An Automated Identification of Forward Looking Statements on Financial Metrics in Annual Reports

Research Paper

Khanh Le Nguyen[1], Diana Hristova[1]

[1] Berlin School of Economics and Law, Department of Business and Economics, Berlin, Germany
{khanhle.nguyen,diana.hristova}@hwr-berlin.de

Abstract. This paper presents a three-phase Decision Support System (DSS) designed to automatically extract, analyze and represent forward-looking information on financial metrics from corporate 10-K reports (10Ks). It was evaluated with S&P 500 company 10-Ks. Phase I uses Natural Language Processing and rule-based techniques to extract relevant sentences, achieving 94% accuracy. Phase II predicts future metric growth using Random Forest and FinBERT models. Random Forest outperformed FinBERT in the evaluation, demonstrating that simpler models can be more effective. Phase III enhances transparency and readability for the user with Explainable and Generative AI, illustrating expected growth and underlying reasoning. Generative summaries achieved an average rating of 3.69 in the evaluation for their high factual consistency and readability. Our DSS transforms unstructured narrative disclosures into actionable, metric-level insights, empowering investors and analysts to make more expert-informed financial decisions.

Keywords: *forward-looking statements, 10-K, financial performance prediction, XAI, GenAI*

1 Introduction

10-K reports (10Ks) are annual reports required by the US Securities and Exchange Commission (SEC) for publicly traded companies. They are an important source for investors to indicate and predict the company's financial performance (Beyer et al., 2010). Aside from the parts where companies disclose their numbers and metrics, in a standard 10-K, the narrative section stands for approximately 80% of the content (Lo, Ramos and Rogo, 2017). Moreover, 10-Ks have grown in length and complexity in the past years, making manual information extraction more difficult (Dyer, Lang and Stice-Lawrence, 2017). Thus, there has been a rising need to automatically identify narrative sections and extract useful information from them (Hsieh and Hristova, 2022).

20th International Conference on Wirtschaftsinformatik,
September 2025, Münster, Germany

Additionally, the SEC has emphasized the importance of forward-looking statements (FLS) in the reports. FLS encompass future projections, outcomes, circumstances, or prospective occurrences that companies envision or foresee. They often contain expressions such as "believe", "anticipate" and "expect". FLS may be of different nature, from hypothetical scenarios to specific statements about future performance. Among those, *metric-related* FLS, defined as statements relating to future financial performance, have been proven to be more informative to investors (Hussainey and Walker, 2009).

Prior research has attempted to automate FLS detection (Tao, Deokar and Deshmukh, 2018; Glodd and Hristova, 2023) but typically treats all FLS types equally or stops at classification. Other studies, such as Li (2010), attempt general financial performance prediction but often ignore explainability and the needs of end-users, such as analysts. In practice, analysts require transparent, metric-specific predictions grounded in actual corporate language, ideally as part of an interactive tool to support their decisions.

Our work contributes to the information systems and finance field by operationalizing a transparent, automated Decision Support System (DSS) that bridges financial text analytics and practical decision-making. This aligns with research showing explainability reduces information asymmetry and fosters system adoption (Weber, Carl and Hinz, 2024), and with the broader direction of digital investment support systems automating data processing and assisting user decisions (Gomber et al., 2018).

Thus, in this paper, we aim to answer the question: *"How can metric-related FLS in 10Ks be automatically extracted and translated into transparent, metric-specific performance forecast to support financial decision-making?".* To this purpose, we propose a three-phase DSS, extending the works by Glodd & Hristova (2023) and Li (2010). In Phase I, we extract metric-related FLS from 10-Ks with modern Natural Language Processing (NLP) techniques. In Phase II, we use these metric-related FLS to predict the future growth of those metrics, thus providing analysts and investors with a tool to support investment decisions. Finally, in Phase III, we follow Glodd & Hristova (2023) and apply an Explainable AI (XAI) approach to understand the reasoning behind the models from Phase II. Additionally, we also utilize Generative AI (GenAI) to summarize the metric-related FLS parts of the reports (in the following, generative summaries) and add a human-like grasp of these lengthy documents, thus facilitating transparency and trust towards our models (Gopalakrishnan, 2024).

2 Related work

In this section, we first present the literature on FLS in 10-Ks, followed by that on financial metrics' information extraction. 10-Ks are obligatory, highly standardized reports that consist of four parts and 15 items. They are publicly available on the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) database. As mentioned above, FLS play an important role in 10-Ks. Athanasakou & Hussainey (2014) found that managers tend to be more forthcoming about future performance in their reports during years when the company raises debt or reports poor earnings news in the financial statements. Bozanic et al. (2018) showed that a higher number of FLS without

quantitative statements about earnings are issued in times of higher uncertainty. This implies that companies use FLS strategically and thus it is crucial to extract them from the reports. Due to the length, complexity, standardization and digital availability of 10-Ks, research has focused on increasing the automation of FLS extraction.

Initially, this was done mostly with a rule-based approach (Li, 2010; Muslu et al., 2015). However, this requires an iterative and costly definition of the rules and may generate many false positives (Tao, Deokar and Deshmukh, 2018). Thus, Tao et al. (2018) proposed to combine it with machine learning (ML) models. The authors manually reviewed the results from the rule-based approach and used the generated dataset to train both traditional ML models (e.g., Support Vector Machine (SVM)) and deep learning ones (e.g., Long Short-Term Memory network (LSTM)). Based on the extracted FLS, they derived different features such as FLS topics, sentiment, quantity, and word frequencies and successfully used those for IPO valuation prediction.

Glodd & Hristova (2023) assessed the idea further with an NLP DistilBERT model, which they compared with methods such as SVM and LSTM. DistilBERT is a lighter version of the BERT model that has shown impressive results in many areas. Thus, unsurprisingly, DistilBERT outperformed other models, highlighting it as the best tool to classify FLS. Similar to Tao et al. (2018), Glodd & Hristova (2023) used the quantity and sentiment of the extracted FLS with a Random Forest Regressor (RFR) for stock price growth prediction. A RFR is a very powerful approach, known for its efficiency, interpretability and good performance with non-linear data relationships.

The above works focus on FLS addressing all topics, while, as mentioned above, literature has shown the importance of *metric-related* FLS. Without focusing on FLS, Kamaruddin et al. (2009) developed a rule-based approach to extract sentences from financial statements referring to "…net profit/loss, share capital and total assets…" (p. 1). According to the authors, using pattern-matching techniques with the list of predefined financial metrics can help extract performance-relevant information from the text without considering the arrangement and structure of language elements (syntactic information). Das et al. (2017) utilized ML techniques to automate sentence labelling in the categories of "Accounting", "Cost", "Employee", "Financing", "Sales", "Investments", "Operations", "Profit", "Regulations" and "Irrelevant" (p. 1). They employed an algorithm combining SVM with a modified label propagation approach.

Li (2010) combined both financial metrics' and FLS extraction using ML. FLS were identified with a rule-based approach and following this, the author manually categorized the content of 30.000 FLS into twelve pre-defined categories and four different sentiments. These are used as the training data for a Naïve Bayesian classifier. Li (2010) examined the relationships between the results from the classifier (sentiment for a given category) and different company characteristics, such as current earnings, contemporaneous stock returns and accruals using regression analysis. Additionally, the relationship between the sentiment of FLS and future earnings was analyzed, following the same approach. Although this work is impressive and explores an idea of great potential, it does not consider the latest developments on FLS extraction, described above.

In this paper, we aim to close this gap by developing an automated and transparent three-phase DSS for the extraction and analysis of metric-specific FLS from 10-Ks. In Phase I, we extract FLS on financial metrics using rule-based methods in combination

with modern NLP models, thus extending the works by Li (2010) and Glodd & Hristova (2023). In Phase II, the results are applied for future metric performance prediction, following Tao et al. (2018) and focusing on the context of FLS, rather than their sentiment and quantity. Finally, in Phase III, similar to Glodd & Hristova (2023), we examine the most important features for each metric, followed by summarization of the metric-related FLS for each company per year. All in all, our core contribution is a validated and explainable DSS that bridges advanced financial NLP with real-world forecasting needs. It empowers analysts and investors to gain actionable, metric-specific insights from dense corporate disclosures, improving decision-making through automation, transparency, and trust. In the next section, we present our methodology.

3 Methodology

Our three-phase DSS is shown in Figure 1, with Phases in roman numerals and the underlying steps under these Phases in Arabic numbers. The DSS requires the raw 10-Ks as input. Previous literature has analyzed Items 1A, 3, 7 and 7A for FLS, in which management typically discusses the company's outlook (Muslu et al., 2015; Tao, Deokar and Deshmukh, 2018; Glodd and Hristova, 2023). However, Item 3 seems to have a significantly lower number of FLS than the others (Glodd & Hristova, 2023). Thus, we only focus on Items 1A, 7 and 7A in this research. In the following, we describe each of the three phases in detail.

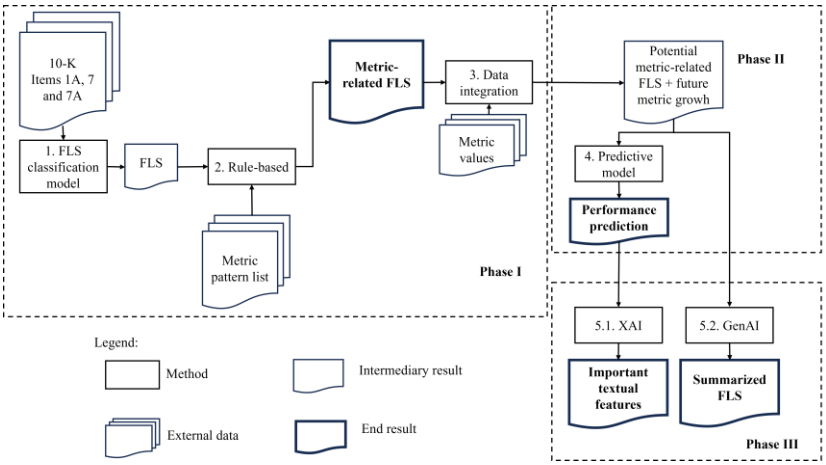

Figure 1. Approach to predict performance based on metric-related FLS

3.1 Phase I – Metric-related FLS extraction

In Step 1, we apply a FLS classification model to the reports, similar to Glodd & Hristova (2023). The result is a set of FLS, from which we would identify metric-related ones in the next step. To determine the set of relevant metrics, we analyzed the literature

on metric predictive power. Bataineh & Rababah (2016) and Dechow (1994) showed the relevance of *net income* (*INCOME*) for future performance. Tao et al. (2018) and Kumar (2017) additionally revealed the significance of Earnings per Share (*EPS*), which is income over the number of common shares. Huang et al. (2015) argued that *revenue* (*REV*, also known as gross income) holds a higher informative value, especially for companies adopting a growth-oriented approach. REV is also the basis for Earnings Before Interest and Taxes (*EBIT*) (Siegel & Shim, 2000) which is a key performance indicator (Velimirović et al., 2011), widely employed in profitability calculations (Ecobici, 2016). Finally, Dechow (1994) empirically examined the predictive capabilities of cashflows, revealing their aptitude for forecasting short-term financial performance. Also, Li (2010), mentioned above, considered cashflows from operations, investment and financing activities (*OCF*, *ICF* and *FCF*), in addition to INCOME, REV, and cost, reflected in EBIT. Thus, in this paper, we focus on INCOME, EPS, REV, EBIT, OCF, ICF and FCF.

In Step 2, we apply a rule-based approach to identify relevant FLS sentences for each of those metrics. A rule for a given metric essentially defines the terms that need to be contained in a sentence for it to be assigned to that metric. Note that since there is not always a common agreement on terminology, financial texts may use different terms for the same metric. Thus, to derive the list of terms for each metric, we analyzed the literature for mentions of the above seven metrics and their synonyms. This led to an initial list of terms (which we call patterns), which was then reviewed and adjusted, resulting in the final list in Table 1.

Table 1. Metric pattern list

Metric	Final pattern list	Example references
REV	Gross income, gross earnings, gross profit, proceeds, takings, receipts, top line	Huang et al., 2015; Huang and Hairston, 2023; Irwin, 2011; Piracha and Moore, 2016
INCOME	Profit, earnings, bottom line, profitability, net income (without *gross* or *operating*)	Bataineh and Rababah, 2016; Halley and Little, 1999; Nishikawa et al., 2016
EBIT	EBIT, operating profit, operating income, earnings before interest + [*e.g., and taxes*]	Ecobici, 2016; Masihabadi et al., 2015; Siegel and Shim, 2000; Silva et al., 2022; Velimirović et al., 2011
OCF	Cash (flow) + operating (activities)/operations OR CFFO	Chen et al., 2012; Dechow, 1994; Millianto and Bangun, 2021; Syaputri, 2019; Wang and Hussainey, 2013
ICF	Cash (flow) + investing (activities)/investment(s)	Dechow, 1994; Varshney and Jain, 2016
FCF	Cash (flow) + financing/funding (activities)	Dechow, 1994; Omag, 2016; Soelehan, 2012; Varshney and Jain, 2016
EPS	Earnings Per Share or EPS	Kumar, 2017; Tao et al., 2018

3.2 Phase II – Performance prediction

In Step 3, we integrate the generated metric-related FLS with the metric values, which are extracted from the SEC's EDGAR via the EDGAR's API. Note that the result from Phase I consists of separate sentences within the reports. Thus, for a given report and metric we could have multiple such sentences. To generate one text per report (and therefore one predicted metric value), we put all FLS for a given metric and report together. For a report rep_T published for the period $[T-1, T]$, this combined text is matched with the future metric growth, calculated as:

$$growth(rep_T) = \frac{value(T,T+1) - value(T-1,T)}{value(T-1,T)} \qquad (1)$$

where $value(T-1, T)$ is the metric value for the same period as rep_T and $value(T, T+1)$ is the metric value for the next reporting period. We consider the difference in the metric values and not just $value(T, T+1)$, because FLS often refer to statements such as "We expect an increase in *metric xxx* during the next year.". Moreover, we focus on growth and not on the absolute difference, because different metrics are measured in different magnitudes, making comparability and estimation difficult. This comes with the additional advantage of interpretability. Also, literature has shown that deep learning models, such as FinBERT, which was trained on financial texts including 10-Ks (Huang, Marquardt and Zhang, 2015), generate more stable results with scaled data. To reduce noise from extreme values in the metric values, we remove outliers based on the interquartile range (IQR) method. Specifically, we excluded samples where the metric growth exceeded 1.5× the IQR above the third quartile or below the first quartile. This ensured that extreme cases do not disproportionately affect model training.

In Step 4, we estimate two predictive models for future metric growth. The prediction task is defined as a supervised regression problem: for each company-year and financial metric, the model predicts the growth of that metric from the current to the next reporting period, using the forward-looking statements (FLS) disclosed in the current year as input. We train a RFR (as in Glodd and Hristova, 2023) and a FinBERT Regressor (FBR). For the RFR, the combined FLS text per report and metric is converted into vectors of lemmatized word frequencies, including 1-, 2-, and 3-grams. Preprocessing removes stop words, punctuation, numbers, and short tokens to reduce noise. This results in a clean bag-of-words representation that captures contextual signals more effectively than prior work focused solely on FLS quantity and sentiment (e.g., Glodd & Hristova, 2023). The FBR takes the combined FLS text as input, tokenized and processed through FinBERT, a pre-trained language model tuned to financial texts. While RFR is simple, fast, and interpretable, FBR allows us to evaluate whether a deep, domain-specific NLP model can improve predictive performance.

This completes the description of Phase II. For each of the metrics, the result is a set of two predictive models (here RFR and FBR) that take metric-related FLS as input and predict the future metric growth. Note that also other models can be applied. In

Phase III, we develop approaches for explaining the results to the decision maker, thus facilitating transparency and trust.

3.3 Phase III – Model explainability

XAI has advanced rapidly alongside increasingly complex NLP models. In finance, XAI helps reduce information asymmetry by showing how predictions are made, thereby fostering user trust and adoption (Weber, Carl and Hinz, 2024). We use SHAP (SHapley Additive exPlanations), based on Shapley values from game theory, to explain the outputs of our models in Phase II (Lundberg and Lee, 2017; Molnar, 2022). SHAP supports model validation and helps analysts identify language patterns that drive predictions, enabling qualitative judgment alongside quantitative output.

Furthermore, to enhance the readability and explainability of our solution, we apply the GenAI model Gemini 1.5 Flash model (Gemini) to generate summaries of the company's metric-related FLS. Gemini is chosen for its efficiency and accuracy in processing large volumes of text data, as well as the capabilities to summarize complex texts while retaining important information (Rane, Choudhary and Rane, 2024). For each report and each financial metric, we use all metric-related FLS as input for the summary, providing a high-level overview of the company's outlook related to a specific metric in a certain year. However, it is crucial to acknowledge the inherent risk of hallucination in GenAI models, where the model might produce plausible but incorrect information from the provided input. Thus, we pay more attention to the prompt, or task-specific instruction, to guide the GenAI model to extract relevant and accurate information, reducing hallucination (Sahoo et al., 2025). This can be used by analysts and investors not only to gain more knowledge about the company's plans and goals but also to compare the model's quantitative predictions with the management's qualitative outlook, thereby facilitating the identification of potential model discrepancies or misleading announcements. It is important to note, however, that there would be no explicit, direct connection between the output of our predictive models in Phase II and the generative summaries produced in Phase III. The summaries are based solely on the FLS text itself, providing management's narrative.

This completes Phase III and therefore our methodology. We present a three-phase DSS that enables analysts and investors to select a company and year, and receive: 1) a list of the FLS related to key financial metrics (Phase I), 2) predicted growth values for those metrics in the following period (Phase II), and 3) the most important model features and concise summaries of the underlying disclosures (Phase III), allowing for a comprehensive understanding and critical validation of the forecasts against management's own narratives. This integrated pipeline supports transparent, data-driven, and expert-informed financial decision-making. In the next section, we evaluate our approach.

4 Results and discussion

For the evaluation, we extracted 10-Ks from SEC's EDGAR for the period from 2015 to 2022 and the S&P 500 companies. We begin with 2015 because all taxonomies (official names of the financial metrics on the EDGAR database) are valid from 2015. Additionally, 2022 is the latest filing year at the time of the analysis. This results in 3,470 reports, restricted to Items 1A, 7 and 7A, as mentioned above.

4.1 Phase I – FLS and metric extraction

In Step 1, we applied the FLS-DistilBERT model described in Glodd & Hristova (2023) to the sentences from the 10-Ks to extract FLS, resulting in 46,534 sentences. The number of FLS per 10-K ranges from just one to as many as 96 sentences. However, most filings contain between 6 and 16 FLS, suggesting a moderate but consistent presence of forward-looking content across reports. Next, in Step 2, we applied the rule-based approach to the extracted FLS to identify sentences discussing financial metrics, or metric-related FLS. For this, we used the final list in Table 1, together with the Python package spaCy. Step 2 resulted in 39,638 sentences addressing the seven metrics above. Out of them, 51% talk about INCOME, while 22% and 21.5% mention EBIT and REV, respectively. Each of the other metrics represents less than 5% of the data with the three cashflow metrics being below 1%.

Before proceeding, we randomly selected 100 extracted sentences for each metric and manually assigned the true metric, if any, addressed in them. For metrics that did not have more than 100 sentences (i.e., the three cashflow metrics), we reviewed all metric-related FLS. Out of the 524 extracted sentences, only one sentence could not be assigned to any metric and 94% of the remaining sentences were assigned correctly through the metric pattern list. Additionally, to determine the false negatives, we randomly extracted 100 FLS that were not assigned to any metric. Out of them, 93 were actually no matches (i.e. true negatives), while the incorrect false negatives mostly discussed OCF.

Table 2. Performance of metric extraction method

Metric	Precision	Recall	F1-score
EBIT	0.95	0.95	0.95
EPS	0.99	0.96	0.98
ICF	0.94	0.99	0.96
FCF	0.71	0.85	0.77
OCF	0.86	0.46	0.60
INCOME	0.88	0.89	0.88
REV	0.96	0.96	0.96
No match	0.93	0.99	0.96

Table 2 provides more details about the performance in terms of precision, recall and F1-score. Precision measures the correctly assigned sentences out of all sentences assigned to a metric. Recall stands for the correctly assigned sentences out of all sentences

truly corresponding to a metric. Since both precision and recall matter, F1-score is a harmonic average of the two. We see that EPS has the best results, with an F1-score of 0.98. Other well-performing metrics are REV, ICF, EBIT and INCOME with F1-scores of more than 0.80. Both FCF and OCF have lower F1-scores. Together with ICF, they have few extracted FLS and are thus removed from the list of metrics, leaving four metrics for the next phase. This completes Phase I, resulting in a list of metric-related FLS.

4.2 Phase II – Performance prediction

In Step 3, we integrated the FLS datasets with the metric values from the EDGAR database by merging the two datasets on the company's central index key – unique key in SEC's system to identify corporations which have filed with the SEC – and the year of the financial report. We cleaned the metric data for outliers and missing values. Also, as mentioned above in Section 3.2, the FLS texts are cleaned of punctuation, multiple white spaces, stop and short words, and numbers as input for RFR, using Gensim Python package. All these add noise to the estimation for RFR. On the other hand, FBR can handle contextual languages and thus does not need cleaned text. The resulting number of data points for each FLS dataset is shown in Table 3. As observed, INCOME and REV have more FLS than EBIT and EPS. Note that, as discussed above, FLS are aggregated per report and metric and so the total number differs from the result mentioned in Step 2.

Table 3. Number of data points after Step 3

	Number of data points
EBIT	445
EPS	508
INCOME	2,753
REV	1,072

In Step 4, we estimated the two predictive models for future metric growth. We split the datasets into training and test sets (80/20, with 80% as training data), same for both models. For FBR (FinBERT Regressor as a reminder), we followed a standard approach, setting the epochs to avoid overfitting. We additionally used a 5-fold cross-validation to help compensate for the smaller datasets. For RFR (Random Forest Regressor as a reminder), we applied GridSearch, again with a 5-fold cross-validation. 5-fold was chosen after testing different values for the best trade-off between over- and underfitting. The GridSearch for RFR is conducted by varying the minimum samples in a leaf with 3, 4, 5, 10, and 15. Also, we varied the number of trees from 5, 10, 50, then 100 to 500 with a step of 100. We tested the three possible max_features in the sklearn Python package. Finally, we extracted both single words, 2-grams, and 3-grams as independent variables.

To assess the contribution of our models, we estimated a linear regression (LR) using the current metric value as an independent variable and the future growth (as in the RFR and FBR models) as a target variable. We cleaned and split the data in the same

way as for RFR and FBR. Table 4 gives an overview of the performance of the three models on the test set, using the mean squared error (MSE) and the mean absolute error (MAE). MSE is a standard error measure in the literature (Botchkarev, 2019) emphasizing larger errors, while MAE is easier to interpret as the average absolute deviation to the true values. Both are dependent on the scale of the target variable. **Bold** values in Table 4 are the best ones for a given metric.

Table 4. MSE and MAE of RFR, FBR and LR on test set

	MSE			**MAE**		
Model	**RFR**	**FBR**	**LR**	**RFR**	**FBR**	**LR**
EBIT	**0.01**	0.04	0.02	**0.10**	0.13	0.11
EPS	**0.11**	0.12	0.58	**0.25**	**0.25**	0.58
INCOME	**0.19**	0.35	0.69	**0.32**	0.45	0.62
REV	**0.01**	0.02	0.05	**0.09**	0.10	0.17

The findings reveal that RFR models outperform the rest in all aspects. Thus, more complex, state-of-the-art models are not always better. Additionally, both RFR and FBR are mostly better than LR, which demonstrates the contribution of our approach. An exception is the FBR model for EBIT, which could be due to the low number of data points. In terms of MAE, we could see good results for EBIT and REV and worse ones for EPS and INCOME, possibly due to metric values' variance as discussed below.

This completes Phase II. The estimated predictive models can be used by investors and analysts to determine future metric growth, based on the FLS in published 10-Ks. Thus, it is crucial to explain their justification and therefore generate trust. This is the task of Phase III.

4.3 Phase III – Model explainability

To improve transparency and explainability, we initially applied SHAP to analyze the most important features of the RFR models. While some patterns aligned with financial logic, such as positive effects from mentions of increased earnings or operating income, and negative signals from margin pressure or debt reliance, many predictions were shaped by a broader combination of textual cues, which SHAP failed to capture. Thus, we moved on to the generative summaries, which capture these complexities more intuitively and provide more insights into the company's expected growth.

To evaluate summary quality, one author randomly selected 20 summaries per financial metric and rated them on a 4-point scale based on four criteria: (1) factual consistency with the original FLS (no hallucination), (2) correct interpretation of financial content, (3) structured presentation, and (4) conciseness and readability. Missing any of these elements led to a lower score.

We began with a zero-shot prompt ("Summarize this text") but found it produced vague, unfocused summaries. To improve relevance and factuality, we adopted a System 2 attention prompting strategy (Weston and Sukhbaatar, 2023; Sahoo et al., 2025), instructing the model to extract only (1) metric-related FLS (EPS, EBIT, INCOME,

REV), (2) associated strategies, and (3) risks or assumptions, while explicitly excluding irrelevant information and staying within 1000 tokens. Additionally, we set the temperature to 0 to further reduce hallucinations and enforce output determinism. This layered approach was designed to boost objectivity and factual consistency while producing summaries that are concise and decision-relevant.

As a result, 93.75% were rated 3 or 4, with 75% achieving the highest score, yielding an average score of 3.69. Note that combined FLS in a report can discuss multiple metrics, thus a summary of the FLS can include expected growth of many metrics. One positive example of a summary rated 4, which includes clear growth tendency, the underlying strategies and implicit risks is: *"The company anticipates increased revenue and earnings through leveraging IMS Health's data assets and capabilities to accelerate clinical trials […] No explicit risks beyond foreign exchange rate volatility are mentioned."* In contrast, a noteworthy example (also rated 4) illustrating a factually-consistent reflection of vague original FLS is: *"The company acknowledges that there are no assurances of success in managing expanded operations and realizing expected growth in earnings, operating efficiencies, and cost savings. No specific financial metrics or associated growth strategies are detailed in the provided text."* Lower-rated outputs were typically too long or repetitive, which can affect readability. The best qualitative summaries were associated with Revenue (averaged 3.90) and EBIT (3.80), followed by EPS (3.55) and Net Income (3.50). This completes Phase III. In the following, we summarize the evaluation results and discuss their implications.

4.4 Discussion

Our evaluation in Phase I showed that the rule-based approach for metric-related FLS extraction was largely successful, with an average of 94% accuracy. However, performance for cashflow metrics (OCF, ICF, FCF) was limited due to inconsistent terminology and vague phrasing. These statements often lacked explicit links to accounting categories, reducing precision. For example, a sentence like *"We currently intend to finance the cash portion of the merger consideration"* mentions financing and cash but does not clearly refer to FCF. As a result, these metrics were excluded from later phases. Future work could improve coverage through expanded synonym lists and domain experts.

During Phase II, we demonstrated the contribution of our approach, by achieving better predictive performance than the baseline model. This implies that the extracted FLS contain valuable information about future metric growth. Contrary to expectations, more modern models are not always better. This could be due to the short FLS texts and small datasets. There are performance differences between the metrics, with REV and EBIT showing the best results, followed by EPS and INCOME. Thus, there is no clear dependency on the size of the dataset. However, INCOME and EPS have a higher variance than EBIT and REV. Therefore, future research should consider splitting companies into income-segments.

In Phase III, we added an interpretability layer to our DSS. While SHAP's token-level attributions were limited, generative summaries offered a more intuitive and user-relevant alternative. After prompt refinement, the summaries were factually consistent

and well-readable (average score 3.69 out of 4 on a sample dataset). Lower-rated summaries were often overly long due to prompt-driven repetition, which future work could address through more concise prompting. Overall, their readability and consistency reinforce their value as a transparent explanation layer within the DSS.

Our research design enables an automated and transparent DSS that streamlines 10-K analysis. By automatically extracting and analyzing metric-related FLS and historical metric data, the DSS predicts future metric growth. Decision makers can gain insight into qualitative summaries of the FLS and quantitative metric projections, along with the key features driving those predictions. Transparency allows users to assess model performance and provide feedback for model improvement, fostering trust. Ultimately, our approach enhances the efficiency, objectivity, and trustworthiness of financial decision-making.

5 Conclusion

This paper presents an automated and transparent DSS for extracting and analyzing metric-related FLS in 10-Ks. Our three-phase methodology combines a rule-based approach and modern NLP methods for FLS extraction, ML for metric growth prediction, and explainability through SHAP and generative summaries. The system translates unstructured narrative disclosures into transparent, metric-specific forecasts to support data-driven financial analysis.

We evaluated the approach on 10-Ks from S&P 500 firms using the EDGAR database. Phase I achieved 94% accuracy in FLS extraction, with F1-scores above 0.80 for EPS, REV, EBIT, and INCOME. In Phase II, the RFR model showed the strongest performance, particularly for REV and EBIT. While it also outperformed alternatives for EPS and INCOME, higher error levels reflected greater variability in these metrics. In Phase III, we used SHAP to identify influential features, though their interpretability was limited. Generative summaries provided a clearer alternative, with great consistency and readability, and an average score of 3.69 out of 4. Together, SHAP and summaries enhance the transparency and trustworthiness of our DSS.

The core contribution of this work lies in integrating advanced NLP and financial forecasting into a usable DSS that supports real-world decision-making. It goes beyond isolated model development by providing a pipeline that can be adopted, audited, and extended in practice.

Current limitations include our focus on S&P 500 companies and the challenge of extracting metric-related FLS due to vague or inconsistent phrasing, particularly for cashflow metrics, which may require a more dynamic extraction approach. While generative summaries enhance interpretability, the inherent risk of GenAI hallucination remains. Critically, the current system lacks a direct link between quantitative predictions and qualitative summaries. Future work should expand the dataset, integrate macroeconomic variables, and focus on explicitly linking predictive model outputs with generative summaries, potentially using SHAP-based insights, to create a more cohesive and interpretable DSS.

References

Athanasakou, V. and Hussainey, K. (2014) 'The perceived credibility of forward-looking performance disclosures', *Accounting and Business Research*, 44(3), pp. 227–259. Available at: https://doi.org/10.1080/00014788.2013.867403.

Bataineh, A. and Rababah, A. (2016) 'Comprehensive Income and Net Income, Which is more powerful in predicting Future Performance', 6, pp. 114–120.

Beyer, A. et al. (2010) 'The financial reporting environment: Review of the recent literature', *Journal of Accounting and Economics*, 50(2), pp. 296–343. Available at: https://doi.org/10.1016/j.jacceco.2010.10.003.

Botchkarev, A. (2019) 'Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology', *Interdisciplinary Journal of Information, Knowledge, and Management*, 14, pp. 045–076. Available at: https://doi.org/10.28945/4184.

Bozanic, Z., Roulstone, D.T. and Van Buskirk, A. (2018) 'Management earnings forecasts and other forward-looking statements', *Journal of Accounting and Economics*, 65(1), pp. 1–20. Available at: https://doi.org/10.1016/j.jacceco.2017.11.008.

Chen, Y.-R., Cheng, A. and Huang, Y.-L. (2012) 'Value of Cash Holdings: The Impact of Cash from Operating, Investment and Financing Activities'. Rochester, NY. Available at: https://doi.org/10.2139/ssrn.1985476.

Das, A., Mehta, S. and Subramaniam, L.V. (2017) 'AnnoFin–A hybrid algorithm to annotate financial text', *Expert Systems with Applications*, 88, pp. 270–275. Available at: https://doi.org/10.1016/j.eswa.2017.07.016.

Dechow, P.M. (1994) 'Accounting earnings and cash flows as measures of firm performance: The role of accounting accruals', *Journal of Accounting and Economics*, 18(1), pp. 3–42. Available at: https://doi.org/10.1016/0165-4101(94)90016-7.

Dyer, T., Lang, M. and Stice-Lawrence, L. (2017) 'The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation', *Journal of Accounting and Economics*, 64(2), pp. 221–245. Available at: https://doi.org/10.1016/j.jacceco.2017.07.002.

Ecobici, M.L. (2016) 'Indicators of Financial Analysis Employed in Quantifying the Financial Performance of a Company'.

Glodd, A. and Hristova, D. (2023) 'Extraction of Forward-looking Financial Information for Stock Price Prediction from Annual Reports Using NLP Techniques'.

Gomber, P. et al. (2018) 'On the Fintech Revolution: Interpreting the Forces of Innovation, Disruption, and Transformation in Financial Services', *Journal of Management Information Systems*, 35(1), pp. 220–265. Available at: https://doi.org/10.1080/07421222.2018.1440766.

Gopalakrishnan, K. (2024) 'TEXT SUMMARIZATION USING GENERATIVE AI: A CASE STUDY IN BANKING INDUSTRY', *Journal of Artificial Intelligence and Machine Learning*, Volume 3, Issue 1, pp. 1–7.

Halley, M.D. and Little, A.W. (1999) 'Net one, net two: the primary care network income statement.', *Journal of the Healthcare Financial Management Association*, 53(10), pp. 61–63.

Hsieh, H.-T. and Hristova, D. (2022) 'Transformer-based Summarization and Sentiment Analysis of SEC 10-K Annual Reports for Company Performance Prediction', in *55th Hawaii International Conference on System Sciences, HICSS 2022, Virtual Event / Maui, Hawaii, USA, January 4-7, 2022*. ScholarSpace, pp. 1–10.

Huang, R., Marquardt, C.A. and Zhang, B. (2015) 'Using Sales Revenue as a Performance Measure'. Rochester, NY. Available at: https://doi.org/10.2139/ssrn.2636950.

Huang, T.-C. and Hairston, S. (2023) ‘Analyst Revenue Forecasts and Firm Revenue Misstatements’, *European Accounting Review*, 32(2), pp. 379–414. Available at: https://doi.org/10.1080/09638180.2021.1983447.

Hussainey, K. and Walker, M. (2009) ‘The Effects of Voluntary Disclosure and Dividend Propensity on Prices Leading Earnings’, *Accounting and Business Research*, 39. Available at: https://doi.org/10.1080/00014788.2009.9663348.

Irwin, D.A. (2011) ‘Revenue or Reciprocity? Founding Feuds over Early U.S. Trade Policy’, in D.A. Irwin and R. Sylla (eds) *Founding Choices: American Economic Policy in the 1790s*. University of Chicago Press, p. 0. Available at: https://doi.org/10.7208/chicago/9780226384764.003.0004.

Kamaruddin, S.S. et al. (2009) ‘Automatic extraction of performance indicators from financial statements’, in *2009 International Conference on Electrical Engineering and Informatics*. Bangi, Malaysia: IEEE, pp. 348–350. Available at: https://doi.org/10.1109/ICEEI.2009.5254714.

Kumar, P. (2017) ‘Impact of Earning per Share and Price Earnings Ratio on Market Price of Share: a Study on Auto Sector in India’, *Int. J. Res. Granthaalayah*, 5(2), pp. 113–118. Available at: https://doi.org/10.29121/granthaalayah.v5.i2.2017.1710.

Li, F. (2010) ‘The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach’, *Journal of Accounting Research*, 48, pp. 1049–1102. Available at: https://doi.org/10.1111/j.1475-679X.2010.00382.x.

Lo, K., Ramos, F. and Rogo, R. (2017) ‘Earnings management and annual report readability’, *Journal of Accounting and Economics*, 63(1), pp. 1–25. Available at: https://doi.org/10.1016/j.jacceco.2016.09.002.

Lundberg, S.M. and Lee, S.-I. (2017) ‘A Unified Approach to Interpreting Model Predictions’, in *Advances in Neural Information Processing Systems*. Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (Accessed: 31 October 2023).

Masihabadi, A. et al. (2015) ‘The Relationship between Earnings before Interest and Taxes and Operating Cash Flow and Stock Return under the Condition of Information Asymmetry in Abadan and Arak Petrochemical Companies through Markov-Switching Approach’, *Marketing and Branding Research*, 2(1), pp. 74–88. Available at: https://doi.org/10.33844/mbr.2015.60316.

Millianto, L. and Bangun, N. (2021) ‘Pengaruh Operating Cash Flow, Investment Activities, Leverage terhadap Corporate Cash Holding’, *Jurnal Ekonomi*, 26(11), pp. 470–493. Available at: https://doi.org/10.24912/je.v26i11.788.

Molnar, C. (2022) *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd edn. Available at: christophm.github.io/interpretable-ml-book/ (Accessed: 14 December 2023).

Mousa, G.A., Elamir, E.A.H. and Hussainey, K. (2022) ‘Using machine learning methods to predict financial performance: Does disclosure tone matter?’, *International Journal of Disclosure and Governance*, 19(1), pp. 93–112. Available at: https://doi.org/10.1057/s41310-021-00129-x.

Muslu, V. et al. (2015) ‘Forward-Looking MD&A Disclosures and the Information Environment’, *Management Science*, 61(5), pp. 931–948. Available at: https://doi.org/10.1287/mnsc.2014.1921.

Nishikawa, I., Kamiya, T. and Kawanishi, Y. (2016) ‘The Definitions of Net Income and Comprehensive Income and Their Implications for Measurement’, *Accounting Horizons*, 30(4), pp. 511–516. Available at: https://doi.org/10.2308/acch-51544.

Omag, A. (2016) 'Cash Flows from Financing Activities: Evidence from the Automotive Industry', *International Journal of Academic Research in Accounting, Finance and Management Sciences*, 6(1), pp. 115–122.

Piracha, M. and Moore, M. (2016) 'Revenue-Maximising or Revenue-Sacrificing Government? Property Tax in Pakistan', *The Journal of Development Studies*, 52(12), pp. 1776–1790. Available at: https://doi.org/10.1080/00220388.2016.1153076.

Rane, N., Choudhary, S. and Rane, J. (2024) 'Gemini or ChatGPT? Efficiency, Performance, and Adaptability of Cutting-Edge Generative Artificial Intelligence (AI) in Finance and Accounting', *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.4731283.

Sahoo, P. *et al.* (2025) 'A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications'. arXiv. Available at: https://doi.org/10.48550/arXiv.2402.07927.

Siegel, J.G. and Shim, J.K. (2000) *Dictionary of accounting terms*. 3rd ed. Hauppauge, N.Y: Barron's Educational Series.

Silva, C.M.D. da, Oliveira, L.M. and Gonçalves, M.A. (2022) 'Uso de Earnings Before Interest, Taxes and Amortization (EBITDA): Estudo Bibliométrico / Use of Earnings Before Interest, Taxes and Amortization (EBITDA): A Bibliometric Analysis', *Revista FSA (Centro Universitário Santo Agostinho)*, 19(9), pp. 79–99.

Soelehan, A. (2012) 'Effect Analysis Of Financing Activities Cash Flows of the Company. Case studies on the PT. Indosat, Tbk and PT. Telekomunikasi Indonesia, Tbk.', *Jurnal Ilmiah Akuntansi dan Manajemen Ranggagading*, 12(2), pp. 163–175.

Syaputri, D.C. (2019) 'Operating Activities Cash Flow Effect on The Performance Of The Market: (Study In Industrial Goods Manufacturing Sector Consumption Listed in BEI 2016-2017)', *Progress Conference*, 2(2), pp. 142–151.

Tao, J., Deokar, A.V. and Deshmukh, A. (2018) 'Analysing forward-looking statements in initial public offering prospectuses: a text analytics approach', *Journal of Business Analytics*, 1(1), pp. 54–70. Available at: https://doi.org/10.1080/2573234X.2018.1507604.

Varshney, N. and Jain, M. (2016) *CASH FLOW STATEMENT OF BANK OF BARODA AND SYNDICATE BANK: A COMPARATIVE ANALYSIS OF OPERATING, INVESTING AND FINANCING ACTIVITIES*. Working paper 2016-09–10. Voice of Research. Available at: https://econpapers.repec.org/paper/vorissues/2016-09-10.htm (Accessed: 30 September 2023).

Velimirović, D., Velimirović, M. and Stanković, R. (2011) 'Role and importance of key performance indicators measurement', *Serbian Journal of Management*, 6(1), pp. 63–72. Available at: https://doi.org/10.5937/sjm1101063V.

Wang, M. and Hussainey, K. (2013) 'Voluntary forward-looking statements driven by corporate governance and their value relevance', *Journal of Accounting and Public Policy*, 32(3), pp. 26–49. Available at: https://doi.org/10.1016/j.jaccpubpol.2013.02.009.

Weber, P., Carl, K.V. and Hinz, O. (2024) 'Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature', *Management Review Quarterly*, 74(2), pp. 867–907. Available at: https://doi.org/10.1007/s11301-023-00320-0.

Weston, J. and Sukhbaatar, S. (2023) 'System 2 Attention (is something you might need too)'. arXiv. Available at: https://doi.org/10.48550/arXiv.2311.11829.