ProMiNet: Prototype-based Multi-View Network for Interpretable Email Response Prediction

Yuqing Wang[*]
UC Santa Barbara
Santa Barbara, CA 93106
wang603@ucsb.edu

Prashanth Vijayaraghavan[*]
IBM Research
San Jose, CA 95120
prashanthv@ibm.com

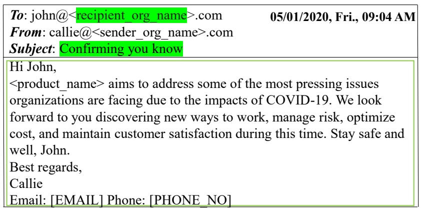
Ehsan Degan
IBM Research
San Jose, CA 95120
edehgha@us.ibm.com

Abstract

Email is a widely used tool for business communication, and email marketing has emerged as a cost-effective strategy for enterprises. While previous studies have examined factors affecting email marketing performance, limited research has focused on understanding email response behavior by considering email content and metadata. This study proposes a **Pro**totype-based **M**ulti-v**i**ew **Net**work (ProMiNet) that incorporates semantic and structural information from email data. By utilizing prototype learning, the ProMiNet model generates latent exemplars, enabling interpretable email response prediction. The model maps learned semantic and structural exemplars to observed samples in the training data at different levels of granularity, such as document, sentence, or phrase. The approach is evaluated on two real-world email datasets: the Enron corpus and an in-house Email Marketing corpus. Experimental results demonstrate that the ProMiNet model outperforms baseline models, achieving a $\sim 3\%$ improvement in F_1 score on both datasets. Additionally, the model provides interpretability through prototypes at different granularity levels while maintaining comparable performance to non-interpretable models. The learned prototypes also show potential for generating suggestions to enhance email text editing and improve the likelihood of effective email responses. This research contributes to enhancing sender-receiver communication and customer engagement in email interactions.

1 Introduction

With the ever-increasing volume of emails being exchanged daily, email communication remains a cornerstone of business interactions and an effective means of content distribution. As the primary communication tool for organizations and individuals alike, email marketing has maintained its popularity over the years, evolving and expanding alongside advancements in technology. This form of marketing enables businesses to tailor targeted messages to customers based on their preferences, leveraging the quick, easy, and cost-effective nature of email communication. In this context, predicting customer response behavior in email marketing campaigns becomes crucial for optimizing customer-product engagements and enhancing communication efficiency between senders and recipients. Consider the example email shown in Figure 1, where various factors such as the email's contents (subject and body) and the recipient's organization, can influence the likelihood of receiving a response. Therefore, understanding the impact of these factors and their correlation with email response behavior is paramount. Research (Kim et al., 2016) has shown that a single word can make a substantial difference in how a text is interpreted. This insight applies to our email response prediction task, making it essential to address this challenge. The likelihood of an email receiving a response can be influenced by various factors, including the use of power words or phrases, the persuasiveness of the text, and alignment with client preferences. Given the sensitivity of words or phrases in our task, we need methods to extract both the structural and semantic information from email text to develop an effective prediction model. Recently, there have been efforts to study different explanation techniques for text clas-



To: john@<recipient_org_name>.com **05/01/2020, Fri., 09:04 AM**
From: callie@<sender_org_name>.com
Subject: Confirming you know

Hi John,
<product_name> aims to address some of the most pressing issues organizations are facing due to the impacts of COVID-19. We look forward to you discovering new ways to work, manage risk, optimize cost, and maintain customer satisfaction during this time. Stay safe and well, John.
Best regards,
Callie
Email: [EMAIL] Phone: [PHONE_NO]

Figure 1: Sample Email with relevant contents.

[*] These authors contributed equally to this work.

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 202–215
December 6-10, 2023 ©2023 Association for Computational Linguistics

sification. These methods typically fall into two categories: post-hoc explanation methods (Madsen et al., 2021) and self-explaining approaches (Alvarez Melis and Jaakkola, 2018). Post-hoc explanations use an additional explanatory model to provide explanations after making predictions, while self-explaining approaches generate explanations simultaneously with the prediction. However, post-hoc explanations may not accurately reveal the reasoning process of the original model (Rudin, 2019), making it preferable to build models with inherent interpretability. In this work, we propose PROMINET, a novel interpretable email response prediction model that integrates semantic and structural information from email data. PROMINET utilizes prototype learning, a form of case-based reasoning, to make predictions based on similarities to representative examples (prototypes) in the training data. Unlike existing prototype-based architectures, PROMINET provides explanations from multiple perspectives: semantic (using transformer-based models) and structural (using graph-based dependency parsing with GNN). By leveraging a multi-branch network, PROMINET offers holistic explanations at different levels, including document-level, sentence-level, and phrase-level prototypes. We conduct quantitative analyses and ablation studies using two real-world email datasets: the Enron corpus and the in-house email marketing corpus. Our PROMINET model achieves superior performance and offers explanations that simulate potential edits, resulting in improved response rates.
Contributions: The key contributions of this work are summarized as follows:

- We present PROMINET, the inaugural method for interpretable email response prediction. By combining transformer-based models and dependency graphs with GNN, our approach captures semantic and structural information at various granularities.

- We conduct extensive experiments on real-world email corpora. PROMINET outperforms the strongest baselines on both the Enron and Email Marketing corpus.

- Simulation experiments demonstrate the effectiveness of learned prototypes in generating email text editing suggestions, leading to a significant enhancement in the overall email response likelihood. These results indicate promising avenues for further research.

2 Related Work

2.1 Email Response Prediction

Researchers have used machine learning methods to improve email efficiency by predicting email responses. Previous work includes predicting email importance and ranking by likelihood of user action (Aberdeen et al., 2010), classifying emails into common actions – read, reply, delete, and delete-WithoutRead (Di Castro et al., 2016), and characterizing response behavior based on various factors (On et al., 2010; Kooti et al., 2015; Qadir et al., 2016) including time, length, and conversion, temporal, textual properties, and historical interactions. Our work differs from previous studies by considering both semantic and structural information in email response prediction and developing an interpretable model.

2.2 Explainability in Text Classification

Model explainability has gained significant attention with different explainability methods categorized into post-hoc or self-explaining. Post-hoc methods (Ribeiro et al., 2016; Simonyan et al., 2013; Smilkov et al., 2017; Arras et al., 2016) separate explanations from predictions, while self-explaining methods (Bahdanau et al., 2014; Rajagopal et al., 2021) generate explanations simultaneously with predictions. Drawing from previous studies (Sun et al., 2020; Ming et al., 2019), our work falls into the self-explainable category, providing explanations through prototypes. Prototype-based networks make decisions based on the similarity between inputs and selected prototypes. Originally used for image classification (Chen et al., 2019), several methods (Ming et al., 2019; Hong et al., 2020; Pluciński et al., 2021) have been adapted for text classification, where a similarity score is used to learn prototypes, that represent the characteristic patterns in the data. These prototypes serve as exemplars or representative instances from the dataset. However, these models providing unilateral explanations have limitations as they lack granularity, provide an incomplete picture, have limited coverage, and reduced interpretability. In contrast, granular prototypes produced by our PROMINET offer a more nuanced and interpretable approach to understanding email data.

3 Problem Setup

We tackle the interpretable email response prediction problem as a self-explainable binary classification task. Given a training set $\mathcal{D}$ with email texts x_i and binary response labels $y_i \in \{0, 1\}$, our goal is to predict the likelihood of receiving a response while providing insights into the decision process. The labels indicate whether an email received a response (1) or not (0), which could include clicks, views, or replies. To enhance interpretability, we learn latent prototypes at the document, sentence, and phrase levels, mapping them to representative observations in the training set. These prototypes serve as classification references and analogical explanations for the model's decisions.

4 Methodology

In this section, we introduce PROMINET model, that incorporates multi-view representations and prototype layers to develop a self-explainable email response prediction model. Our architectural choices prioritize two key factors: accuracy and interpretability. To ensure accurate email response predictions, our model leverages features derived from the email subject, body, and recipient information. It does so by employing a multi-view architecture that captures the interplay between different factors. The model extracts both structural and semantic information to comprehend the valuable cues pertaining to email persuasiveness and engagement. Moreover, our model is designed to be interpretable, offering insights into decision-making at various levels. Using the information from the multi-view representations, the model achieves interpretability through granular latent prototypes that serve as explanations for predictions. By considering both accuracy and interpretability, the model aims to strike a balance between making accurate predictions and providing transparent reasoning. In our PROMINET model, we incorporate two main views, namely the Semantic view and the Structural view, to achieve our goal. We acquire embeddings at the document, sentence, and phrase-level by employing different components described in the subsequent subsections.

4.1 Semantic View

The Semantic view focuses on capturing features at both the document-level and sentence-level from email data. To extract document-level features, we employ a document encoder (f^D) that considers the interaction between different elements such as the email subject (S), body/content (C), recipient organization (O), and their interests (E). These elements are separated by a special token ($[SEP]$), and we prepend the email with a token ($[CLS]$). By utilizing a pre-trained transformer-based encoder, the email data is transformed into token-level representations, where the $[CLS]$ token representation serves as the document-level embedding, e^D. For sentence-level features, a similar transformer-based sentence encoder (f^S) is used to process each sentence within the email body. We add special tokens ($[CLS]$ and $[SEP]$) at the beginning and end of each sentence respectively. We denote the sentence-level embedding as e^S.

4.2 Structural View

The structural view emphasizes the importance of specific phrases in email engagement by examining the relationships between tokens or phrases within email sentences. By employing dependency parsing on the sentences, we create a graphical representation known as a dependency graph. The dependency graph comprises nodes representing tokens and links representing dependency relationships. These relationships are expressed as triples: (v_{dep}, $< rel >$, v_{gov}), where v_{dep} and v_{gov} denote the dependent and governing tokens, respectively; $< rel >$ refers to the dependency relationship between the tokens. To obtain phrase-level embeddings e^P, we extract dependency subgraphs from the sentences, focusing on dependencies like nominal subject (nsubj) and direct object (dobj) relative to the 'ROOT' token. Utilizing a graph encoder (f^P), we generate embeddings for each dependency subgraph, effectively capturing the structural information they convey.

4.3 Prototype Layers

In our approach, we utilize a prototype layer p consisting of three sets of prototypes: $p^D \in \mathcal{R}^{j \times d}$ for latent document prototypes, $p^S \in \mathcal{R}^{k \times d}$ for sentence prototypes, and $p^P \in \mathcal{R}^{m \times d}$ for phrase prototypes, where d is the dimension of the prototype embeddings (set identical to the dimensions of the output representations from the encoders) and j, k, m refers to the number of prototypes associated with each granularity level. To guarantee effective representation of each class through learned prototypes at varying levels of granularity, we assign a fixed number of prototypes to each class. These prototypes are learned during the training process and represent groups of data instances, such as documents, sentences, or phrases, found in the training set. For each granularity level g,

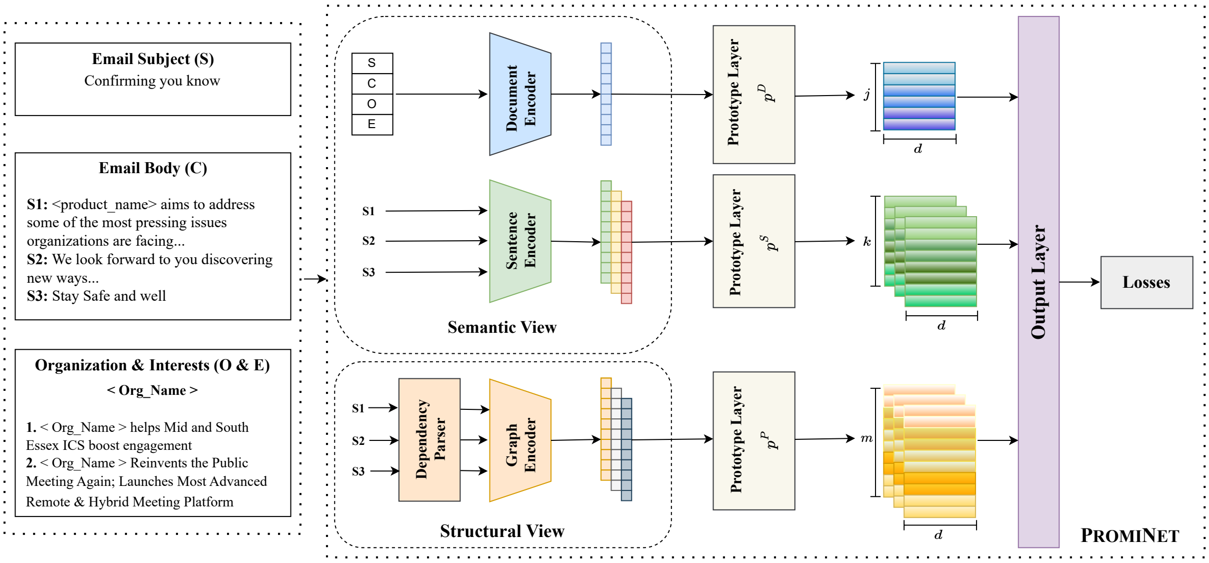

Email Subject (S)
Confirming you know
Email Body (C)
S1: <product_name> aims to address some of the most pressing issues organizations are facing...
S2: We look forward to you discovering new ways...
S3: Stay Safe and well
Organization & Interests (O & E)
< Org_Name >
1. < Org_Name > helps Mid and South Essex ICS boost engagement
2. < Org_Name > Reinvents the Public Meeting Again; Launches Most Advanced Remote & Hybrid Meeting Platform
S
C
O
E
Document Encoder
S1
S2
S3
Sentence Encoder
Semantic View
S1
S2
S3
Dependency Parser
Graph Encoder
Structural View
Prototype Layer
p^D
Prototype Layer
p^S
Prototype Layer
p^P
j
d
k
d
m
d
Output Layer
Losses
PROMINET

Figure 2: Illustration of our PROMINET model. The model consists of an encoder and a prototype layer for each granularity g (document (D), sentence (S) and phrase (P)) with two different views – semantic & structural.

which can be either document (D), sentence (S), or phrase (P), the layer calculates the similarity between the granularity-specific embedding (e^g) and each trainable prototype. Formally,

$$sim(p_i^g, e^g) = \log\left(\frac{||p_i^{(g)} - e^g||_2^2 + 1}{||p_i^g - e^g||_2^2 + \epsilon}\right) \quad (1)$$

Here, p_i^g represents the i^{th} prototype for granularity g, and it has the same dimension as the embedding (e^g). The similarity score decreases monotonically as the Euclidean distance $||p_i^g - e^g||_2$ increases, and it is always positive. For numerical stability, we set ϵ to a small value, specifically $1e-4$. We denote the computed similarity for each granularity level g as $\mathcal{S}^g$.

4.4 Output Layer

Finally, our model's output layer, denoted as c, includes a fully connected layer followed by a softmax layer to predict the likelihood of an email receiving a response. The prediction is determined by the weighted sum $\mathcal{S}^D + \lambda_1\mathcal{S}^S + \lambda_2\mathcal{S}^P$, which involves averaging the scores at the sentence and phrase levels with their weights denoted by λ_1 and λ_2, respectively.

4.5 Learning Objectives

We introduce different loss functions that ensure accuracy and interpretability. For accuracy, we have cross entropy loss:

$$L_{ce} = \frac{1}{n}\sum_{i=1}^{n} CE(c \circ p \circ f(x_i), y_i) \quad (2)$$

where the output layer c combines the information captured by different encoders (f) and prototype layers (p) from multiple views at different granularity levels. Drawing ideas from previous studies (Zhang et al., 2022; Ming et al., 2019), we introduce additional losses for prototype learning including: (a) diversity loss (L_{div}) that penalizes prototypes that are too similar to each other, (b) clustering loss (L_{cls}) that ensures that each embedding (text or graph) is close to at least one prototype of its own class and (c) separation loss (L_{sep}) encourages each embeddings to be distant from prototypes not of its class. Formally,

$$L_{div} = \sum_{k=1}^{C} \sum_{\substack{q \neq r \\ p_q^g, p_r^g \in p}} \max(0, \cos(p_q^g, p_r^g) - \theta) \quad (3)$$

$$L_{cls} = \frac{1}{n}\sum_{i=1}^{n} \min_{q: p_q^g \in p_{y_i}^g} ||f^g(x_i) - p_q^g||_2^2 \quad (4)$$

$$L_{sep} = -\frac{1}{n}\sum_{i=1}^{n} \min_{q: p_q \notin p_{y_i}^g} ||f^g(x_i) - p_q^g||_2^2 \quad (5)$$

where n is the total number of samples, C is the number of classes, θ is the threshold of cosine similarity, and $\cos(\cdot,\cdot)$ measures the cosine similarity, $p_{y_i}^g$ represents the set of prototypes belonging to class y_i for granularity g. Finally, we use L_1 regularization as the sparsity loss (L_{spa}) to the output layer weights. The overall objective is:

$$\mathcal{L} := L_{ce} + \alpha L_{div} + \beta L_{cls} + \gamma L_{sep} + \delta L_{spa} \quad (6)$$

where $\alpha, \beta, \gamma, \delta$ are the loss coefficients.

4.6 Prototype Projection

For improved interpretability, we project the latent prototypes onto the closest emails, sentences, or phrases from the training data. Each prototype's abstract representation is substituted with the nearest latent email, sentence, or phrase embedding in the training set that corresponds to its respective class of interest, measured by Euclidean distance. This conceptual alignment of prototypes with samples from the training set offers an intuitive and human-understandable interpretation of the prototypes associated with each class.

5 Experimental Setup

5.1 Datasets

Our framework is evaluated on two email datasets: the Enron corpus[1] and the email marketing corpus. The Enron dataset, collected by the CALO Project, consists of $\sim 500k$ emails from around 150 Enron Corporation employees. The email marketing corpus contains $\sim 400k$ email data, including response details such as clicks, views, and replies from vendors. These emails were part of an email marketing program and only a subset of these emails get responded to. In order to handle the data imbalance, we perform a random sampling to create a balanced split and conduct experiments over 5 runs. The dataset statistics and our other experimental settings for both the datasets are included in Appendix A, C.

5.2 Baselines

To demonstrate the effectiveness of our method, we compare our proposed ProMiNet with transformer-based pretrained masked language models such as BERT-base (Kenton and Toutanova, 2019), DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), autoregressive language model like XLNet (Yang et al., 2019), graph neural network-based TextGCN (Yao et al., 2019) that operates over a word-document heterogeneous graph, and prototype learning-based (ProSeNet (Ming et al., 2019) and ProtoCNN (Pluciński et al., 2021)) methods that learns to construct prototypes forsentences or phrases.

Methods	Enron	Email Marketing
BERT-base	$83.9_{\pm 2.9}$	$78.9_{\pm 2.5}$
DistilBERT	$79.3_{\pm 2.6}$	$73.6_{\pm 2.7}$
RoBERTa	$85.2_{\pm 3.0}$	$79.5_{\pm 2.9}$
XLNet	**85.6**$_{\pm 3.4}$	**80.2**$_{\pm 3.6}$
TextGCN	$80.9_{\pm 3.7}$	$74.1_{\pm 3.4}$
ProSeNet	$82.1_{\pm 3.0}$	$73.6_{\pm 3.2}$
ProtoCNN	$83.6_{\pm 3.8}$	$73.3_{\pm 3.6}$
ProMiNet Variants		
BERT + GCN	$84.6_{\pm 3.3}$	$81.1_{\pm 3.6}$
BERT + GAT	$85.2_{\pm 2.9}$	$81.2_{\pm 2.6}$
RoBERTa + GCN	$87.8_{\pm 2.8}$	**83.1**$_{\pm 3.2}^{*}$
RoBERTa + GAT	$87.4_{\pm 3.4}$	$82.6_{\pm 3.4}$
XLNet + GCN	$88.2_{\pm 3.2}$	**83.1**$_{\pm 3.6}^{*}$
XLNet + GAT	**88.6**$_{\pm 3.3}^{*}$	$82.6_{\pm 3.4}$
Improvement (%)	3.50	3.62

Table 1: Evaluation results on two email corpus. We report the weighted F_1 score (%) & SD based on 5 runs. Our method achieve statistically significant improvements over the closest baselines ($p < 0.01$).

5.3 Metrics

We calculate both the macro F_1 and the weighted F_1-score to evaluate the performance of the proposed models in the context of email response prediction on both datasets. Nevertheless, we prioritize the weighted F_1-score as our primary evaluation metric due to the balanced class distributions in our data splits. Additionally, we present the mean and standard deviations of the F_1-score across five runs in Section 6. Finally, we also perform a statistical analysis to assess the significance of the differences in F_1-scores between our proposed method and the nearest baselines using a paired t-test.

6 Results & Discussion

6.1 Overall Performance Comparison

Table 1 summarizes our evaluation results. ProMiNet consistently achieves the best performance on both datasets. Specifically, using XLNet encoder for texts and GAT encoder for dependency graphs, our model improves the weighted average F_1 score by 3.50% for the Enron corpus. Similarly, with RoBERTa/XLNet encoder for texts and GCN encoder for dependency graphs, ProMiNet improves the weighted average F_1 score by 3.62% for the Email Marketing corpus. Compared to the other transformer-based models, ProMiNet

[1] https://www.cs.cmu.edu/~enron/

demonstrates performance improvements, indicating that incorporating dependency graphs enhances word connections and contextual meaning, leading to better overall performance. PROMINET outperforms TextGCN significantly, suggesting that considering local information, such as word order, in addition to global vocabulary information is crucial for accurate classification. Moreover, PROMINET surpasses prototype learning methods (ProSeNet and ProtoCNN), highlighting the importance of learning prototypes that capture both semantic and structural aspects.

Methods	Enron	Email Marketing
XLNet	85.6$_{\pm 3.4}$	80.2$_{\pm 3.6}$
GAT	81.4 $_{\pm 3.1}$	78.1$_{\pm 2.9}$
XLNet + GAT	**88.2**$_{\pm 2.8}$	**81.8**$_{\pm 3.0}$
XLNet + Prototypes	86.2$_{\pm 2.9}$	80.8$_{\pm 3.1}$
GAT + Prototypes	82.9$_{\pm 3.2}$	78.3$_{\pm 3.6}$
XLNet + GAT + Prototypes (PROMINET)	**88.6**$^{*}_{\pm 3.3}$	**82.6**$^{*}_{\pm 3.4}$

Table 2: Investigation of the Impact of Various Components in PROMINET on Both Datasets. We analyze different model variants to assess the influence of semantic and structural views, as well as prototype layers.

6.2 Ablation Study

In our ablation experiments[2], we scrutinize the influence of various model components in PROMINET. The amalgamation of XLNet, GAT, and prototype learning demonstrates the highest performance, underscoring their complementary attributes. Prototype-based models exhibit comparable performance to their non-interpretable counterparts. Moreover, the fusion of XLNet with prototype learning surpasses the combination of GAT and prototype learning, highlighting the significance of semantic information in text comprehension. This experiment not only illustrates the superiority of multi-view representations derived from both semantic and structural perspectives over models relying on embeddings from a single view, but also showcases that when coupled with prototype learning, PROMINET achieves the highest performance. We present comprehensive ablation studies that assess the impact of factors such as the number of prototypes, sensitivity to weights (λ_1, λ_2), the contribution of various email metadata, and detailed error analyses. For further information, please refer to Appendix D.

6.3 Explanations for Prediction

6.3.1 Case study

Figure 3 illustrates the reasoning process of PROMINET using an input example from the test set in the Email Marketing corpus. It showcases the most similar prototypes at the document, sentence, and phrase levels. The selected prototypes, along with their original labels, provide evidence for why the input example is classified as "negative". Two key observations emerge from the analysis: (a) All the learned prototypes associated with the input have the label "negative", consistent with the prediction of the input example; (b) The document-level prototypes exhibit similar topics to the input example, such as event invitations and basic event introductions. The sentence-level and phrase-level prototypes share similarities in terms of client interests, patterns, and grammatical relationships. We present similar analysis for an example from the Enron corpus in Appendix E.

6.4 Suggest Edits based on Prototypes

We utilize the attention mechanism from GAT to identify key phrases and important dependency relationships[2] that contribute to the prediction. The words in a sentence are categorized into different types, such as nouns, verbs, adjectives, and adverbs. Since adjectives and nouns usually form key phrases, which are crucial, we focus on nouns and related words, considering their attention scores. Additionally, we use layer integrated gradients (LIG) (Sundararajan et al., 2017) and transformer-based embeddings to determine the importance of words in a sentence. After extracting the top-1 keyword/top-1 keyphrase for each sentence, we substitute keywords/keyphrases associated with prototypes with the label “positive" (i.e., emails with response) for the keywords and key phrases of sentences in the test set with the label “negative" (i.e., emails without response) to investigate whether there is a possibility to improve the ratio of “positive" labels, that is, to improve the overall response rate. Here, the selected prototype emails share similar topics with the email to be edited. Otherwise, we randomly choose a prototype with response for edits. For an email, there are a few positions we consider editing: (1) email subjects; (2) email opening sentence/greeting (e.g., I hope you

[2] Please refer to the Appendix for additional analyses, including information on Datasets, Hyperparameters, Ablation studies, Visualization, and Limitations.

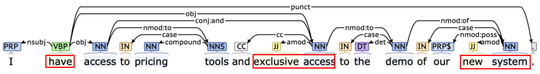
		Importance Score (Similarity * Weight)	Prototype Label
Input Text	**S1:** Hi Jorge, you may have heard that <company>'s conference is entirely virtual this year and free with everything going on in the world right now. **S2:** It may be a welcome distraction to advance your expertise and learn something new. **S3:** We think we can help here. **S4:** There are some particular <product team> topics we'll discuss that you may be interested in supporting cyber resiliency monitor your infrastructure with storage insights. **S5:** Take a look at the Think site to see additional topics and register for your free pass.		
Top-2 Prototypes (email-level)	Hi Andrew, I hope your New Year is off to a great start. It's me again from the <product> team. Come to learn about converges IDA and <company>'s hybrid cloud approach while sampling some delicious whiskey wings. Interested join us Wednesday February 17th at 2pm? You do not want to miss this great dialogue and food seating is limited, so reserve you spot now. If you have any questions, feel free to reach out.	**3.28 * 0.46 = 1.51**	**Negative**
	Hello Delli, I'm reaching out to invite you to <event>. At <event>, you'll have the chance to directly engage with world-class experts, industry leaders and peers gain insights guidance and valuable connections your business needs and learn how groundbreaking technologies like hybrid cloud and AI can positively impact your business.	**3.06 * 0.35 = 1.07**	**Negative**
Prototypes (sentence-level)	You're interested in taking advantage of fast low-cost storage or needing a solution that can grow according to your requirements. ⇢ **S4**	**2.50 * 0.29 = 0.73**	**Negative**
	Again, I invite you to take a look at the short video and supporting materials. ⇢ **S5**	**2.34 * 0.68 = 1.59**	**Negative**
Prototypes (dependency tree)	Text: I have access to pricing tools and exclusive access to the demo of our new system. ⇢ **S2**  punct obj conj:and nmod:to case compound nsubj obj cc amod nmod:to case det nmod:of case nmod:poss amod PRP VBP NN IN NN NNS CC JJ NN IN DT NN IN PRP$ JJ NN . I have access to pricing tools and exclusive access to the demo of our new system .	**1.98 * 0.42 = 0.83**	**Negative**
	Prediction: **Negative** Gold Standard: **Negative**		

Figure 3: Example inputs and PROMINET prototypes for Email Marketing corpus. While classifying the input as negative (no response), the labels of prototypes are also negative. Due to space constraint, we only show a few prototypes with the largest weights.

are doing well); (3) main contents of the email; (4) closing sentence (e.g., best regards). In our experience, we observe that using prototype-based edits of email subjects and main contents bring significant improvement of the overall email response rate on both datasets. For instance, the model captures the importance of creating a sense of urgency that improves the likelihood of receiving a response. A sentence from an email labeled as "negative" turns "positive" when the sentence containing a phrase "register for your free pass" is replaced with a prototype-based phrase "get your free pass before the offer expires". Investigations on the impact of suggested edits on the effectiveness of our models are detailed in Appendix D.

7 Conclusion

In this study, we introduced PROMINET, a Prototype-based Multi-view Network that incorporates semantic and structural information from email data for interpretable email response prediction. PROMINET compared inputs to representative instances (prototypes) in the latent space to make predictions and offers prototypical explanations at the document, sentence, and phrase levels for enhanced human understanding. The evaluation on real-world email datasets demonstrates that PROMINET outperforms baseline models, achieving a significant improvement of approximately 3% in F_1 score on both the Enron corpus and the Email Marketing corpus. Our research contributes to enhancing sender-receiver communication and customer engagement in email interactions, filling a gap in understanding by considering email content and metadata. Future research directions involve addressing limitations such as time and historical interactions, handling unseen scenarios, improving interpretability, and balancing personalized content with prototypical information. These advancements will further propel the usage AI technqiues in email marketing and communication.

Ethics Statement

For this research, we utilized two distinct datasets. One of them comprises a publicly available collection, while the other involves IBM's internal email marketing corpus. It's important to note that we exclusively employed anonymized training data from the latter [3], ensuring the removal of any personally identifiable information. Furthermore, our methodology aims to enhance the interpretability of the email response prediction system, providing insights into the model's decision-making process

[3] Although anonymized data was utilized for training and evaluation, in this paper, we have incorporated randomly generated names in email samples for the purpose of visualization and enhanced comprehension.

at different levels of granularity without compromising proprietary or sensitive information. However, a noteworthy concern arises regarding the potential influence on user sentiments and actions in subtle ways, which could be interpreted as coercion. In such scenarios, the explanations provided through prototypes may inadvertently reveal biases or problematic training scenarios. This underscores the need for stringent guidelines and explainability, particularly in sensitive real-world contexts, to ensure that the model's predictions do not exert any harmful or ethically questionable influences on user decision-making. It's important to acknowledge that these risks are not unique to our methodology, but rather, they are pertinent to various AI techniques. This emphasizes the necessity for a consistent and vigilant review process and update of ethical standards and practices.

References

Douglas Aberdeen, Ondrey Pacovsky, and Andrew Slater. 2010. The learning behind gmail priority inbox.

David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Explaining predictions of non-linear classifiers in nlp. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 1–7.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Dotan Di Castro, Zohar Karnin, Liane Lewin-Eytan, and Yoelle Maarek. 2016. You've got mail, and here is what you could do with it! analyzing and predicting actions on email messages. In *Proceedings of the ninth acm international conference on web search and data mining*, pages 307–316.

Dat Hong, Stephen S Baek, and Tong Wang. 2020. Interpretable sequence classification via prototype trajectory. *arXiv preprint arXiv:2007.01777*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Joon Hee Kim, Amin Mantrach, Alejandro Jaimes, and Alice Oh. 2016. How to compete online for news audience: Modeling words that attract clicks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1645–1654.

Farshad Kooti, Luca Maria Aiello, Mihajlo Grbovic, Kristina Lerman, and Amin Mantrach. 2015. Evolution of conversations in the age of email overload. In *Proceedings of the 24th international conference on world wide web*, pages 603–613.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2021. Post-hoc interpretability for neural nlp: A survey. *arXiv preprint arXiv:2108.04840*.

Yao Ming, Panpan Xu, Huamin Qu, and Liu Ren. 2019. Interpretable and steerable sequence learning via prototypes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 903–913.

Byung-Won On, Ee-Peng Lim, Jing Jiang, Amruta Purandare, and Loo-Nin Teow. 2010. Mining interaction behaviors for email reply order prediction. In *2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 306–310. IEEE.

Kamil Pluciński, Mateusz Lango, and Jerzy Stefanowski. 2021. Prototypical convolutional neural network for a phrase-based explanation of sentiment classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 457–472. Springer.

Ashequl Qadir, Michael Gamon, Patrick Pantel, and Ahmed Hassan. 2016. Activity modeling in email. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1452–1462.

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H Hovy, and Yulia Tsvetkov. 2021. Selfexplain: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 836–850.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on* *knowledge discovery and data mining*, pages 1135–1144.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Navdeep S Sahni, S Christian Wheeler, and Pradeep Chintagunta. 2018. Personalization in email marketing: The role of noninformative advertising content. *Marketing Science*, 37(2):236–258.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Zijun Sun, Chun Fan, Qinghong Han, Xiaofei Sun, Yuxian Meng, Fei Wu, and Jiwei Li. 2020. Self-explaining structures improve nlp models. *arXiv preprint arXiv:2012.01786*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Cheekong Lee. 2022. Protgnn: Towards self-explaining graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9127–9135.

A Dataset Details

Summary statistics of the datasets are shown in Table A1. Here, the striking difference in the ratio of “response" and “no response" samples between two email corpus is due to different email intents. The Enron corpus is pertaining to personal communication while the Email Marketing corpus is used for focused marketing campaigns.

Datasets	Enron	Email Marketing
Total	497,465	404,167
Response	270,309	57,607
No Response	227,156	346,560

Table A1: Statistics of the datasets.

A.1 Enron Corpus

The Enron email dataset does not have explicit “response" and “no response" classes. Since “reply" and “forward" email threads appear in the original corpus, we categorize a single email as “response" as long as the original email contains “reply" or “forward" tag and extract only the portion of the email after the “reply" or “forward" tag. Otherwise, we categorize the single email as “no response" and keep the entire email text.

A.2 Email Marketing Corpus

We obtained this in-house corpus for research purposes. This dataset contains response information from clients in the form of clicks, views or replies. We label a single email as “response" as long as the original email is clicked, viewed, or replied at least once. Otherwise, we label the email as “no response".

B Experimental Settings

At encoder layer f, we use three variants of BERTs for text embedding, i.e., BERT-base, RoBERTa, and XLNet. Meanwhile, we use two variants of GNNs for subgraph embedding, i.e., GCN and GAT. Since the dataset is skewed, we perform a random downsample to create a balanced split and conduct experiments over 5 runs. We adopt the AdamW optimizer with a weight decay of 0.1. The hyperparameter search space for both datasets is included in Table A2. We perform random search for hyperparameter optimization. All of the experiments are conducted on four NVIDIA Tesla P100 GPUs.

C Hyperparameter Search Space

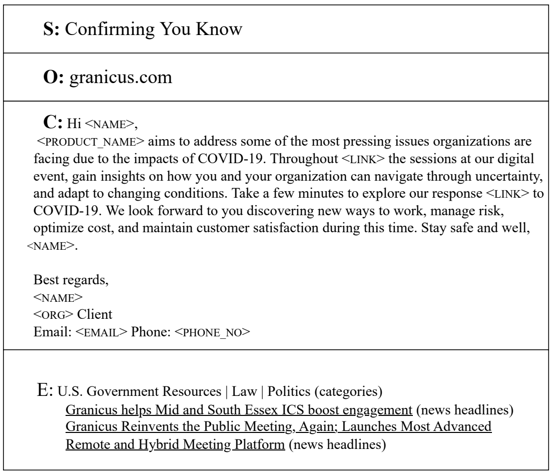
Hyperparameters	Search Space
Batch size	[16, 32, 64, 128]
Learning rate	[1e$-$5, 2e$-$5, 5e$-$5]
Class weight for l_{ce}	[0.2, 0.3, 0.4, 0.5]
j, k, m	[6, 10, 20, 30, 40, 50]
θ	[0.2, 0.3, 0.4]
α	[0.001, 0.005, 0.01, 0.015, 0.02]
β	[0.005, 0.01, 0.02, 0.05, 0.1]
γ	[0.001, 0.005, 0.01, 0.015, 0.02]
δ	[0.001, 0.005, 0.01, 0.015, 0.02]
λ_1	[0.1, 0.3, 0.5, 0.7, 0.9]
λ_2	[0.1, 0.3, 0.5, 0.7, 0.9]

Table A2: Hyperparameter search space of PROMINET on both datasets.

D Ablation Studies

D.1 Effect of Email Components

We study the contribution of different email components as text inputs to the model's performance. In this study, we consider the subject, body text, and recipient's email organization as email composition components. Additionally, we utilize the AYLIEN news API[4] to extract the interests of organizations. Our assumption is that the intent of an email may be associated with the recipient's organization's topic of interest. The API extracts news categories and headlines associated with the organization. For the Enron corpus, we evaluate the influence of the subject and body text only since all the recipients' email organizations in this corpus are from Enron. In the Email Marketing corpus, there are a considerable number of email recipient organizations for which the API is unable to extract interest information. In such cases, we leave the interests unknown. However, the goal of this experiment is to estimate the extent to which the interest information can boost our prediction performance. An example email from the Email Marketing corpus that contains all the pieces of information is provided in Figure A1. This example helps in understanding the information contained in each part of the email before feeding it to the model. Based on the results presented in Table 1, we evaluate the contributions of email components using the PROMINET setting (XLNet + GAT) for the Enron corpus and the PROMINET setting (XLNet + GCN) for the Email Marketing corpus. We summarize the experimental results in Table A3 and make the following observations: The introduction of organization interests in the Email Marketing corpus shows marginal improvements in performance, confirming our assumption that there is an association between the intent of the sending email and the interests of the recipient's organization. The high standard deviation in performance when incorporating organization interests can be attributed to incomplete information. Despite these limitations, we observed some marginal improvement. A more in-depth analysis with complete information could yield significantly better results, but such investigation is beyond the scope of this paper and can be pursued in future research. When considering individual components of an email, the model's performance using body text as input outperforms the performance when using only the subject or email organization information. This finding highlights the significance of body texts in predicting email responses. The best performance is achieved when incorporating all three components—the subject, body text, and recipient's email organization. This indicates that each piece of information is valuable and contributes to performance gains. Overall, these observations emphasize the importance of considering multiple components and organization interests in improving the performance of email response prediction models.



S: Confirming You Know

O: granicus.com

C: Hi <NAME>,
<PRODUCT_NAME> aims to address some of the most pressing issues organizations are facing due to the impacts of COVID-19. Throughout <LINK> the sessions at our digital event, gain insights on how you and your organization can navigate through uncertainty, and adapt to changing conditions. Take a few minutes to explore our response <LINK> to COVID-19. We look forward to you discovering new ways to work, manage risk, optimize cost, and maintain customer satisfaction during this time. Stay safe and well, <NAME>.

Best regards,
<NAME>
<ORG> Client
Email: <EMAIL> Phone: <PHONE_NO>

E: U.S. Government Resources | Law | Politics (categories)
Granicus helps Mid and South Essex ICS boost engagement (news headlines)
Granicus Reinvents the Public Meeting, Again; Launches Most Advanced Remote and Hybrid Meeting Platform (news headlines)

Figure A1: An example email from Email marketing corpus that contains subject (S), content (C), organization (O), and interests (E).

D.2 Error Analysis of Transformer-based/GNN Models

Analyzing the error patterns of our Transformer-based/GNN models allows us to demonstrate the benefits provided by our PROMINET model. We

[4] https://aylien.com/

focus on qualitatively examining email samples that are correctly classified by PROMINET but misclassified by other baseline models. For example, the sample email input provided in Figure 3 was misclassified by models that do not jointly model semantic and structural prototypes for response prediction. Additional analyses on models that solely utilize either semantic or structural prototypes are provided in Appendix D.2.1.

D.2.1 Effects of Using only BERT/GNN

When using a combination of a transformer-based model and prototype learning, the input shown in Figure 3 is associated with the following top-2 email-level prototypes:

1. **Prototype 1**: "Hi Phil, I hope this email finds you well. Just a quick line inviting you to attend <company_name>'s online launch event storage made simple for all. During this event, you'll see how we are revolutionizing the entry enterprise storage space. If aah pharmaceuticals is challenged to deliver more with less budget it will be well worth your time attending."

2. **Prototype 2**: "Hi John, hope this message finds you doing well today. My name is Nicholas Tompkins with <product_name>, reaching out to personally invite you to an upcoming event. Did you know <company_name> technology is simple innovative flexible fast and infused by AI. In this session, you will learn how we co-create solutions with you using flash systems virtualization, data protection cyber resiliency and business continuity strategies."

The most similar prototypes mapped to S4 and S5 are as follows:

- Prototype mapped to S4: "Hi Jack, do you have the need to refresh or add additional storage to your environment?"

- Prototype mapped to S5: "Click here to register."

The majority of prototypes associated with the input are labeled as "positive." However, the true label of the input is "negative." It is possible that although BERT captures the contextual information of an email, its ability to analyze dependencies and determine the grammatical structure of sentences is limited. Grammatical structures play a crucial role in enhancing sentence clarity and governing how words can be combined to form coherent sentences.

Email	Enron	Email Marketing
Components	PROMINET (XLNet + GAT)	PROMINET (XLNet + GCN)
S	80.2 ± 3.6	76.9 ± 3.2
O	—-	73.4 ± 2.9
C	85.1 ± 3.2	80.1 ± 3.4
$S + O$	—-	78.4 ± 3.6
$S + C$	**88.6 ± 3.3**	82.6 ± 3.3
$O + C$	—-	81.8 ± 3.7
$S + O + C$	—-	**83.1 ± 3.6**
$S + O + C + E$	—-	**83.5 ± 4.1**[*]

Table A3: Effects of different email components as inputs of PROMINET on both datasets. The performance is evaluated via weighted average F_1 score (%). Experiments are conducted with 5 random initializations. The results are shown in the format of mean and standard deviation. Here, *S*, *O*, *C*, and *E* represent subject, organization, body text and organization interests, respectively.

When employing a combination of GNN (Graph Neural Network) and prototype learning, we observe that the most similar prototype mapped to S2 can be seen in Figure 3. However, there is a discrepancy between the label assigned to the prototype ("positive") and the label assigned to the sentence ("negative"). This mismatch suggests that the GNN component might lack the necessary information on text semantics to fully comprehend the content of the text.

This observation highlights the importance of investigating the interpretability capability of individual model components. In this case, it verifies the effectiveness of combining a transformer-based model, which excels at capturing contextual information, with a GNN, which is adept at capturing grammatical structures. The combination of these two components allows them to mutually influence and complement each other, resulting in a more comprehensive understanding of the input text.

D.3 Effect of Suggested Edits

We also investigate the impact of suggested edits on the effectiveness of our models. To simulate this, we conduct experiments where we make edits to the emails and observe the resulting changes in the ratio of "positive" labels. Table A4 presents the ratios of original "negative" emails that are predicted as "positive" after making edits under different situations on both datasets. We employ a combination of XLNet and GAT for predictions and find that appropriate edits to email subjects and main content

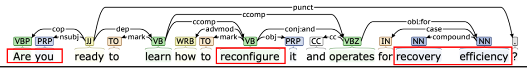
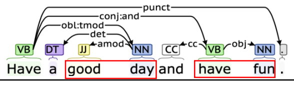
	Weights	Prototypes	
Email Marketing Corpus	0.45	Hello Iris, it's Ann Marie from the <product> team. I hope you are keeping well. I'm just following up on my previous email to advise that the storage assessment is still available to you and your team, but that we also have a live demo I can take you through on <company>'s new storage all <product_name> this year. <company> have released some incredibly affordable all <product_name>. If you are in the market for an upgrade on your current infrastructure or are curious to see how the new technology works, I really think it would be worth taking a look at.	p^D
	0.64	Following my previous email, I wanted to make sure you're aware of some of the changes we've made to our portfolio and how that may impact your systems as a storage contact for <company>.	p^S
	0.33	 punct ccomp ccomp cop nsubj dep mark advmod mark obj conj:and cc obl:for case compound VBP PRP JJ TO VB WRB TO VB PRP CC VBZ IN NN NN . Are you ready to learn how to reconfigure it and operates for recovery efficiency ?	p^P
Enron Corpus	0.61	Dear Brad, it was great to speak to you today. I have provided the instructions to upgrade ICE to version 737. Please forward these instructions to your IT Department. We are strongly recommending that you upgrade. It involves some major changes. Please call me if you have any further questions. Thanks for your assistance.	p^D
	0.21	We have delivered an electronic ticket to the airlines notifying them of your purchase.	p^S
	0.24	 punct conj:and obl:tmod det amod cc obj VB DT JJ NN CC VB NN . Have a good day and have fun .	p^P

Figure A2: Visualization of three types of prototypes (i.e., document (p^D), sentence (p^S), phrase-level (p^P)) learned from the PROMINET model on Enron Corpus and Email Marketing corpus.

lead to improvements in the overall email response rate for both datasets. These improvements signify the potential of using prototypes to enhance the likelihood of generating favorable email responses.

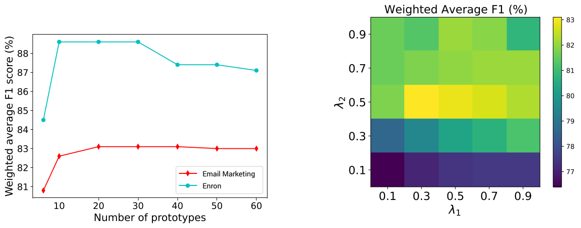
Editing Positions	Enron	Email Marketing
Subjects	1.4 ± 0.2	2.1 ± 0.3
Open sentence	0.3 ± 0.0	0.9 ± 0.1
Main contents	1.9 ± 0.3	3.8 ± 0.5
Closing sentence	0.3 ± 0.0	0.4 ± 0.1

Table A4: Drop ratio of "negative" labels after making edits on both datasets.

D.4 Effect of hyperparameters in PROMINET

We conducted a study to examine the impact of certain hyperparameters on model performance, specifically focusing on the number of prototypes and the addition weights.

Number of Prototypes (*j*,*k*,*m*): Figure A3a illustrates the relationship between the number of prototypes and the model performance, measured by the weighted average F_1 score, for both datasets. We observed that increasing the number of prototypes initially led to a significant improvement in performance. However, once the number of prototypes surpassed 20, the performance gains became less prominent, and in some cases, adding more prototypes even resulted in slightly worse performance. This phenomenon can be attributed to the increased complexity of the model, making it more challenging to train and comprehend. It demonstrates the trade-off between performance and interpretability. The optimal number of prototypes was found to be 20 for the Email Marketing corpus and 10 for the Enron corpus, as the model performance peaked at these values.



(a) Effects of j, k, m.

(b) Effects of λ_1, λ_2.

Figure A3: Hyperparameter effects on performance.

Addition Weight ($\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}$): The addition weights, λ_1 and λ_2, control the training balance among the three branches in our model. Figure A3b presents the performance variations on the Email Marketing corpus when different combinations of λ_1 and λ_2 were used. The results demonstrate that the best performance was achieved when λ_1 was set to 0.3 and λ_2 was set to 0.5 in PROMINET.

By investigating these hyperparameters, we gain insights into their effects on model performance, enabling us to optimize the performance and interpretability of our PROMINET model.

E Case Study

In Figure A4, we can examine the selected prototypes and their original labels, which serve as evidence for why the input example has been classified as positive. We can make two key observations:

- The majority of prototypes associated with the input have the label "positive", which aligns with the prediction of the input example being "positive".

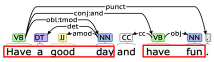
		Importance Score (Similarity * Weight)	Prototype Label
Input Text	**S1:** Dorie and Michelle, this will update you on the latest developments. **S2:** Bob and I are working on finding a solution that resolves all the issues which get complex due to the fact that not only has the Hilton made claims against Enron but also against Event Resources in the bankruptcy proceeding. **S3:** In other words, even if we succeed in facing down the Hilton on its claims directly against us, the Hilton may be able to recover some percentage of those claims directly from Event Resources in the bankruptcy proceeding.		
document-level	The consumer advocates strongly feel that the generators have to take a hair cut as part of the solution as well. The governor will announce later this afternoon a framework solution identical to that we have reported previously a state purchase of transmission assets and an issuance of bonds by the utilities but with state support through the DWR. However, it remains unclear whether the framework will be acceptable to all parties.	**2.26 * 0.54 = 1.22**	**Negative**
	Governor Davis is committed to solving the California energy crisis by developing consumer driven solutions. Protecting customers from short term market aberrations. Continuing to expand the consumers ability to choose. We are missing thoughtful Orderly Process.	**2.03 * 0.48 = 0.97**	**Positive**
sentence-level	I would like to keep everyone updated with changes on the floor.····→**S1**	**1.78 * 0.29 = 0.52**	**Positive**
	Transwestern had met with the AQB over this issue in 1996 and assumed that the issue had been resolved..····→**S5**	**1.96 * 0.33 = 0.65**	**Negative**
phrase-level	Text: : Have a good day and have fun.····→**S1**  punct conj:and obl:tmod det amod cc obj VB DT JJ NN CC VB NN . Have a good day and have fun.	**1.46 * 0.20 = 0.829**	**Positive**
	Prediction: **Positive** Gold Standard: **Positive**		

Figure A4: Example inputs and PROMINET prototypes for Enron corpus. While classifying the input as positive (response), the majority of the prototype labels are also positive. Due to space constraint, we only show a few prototypes with the largest weights.

- The prototypes at the document-level share similar topics with the input example, specifically related to problem-solving. At the sentence-level, both S1 and its corresponding prototype discuss update notifications, while S2 and its prototype exhibit similar patterns. In terms of phrase-level prototypes, phrases extracted by S1 and its prototype share similar grammatical relationships, such as nominal subject (nsubj), adjectival modifier (amod), coordination (cc), and so on.

F Prototype Visualization

We provide prototype visualization, where each prototype is mapped to the latent representation of the most similar email in the training set. This mapping is facilitated by assigning static index numbers to each email or sentence from the same email during the model training phase. These index numbers enable us to visualize the prototypes later on. Figure A5 showcases some learned prototypes in a human-readable form for both datasets. The weight assigned to each prototype is derived from the fully connected layer. This diversity in different types of prototypes enhances our ability to provide explanations for prototype-based predictions.

G Limitations

This work has several limitations that should be acknowledged. Firstly, the focus of this study is primarily on the text aspects of email data, disregarding factors such as time and historical interactions with customers. While this approach is suitable for the prediction task at hand, it overlooks potentially valuable contextual information that could impact email response behavior. Additionally, while prototypes are useful for the intended use case, there may be unseen scenarios or outliers that cannot be accurately mapped to examples in the training set, posing a challenge in dealing with such cases. Exploring alternative approaches to enhance interpretability and present explanations in a more user-friendly manner is an avenue for future research. Furthermore, the prototype-based suggestion of edits presented in this work is a simulation experiment and may not capture the exact dynamics of real-time scenarios. The proposed shortcuts for improving model performance should be carefully considered to ensure alignment with actual email interactions. Lastly, using prototypical information in email composition runs the risk of generating templated emails with reduced personalization, even though personalization is known to be beneficial in email marketing (Sahni et al., 2018). Thus, addressing these limitations and exploring

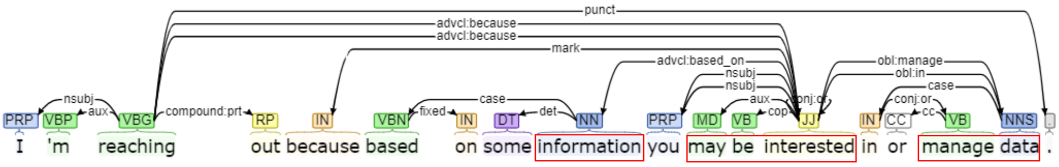


Figure A5: Most similar dependency subgraph prototype associated with S2 of input example in Figure 3 using only GNN.

these areas of improvement could be the scope of future research.