Baltic J. Modern Computing, Vol. 9 (2021), No. 2, pp. 180–194
https://doi.org/10.22364/bjmc.2021.9.2.03

Natural Language Based Posting Account Classification

Zigmunds BEĻSKIS[1], Marita ZIRNE[1], Viesturs SLAIDIŅŠ[1], Mārcis PINNIS[1,2]

[1] Tilde, Vienibas gatve 75A, Riga, Latvia, LV-1004
[2] Faculty of Computing, University of Latvia, Raiņa blvd 19-125, Riga, Latvia, LV-1586

{zigmunds1belskis|marita.zirne|viesturs.slaidins|marcis.pinnis}@tilde.lv

Abstract. Any process automation is meaningful when minimizing or excluding human involvement compared to the manual process. Operational accuracy plays a key role in accounting, so improving the accuracy of the results of the automation process will increase the willingness of accountants to use these solutions in their daily lives. This paper examines methods for the automation of the assignment of account codes to posting entries using natural language processing-based classification methods. We find that using textual comment features in the machine learning process allows to achieve debit account classification precision that is by up to 2.56 percentage points higher compared to previous work, reaching an overall classification precision of 94.78%.

Keywords: posting account classification, automation of accounting, machine learning, natural language processing

1 Introduction

The invoice processing procedure is a common business process in organizations. There are several steps involved in handling of invoices. One of the steps is invoice posting in the general ledger of an accounting system. Invoice posting is a manual job that is carried out by accountants and because of its monotonous nature it has a high level risk of errors attached to it. The accountant's task is to assign several parameters for each posting, such as credit and debit accounts, amount, currency, etc. Since each transaction has multiple life cycle stages, an error in one stage leads to an error not only in the next stage but can also affect other transactions. Typically, it is very complicated and time-consuming to trace back in which step an error was made.

One of the solutions to avoid or reduce errors in the posting process is to use templates. However, in practice, not all accountants are using templates. The reasons for this situation may vary, for example, some accountants do not know about such functionality due to the complexity of the accounting system or do not prepare these templates because of the lack of time.

Another solution to reduce errors in the posting process is the automation of the process or certain parts of it. In the process of automation, the accountant shall be offered the accounts used in invoices of a similar type and shall not be required to prepare specifically for this process, as the system has done the preparatory work to find the most relevant accounts.

In a recent work, Beļskis et al. (2020) showed that it is possible to reach relatively high accuracy of 93% in debit account classification by using features that allow identifying parties (buyers and suppliers) involved in a financial document and provide information of just numeric features (document and posting entry amounts). Although such a limited feature set allows achieving good results, thereby allowing the method to be used practically, they also showed that there is a drop in performance (e.g., to 79% when using the averaged perceptron classifier) when posting parties are unknown to the pre-trained classification model. To counteract the drop in performance, Beļskis et al. (2020) supplemented training data with training examples where suppliers and/or buyers were masked. This allowed increasing the classification precision to 87% (for the averaged perceptron classifier). However, there is also textual information present in many financial documents, such as row and document comments. These comments are made by accountants and can describe the type of transaction or justify the choice of a journal entry, or provide other important information that could help understand the posting principles. Therefore, in this work, we put forth two hypotheses:

- Comment fields describing the document and posting entries contain relevant information for posting account classification.
- Comment fields as additional features can improve posting account classification beyond prior work.

It may be that information that accountants use for making a decision about posting cannot be found in the accounting system and requires external knowledge. However, testing the first hypothesis will allow us to identify whether accountants at least to some extent document information that allows making a correct posting decision in the textual comment fields. To test the hypothesis, we will train posting account classifiers using features derived from only the textual comment fields. We will also investigate what characteristics (or tendencies) can be observed in textual comment fields that allow achieving higher or lower classification results.

The paper is further structured as follows: Section 2 discusses related work on text classification, Section 3 describes the methodology used in the paper, Section 4 describes the experimental setup (i.e., the data and the machine learning classification methods used in our experiments), Section 5 provides analysis of textual comment feature relevance for posting account classification. It also provides empirical proof for the first hypothesis. Section 6 documents the overall posting account classification results and provides empirical proof for the second hypothesis. Finally, Section 7 concludes the paper.

2 Related Work

As shown by Fisher et al. (2016) classification of financial documents has been an actively researched field. However, closest to our use case is related work on posting entry

classification (e.g., Agarwal et al. (2019), Beļskis et al. (2020), Bergdorf (2018) etc.) and financial transaction classification (e.g., García-Méndez et al. (2020), Bengtsson and Jansson (2015), Mateush et al. (2018), etc.).

Agarwal et al. (2019) propose an automated financial transaction system that uses a keyword spotting mechanism (with the help of synonym tables) for automatic posting of financial transactions. Although theoretically possible, this method cannot be applied for posting account classification of human-created posting entries, because accountants do not limit their language to a fixed vocabulary and that vocabulary may differ from buyer to buyer. However, the author method is a viable solution in fully automated scenarios where human intervention is not possible or would be too costly.

Beļskis et al. (2020) described supervised machine learning methods that can effectively predict debit and credit accounts for posting entries. They concluded that the best methods were decision tree, feed-forward neural network and averaged perceptron. They reported that they achieved a credit account classification precision of 98% and a debit account classification precision of 93%. Different from their work, we focus on analysing the usefulness of textual comment features in posting account classification. We also show that when combining textual comment features with the productive features identified by Beļskis et al. (2020), we can improve classification results further.

Bergdorf (2018) compared rule based methods, namely, FURIA and MODLEM, with machine-learning-based methods, namely, random forest and support vector machine classifiers, and concluded that the random forest classifier achieves a minor improvement over rule-based methods. Compared to our work, Bergdorf (2018) experiment using data that differs from our examined use case. Although the data represents double entry bookkeeping posting entries, it does not account for the destination of the money flow, which consequently allows to use single classification model for both debit and credit account classification. However, their dataset allows to train only company-specific models. As the accounts used by Bergdorf (2018) are specific to Sweden, just as ours are specific to Latvia, it could have been possible to investigate a cross-company model, but it was not considered by the authors.

Bengtsson and Jansson (2015) described work on classification of financial transactions with several machine learning classifiers. They found that for their use case a simple rule-based method is suited best. However, they observe cases where there is only one classification entry for each financial transaction, however in our use case financial documents typically have two or more posting entries per document. They also consider only a scenario where parties that are involved in a transaction are known, however, we analyse whether the classification methods can be applied to posting entries of unknown parties.

Although not directly related to posting account classification, García-Méndez et al. (2020) proposed a method for financial transaction classification that uses a classifier that combines a similarity-based method for transactions, which are highly similar to transactions previously observed in training data, and a support vector machine classifier for unseen (or not similar) transactions. Our work differs from the work by García-Méndez et al. (2020) by using a significantly larger data set (over 1.7 million entries compared to just 30.8 thousand entries) and also a considerably larger class set (210 compared to 15). Our data comes also from a production accounting system that fea-

tures many buyers. Different buyers tend to use similar or identical comments for different debit accounts. For instance, the row comment with the most debit accounts (21 in total) is “materiali” (in English: materials) and it is featured in 2,535 posting entries of the training data. It is also the document comment with the most debit accounts (62 in total) and it is featured in 36,137 posting entries. Therefore, a rule-based text similarity method is not applicable in our analysed use case.

Mateush et al. (2018) created a “hybrid classifier” by combining machine-learning-based methods, user-defined rules, and manually assigned user labels. Their goal was to classify card payments and account payments according to industries. As the data used in the paper is cross-country, the classification target could not have been account codes as each country has a unique set of natively used account codes. Vectorized text features were used in the paper but there is no information as how the data was vectorized as well as the impact of the used text features.

3 Methods for Posting Account Classification

In this work, we follow the findings from Beļskis et al. (2020) and extend the best-performing posting account classification methods with support for textual features. More specifically, we focus on the averaged perceptron (Collins, 2002) and deep neural network classifiers, which achieved the best results in the previous work by Beļskis et al. (2020). As baseline classification methods, we use three types of majority classifiers: ZeroR, which does not use any features, OneR, which uses the majority class for each supplier and falls back to ZeroR if supplier information is missing, and a majority classifier (named TwoR further), which uses the majority class for each buyer and supplier pair and falls back to the respective OneR or the ZeroR classifier if either buyer or supplier information is missing. The ZeroR classifier sets the lower bound classification performance, which should be reached by machine-learning-based classifiers that use productive features. The OneR and TwoR classifiers act as stronger rule-based baselines.

We test the first hypothesis by training posting account classifiers using two classification methods: the ZeroR (or majority class) classifier to show the baseline classification performance without using any features, and the averaged perceptron classifier when using three types of features - row comment, document comment, and combined row and document comment features.

To test the second hypothesis, we compare posting account classifiers using different feature sets with and without textual comment features. Classifiers are trained using all three baseline classifiers and the averaged perceptron and feed forward neural network classifiers that were identified in previous work as best-performing for posting account classification.

4 Experimental Setup

4.1 Data for Posting Account Classification

The data that is used in our experiments comes from a real-world accounting system, the same system the data for Beļskis et al. (2020) came from. In comparison with Beļskis

Table 1. Example of posting entries from a financial document (the first row corresponds to goods; the second row corresponds to taxes)

Buyer		Supplier		Posting-specific data					Document-specific data				
Nº	NACE	Nº	NACE	Rel. sum	Sum	Rev. VAT	Debit acc.	Row comment	Year	Sum	Doc. series	Doc. comment	Curr.
X	4719	X	4643	1	92.2	1	7110	Saņemta prece	2019	111.56	VT	Saņemta prece, sw	EUR
X	4719	X	4643	0.21	19.36	1	5721	Priekš-nodoklis	2019	111.56	VT	Saņemta prece, sw	EUR

et al. (2020), the data contains four additional months of data and three additional features - document series, document comment, and row comment. An example of two posting entries from a single financial document (with anonymised parties) from the dataset is given in Table 1. As in this work we are interested in analysing the usefulness of textual comment fields, we discarded entries that did not have either document or row comments. Out of 2,349,267 entries:

- 1,625,892 entries contained document comments;
- 619,649 entries contained row comments;
- 1,790,349 entries contained row comments or document comments.

The document comment field on average contains 23.26 symbols (37,821,878 total) and 2.82 words (4,587,651 total). The row comment field on average contains 18.74 symbols (11,614,341 total) and 2.42 words (1,501,644 total). This shows that comments that are added to documents by accountants are relatively short and concise.

The most frequent words in row and document comment fields are listed in Table 2. We can see that there is a clear difference between the most frequent words featured in row and documment comments - row comments feature words that identify posting entries for taxes (e.g., "*PVN*" (VAT), "*priekšnodoklis*" (input tax), "*21*"). The document comments, on the other hand, typically do not feature tax related information. Since Beļskis et al. (2020) found that the inability of differentiating between the main and tax amount entries caused the majority of misclassifications, we see that row comments could assist in mitigating this issue. Both row and documment comments feature information about various products (e.g., "*preču*" (of goods)), services (e.g., "*noma*" (rent)), and actions (e.g., "*iegrāmatots*" (is posted), "*iepirkšana*" (purchase)). We believe that such information can be helpful in differentiating between non-tax-related posting accounts.

Data for posting account classification model training and evaluation were split as follows: 1,680,349 entries were used for training, 10,000 entries were used for validation (e.g., early stopping and monitoring of training progress), and 100,000 entries were used for evaluation. Data were randomised before splitting them in the three data sets.

In our experiments, we use seven different feature sets. To test the first hypothesis, we use three feature sets that consist of just row comment features, document comment features, and combined row and document comment features. To test the second hypothesis, we start with the best-performing feature set as identified by Beļskis et al. (2020). This feature set uses the following 10 feature categories: year of the financial

Table 2. Most frequent tokens in the row and document comment fields

Row comments			Document comments		
Word	**Count**	**English translation**	**Word**	**Count**	**English translation**
pvn	153,917	VAT	materiāli	157,549	materials
iegrāmatots	85,770	is posted	prece	133,721	goods
preču	70,547	of goods	preču	126,915	of goods
iepirkšana	57,326	purchase	pakalpojumi	125,033	services
prece	35,350	goods	saņemta	110,614	received
pakalpojumi	33,553	services	preces	73,820	goods
materiāli	22,249	materials	iegāde	71,351	acquisition
nr	21,811	No	nr	61,902	No
eur	20,132	EUR	izdevumi	59,516	expenses
priekšnodoklis	18,942	input tax	par	58,014	for
saņemta	16,524	received	pārdošanai	55,943	for selling
21	14,657	21	noma	49,585	rent
izdevumi	14,330	expenses	saņemšana	46,507	receival
1	12,967	1	sw	44,178	sw
par	11,562	for	2018	41,510	for
noma	10,878	rent	saņemšana1	40,285	receival1
rēķ	10,511	invoice (abbr.)	sia	39,573	LLC
2018	9,826	2018	2019	37,352	2019
r9	9,181	r9	celtniecības	36,556	of construction

document, registration number of the buyer, NACE code of the buyer, registration number of the supplier, NACE code of the supplier, proportion of the row amount in respect to the document amount, document amount, document amount rounded to hundreds, reverse value added tax indicator, and currency. Then, for the fifth feature set we add row amount and row amount rounded to hundreds. We continue by adding row and document comment features in the sixth feature set. Finally, for the seventh feature set, we add document series as an additional feature.

4.2 Classification Methods

For classification, we use three rule-based classifiers: ZeroR, OneR, and TwoR, and two machine-learning-based classifiers: an averaged perceptron classifier and a feed-forward neural network classifier. For the ZeroR, OneR, and TwoR classifiers, we use an in-house implementation[3].

The averaged perceptron classifier is an extended version of the implementation by Beļskis et al. (2020), which in turn is based on the implementation by Pinnis (2018). We extend the classifier with additional features for document comments, row comments, document series, as well as separate document amount and row amount features (the previous work considered only document amount and relative proportion features). For document and row comments, we perform the following pre-processing steps: 1)

[3] The source code can be found online at:
https://github.com/tilde-nlp/averaged-perceptron-classifier

all text is lower-cased, 2) stop-words are discarded, 3) all words are stemmed using a rule-based stemming algorithm for Latvian. Word features from document and row comments are treated as separate features. Different from previous work, the averaged perceptron classifier models are trained till convergence and not for a fixed number of epochs.

The feed-forward neural network is implemented using the *pytorch* machine learning framework (Paszke et al., 2019) different from previous work by Beļskis et al. (2020) who used the *scikit-learn* (Pedregosa et al. (2011)) implementation. We use a different implementation because the comment features increase the feature set significantly and training a feed-forward neural network model on CPU becomes infeasible when adding comment features. Apart from that, the model architecture is the same. All textual features were *n-hot* encoded using the *CountVectorizer* from *scikit-learn* with the binary output parameter set to true. All numeric features were one-hot encoded using the OneHotEncoder from scikit-learn. Model hyperparameters were as follows:

- ReLU activation function
- Binary cross-entropy loss function
- 10% dropout
- 1000 hidden neurons
- 1 hidden layer
- maximum of 100 epochs (convergence was reached at approximately 33 epochs)
- Gradient clipping with a value of 1
- Sigmoid output normalization

5 Analysis of Textual Comment Feature Relevance for Posting Account Classification

To test the first hypothesis, we trained posting account classifiers using just the textual comment features and compared the quality to the rule-based baselines. The results in Table 3 show that the classifiers that were trained using just document comment features and both row and document comment features (see second column) surpass the quality of the ZeroR (or majority) classifier. The classifier that was trained using just row comment features did not surpass the baseline, because only 34.7% of all entries in the test set contain row comments. However, if we combine the comment-based classifiers with ZeroR for empty entries (see third column), we clearly see that all comment-based classifiers exceed the baseline ZeroR classifier. Furthermore, we see that the comment-based classifiers that use row comment features and the combined row and document comment features surpass also the two stronger rule-based baseline models that use supplier (OneR) and both supplier and buyer (TwoR) information. This allows us to conclude that comments that are added by accountants to financial documents contain relevant information for posting account classification.

The data set used for training and evaluation consists of data from a production system. This allows us to analyse also whether textual comment features allow performing better classification for individual buyers and to identify what characteristics are evident for buyers for which classification results are higher or lower. For this, we investigated

the data from the 20 most frequent buyers. The individual buyer classification results are given in Table 4. We compared the best comment-based classifier (column three) and the ZeroR classifier (column four). The ZeroR classifier was trained separately for each buyer using only the data from the particular buyers. The table also lists the respective buyer use of row comment fields (column six) and document comment fields (column seven). We found the following characteristics:

- The results show that buyer 1 provides very useful information in row comments. Upon closer inspection, it is evident that the buyer's accountants consistently differentiate between tax-related entries and other entries in row comments. A similar trend is evident for buyers 2, 3, 6, and 9. However, because row comments are not present for all financial documents, the comment-based classifier cannot always differentiate between tax-related and other entries.
- Some accountants copy document comments in row comments. This may hinder differentiating between multiple posting entries of a single financial document. This is especially evident for buyers 4, 14, and 16. However, the accountants of buyer 4 added row comments only for non-tax-related entries, which has allowed the comment-based classifier to differentiate between tax-related and other entries, therefore achieving considerably higher precision compared to the ZeroR classifier. Similar use of row comments is evident for buyer 10, whose accountants use row comments to indicate non-tax-related entries and approximately half of such row comments are identical to document comments.
- Accountants of buyer 15 add row comments for entries for the value added tax. Although this allows differentiating between tax-related and other entries, the value added tax related posting account is the most frequently used and document comments are not informative enough to differentiate between different non-tax-related entries. Therefore, the comments for buyer 15 did not help the comment-based classifier to achieve higher results than the ZeroR classifier.
- Row comments, although present for 48% of entries, were mainly not beneficial also for buyer 20 as accountants used the row comment field to store document numbers instead of relevant information.
- For some buyers (e.g., buyers 5, 8, 12, and 18), the ZeroR precision is very high, which means that these buyers tend to use mainly one posting account for the ma-

Table 3. Classification results when using comment features in comparison with baseline classifiers

	First class for empty entries	**ZeroR for empty entries**
Averaged perceptron classifier		
Document and row comment features	65.79%	66.03%
Document comment features	51.93%	53.36%
Row comment features	29.17%	62.60%
Baseline classifiers		
ZeroR classifier		49.69%
OneR classifier		56.67%
TwoR classifier		61.74%

Table 4. Classification precision for most frequent buyers and statistics of their use of comment fields (P stands for precision)

Buyer	Entries	P (perceptron)	P (ZeroR)	Diff.	Entries with row comm.	Entries with doc. comm.	Doc. comm. equal to row comm.
1	7,893	99%	60%	39%	100%	11%	2%
2	7,288	80%	63%	16%	34%	100%	7%
3	4,930	55%	45%	9%	9%	100%	0%
4	3,299	83%	41%	41%	47%	100%	42%
5	2,907	98%	98%	0%	0%	100%	0%
6	2,568	88%	51%	37%	39%	100%	0%
7	2,162	74%	63%	11%	3%	100%	0%
8	1,533	92%	91%	1%	4%	100%	0%
9	1,273	79%	56%	23%	52%	100%	16%
10	1,169	84%	49%	35%	44%	100%	20%
11	1,073	55%	46%	9%	3%	99%	1%
12	981	97%	98%	-1%	0%	100%	0%
13	938	52%	46%	6%	2%	100%	0%
14	867	63%	30%	33%	100%	100%	88%
15	848	52%	53%	-1%	19%	100%	1%
16	846	57%	43%	14%	98%	100%	70%
17	834	44%	50%	-6%	0%	100%	0%
18	824	98%	94%	4%	100%	100%	3%
19	817	41%	51%	-10%	1%	100%	0%
20	799	52%	48%	3%	48%	100%	0%

jority of financial documents. Nevertheless, the comment-based classifier is able to match and for buyer 18 even outperform the ZeroR classifier.

– We also see that for buyers whose accountants in general do not enter row comments (e.g., buyers 5, 12, 17, and 19), the classification precision of the comment-based classifier is subpar to ZeroR, thereby showing the importance of relevant row comments.

The analysis showed that there is no unified or standard approach how row comments are filled by accountants. Some accountants ignore these fields, some copy document comments or document numbers, some provide information relevant to posting entries. In cases where row comments contain information relevant to posting entries, this information, as shown by the results in Table 4, can be beneficial for posting account classification.

Next, we analysed whether the textual comment features when combined with other features that are extracted from financial documents, allow improving posting account classification for individual buyers. Table 5 shows that when adding comment features (column six) to the feature set used by Beļskis et al. (2020), the classification precision improves for 14 out of 20 analysed buyers. For 5 buyers the precision did not change (including 3 buyers with a precision of 100% even before adding comment features). Precision slightly dropped for buyer 17 whose accountants, as explained above, do not use row comments for posting entries.

Table 5. Classification precision for most frequent buyers for different feature sets. PF stands for the set of the productive features identified by Beļskis et al. (2020).

Buyer	Entries	Precision			
		ZeroR	Only comments	PF + row amount	PF + row amount + comments
1	7893	60%	99%	100%	100%
2	7288	63%	80%	100%	100%
3	4930	45%	55%	93%	96%
4	3299	41%	83%	91%	95%
5	2907	98%	98%	100%	100%
6	2568	51%	88%	97%	99%
7	2162	63%	74%	85%	89%
8	1533	91%	92%	98%	98%
9	1273	56%	79%	95%	97%
10	1169	49%	84%	96%	99%
11	1073	46%	55%	75%	80%
12	981	98%	97%	99%	99%
13	938	46%	52%	90%	92%
14	867	30%	63%	91%	95%
15	848	53%	52%	93%	95%
16	846	43%	57%	89%	92%
17	834	50%	44%	98%	97%
18	824	94%	98%	97%	99%
19	817	51%	41%	94%	95%
20	799	48%	52%	85%	89%

We were also interested in identifying whether longer comments feature more relevant information for posting account classification. Therefore, we analysed the overall classification precision for posting entries featuring textual comments of different lengths. The analysis results are depicted in Figures 1 (when using a model that was trained using only textual comment features) and 2 (when using a model that was trained using the best-performing feature set). It is evident that shorter (more precise, more laconic) row comments are better suited for posting account classification. Posting entries that have row comments that are only two words long allow achieving the highest classification accuracy. However, document comments allow achieving peak classification accuracy (if we ignore entries without document comments) when they are approximately 10 words long. This was to be expected as document comments typically feature more information than just what may be relevant for posting account classification. If we compare the two figures, we also see that comment length is less indicative of classification precision for the model that was trained using the best-performing feature set. However, also here we see that classification peaks when row comments feature only two words[4].

[4] Although Figure 2 depicts peak precision for comments that are 20 words long, this is achieved for an insignificant number of posting entries - only five entries were 20 words long.

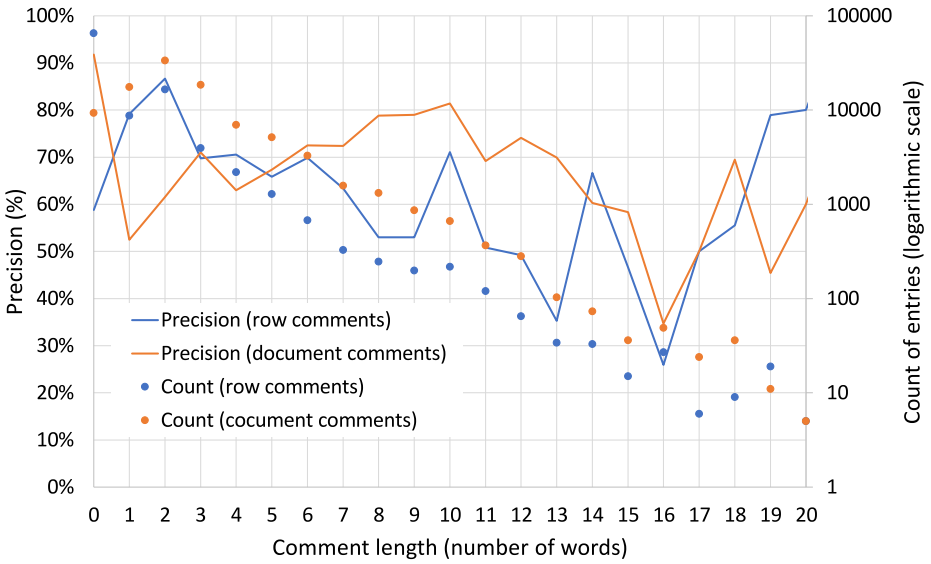

Fig. 1. Posting account classification precision for the model trained using only textual comment features for posting entries featuring different lenghts of textual comments

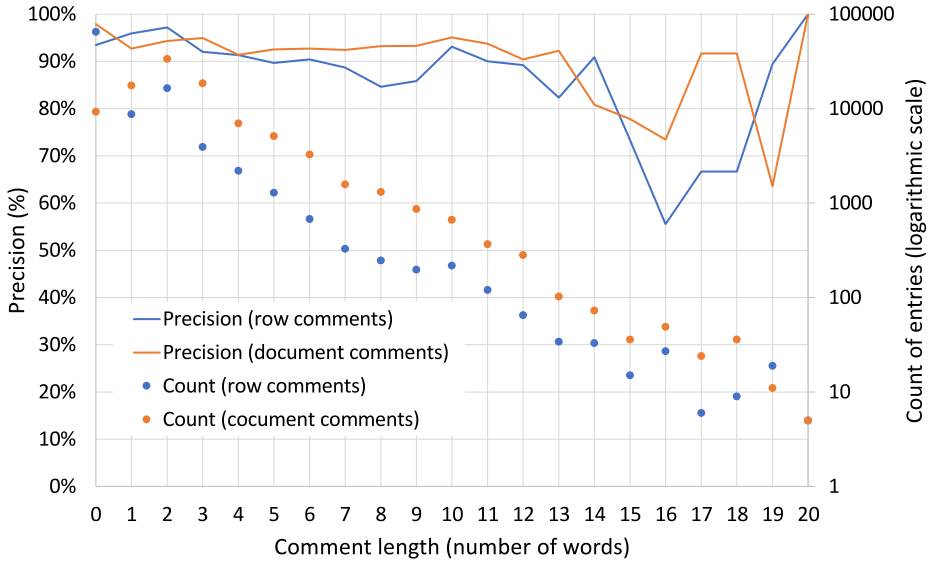

Fig. 2. Posting account classification precision for the model trained using the best-performing feature set for posting entries featuring different lenghts of textual comments

6 Posting Account Classification Results

After having identified that posting account classification can benefit from textual comment features, we trained posting account classification models using the averaged per-

Table 6. Posting account classification results for averaged perceptron and feed-forward neural network classifiers. PF stands for the set of the productive features identified by Beļskis et al. (2020).

Classifier / feature set	Precision
Baseline classifiers	
ZeroR classifier	49.69%
OneR classifier	56.67%
TwoR classifier	61.74%
Averaged perceptron classifier	
PF	91.21%
PF + row amount	91.55%
PF + row amount + comments	93.83%
PF + row amount + comments + document series	94.04%
Feed-forward neural network classifier	
PF	92.22%
PF + row amount	92.06%
PF + row amount + comments	94.74%
PF + row amount + comments + document series	94.78%

ceptron and feed-forward neural network classifiers. For comparison with the work by Beļskis et al. (2020), we include classification results of both neural network-based classifiers using their most productive feature set. To test the hypothesis that textual comment features allow increasing results beyond prior work, we train classifiers with three additional feature sets. More specifically, we incrementally add row amount features, textual comment features, and the document series feature to the productive feature set established in related work. The classification results are given in Table 6. The results show that the addition of row ammount features increases the classification precision for the averaged perceptron classifier, but lowers classification results for the feed-forward neural network classifier. The addition of textual comment features increases classification precision by 2.28 and 2.68 percentage points for the averaged perceptron classifier and feed-forward neural network classifier respectively. This constitutes an increase of up to 2.52 percentage points over previous work, thereby validating the second hypothesis. The best classification results were achieved when including in the feature set also document series features. This constitutes a cumulative increase of up to 2.56 percentage points over previous work.

Finally, we analysed whether textual comment features help maintaining classification precision when posting parties are unknown to the pre-trained classification model. For this, we trained additional classification models with the previously established feature sets. We also generated for each posting entry synthetic alternatives with masked buyer, supplier, and both registration numbers, thereby effectively quadrupling the training data amount. Then, we evaluated the models using the evaluation data in three scenarios: 1) with buyer and supplier information available, 2) with buyer information masked (unavailable during classification), and 3) with supplier information masked. The results of this evaluation are provided in Table 7. It is evident that synthetic data allows training models that are more robust and achieve higher classification precision

Table 7. Posting account classification results for the averaged perceptron classifier trained using four different feature sets when masking buyer and supplier information and using synthetic data. PF stands for the set of the productive features identified by Beļskis et al. (2020).

	PF	+ row amount	+ comments	+ document series
Buyer and supplier not masked				
1 Without synthetic data	91.21%	91.55%	93.83%	94.04%
2 With synthetic data	90.73%	91.05%	93.57%	93.78%
Buyer masked (not known)				
3 Without synthetic data	81.00%	81.73%	89.15%	92.84%
4 With synthetic data	88.68%	88.90%	92.84%	93.10%
Difference $(4-1)$	-2.53	-2.65	-0.99	-0.94
Supplier masked (not known)				
5 Without synthetic data	84.33%	85.28%	91.46%	92.09%
6 With synthetic data	87.90%	88.38%	92.87%	93.27%
Difference $(6-1)$	-3.31	-3.17	-0.96	-0.77

when buyer and supplier information is unknown. We also see that comment features allow maintaining a comparable classification quality even with masked buyer or supplier information. The quality reduces by less than one percentage point compared to a loss of 2.53 to 3.31 percentage point when not using textual comment features. This shows that the models with textual comment features may be better suited when classifying data for new (or unknown) buyers and suppliers.

7 Conclusion

In this paper, we analysed the usefulness of textual comment fields in posting account classification. We showed that comment fields from the production accounting system do contain relevant information for posting account classification and that usage of textual comment fields together with other features that are extracted from posting entries can boost posting account classification results. We achieved the highest posting account classification precision of 94.78% when using the feed-forward neural network classifier, which is 2.56 percentage points higher compared to previous work.

We also analysed characteristics of textual comment fields that allowed achieving higher classification results. Our analysis of the most frequent buyers in our data set showed that posting entries that contained concise row comments (approximately two words long) that describe the posting entry can allow performing posting account classification with a precision of up to 99%. We also showed that if accountants do not use row comments or store irrelevant information (such as document numbers or copies of document comments) in row comments, textual comment features may have no beneficial effect in posting account classification. This shows that if the posting process is to be automated, accountants have to be instructed on best practices in how to provide relevant information for posting account classification.

Finally, we also showed that textual comments can help training more robust posting account classification models for situations when processing data from unknown buyers or suppliers. We measured a reduction in precision of less than one percentage

point when masking buyers or suppliers in the evaluation data. However, in future work this finding should be validated when using data from truly unknown (new) buyers or suppliers and not synthetic data.

Source code for training of the averaged perceptron and neural network-based classifiers will be open-sourced and shared on GitHub upon publication of the paper.

Acknowledgements

The research has been supported by the ICT Competence Centre (www.itkc.lv) within the project "2.6. Research of artificial intelligence methods and creation of complex systems for automation of company accounting processes and decision modeling" of EU Structural funds, ID n° 1.2.1.1/18/A/003.

References

Agarwal, S., Mukherjee, P., Chakraborty, B., Nandi, D. (2019). A novel automated financial transaction system using natural language processing, *International Conference on Advanced Machine Learning Technologies and Applications*, Springer, pp. 535–545.

Bengtsson, H., Jansson, J. (2015). *Using classification algorithms for smart suggestions in accounting systems*, Master's thesis, Department of Computer Science & Engineering, Chalmers University of Technology.

Bergdorf, J. (2018). Machine learning and rule induction in invoice processing: Comparing machine learning methods in their ability to assign account codes in the bookkeeping process.

Beļskis, Z., Zirne, M., Pinnis, M. (2020). Features and methods for automatic posting account classification, *Proceedings of 14th International Baltic Conference on Databases and Information Systems*, (in press).

Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms, *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, Association for Computational Linguistics, pp. 1–8.

Fisher, I. E., Garnsey, M. R., Hughes, M. E. (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research, *Intelligent Systems in Accounting, Finance and Management* **23**(3), 157–214.

García-Méndez, S., Fernández-Gavilanes, M., Juncal-Martínez, J., González-Castaño, F. J., Seara, Ó. B. (2020). Identifying banking transaction descriptions via support vector machine short-text classification based on a specialized labelled corpus, *IEEE Access* **8**, 61642–61655.

Mateush, A., Sharma, R., Dumas, M., Plotnikova, V., Slobozhan, I., Übi, J. (2018). Building payment classification models from rules and crowdsourced labels: A case study, *International Conference on Advanced Information Systems Engineering*, Springer, pp. 85–97.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library, *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., pp. 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**, 2825–2830.

Pinnis, M. (2018). Latvian Tweet Corpus and Investigation of Sentiment Analysis for Latvian, *Human Language Technologies – The Baltic Perspective - Proceedings of the Seventh International Conference Baltic HLT 2018*, IOS Press, Tartu, Estonia, pp. 112–119.

Received October 26, 2020 , revised March 15, 2021, accepted April 19, 2021