Using Topic Modelling to Improve Prediction of Financial Report Commentary Classes

Karim El Mokhtari, Mucahit Cevik(✉), and Ayşe Başar

Data Science Lab, Ryerson University, Toronto, Canada
{elmkarim,mcevik,ayse.bener}@ryerson.ca
https://www.ryerson.ca/dsl

Abstract. We consider the task of predicting the class of commentaries associated with financial discrepancies between actual and estimated sales data. Such analysis of the financial data is helpful in meeting targets and assessing the overall performance of the company. While generating a commentary and its associated class is the task of an analyst, these manual operations might be erroneous and as a result, might lead to a diminished performance for the employed prediction model due to wrong class labels. Accordingly, we propose using topic modelling, namely Latent Dirichlet Allocation (LDA), for automated extraction of the classes of the commentaries. In addition, we use feature selection strategies to improve the accuracy of the prediction models. Our analysis with various time series classification methods points to improved performance due to LDA and feature selection.

Keywords: Time series classification · LDA · NLP

1 Introduction

On an annual basis, companies make budgets to set targets for sales, revenues and expenses. Financial reports that are generated through various transactions are then used to understand the discrepancies between actual performance and financial forecast. These reports are generated from different sources including the daily sales transactions, inventories, cash flows, and supplier transactions. Financial analysts within the company examine the variances between projected outcomes and the actuals in certain time intervals (e.g. weekly, monthly, or quarterly), and provide management with insights. Accordingly, it is determined whether some adjustments are needed to be made to the targets or some other actions are needed to be taken in various departments to meet these targets.

The analyst report usually involves a summary information that relates each variance with a short commentary that serves as a baseline for top management to take immediate actions. While this process relies on the analyst's knowledge of the business areas, the data coming from different silos of the company might

C. Goutte and X. Zhu (Eds.): Canadian AI 2020, LNAI 12109, pp. 201–207, 2020.
https://doi.org/10.1007/978-3-030-47358-7_19

lead to erroneous assessments. As such, there is a need to learn from existing commentaries and build a prediction model that helps the automatic generation of the commentaries and their associated classes.

In our previous work [5], we tried to predict the commentaries using an encoder/decoder structures that predict jointly from the variance dataset and the expert commentaries. However, the reduced set of commentaries and a high variety among commentaries in terms of writing and wording, made it hard to predict meaningful commentaries. Later, we simplified the problem and aimed to predict the category of the commentary rather than trying to generate the original text [6]. Our analysis with various machine learning models including Support Vector Machines (SVMs), Random Forests (RFs) and Long Short Term Memory Neural Networks (LSTMs) indicated a positive predictive performance.

We note that, while generating a commentary and its associated class is the task of an analyst, such manual operations might be erroneous and as a result might lead to a diminished performance for the employed prediction model due to wrong class labels. Accordingly, we propose applying topic modelling, namely Latent Dirichlet Allocation (LDA) on the commentaries dataset in order to cluster commentaries based on their similarity, and then predict the topic of the commentary rather than the full text. The model learns from the variance dataset and predicts the topic, which helps automating the overall process.

Topic modelling aims to identify abstract or hidden structures of the text bodies which are called topics. There are various topic modelling techniques in the literature such as Latent Semantic Analysis (LSA) [3], probabilistic latent semantic analysis (pLSA) [7] and LDA [2]. While LSA is a simple dimensionality reduction technique with SVD, pLSA and LDA consider a statistical approach that builds a probabilistic language model from documents considered as a mixture of topics. Unlike pLSA that ignores how the topic mixture is generated for a document [9], LDA uses the Dirichlet distribution to determine this topic distribution. Liu et al. [10] used sentiment as topics and applied pLSA to model weblogs entries and predict product sales performance. LDA was successfully applied for Pseudo Relevance Feedback as well, where Miao et al. [11] introduced a topic space to evaluate the reliability of each candidate feedback document, then they used the reliability scores to adjust the weights of terms. LDA was also applied in identifying medical prescription patterns [12] and modelling the correlations of news items with stock price movements [8].

2 Methodology

2.1 Topic Modelling

Data is provided by our partner company in two datasets that include 34 months, from January 2016 to October 2018. The first dataset, COM (see Table 2), contains the commentaries written by the analyst for every customer and brand. The second dataset, VAR (see Table 1), includes the difference between forecasts and actuals in millions of dollars for every customer and brand.

Table 1. The VAR dataset records the discrepancy between forecast and actuals by customer and brand in Millions of CAD. Sample for one customer.

Customer 1				
	Jan. 2016	Feb. 2016	...	Sep. 2018
B1	+0.10	−0.01	...	+0.05
B2	−0.08	−0.05	...	+0.12
...	...	...	...	...

Table 2. Commentaries are written monthly for brands with discrepancies higher than 0.2 Millions CAD. Sample commentaries for January 2016

January 2016		
Brand	Var.	Commentary
B3	+0.50	Customer10: caused by over delivery
B15	+0.63	Customer25: declining faster than seen in the market
...	...	...

The analyst labelled each commentary with one of the following five labels: 1. "Promo" when a promotion did not perform as expected, 2. "SP&D" related to special offers including multiple products sold as a package, 3. "POS" related to sales that showed unexpected highs or lows at retail stores, 4. "Other" for rare topics, and 5. "NoComm" when no commentary was provided. The process of labelling commentaries is challenging as it often happens that a commentary has a main topic and one or two secondary topics. Further, it leads to an imbalanced dataset having NoComm as a majority class and Other as a minority class. We apply undersampling for NoComm and oversampling for the remaining classes to balance the labels' occurrences.

We process the VAR dataset in a way that for every commentary emitted for a customer and a brand in any given month, we associate a time series of variances (i.e. *forecasts* − *actuals*) recorded for the customer and the brand for all previous 12 months. Therefore, we build a new dataset VAR-COM where each row is associated with a brand b, a customer c and a month m. The input is in the form of a vector of 13 elements that represents the variance of the brand/customer pair (b, c) during the month m and the 12 preceding months: $\{v_{m-12}^{(b,c)}, \ldots, v_{m-1}^{(b,c)}, v_m^{(b,c)}\}$. The output is a one-hot-encoded representation of the label of the commentary generated by the expert for the tuple (b, c, m).

In addition, we pre-process all commentaries by converting them to lowercase then removing numbers, punctuation signs and common English stop-words. We also build a context-related list of stop-words such as customer name, product description and months. The comment after pre-processing is supposed to contain only useful keywords. We later apply lemmatizing and stemming to transform all word variants to standardized roots. Then, we apply *TF-IDF* to the resulting commentaries according to the formula $\textit{TF-IDF}(t, c) = TF(t, c) \cdot \log\left(\frac{N}{DF(t)}\right)$ where t denotes a term in a commentary c, $TF(t, c)$ is the frequency of the term t in the commentary c, N is the number of non-empty commentaries in the dataset, and $DF(t)$ the document frequency defined as the

number of documents containing the term t. Therefore, *TF-IDF* associates a higher score to the relevant words. Every commentary is converted into a set of tokens associated with their *TF-IDF* scores. We choose to use TF-IDF scoring instead of embedding vectors as the vocabulary size and the number of commentaries are relatively low (around 521 commentaries and 441 unique words after pre-processing).

Finally, we apply LDA to all commentaries by using three topics. The one with the highest probability is considered as the main topic of the commentary. It may happen that all words in a commentary are removed in the pre-processing phase, in this case, we associate a label "General" to the commentary. Empty commentaries are labelled with NoComm. Consequently, LDA results in five labels, which is equal in number to the labels provided by analysts, but different in nature as they are produced by clustering (i.e. LDA). Each commentary is assigned one of the following labels: "Topic1", "Topic2", "Topic3", "General" or "NoComm". We experiment with a different number of topics as well.

2.2 Learning Model

We train five models, each one is based on a different classification approach: RF, Gradient Boosting (XGB), SVM, LSTM and one-nearest neighbour dynamic time warping (1-NN DTW) [1]. Previously, we observed that not all months in the dataset are important in the time series [4], thus we consider feature selection based on RF feature importance by keeping 4 months out of 13 months, namely v_{m-12}, v_{m-11}, v_{m-1} and v_m where m is the month the commentary is issued.

3 Results

The commentaries dataset contains 2607 data points where only 521 are associated with commentaries. The average number of words in a commentary is nine. After pre-processing this number is reduced to four. Around 6% of the commentaries ended up being empty after pre-processing. We carry out three different experiments: (1) with analyst and LDA labels using 3 topics and no feature selection, (2) with analyst and LDA labels using 3 topics and with feature selection, (3) with analyst and LDA labels using 2 topics and with feature selection. After splitting the dataset into a train and test sets with an 80/20 ratio, we run all models 25 times with random oversampling and downsampling. We report the F1-score macro resulting from each run.

Figure 1 indicates that there is no significant improvement when learning with the original labels and the ones generated by LDA without feature selection. We note that, after feature selection, results are improved for XGB and SVM. With 2 topics, LDA-based models perform better than with the original labels as shown in Fig. 2. This can be explained by the fact that the number of labels with LDA is smaller, however, from a practical perspective, such results are still interesting for the consumer goods company as the model is able to predict the commentary topic and provide insightful keywords for every topic. Figure 3 illustrates the

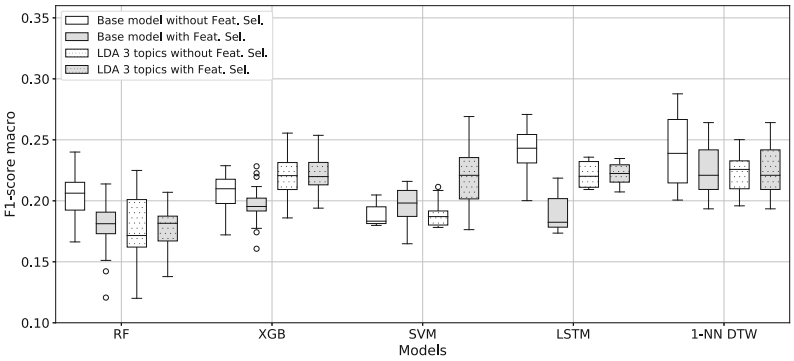

Fig. 1. F1-scores for different models before and after applying LDA using 3 topics (i.e. 5 labels), with and without feature selection.

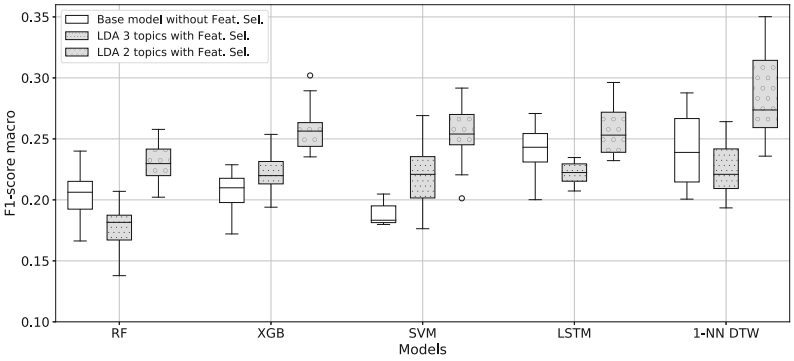

Fig. 2. F1-scores for considering different number of topics selected by LDA.

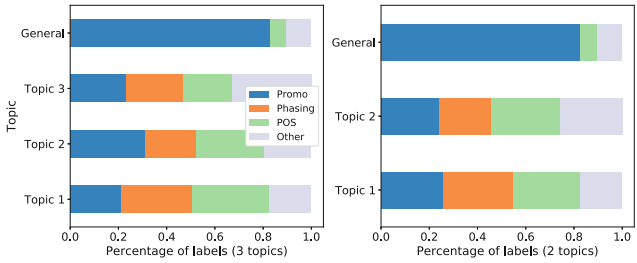

Fig. 3. Percentage of each expert label in LDA topics.

distribution of each analyst label by LDA topic. If we discard the "General" topic, the figure does not reveal a clear association between topics and labels which may suggest that the labels may not be related to specific keywords as was assumed in this analysis. Alternatively, this might be considered as evidence for data quality issues, if the expectation is to have some correlation between the labels and the associated commentaries.

4 Conclusions and Future Work

In this paper, we incorporate LDA into the classification of financial time series data into categories of commentaries. We examine the impact of feature selection on prediction quality as well. Our analysis with various time series classification methods shows that LDA in combination with feature selection leads to improved results. Moreover, automated identification of class labels from commentaries enables experimenting with different numbers of class labels, which further improves the predictive performance. 1-NN DTW performed the best in predicting labels after applying LDA. However, no significant correlation is observed between analysts' labels and LDA topics. Future work includes acquiring more data to test out our approaches as well as employing an active learning mechanism to guide the expert in manually labelling the instances so that number of manually labelled instances and the associated errors will be reduced.

Acknowledgement. This work is supported by Smart Computing For Innovation (SOSCIP) consortium, Toronto, Canada.

References

1. Bagnall, A., Lines, J., Bostrom, A., Large, J., Keogh, E.: The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Min. Knowl. Disc. **31**(3), 606–660 (2016). https://doi.org/10.1007/s10618-016-0483-9
2. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. JMLR **3**, 993–1022 (2003)
3. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)
4. El Mokhtari, K., Higdon, B., Başar, A.: Interpreting financial time series with SHAP values. In: Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering, pp. 166–172 (2019)
5. El Mokhtari, K., Maidens, J., Bener, A.: Predicting commentaries on a financial report with recurrent neural networks. In: Meurs, M.-J., Rudzicz, F. (eds.) Canadian AI 2019. LNCS (LNAI), vol. 11489, pp. 531–542. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-18305-9_56
6. Peachey Higdon, B., El Mokhtari, K., Başar, A.: Time-series-based classification of financial forecasting discrepancies. In: Bramer, M., Petridis, M. (eds.) SGAI 2019. LNCS (LNAI), vol. 11927, pp. 474–479. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-34885-4_39
7. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57 (1999)
8. Asadi Kakhki, S.S., Kavaklioglu, C., Bener, A.: Topic detection and document similarity on financial news. In: Bagheri, E., Cheung, J.C.K. (eds.) Canadian AI 2018. LNCS (LNAI), vol. 10832, pp. 322–328. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-89656-4_34
9. Lee, S., Baker, J., Song, J., Wetherbe, J.C.: An empirical comparison of four text mining methods. In: 2010 43rd Hawaii International Conference on System Sciences, pp. 1–10, January 2010. https://doi.org/10.1109/HICSS.2010.48

10. Liu, Y., Huang, X., An, A., Yu, X.: ARSA: a sentiment-aware model for predicting sales performance using blogs. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 607–614 (2007)
11. Miao, J., Huang, J.X., Zhao, J.: TopPRF: a probabilistic framework for integrating topic space into pseudo relevance feedback. ACM TOIS **34**(4), 1–36 (2016)
12. Park, S., Choi, D., Kim, M., Cha, W., Kim, C., Moon, I.C.: Identifying prescription patterns with a topic model of diseases and medications. J. Biomed. Inform. **75**, 35–47 (2017)