future years or, in case of the approach by Tao et al. (2018), the

Table 2: Accuracy (acc) and recall (rec) of FLS classification with inclusive wordlists only.

	acc D_1	rec D_1	acc D_2	rec D_2
wl-Li	0.67	0.60	0.68	0.70
wl-At	0.68	0.66	0.62	0.74
wl-Mu	0.64	0.45	0.71	0.59
wl-Ta	0.64	0.45	0.71	0.59
wl-all	0.71	0.78	0.60	0.87

use of wordlists represented only a part of the FLS detection approach.

The last series of experiments, assessment of the effect of embeddings-based extensions of wordlists, was done only with one original wordlist - the one proposed by Li (2010). Again, both only the inclusive and the inclusive/exclusive options were experimented with. The word vector representations were learned with the fastText approach (Bojanowski et al., 2016) in the ClowdFlows3[1] prototype online tool for data analysis (parameters for learning the fastText model and neighbors selection: *vector size*=20, *context window size*=5, *minimal word occurences*=5, *distance threshold*=0.9). For each of the words in the original wordlist, the original word and five of the neighboring words from the vector space were included in the extended wordlist. The word neighbors were post-processed as follows: (I) any punctuation character at the start or the end of the word was removed, (II) any words that are considered English stop-words by the NLTK language toolkit[2] were removed.

The exclusive wordlist from Li (2010) includes also some bi-grams that are combinations of words: 'expected','anticipated','forecasted','projected','believed' that are preceded with each of the following auxiliary verbs: 'was' , 'were', 'had' and 'had been'. To obtain the corresponding embedding-based neighbors of these terms, we first calculated the neighbors of the words without the auxiliary verbs and then added all the combinations with auxiliary verbs to all the resulting word neighbors.

The resulting extended wordlists are provided in Appendix A.

4 Results and findings

Results of the assessment for inclusive wordlists are presented in Table 2 in terms of accuracy and recall of the FLS class. The recall might be more of interest if the aim of FLS detection is to analyse FLS contents or pre-filtering. For estimation of the amount of FLS the more relevant measure is accuracy, but it needs to be considered carefully in case of unbalanced datasets such as D_1 and D_2. From Table 2 we can see that on D_1 the best individual wordlist results are obtained with wl-At and that the merged wordlist yields better results as any of the individual approaches in terms of both performance measures. This is not the case on D_2, which has more non-FLS sentences. On D_2 these two approaches are better in terms of recall, but worse than others in terms of accuracy.

Addition of excluding wordlists into consideration slightly reduced all the recalls, with profound effect mostly in case of the merged wordlist. In such a setting, the combined wordlist did not outperform individual ones on any of the two datasets as it for example performs worse than wl-Li with respect to both measures on both datasets.

What we can draw from the first two experiments is that a combination of individual wordlists is not necessarily beneficial, particularly not for the case of considering also exclusive keywords.

Experimental assessment of the embeddings-based extensions of a wordlist are presented in Table 4. The extended inclusive wordlist improves recall, but in D_2 at the expense of accuracy. In comparison with the wordlist extension approach of merging the wordlist with other proposed ones, the

Table 3: Accuracy (acc) and recall (rec) of FLS classification with inclusive and exclusive wordlists.

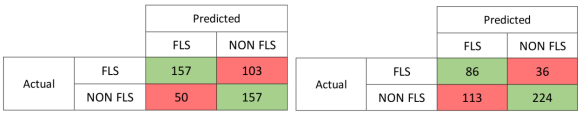
	acc D_1	rec D_1	acc D_2	rec D_2
wl-Li	0.67	0.60	0.68	0.70
wl-At	0.68	0.66	0.62	0.74
wl-Mu	0.63	0.41	0.71	0.51
wl-Ta	0.63	0.40	0.71	0.51
wl-all	0.65	0.58	0.63	0.65

Table 4: Accuracy of FLS detection with embeddings-based extensions of the wordlists by Li (2010). Use of extension is denoted by e(), in stands for the use of the inclusive and ex for the use of the exclusive list.

	acc D_1	rec D_1	acc D_2	rec D_2
in	0.67	0.60	0.68	0.70
e(in)	0.68	0.69	0.59	0.78
e(in) ex	0.68	0.69	0.59	0.78
e(in) e(ex)	0.63	0.59	0.60	0.64

[1]ClowdFlows3 homepage: https://cf3.ijs.si/ The used workflow is available at: https://cf3.ijs.si/workflow/223

[2]https://www.nltk.org/

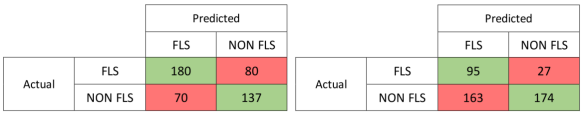


		Predicted	
		FLS	NON FLS
Actual	FLS	157	103
	NON FLS	50	157

		Predicted	
		FLS	NON FLS
Actual	FLS	86	36
	NON FLS	113	224

Figure 1: Contingency matrix for original inclusive wl-Li on D_1 (left) and D_2 (right).

extension with embeddings performs worse than the merged wordlist for both measures on both datasets.



		Predicted	
		FLS	NON FLS
Actual	FLS	180	80
	NON FLS	70	137

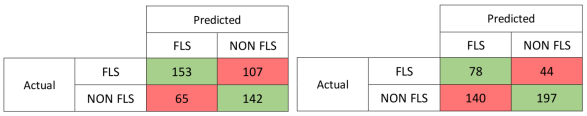
		Predicted	
		FLS	NON FLS
Actual	FLS	95	27
	NON FLS	163	174

Figure 2: Contingency matrix for extended inclusive wl-Li on D_1 (left) and D_2 (right).

Consideration of exclusive keywords was expected to compensate for some of the accuracy lost on D_2 due to potentially too wide reach of the inclusive keyword extensions, but consideration of the original exclusive keywords did not have an effect on results (a single sentence was classified differently in D_1), while use of an embedding-extended exclusive wordlist caused more non-FLS sentences to be correctly classified and vice-versa for the FLS (for details see Figures 1 to 3). This caused slight changes in accuracy in line with the class distributions of the two datasets. Most importantly, overall the approach with both the extended inclusive and extended exclusive wordlist in all aspects performed worse than the approach with original state of these two wordlists (for comparison see Table 3).

Our study is preliminary and we intend to conduct more experiments on larger datasets, but the current results indicate that straightforward extensions of wordlists with vector-space representation neighbors might not be suitable for FLS detection. In most experimental settings this holds also for extensions of wordlists by merging them together, although by a lesser extent.

This does not mean that such approaches cannot improve FLS detection, but it indicates that it might be necessary to go beyond a simple automated word vector neighbor extension and that such methodological improvements would be sensible already before further experimentation.



		Predicted	
		FLS	NON FLS
Actual	FLS	153	107
	NON FLS	65	142

		Predicted	
		FLS	NON FLS
Actual	FLS	78	44
	NON FLS	140	197

Figure 3: Contingency matrix for extended inclusive and extended exclusive wl-Li on D_1 (left) and D_2 (right).

Acknowledgements

This paper is supported by the project Quantitative and qualitative analysis of the unregulated corporate financial reporting (No. J5-2554), which was financially supported by the Slovenian Research Agency. The paper was supported also by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The authors acknowledge also the financial support from the Slovenian Research Agency for research core funding for the programme Knowledge Technologies (No. P2-0103). For access to the dataset of labeled forward looking sentences we thank the Faculty of Economics of the University of Ljubljana.

References

Vasiliki Athanasakou and Khaled Hussainey. 2014. The perceived credibility of forward-looking performance disclosures. *Accounting and business research*, 44(3):227–259.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Feng Li. 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.

Tim Loughran and Bill McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Chiara Mio, Pier Luigi Marchini, and Alice Medioli. 2020. Forward-looking information in integrated reports: Insights from "best in class". *Corporate Social Responsibility and Environmental Management*, 27(5):2212–2224.

Volkan Muslu, Suresh Radhakrishnan, KR Subramanyam, and Dongkuk Lim. 2015. Forward-looking md&a disclosures and the information environment. *Management Science*, 61(5):931–948.

Senja Pollak, Andraž Repar, Matej Martinc, and Podpečan Vid. 2019. Karst exploration: extracting terms and definitions from karst domain corpus. In *Proceedings of eLex19*, pages 934–956, Sintra, Portugal.

Jie Tao, Amit V Deokar, and Ashutosh Deshmukh. 2018. Analysing forward-looking statements in initial public offering prospectuses: a text analytics approach. *Journal of Business Analytics*, 1(1):54–70.

Špela Vintar, Larisa Grčić Simeunović, Matej Martinc, Senja Pollak, and Uroš Stepišnik. 2020. Mining semantic relations from comparable corpora through intersections of word embeddings. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 29–34, Marseille, France. European Language Resources Association.

A Embedding-based extensions

Extension of the inclusive part of wl-Li. The words from the original wordlist are in bold, followed by up to five extensions (less, if removed as stop-words or duplicates with extensions of preceeding original words):
will accordingly furthermore **should** relied regarded ultimate context **can** frequently unreliable predicate producibility problem **could** harm harmed adverse **may** even substantial us **might** materialize pursued occur difficult **expect** expand effectively continue expansion **anticipate** profitable **believe** proactively history believes regularly **plan** plans sponsors sponsor **hope** hopes success perspectives identify teamwork **intend** intends **seek** seeking stop decide **project** progress feasibility projects predevelopment **forecast** quarter-to-quarter profitability forecasting forecasts **objective** objectively objectivity maximize **goal** toward targeting driving striving excellence

Extension of the exclusive part of wl-Li. The words from the original wordlist are in bold, followed by up to five extensions (less, if removed as stop-words or duplicates with extensions of preceeding original words):
undersigned, undersigned's, duly, thereunto, countersigned, **herein**, reference, referenced, **hereinafter**, hereinabove, mean, indicated, **hereof**, TAA, **hereon**, henceforth, **hereto**, confirms, **theretofore**, grantor, asserted, party, deemed, obligated, **therein**, documents, **thereof**, therefor, **thereon**, **expected**, differences, future, reversals, different, **was expected**, was differences, was future, was reversals, was different, **were expected**, were differences, were future, were reversals, were different, **had expected**, had differences, had future, had reversals, had different, **had been expected**, had been differences, had been future, had been reversals, had been different, **anticipated**, negative, forecast, unanticipated, results, **was anticipated**, was negative, was forecast, was Unanticipated, was results, **were anticipated**, were negative, were forecast, were Unanticipated, were results, **had anticipated**, had negative, had forecast, had Unanticipated, had results, **had been anticipated**, had been negative, had been forecast, had been Unanticipated, had been results, **forecasted**, magnified, imbalance, movements, variability, fluctuation, **was forecasted**, was magnified, was imbalance, was movements, was variability, was fluctuation, **were forecasted**, were magnified, were imbalance, were movements, were variability, were fluctuation, **had forecasted**, had magnified, had imbalance, had movements, had variability, had fluctuation, **had been forecasted**, had been magnified, had been imbalance, had been movements, had been variability, had been fluctuation, **projected**, projecting, **was projected**, was projecting, **were projected**, were projecting, **had projected**, had projecting, **had been projected**, had been projecting, **believed**, likelihood, verified, mistaken, livelihood, **was believed**, was likelihood, was verified, was mistaken, was livelihood, **were believed**, were likelihood, were verified, were mistaken, were livelihood, **had believed**, had likelihood, had verified, had mistaken, had livelihood, **had been believed**, had been likelihood, had been verified, had been mistaken, had been livelihood, **shall**, hereunder,