Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Scalable Reputation Management: A Multi-Task Prompting Approach Using Fine-Tuned PLMs for Sentiment and Topic Classification

Paulo R. S. da Costa Jr. [1,2], **Matheus Utino** [1,3], **João Silva-Leite** [2], **Lyncoln S. de Oliveira** [2], **Rodrigo S. Monteiro** [2], **Rodrigo Dias** [1], **Daniel de Oliveira** [2], **Paulo Mann** [4], **Marcos Bedo** [2]

[1]Wikki Brasil – Parque Tecnológico – Rio de Janeiro/RJ – Brazil
[2] Institute of Computing, Fluminense Federal University – Niterói/RJ – Brazil
[3] Institute of Mathematics and Computer Science, University of São Paulo – São Carlos/SP – Brazil
[4] Institute of Computing, Federal University of Rio de Janeiro – Rio de Janeiro/RJ – Brazil

Abstract

Social media platforms provide a direct, real-time channel for companies to engage with their audiences, making the management of corporate reputation on social media a pivotal factor for organizational success. Reputation is shaped by factors such as the relationship with the audience, crisis communication, public sentiment, and the topics most frequently associated with the company. Although reputation is often viewed as an intangible asset, it can be quantified and monitored through various metrics. Advances in AI and Pretrained Language Models (PLMs) have automated tasks such as sentiment analysis, sentiment strength assessment, and topic classification, creating new opportunities to manage reputation more efficiently. However, current PLM solutions typically address these tasks individually and require specialized training for each company, limiting scalability and flexibility. To address these challenges, we propose a novel system framework to assist public relations firms by leveraging PLMs for automatic classification. We evaluated this approach through experiments on four companies, comparing zero-shot and fine-tuned models in task-specific and multi-task configurations. Additionally, we explored the transfer of knowledge between client companies using fine-tuned models. Results indicate that multi-task, multi-company fine-tuned PLM models offer simpler system management with competitive performance compared to highly specialized models.

Introduction

Social media platforms provide a ubiquitous and accessible communication tool, enabling real-time interaction between users and organizations (Al Olaimat et al. 2022). Companies are increasingly leveraging these platforms to engage with their audiences, fostering more direct and efficient communication (Wang, Cheng, and Sun 2021). As a result, social media has become an important component of corporate communication strategies, offering businesses opportunities to build brand loyalty, address customer concerns, and promote products or services. This shift toward digital interaction highlights the growing importance of understanding and optimizing company-audience dynamics on these platforms.

A company's reputation is shaped by several factors, including the quality of its relationships with the audience (Wang, Cheng, and Sun 2021; Al Olaimat et al. 2022), effective crisis communication on social media (Civelek, Çemberci, and Eralp 2016), public sentiment towards the company, and the prevalent topics discussed in relation to it (Doorley and Garcia 2015).

Together, these elements influence stakeholder perceptions and can have a significant impact on a company's market position, customer loyalty, and overall business performance. While reputation may initially appear to be an intangible asset, it can be quantified and its performance tracked through various metrics and analytical methods (Doorley and Garcia 2015). This measurability allows companies to objectively assess, monitor, and strategically manage their reputational standing over time. The benefits of maintaining a consistent, strong reputation are far-reaching: businesses with solid reputations attract top talent, negotiate more favorable terms with suppliers, gain competitive advantages in the marketplace, receive free social media coverage, and are often able to command premium prices for their services (Doorley and Garcia 2015).

In this study, we address the challenge of supporting the corporate reputation measurement pipeline by simultaneously tackling the following tasks: *(i)* sentiment analysis, *(ii)* sentiment strength assessment, and *(iii)* classification of topics discussed on social media platforms. These tasks provide businesses with valuable insights into public perception. Traditionally, Public Relations (PR) practitioners relied on manual labeling to accurately assess sentiment and categorize discussions, often due to concerns about data quality. However, advancements in Artificial Intelligence (AI) and Pretrained Language Models (PLMs) have transformed this process, enabling automated analysis with high accuracy, efficiency, and data-driven scalability.

Despite such advancements, current practices in automated sentiment analysis and topic classification for reputation management face several constraints. Many existing PLM solutions typically address only a single task at a time (Kim et al. 2023; Gao, Ghosh, and Gimpel 2023; Hu et al. 2024; Zhang et al. 2024), and those that handle multiple tasks (Wang et al. 2023; Liu et al. 2023b; Xin et al. 2024; Shen et al. 2024) still require highly specialized training regimes tailored to specific companies or contexts. This complexity makes it difficult to reuse models across multiple companies for various tasks, limiting scalability and flexibility. As a result, PR practitioners and Machine Learning Op-

erations (MLOps) professionals must rely on their expertise to adapt solutions for new clients, often experimenting and fine-tuning several combinations along the way. Key limitations include the challenges of transferring knowledge between clients and the lack of comprehensive system designs that incorporate machine learning models specifically for social media analysis in the PR industry.

We propose a novel approach that uses PLMs for the automatic classification of sentiment analysis, sentiment strength, and topic classification to overcome these limitations with a single request. Instead of relying on prompt learning, which introduces complexity for multi-task approaches, complicates system design, and lacks intuitive interpretability (Wang et al. 2023; Hu et al. 2024), we use discrete prompting. This method employs straightforward textual prompts that are easily understood by both MLOps and PR professionals, simplifying the system design while maintaining flexibility.

To evaluate our approach, we conducted a three-step experiment. In the first step, we addressed the tasks individually through zero-shot learning and fine-tuning of task-specific models. In the second step, we tested a multi-task prompt to assess both zero-shot and fine-tuned company-specific models. Finally, in the third step, we explored how existing knowledge from current client companies could be applied by testing a single and global fine-tuned model across new companies.

To test our proposal in a real-world scenario, we implemented a system framework designed to handle large-scale data ingestion from multiple social media platforms through a custom-built Data Warehouse. This pipeline ensured consistent data transformation, facilitated the identification of bottlenecks, and enabled automated classification for topics and sentiment-related tasks using fine-tuned PLMs. As a result, we were able to conduct a more accurate cost analysis that mirrors a large-volume PR environment. The results indicate that a single, globally fine-tuned model covering all companies and tasks offers simpler management, lower costs, and competitive performance across the board.

The remainder of this paper is structured as follows: Section 2 presents the main related work. Section 3 introduces our framework and dataset. Section 4 provides the evaluation and discusses the key findings regarding performance. Finally, Section 5 outlines future work.

Background and Related Work

Public Relations and Social Media. PR companies manage their clients' reputations and communication strategies, aiming to build and sustain a positive public image while tackling challenges such as crisis management, media relations, brand positioning, and audience engagement. Social media platforms have revolutionized this process by providing dynamic environments where reputation is continuously developed and managed in real-time (Batrinca and Treleaven 2015). PR professionals still rely on traditional media for tasks like crisis management (Taylor and Kent 2010; Verhoeven et al. 2014; Kent and Taylor 2016) but the rise of social media analytics (Batrinca and Treleaven 2015; Stieglitz et al. 2018) has enabled companies to enhance communication, boosting audience engagement through strategic monitoring and direct interaction with stakeholders (Stieglitz et al. 2018; Andreotta et al. 2019; Valentini 2015).

The analysis of interactions between users and companies on social platforms relies on a vast amount of both structured and unstructured data, including text, images, and videos. This data offers insights into public opinion and behavior towards companies (Verhoeven et al. 2014; Stieglitz et al. 2018). For example, users may engage with companies by directly tagging them with their official profile names or indirectly mentioning the company within the content they share, whether textual or visual.

Such interactions, referred to in this paper as **mentions**, provide a common resource that companies often transform into valuable information. Merely tracking the volume and frequency of social media mentions is insufficient for extracting strategic insights. PR companies rely on content analysis and annotation to gain more refined information, such as identifying the topics of discussion, the sentiment expressed, and the context surrounding each mention. These in-depth analyses enable PR firms to develop comprehensive key performance indicators, dashboards, and periodic reports for clients, providing valuable support for decision-making strategies (Stieglitz et al. 2018). Thus, PR companies track mentions along with their associated refined information to understand how audiences interact with their clients (Batrinca and Treleaven 2015). However, the manual nature of this process limits the scale of insights that can be generated. To address this, automated text analytics and sentiment analysis tools have been designed to process large volumes of online mentions more efficiently (Batrinca and Treleaven 2015; Stieglitz et al. 2018; Andreotta et al. 2019). In particular, AI-powered solutions can categorize discussions by topic, sentiment, and identify influential voices on a much larger scale than manual processes, which allows PR professionals to respond quickly to shifts in brand perception (Andreotta et al. 2019).

Accordingly, this study focuses on optimizing the labeling and analysis of three key PR tasks for social media: *(i)* sentiment analysis, *(ii)* sentiment strength, and *(iii)* topic classification. Sentiment analysis involves determining whether a piece of text expresses a positive, negative, or neutral opinion (Balaji, Annavarapu, and Bablani 2021). Sentiment strength refers to the intensity of the expressed sentiment, which can be categorized into levels such as weak, mild, or strong—whether positive or negative (Batrinca and Treleaven 2015). Topic classification involves grouping social media posts into meaningful categories, such as "health" or "education". We also observe that when PR companies offer this service to clients, they typically establish a predefined list of themes. However, this list may need to be updated over time as new events emerge or the scope of monitoring evolves (Stieglitz et al. 2018).

Text Classification using PLMs. Recent advancements in NLP have introduced pretrained language models such as the GPT family (Brown et al. 2020), BERT (Devlin et al. 2019), T5 (Raffel et al. 2020), and Gemini (Anil et al. 2024), which have transformed tasks like sentiment analysis, sen-

timent strength, and topic classification. These models utilize their extensive training on large volumes of linguistic data and factual information. In this study, we approach all three tasks – sentiment analysis, sentiment strength, and topic classification as text classification problems.

To the best of our knowledge, there is limited literature on tackling text classification problems in the PR domain (Andreotta et al. 2019), which is why we turn to the broader AI literature to explore text classification in other domains. We have observed a growing body of research that examines the use of PLMs in zero-shot and few-shot text classification scenarios across various datasets and contexts (Zhang et al. 2022; Gera et al. 2022; Kim et al. 2023; Hu et al. 2024). In zero-shot learning, PLMs can be conditioned through a textual prompt to achieve the desired outcome, relying solely on their inherent parametric knowledge without the need for fine-tuning on the downstream task (Brown et al. 2020).

An early approach to text classification involves framing the task using cloze-style prediction with masked tokens and a pattern-verbalizer method (Schick and Schütze 2021; Zhang et al. 2022; Gao, Ghosh, and Gimpel 2023; Hu et al. 2024). For example, a movie review like "The movie was amazing!" could be framed as "The movie was amazing! It was [MASK]" and the PLM would be asked to predict the masked token using the masked language modeling classification head. Since the predicted token may not always align perfectly with the expected outcome (positive or negative), previous studies have used verbalizers to map the predicted token to one of the valid classes (positive or negative) (Zhang et al. 2022; Gao, Ghosh, and Gimpel 2023). However, this approach has faced criticism due to limitations with the pattern-verbalizer pair, where small changes in the pattern or verbalizer can cause significant performance degradation (Gao, Ghosh, and Gimpel 2023).

To overcome these limitations, advanced prompt-based methods, such as continuous prompts and prompt tuning, have emerged. These techniques enable models to adapt to new tasks or linguistic patterns without requiring manual updates to verbalizer mappings (Lester, Al-Rfou, and Constant 2021; Liu et al. 2021). However, continuous prompts often suffer from a lack of intuitive interpretability (Hu et al. 2024), may underperform on smaller large language models (LLMs) (Wang et al. 2023), and require PR professionals to fine-tune and manage specific embeddings for each task. While fine-tuning continuous prompts is more cost-effective than fine-tuning an entire PLM, full fine-tuning typically yields better downstream performance and offers greater theoretical expressiveness (Petrov, Torr, and Bibi 2024). Moreover, avoiding storing and managing multiple task-specific embeddings can significantly streamline system architecture. Failing to consider this can lead to more software complexity, hindering system maintenance and degrading performance, especially in high-throughput environments such as PR, where operational efficiency is critical.

Accordingly, we focus on discrete prompting and fine-tuning small commercial PLMs to address these issues while keeping the solution accessible to both data scientists and PR professionals. Discrete prompting relies entirely on textual conditioning of the PLMs to solve downstream tasks, offering greater transparency and interpretability. In our study, we explore two primary approaches to discrete prompting: *(i)* single-task prompting and *(ii)* multi-task prompting. For single-task prompting, we examine the performance of PLMs when prompted with task-specific instructions. This approach allows for targeted optimization of individual tasks but may require separate prompts for each task. It is also the most extensively studied method in the existing literature (Kim et al. 2023; Gao, Ghosh, and Gimpel 2023; Hu et al. 2024; Zhang et al. 2024).

In the context of multi-task prompting, we evaluate the effectiveness of using task-shared prompts, which offer a more versatile and efficient approach. This strategy allows a single prompt to address multiple related tasks for a given input, potentially enhancing the efficiency of PLMs by reducing tokens per minute and requests per minute, as demonstrated in several domains (Wang et al. 2023; Liu et al. 2023b; Xin et al. 2024; Shen et al. 2024). However, this has yet to be verified in public relations systems. Next, we present a new design and strategy for applying multi-task prompting within the field of public relations and monitoring. This is achieved through the development of a comprehensive framework that effectively emulates real-world challenges.

Material and Methods

This section introduces a new strategy for framing the text classification problem using discrete prompting in both single-task and multi-task scenarios, with approaches for zero-shot learning and fine-tuning.

Additionally, we present a comprehensive system framework designed to handle near real-time, large-scale ingestion and classification through discrete prompting. Our proposed framework is illustrated in Figure 1. We first define the text classification problem through discrete prompting, then we detail the dataset collection process and its basic statistics, followed by an exploration of the prompting and system design.

Problem Definition. Let $\{\mathcal{T}_i\}_{i=1}^{N}$ be a set of tasks, each containing a set $\mathcal{T}_i = \{\mathcal{X}_i, \mathcal{Y}_i\}$ of documents $\mathcal{X}_i = \{x_1, x_2, \ldots, x_{|\mathcal{X}_i|}\}$ and a set of textual labels $\mathcal{Y}_i = \{y_1, y_2, \ldots, y_{|\mathcal{Y}_i|}\}$, where each document contains a set of strings. For each $x_k \in \mathcal{X}_i$, there is one and only one associated label y_l. Since the set of documents is the same for every task, we simplify notation by defining a task i as $\mathcal{T}_i = \{\mathcal{X}, \mathcal{Y}_i\}$. The datasets $\mathcal{D}_{train}$, $\mathcal{D}_{val}$, and $\mathcal{D}_{test}$ represent the training, evaluation, and test sets, respectively, obtained from $\mathcal{X}$; thus, $\mathcal{X} = \mathcal{D}_{train} \cup \mathcal{D}_{val} \cup \mathcal{D}_{test}$.

Let $\{t_i\}_{i=1}^{N}$ represent all template strings used to address each task by prompting the large pretrained language model $\mathcal{M}(\theta)$ with parameters θ. Thus, the prompt $p_i^k = t_i(x_k)$ is employed to solve task i for document k, generated by applying the template string t_i to document x_k. The set $\mathcal{P}_i = \{p_i^1, p_i^2, \ldots, p_i^{|\mathcal{X}|}\}$ includes all prompts for addressing task i. Prompts $\mathcal{P}_i^{test}$ are created by applying the template strings to each document in the dataset $\mathcal{D}_{test}$. To address task i, we use the model $\mathcal{M}$ to map each input document to the target labels using a single-task and task-specific

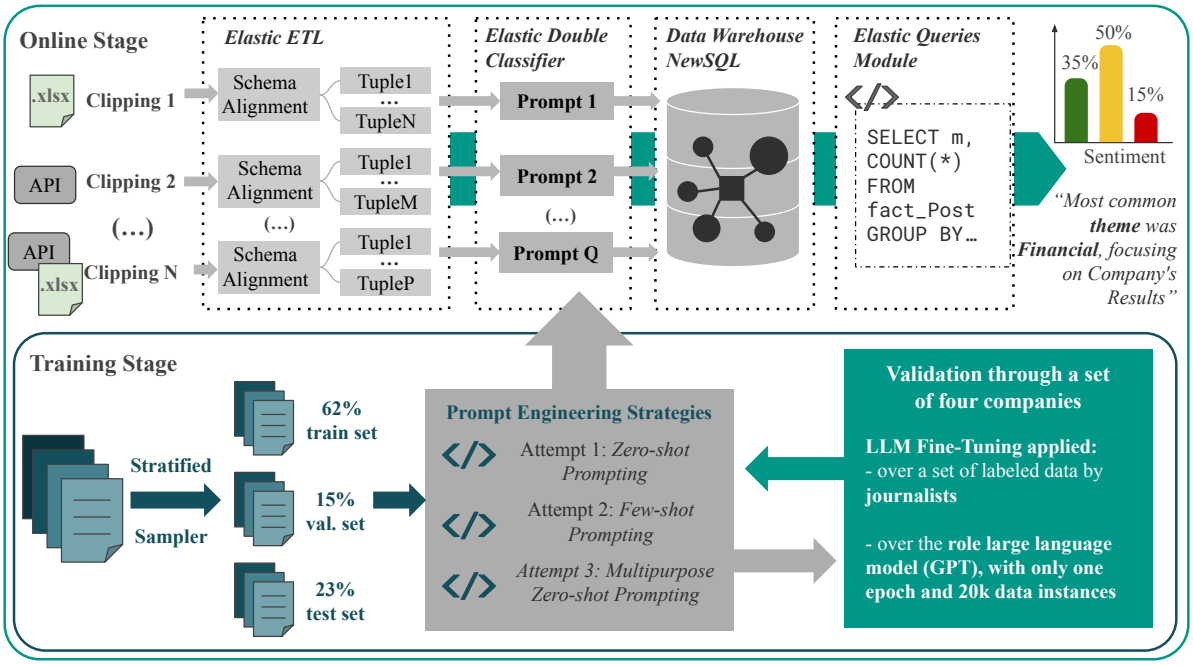

Figure 1: Overall pipeline of the proposed framework, including the training and online stages.

prompt paradigm as follows: $\mathcal{M}(\mathcal{P}_i^{test}, \theta) \rightarrow \mathcal{Y}_i$. In contrast, in a multi-task paradigm, we use a task-shared prompt to unify the representation of various tasks. Here, a single prompt template $\mathcal{P}_{multi}^{test}$ is designed to be applicable across all tasks, allowing the model $\mathcal{M}$ to process input documents from the entire dataset $\mathcal{D}_{test}$ within the same prompt framework, as follows: $\mathcal{M}(\mathcal{P}_{multi}^{test}, \theta) \rightarrow \mathcal{Y}_{multi}$, where $\mathcal{Y}_{multi} = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \ldots \times \mathcal{Y}_N$ and both $\mathcal{Y}_i$ and $\mathcal{Y}_{multi}$ are textually generated.

Prompting model $\mathcal{M}$ to solve new tasks without any concrete pairs (x_k, y_l) of examples in the input is referred to as zero-shot learning. In contrast, prompting with one or more concrete pairs of examples in the input is known as few-shot learning. For both zero-shot and few-shot learning, the parameters θ are fixed. If model performance on a task does not improve with few-shot learning or prompt engineering, model $\mathcal{M}$ can be fine-tuned to enhance performance.

However, the approach to fine-tuning varies depending on whether a task-specific or task-shared prompt is used. In the single-task paradigm (i.*e.*, using task-specific prompts), the model is fine-tuned using prompts $\mathcal{P}_i^{train}$. We denote the model $\mathcal{M}$ fine-tuned with the set of prompts $\mathcal{P}_i^{train}$ as $\mathcal{M}_i^*$. The fine-tuned model $\mathcal{M}_i^*$ learns to generate target labels $\mathcal{Y}_i$ based on the corresponding input prompts $\mathcal{P}_i^{train}$ for each task independently. Consequently, we obtain a set of models $\{\mathcal{M}_1^*, \mathcal{M}_2^*, \ldots, \mathcal{M}_N^*\}$, each optimized for a specific task.

In the multi-task paradigm, a task-shared prompt $\mathcal{P}_{multi}^{train}$ is used across all tasks, allowing the model $\mathcal{M}$ to process input documents from the entire dataset $\mathcal{D}_{train}$ within a unified framework. The model is fine-tuned on prompts $\mathcal{P}_{multi}^{train}$, which is a single template string applied to all documents in $\mathcal{D}_{train}$ for addressing all tasks. This fine-tuning optimizes the parameters θ to capture both shared and task-specific patterns simultaneously. We denote the fine-tuned model in the multi-task paradigm as $\mathcal{M}^*$. Unlike the single-task paradigm, the resulting set of models in the multi-task approach consists of a single model, $\mathcal{M}^*$.

While multi-task fine-tuning can improve generalization and leverage shared knowledge across tasks, it may occasionally degrade performance on individual tasks. In contrast, single-task fine-tuning allows for precise adaptation to each specific task but may require significant computational and data resources, especially when dealing with many tasks, as each task requires its own dedicated model and ongoing maintenance.

Gathering Data from Social Media. We propose a system design to emulate a real-world PR environment that handles multiple data streams from various social media platforms (see Figure 1). However, rather than navigating through a myriad of raw, proprietary APIs for data collection and cleansing, PR systems (and ours) rely on third-party solutions that aggregate, clean, and curate mentions and publications from a specific set of data sources. These solutions are known as *clipping* software, as they also collect information from social media, digital and traditional newspapers, radio, and TV (following proper digitization). Consequently, this software is highly specialized, with different solutions excelling in handling data from specific media types – some are better suited for short texts on social media, others for image-based networks, etc.

The clipping data collection relies on keywords and regular expressions. For example, if a public oil company, *CompanyFoo123*, wants to monitor its reputation, a specialized

Company	Industry/Type	Description
C_1	Mineral Resources	A holding company with mineral extraction brands.
C_2	Consulting	Offers consulting and training for private companies.
C_3	Banking	An investment bank with private investment funds.
C_4	Communication	A communication agency/department.
C_5	Tourism	A tourism agency/department.
C_6	Insurance	A holding of insurance services nationwide.

Table 1: Characteristics of the six companies in the `PRMentions` dataset.

clipping tool can be used to gather information from their Investor Relations releases in specific circles. A second clipping tool can collect mentions of *CompanyFoo123* along with keywords such as "oil," "gas," "ecoenvironment," and "leakage" in general, non-specialized social networks. A third clipping tool can monitor mentions of the company in mainstream press articles regarding its business with governmental institutions. Naturally, these sources are exported to PR solutions in different styles and formats (e.g., API, .json, *scraping*) and must undergo a second conformance pipeline to enhance data quality and compliance.

In this study, we constructed a dataset, named `PRMentions`, using five real-world clipping software tools: two specialized in regular social media, featuring data from Facebook, Instagram, Twitter, LinkedIn, TikTok, and Snapchat; one from a global web link crawler; and two focused on mentions from newspapers, TV, and video streaming platforms, including YouTube. The dataset includes mentions related to six companies from very different sectors, with data spanning from January 2023 to December 2023. The activities and characteristics of the companies are detailed in Table 1. The selection of filter keywords for these companies was collaboratively determined between a PR expert and the company communication department during an initial briefing meeting, with no more than 12 topics selected. This approach ensures that the captured data covers a wide range of relevant business areas, with the chosen keywords and expressions serving as the primary filters throughout the data collection phase.

Data filtering and staging are handled by the clipping software, with mentions also being manually labeled by PR personnel. These experts classify the data across the three key tasks: sentiment analysis, sentiment strength, and text classification. While the labeling *schema* varies among client companies and tasks, certain standards are maintained across all clients. For instance, text classification is flexible and tailored to each client's specific needs and interests, with the number of categories varying accordingly. However, for sentiment analysis, all companies use a binary positive/negative classification system. Sentiment strength is universally measured on a discrete scale from 1 to 10, where 1 represents *very negative* and 10 represents *very positive*. To label our dataset, we requested two different humans to label each mention according to every task, with a third individual serving as a tie-breaker if necessary.

Discrete Prompting. Discrete prompting involves manually crafting prompts using natural language. Here, we applied discrete prompting to classify a single post $x_k \in \mathcal{X}$ using two main approaches: single-task and multi-task. In the single-task approach, we designed separate prompts for each classification task: sentiment analysis, sentiment strength, and text classification. In this case, classifying a single post requires querying the PLM three times for zero-shot classification or using three distinct fine-tuned PLMs, which increases both the RPM and system complexity, including cost and maintenance. To address these limitations, we also explored a multi-task strategy. This approach consolidates the prompts by creating a unified prompt $\mathcal{P}_{multi}$ that handles all three classification tasks simultaneously.

For the experimental methodology, we first conducted a prompt engineering phase following the approach outlined in Liu et al. (2023), where we selected the most effective textual prompts for both the single-task and multi-task approaches in a zero-shot context. Initially, we developed an empirical prompt using the chain of thought technique, which encourages the model to articulate its reasoning process sequentially. Next, we prompted the LLM to generate three additional prompts for each approach. We then conducted preliminary experiments on a selected sample of companies to assess the effectiveness of each prompt, using the F1-score as the evaluation metric (see Appendix).

We also fine-tuned PLMs for different contexts. In the single-task paradigm, each model was specifically adapted to its corresponding classification task—one for sentiment analysis, one for sentiment strength, and one for text classification. For fine-tuning in the single-task paradigm, we prepared a set of input prompts $\mathcal{P}_i^{train}$ and associated classes $\mathcal{Y}_i$ for each task i. In the multi-task paradigm, we prepared a unified set of input prompts $\mathcal{P}_{multi}^{train}$ and associated classes $\mathcal{Y}_{multi}$. In the single-task case, fine-tuning ensures that each model is optimized for its respective task for a company C, resulting in models such as $\{\mathcal{M}_1^*, \mathcal{M}_2^*, \mathcal{M}_3^*\}$. In contrast, the multi-task approach yields a single model $\mathcal{M}^*$ that is fine-tuned to solve all three tasks simultaneously.

Framework Design. The proposed framework is designed to efficiently process multiple data streams from various social media platforms through different clipping software. It employs an Extract, Transform, Load (ETL) process (Online Stage) to maintain data quality, consistency, and timeliness, storing the mentions in a Data Warehouse. This ETL routine is implemented using a custom-built, open-source Python framework, designed to handle a diverse range of data from clipping software. Thus, ETL processes are executed in

parallel in an elastic cloud environment, with each process connected to a classifier that can execute one or more prompts. The outputs of these classifiers are aggregated into a centralized Data Warehouse. The prompt-driven classifier serves as the entry point for experimenting with and testing single- and multi-task approaches (Training Stage). Our framework relies on a NewSQL solution not only to reduce insertion time and enable horizontal scalability but also to facilitate the querying module, which can be expressed in simple SQL or ROLAP fashion. The query results are displayed through interactive dashboards for PR experts.

Simplified Schema and Data Transformation. The ETL process handles data from a variety of clipping sources, standardizing it with the following attributes: *(i)* Publication date; *(ii)* Title of the mention; *(iii)* Original source (e.g., which social network); *(iv)* Author; *(v)* List of interactions (including comments, reactions, shares, reach, audience, and view counts); *(vi)* Ad value; and *(vii)* Original URL of the mention. The ETL stages organize data from different sources into a dataframe abstraction and execute two main types of routines: `transform` functions (e.g., converting text to uppercase, removing accents, and standardizing data types such as dates) and `user functions` for customized data cleansing and imputation.

Schema Alignment and Loading Stage. The ETL process maps the original schema of different data sources into a dataframe abstraction, which is then loaded into the Data Warehouse. This mapping is defined by a YAML file, with parameters that include: *columns*, which specify the data to be extracted to populate the attributes; *functions*, to be applied during insertion; and operations for removing the consumed data instance, preventing its recategorization into another attribute. Multiple functions can be configured per column, with each attribute capable of returning a specific type, such as a unique value, a list, or a dimensional list.

Deterministic and Unique Identifiers. During the loading stage of the ETL process, a deterministic hash is generated from the post's URL to create a unique identifier. This approach offers several advantages, as it eliminates the need to lock database tables during insertion, improving performance and ensuring consistency. Since the hash is deterministic, any collisions would indicate posts with the same URL, enabling entry deduplication at the loading stage without requiring additional procedures. The hashing strategy involves taking the URL and generating a deterministic UUID (Universally Unique Identifier) using Python's `hashlib` and `uuid` libraries. Specifically, the URL is first hashed with the SHA-1 algorithm, and this hash is then used to generate a UUID using the `uuid5` function. This UUID, along with the company ID, serves as the primary key for mentions across all related tables.

Data Transfer Object. The ETL process also includes a Data Transfer Object (DTO) mechanism to efficiently transfer data through the pipeline. The DTOs encapsulate the data, ensuring it is passed smoothly between the various stages of the ETL process. This approach minimizes the risk of data corruption and ensures that the transformation functions can be executed across different datasets.

Parallel Processing. We rely on a cloud environment that deploys separate instances of the ETL process for each portion of input data, ensuring that the data is processed in isolation. Our framework also utilizes LLM services, which provide similar elasticity to our own, with known levels of service. Finally, we take advantage of the scalability of a NewSQL system for data insertion, enabling high-volume data transfers without pooling, so that each data stream is processed independently from the others.

Experimental Evaluation

In this section, we present an extensive evaluation using real-world data. We conducted experiments with both single-task and multi-task paradigms in zero-shot and fine-tuning scenario, which were carried out in three incremental steps: *(i)* zero-shot evaluation for individual companies and tasks, *(ii)* fine-tuning for individual companies and tasks, and *(iii)* fine-tuning the multi-task approach across all companies. This methodology enabled us to highlight the strengths observed in each competing approach.

`PRMentions` statistics. Figure 2 presents the `PRMentions` statistics. From the pool of collected mentions, we selected a stratified sample based on the proportion of classes in each task for each client company. Sampling was applied to each task and company to ensure balanced data representation while controlling the costs associated with PLM usage. After preparing the sample $\mathcal{X}$, we divided it into $\mathcal{D}_{train}$, $\mathcal{D}_{val}$, and $\mathcal{D}_{test}$ sets. These splits were then used for experimenting with discrete prompts and fine-tuning PLMs.

Experimental Setup. We used the latest GPT-4o Mini version (gpt-4o-mini-2024-07-18) via OpenAI's API and Gemini-1.5-Flash, with inference set to a temperature of 0.3 and fine-tuning performed over a single epoch. A stratified sample was employed, consisting of 1,000 posts for training, 1,000 posts for testing, and 200 posts for validation.

Intra-company Experiments. This part of the experiment focuses on fine-tuning and testing data from the same company. We conducted experiments with four selected companies from our pool of organizations. This approach enabled us to evaluate the model's ability to specialize in a particular company's domain, offering insights into how well the fine-tuned models perform when applied to the test dataset split from the same company. Specifically, we carried out zero-shot and fine-tuning experiments for both single-task and multi-task paradigms within this context.

Table 2 (1–5) presents the results for the intra-company experiments. For experiments 1 and 2, we rely solely on the test split for company C_j, as both are zero-shot experiments. Experiment 2 generates a set of prompts, $\mathcal{P}^{test}_{multi}$, containing a single prompt for each example in the dataset $\mathcal{D}^{C_j}_{test}$, while

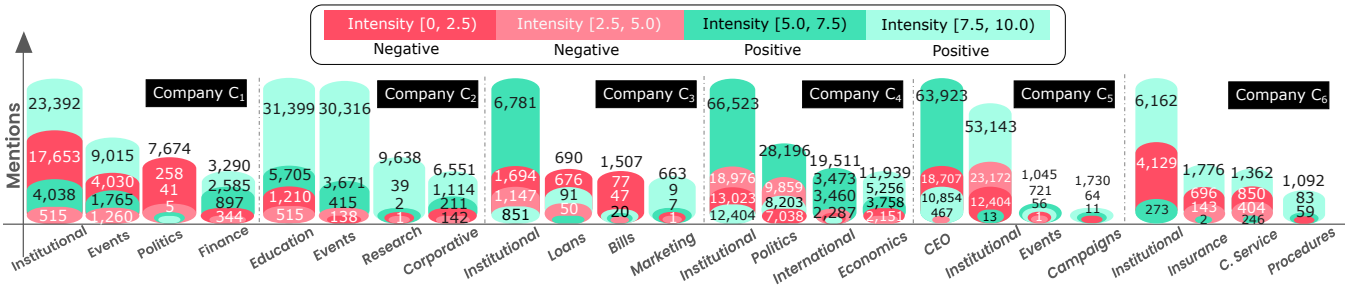

Figure 2: Univariate distribution for top-4 topics regarding their sentiment and intensity across the six examined companies.

	Exp.	Learning Method	Task	Train	Test	# Fine-tunings
Intra-company	Exp. 1	Zero-shot	Single-task	—	$\mathcal{D}_{test}^{C_j}$	—
	Exp. 2	Zero-shot	Multi-task	—	$\mathcal{D}_{test}^{C_j}$	—
	Exp. 3	Fine-tuning	Single-task	$\mathcal{D}_{train}^{C_j}$	$\mathcal{D}_{test}^{C_j}$	18
	Exp. 4	Fine-tuning	Multi-task	$\mathcal{D}_{train}^{C_j}$	$\mathcal{D}_{test}^{C_j}$	6
	Exp. 5*	Fine-tuning	Single-task	$\bigcup_{i=1}^{3} \mathcal{D}_{train}^{C_j}$	$\bigcup_{i=1}^{3} \mathcal{D}_{test}^{C_j}$	6
Inter-company	Exp. 6**	Fine-tuning	Single-task	$\bigcup_{j=1}^{4} \mathcal{D}_{train}^{C_j}$	$\bigcup_{j=1}^{4} \mathcal{D}_{test}^{C_j}$	3
	Exp. 7	Fine-tuning	Single-task	$\bigcup_{i=1}^{3}\bigcup_{j=1}^{4} \mathcal{D}_{train}^{C_j}$	$\bigcup_{i=1}^{3}\bigcup_{j=1}^{4} \mathcal{D}_{test}^{C_j}$	1
	Exp. 8	Fine-tuning	Multi-task	$\bigcup_{j=1}^{4} \mathcal{D}_{train}^{C_j}$	$\bigcup_{j=1}^{4} \mathcal{D}_{test}^{C_j}$	1
Inter-company	Exp. 9	Transfer learning	Single-task	—	$\mathcal{D}_{test}^{C_5}, \mathcal{D}_{test}^{C_6}$	—
	Exp. 10	Transfer learning	Multi-task	—	$\mathcal{D}_{test}^{C_5}, \mathcal{D}_{test}^{C_6}$	—

Table 2: Experiments overview. $\mathcal{D}_{train}^{C_j}$ and $\mathcal{D}_{test}^{C_j}$ represents the train and test splits for company C_j. For experiment 5 (*), single-task prompts are joined for all tasks for a single company. For experiment 6 (**), it is the union of all prompts for task i for all companies $j = \{1, 2, 3, 4\}$, resulting in three different training/test pairs for each task.

experiment 1 produces three distinct prompts – one for each task – for every example in the dataset. For experiment 5, we compile all single-task prompts for all tasks into a unified pool, $\mathcal{P}_1^{C_j} \cup \mathcal{P}_2^{C_j} \cup \mathcal{P}_3^{C_j}$, and perform this consolidation for each train, validation, and test split.

Inter-company Experiments. Given that PR companies often retain existing clients while acquiring new ones, we conducted several experiments to evaluate the model's ability to use knowledge from existing clients and apply it to new clients. Specifically, we assessed how the knowledge gained from fine-tuned models on existing clients could be transferred to new clients. This approach aimed to determine whether new clients could benefit from the insights and models developed for existing clients, thus evaluating the efficiency and effectiveness of knowledge transfer across different organizations. For experiments 6, 7, and 8 in Table 2, we explored various methods of integrating knowledge from all companies, except for C_5 and C_6, which were kept separate for the transfer learning evaluation. In experiment 6, we trained and tested the model using single-task prompts from companies C_{1-4} by aggregating all prompts for each task. This resulted in three fine-tuned models, each specialized in one task and incorporating knowledge from all C_{1-4} companies. Next, we randomly selected two additional companies, C_5 and C_6, from the pool of client organizations, excluding C_1 to C_4. These companies were used to evaluate the transferability of knowledge gained from the fine-tuned models, assessing how well the model's performance generalizes to new contexts without prior specific training. Each model was tested once for each company. In experiment 7, we aggregated all single-task prompts for every task across all companies, resulting in a single model trained with task-specific prompts. In experiment 8, we fine-tuned a model using multi-task prompts for all companies, producing a single model capable of handling multi-task prompts. Each fine-tuned model was then tested across C_{1-4}.

Analysis of Intra and Inter-company Results. To evaluate predictive performance, we use precision, recall, and F1-score metrics. Since PLMs occasionally produce outputs outside the predefined set of labels, we also calculate the percentage of examples that result in out-of-scope or invalid predictions. This metric quantifies the extent of label mismatch regarding the model's reliability for constrained classification. In the intra-company zero-shot experiments (1 and 2 – Tables 3 and 3), the multi-task paradigm outper-

Task	Exp.	Company C_1			Company C_2			Company C_3			Company C_4			Company C_5			Company C_6		
		RE	**PR**	**F1**	**RE**	**PR**	**F1**	**RE**	**PR**	**F1**	**RE**	**PR**	**F1**	**RE**	**PR**	**F1**	**RE**	**PR**	**F1**
Sentiment	1	0.79	0.82	0.80	0.56	0.96	0.68	0.82	0.83	0.81	0.61	0.78	0.61	0.88	0.89	0.88	0.81	0.83	0.82
	2	0.82	0.83	0.82	0.59	0.96	0.71	0.84	0.85	0.84	0.61	0.78	0.62	0.88	0.89	0.88	0.87	0.87	0.87
	3	0.88	0.88	0.87	0.96	0.92	0.94	0.92	0.92	0.92	0.86	0.86	0.86	0.88	0.89	0.88	0.93	0.94	0.93
	4	0.88	0.88	0.87	0.97	0.96	0.96	0.91	0.91	0.91	0.86	0.86	0.85	0.96	0.96	0.96	0.94	0.94	0.94
	5	0.87	0.87	0.86	0.96	0.92	0.94	0.92	0.92	0.92	0.83	0.84	0.82	0.95	0.96	0.95	0.94	0.94	0.93
	6	0.87	0.87	0.87	0.95	0.96	0.95	0.91	0.91	0.91	0.82	0.82	0.82	-	-	-	-	-	-
	7	0.86	0.86	0.86	0.68	0.96	0.78	0.84	0.86	0.84	0.62	0.80	0.63	-	-	-	-	-	-
	8	0.84	0.84	0.83	0.97	0.97	0.97	0.90	0.91	0.90	0.82	0.81	0.81	-	-	-	-	-	-
Intensity	1	0.66	0.69	0.67	0.48	0.84	0.59	0.70	0.74	0.71	0.67	0.80	0.71	0.72	1.00	0.83	0.78	1.00	0.87
	2	0.73	0.70	0.67	0.83	0.82	0.83	0.78	0.76	0.74	0.29	0.80	0.24	0.95	1.00	0.98	0.94	1.00	0.97
	3	0.80	0.77	0.78	0.89	0.81	0.85	0.85	0.85	0.84	0.82	0.81	0.81	1.00	1.00	1.00	1.00	1.00	1.00
	4	0.82	0.80	0.80	0.93	0.92	0.92	0.86	0.85	0.85	0.84	0.83	0.84	1.00	1.00	1.00	1.00	1.00	1.00
	5	0.81	0.78	0.78	0.93	0.93	0.92	0.86	0.86	0.86	0.83	0.81	0.81	1.00	1.00	1.00	1.00	1.00	1.00
	6	0.81	0.81	0.80	0.90	0.90	0.89	0.85	0.86	0.85	0.82	0.82	0.82	-	-	-	-	-	-
	7	0.65	0.78	0.68	0.25	0.88	0.30	0.71	0.81	0.70	0.83	0.82	0.82	-	-	-	-	-	-
	8	0.78	0.79	0.76	0.90	0.89	0.89	0.81	0.81	0.79	0.78	0.79	0.79	-	-	-	-	-	-
Topic	1	0.40	0.49	0.33	0.37	0.62	0.43	0.63	0.70	0.64	0.66	0.69	0.66	0.62	0.83	0.70	0.38	0.58	0.33
	2	0.38	0.53	0.30	0.34	0.60	0.37	0.63	0.68	0.63	0.64	0.67	0.63	0.43	0.77	0.54	0.39	0.58	0.39
	3	0.82	0.81	0.81	0.82	0.82	0.81	0.80	0.79	0.79	0.74	0.76	0.73	0.94	0.92	0.93	0.86	0.85	0.85
	4	0.84	0.84	0.84	0.83	0.82	0.82	0.79	0.79	0.79	0.76	0.76	0.75	0.95	0.92	0.93	0.85	0.85	0.85
	5	0.84	0.82	0.83	0.83	0.83	0.83	0.77	0.77	0.76	0.77	0.77	0.77	0.92	0.90	0.91	0.83	0.83	0.83
	6	0.84	0.83	0.83	0.82	0.82	0.82	0.80	0.80	0.79	0.75	0.76	0.75	-	-	-	-	-	-
	7	0.82	0.86	0.83	0.80	0.81	0.80	0.79	0.81	0.79	0.75	0.75	0.74	-	-	-	-	-	-
	8	0.83	0.83	0.82	0.78	0.80	0.78	0.77	0.78	0.76	0.77	0.77	0.77	-	-	-	-	-	-

Table 3: Performance regarding Precision (PR), Recall (RE), and F1-Score (F1) for the GPT-based evaluation. Cells are color-coded based on their comparison against Gemini-based counterparts: lower (red), equal (orange), or higher (green).

formed in 12 out of 18 cases based on F1-score for both GPT and Gemini. This outcome suggests the multi-task approach not only significantly reduces token usage but also consolidates tasks into a single request, enhancing throughput efficiency. We observed that single-task prompts performed better for topic classification, while multi-task prompts showed clear dominance in sentiment analysis and strength tasks.

This suggests that, while the multi-task paradigm offers advantages, task-specific prompts in the zero-shot paradigm may still be preferable for lower budgets. Additionally, zero-shot experiments for sentiment and sentiment strength analysis showed performance scores closer to their fine-tuned counterparts, in contrast to the larger gap observed in the topic classification experiments. This suggests that sentiment-related tasks can effectively utilize the PLM's parametric knowledge, even in resource-constrained settings, reducing the need for extensive fine-tuning. These observations were consistent for both examined PLMs.

In the intra-company fine-tuning experiments (3–5), all cases showed improved F1-scores compared to the zero-shot experiments. The most significant gains were observed in the topic classification task, which aligns with expectations due to the sensitivity of this task to the specific class names in the list of topics. Zero-shot performance tends to be lower in this area because of the wide variation in topic lists across different client companies, making it difficult for the model to generalize without fine-tuning. Fine-tuning allowed the model to adapt to the unique topics relevant to each client, resulting in substantial improvements. Additionally, experiment 4, which trains a company-specific model to handle all three tasks simultaneously with a multi-task prompt, outperformed in 11 of 18 experiments. These results suggest that the multi-task approach not only improves performance but also provides a more cost-effective solution, primarily due to the reduced need for multiple fine-tuned models. Experiment 4 demonstrates that using a company-specific model to handle all tasks at once improves performance, simplifies system management, and reduces operational costs, making it an efficient choice for PR companies. The potential benefits are further emphasized by an average 38% reduction in token usage when comparing the multi-task approach to the single-task approach.

Further significant improvements could be achieved by developing a single model capable of serving multiple companies and tasks simultaneously. To explore this potential, we conducted experiments 6–8 (as detailed in Table 2). These experiments aimed to evaluate the effectiveness of shared, multi-company, multi-task models in enhancing scalability, reducing costs, and simplifying management by eliminating the need for company-specific models. Despite consolidating tasks and companies, the model maintained high performance across all contexts. As shown in Table 2, the best inter-company models outperformed their intra-company counterparts in 7 out of 12 cases. In the instances where the multi-company model's performance was slightly lower, the decrease was minimal, with a drop

Task	Exp.	Company C_1			Company C_2			Company C_3			Company C_4			Company C_5			Company C_6		
		RE	**PR**	**F1**	**RE**	**PR**	**F1**	**RE**	**PR**	**F1**	**RE**	**PR**	**F1**	**RE**	**PR**	**F1**	**RE**	**PR**	**F1**
Sentiment	1	0.78	0.82	0.80	0.51	0.96	0.64	0.77	0.83	0.77	0.50	0.78	0.52	0.87	0.89	0.88	0.85	0.86	0.86
	2	0.80	0.82	0.80	0.55	0.96	0.67	0.83	0.86	0.83	0.55	0.79	0.56	0.88	0.90	0.89	0.88	0.88	0.88
	3	0.87	0.87	0.87	0.91	0.92	0.91	0.89	0.92	0.90	0.79	0.86	0.82	0.88	0.89	0.88	0.96	0.96	0.96
	4	0.87	0.88	0.87	0.92	0.93	0.92	0.86	0.90	0.88	0.78	0.83	0.80	0.96	0.96	0.96	0.97	0.96	0.97
	5	0.86	0.86	0.86	0.91	0.92	0.91	0.89	0.92	0.90	0.74	0.84	0.79	0.95	0.96	0.96	0.97	0.96	0.96
	6	0.86	0.86	0.86	0.90	0.91	0.90	0.87	0.92	0.89	0.73	0.83	0.78	-	-	-	-	-	-
	7	0.85	0.85	0.85	0.63	0.94	0.75	0.81	0.86	0.83	0.58	0.80	0.67	-	-	-	-	-	-
	8	0.83	0.83	0.83	0.92	0.94	0.93	0.90	0.90	0.90	0.74	0.81	0.77	-	-	-	-	-	-
Intensity	1	0.66	0.69	0.67	0.48	0.84	0.59	0.70	0.74	0.71	0.67	0.80	0.71	0.72	1.00	0.83	0.78	1.00	0.87
	2	0.73	0.70	0.67	0.83	0.82	0.83	0.78	0.76	0.74	0.29	0.80	0.24	0.95	1.00	0.98	0.94	1.00	0.97
	3	0.80	0.77	0.78	0.89	0.81	0.85	0.85	0.85	0.84	0.82	0.81	0.81	1.00	1.00	1.00	1.00	1.00	1.00
	4	0.82	0.80	0.80	0.93	0.92	0.92	0.86	0.85	0.85	0.84	0.83	0.84	1.00	1.00	1.00	1.00	1.00	1.00
	5	0.81	0.78	0.78	0.93	0.93	0.92	0.86	0.86	0.86	0.83	0.81	0.81	1.00	1.00	1.00	1.00	1.00	1.00
	6	0.81	0.81	0.80	0.90	0.90	0.89	0.85	0.86	0.85	0.82	0.82	0.82	-	-	-	-	-	-
	7	0.65	0.78	0.68	0.25	0.88	0.30	0.71	0.81	0.70	0.83	0.82	0.82	-	-	-	-	-	-
	8	0.78	0.79	0.76	0.90	0.89	0.89	0.81	0.81	0.79	0.78	0.79	0.79	-	-	-	-	-	-
Topic	1	0.40	0.49	0.33	0.37	0.62	0.43	0.63	0.70	0.64	0.66	0.69	0.66	0.62	0.83	0.70	0.38	0.58	0.33
	2	0.38	0.53	0.30	0.34	0.60	0.37	0.63	0.68	0.63	0.64	0.67	0.63	0.43	0.77	0.54	0.39	0.58	0.39
	3	0.82	0.81	0.81	0.82	0.82	0.81	0.80	0.79	0.79	0.74	0.76	0.73	0.94	0.92	0.93	0.86	0.85	0.85
	4	0.84	0.84	0.84	0.83	0.82	0.82	0.79	0.79	0.79	0.76	0.76	0.75	0.95	0.92	0.93	0.85	0.85	0.85
	5	0.84	0.82	0.83	0.83	0.83	0.83	0.77	0.77	0.76	0.77	0.77	0.77	0.92	0.90	0.91	0.83	0.83	0.83
	6	0.84	0.83	0.83	0.82	0.82	0.82	0.80	0.80	0.79	0.75	0.76	0.75	-	-	-	-	-	-
	7	0.82	0.86	0.83	0.80	0.81	0.80	0.79	0.82	0.79	0.75	0.75	0.74	-	-	-	-	-	-
	8	0.83	0.83	0.82	0.78	0.80	0.78	0.77	0.78	0.76	0.77	0.77	0.77	-	-	-	-	-	-

Table 4: Performance regarding Precision (PR), Recall (RE), and F1-Score (F1) for the Gemini-based evaluation. Cells are color-coded based on their comparison against GPT-based counterparts: lower (red), equal (orange), or higher (green).

	Learned Co.	C_1			C_2			C_3			C_4		
	Exp.	**S**	**I**	**T**	**S**	**I**	**T**	**S**	**I**	**T**	**S**	**I**	**T**
C_5	3	0.82	0.90	0.79	0.44	1.00	0.76	0.90	0.98	0.72	0.75	0.45	0.74
	4	0.88	0.95	0.69	0.74	0.75	0.69	0.88	0.95	0.69	0.85	0.43	0.68
	5	0.83	0.95	0.79	0.44	0.76	0.67	0.87	0.96	0.73	0.82	0.31	0.58
C_6	3	0.92	1.00	0.18	0.68	1.00	0.31	0.92	1.00	0.26	0.76	0.07	0.26
	4	0.91	1.00	0.16	0.71	0.98	0.35	0.88	0.98	0.28	0.84	0.10	0.37
	5	0.90	1.00	0.16	0.68	0.95	0.33	0.90	1.00	0.25	0.78	0.02	0.32
	Exp.	**6**			**7**			**8**			**ZS-Ref.**		
	Task	**S**	**I**	**T**	**S**	**I**	**T**	**S**	**I**	**T**	**S**	**I**	**T**
C_5		0.80	0.86	0.81	0.88	0.50	0.78	0.76	0.85	0.69	0.88	0.83	0.70
C_6		0.91	1.00	0.17	0.91	0.89	0.20	0.85	1.00	0.18	0.82	0.87	0.33

Table 5: Transfer learning experiments. Columns C_1 to C_4 indicate the training datasets $\mathcal{D}_{train}^{C_j}$ used to train the PLM model, which is subsequently tested on companies C_5 or C_6. Columns S, I, and T represent tasks for sentiment analysis, intensity, and topic classification, respectively. ZS-Ref denotes the best intra-company single-task zero-shot F1 metrics (see Table 3).

of only 1% compared to the best intra-company results. Another significant finding is that experiment 6, which focused on task-specific models where each fine-tuned model addresses a single task individually, produced the best-performing models in 7 out of 12 cases. This suggests that task specialization may lead to more optimized models. The multi-task company-specific model from experiment 8 also achieved 5 out of 12 winning models compared to the best intra-company models. Moreover, when the experiment 8 model outperformed others, the performance increase was sometimes substantial, as observed with companies C_1 and C_4 for the sentiment strength task. Conversely, when its performance was lower, the drop in F1-score was minimal, indicating a robust and versatile approach to multi-task, multi-company modeling.

Transfer Learning. The transfer learning experiments presented in Table 5 over GPT-4o show that knowledge transfer

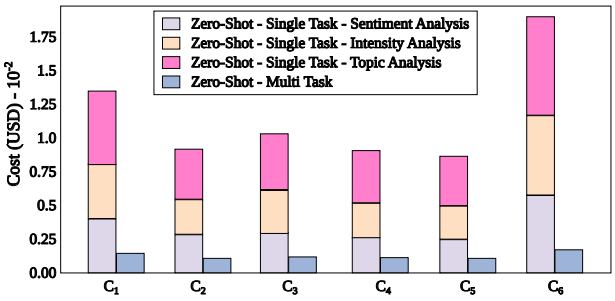

Figure 3: GPT-4o Mini Zero-Shot *vs.* Multi-task costs.

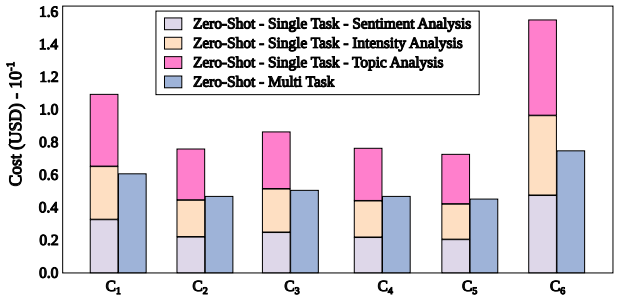

Figure 4: Gemini-1.5-Flash Zero-Shot *vs.* Multi-task costs.

can be effective in scenarios without additional processing. When the PLM is fine-tuned on a single-source company C_j and then tested on C_5 or C_6, performance can vary, sometimes exceeding or falling short of the zero-shot baseline.

This variability suggests potential instability when relying solely on a single-source company for fine-tuning. However, using data from all companies – through either single-task fine-tuning (experiment 6) or multi-company models – generally results in performance that matches or surpasses zero-shot results, especially for more complex tasks such as topic classification.

Despite these improvements, intra-company fine-tuned models still outperform, highlighting the need for further research into utilizing existing knowledge to consistently enhance performance beyond zero-shot.

Results and Discussion. Overall, the fine-tuning performances surpassed the zero-shot counterparts in all comparisons presented in Tables 3 and 4. The results also reveal that the average GPT measures outperformed Gemini in at least 70% of the entries for both sentiment and intensity analyses. However, for topic classification, the performances were nearly identical, with Gemini edging GPT by 51% to 49%. For the cost analysis, we measured the costs associated with the prompts used in all tests for both GPT-4o Mini and Gemini-1.5-Flash through their native APIs and cloud services. Figures 3 and 4 display the accumulated costs, highlighting the significant savings provided by the multi-task approach (approximately $5\times$ for GPT and $1.8\times$ for Gemini). In summary, GPT-4o Mini was more cost-effective than its counterpart, with an average budget reduction of 80%. Results in Figures 5 further suggest that the cost of fine-tuning multi-task prompts is roughly three-quarters of the cost for fine-tuned single-task prompts for both GPT-4o Mini and Gemini-1.5-Flash. This indicates that fine-tuning per company is more expensive to manage and takes longer to amortize compared to zero-shot models. Consequently, if newer versions of zero-shot models continue to improve, fine-tuning may become less advisable in the long term.

As a final observation, we experimented with few-shot learning by including a randomly selected pair (x_k, y_l) for each class $y_l \in \mathcal{Y}_i$ in the prompt. However, this approach did not lead to performance improvements compared to zero-shot. Moreover, the increased input token length proved im-

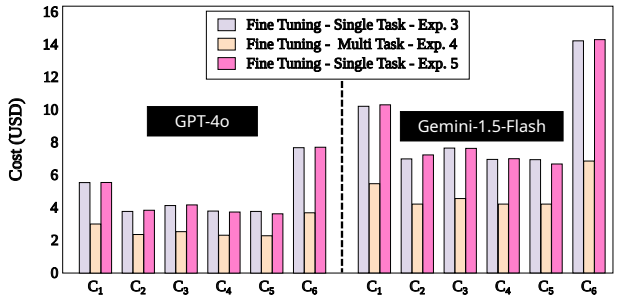

Figure 5: Tuning of GPT-4o Mini and Gemini-1.5-Flash.

practical for large-scale systems. As a result, we did not proceed with few-shot learning in the experimentation.

Conclusions

In this paper, we presented a novel approach to reputation management that utilizes PLMs for automated sentiment analysis, sentiment strength assessment, and topic classification across multiple companies by employing discrete prompting to develop a multi-task strategy. Through extensive experimentation with zero-shot and fine-tuned models, we observed that multi-task and multi-company configurations outperform task-specific and company-specific models in terms of scalability, efficiency, and associated costs.

Our findings suggest that multi-task models not only reduce operational complexity but also deliver competitive performance across diverse tasks with a single prompting request. Additionally, we explored how simulating the process of utilizing knowledge from existing clients in a PR company could aid in onboarding new clients, uncovering significant opportunities for scalable and cost-effective solutions in the PR industry. Our results indicate that performance varies depending on how data is aggregated across companies, whether using single-task or multi-task models. We hope this work will support PR professionals and researchers in advancing reputation management strategies.

Future research will focus on enhancing performance by incorporating Retrieval-Augmented Generation strategies. Additionally, we plan to explore multi-task interpretability by integrating explainable architectures to better understand how models share and specialize features for each task.

Acknowledgements

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), Brazil - Finance Code 001, and by the Carlos Chagas Filho Research Support Foundation of the Rio de Janeiro State (FAPERJ), Brazil (Grants E-26/204.238/2024 and E-26/204.544/2024).

References

Al Olaimat, F.; Habes, M.; Al Hadeed, A. Y.; and Al Jwaniat, M. I. 2022. Reputation management through social networking platforms for PR purposes: A SEM-based study in the Jordan. *Frontiers in Communication*, 7: 1009359.

Andreotta, M.; Nugroho, R.; Hurlstone, M. J.; Boschetti, F.; Farrell, S.; Walker, I.; and Paris, C. 2019. Analyzing social media data: A mixed-methods framework combining computational and qualitative text analysis. *Behavior research methods*, 51: 1766–1781.

Anil, R.; et al. 2024. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Balaji, T.; Annavarapu, C. S. R.; and Bablani, A. 2021. Machine learning algorithms for social media analysis: A survey. *Computer Science Review*, 40: 100395.

Batrinca, B.; and Treleaven, P. C. 2015. Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, 30: 89–116.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Civelek, M. E.; Çemberci, M.; and Eralp, N. E. 2016. The role of social media in crisis communication and crisis management. *International Journal of Research in Business & Social Science*, 5(3).

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Doorley, J.; and Garcia, H. F. 2015. *Reputation management: The key to successful public relations and corporate communication*. Routledge.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gao, L.; Ghosh, D.; and Gimpel, K. 2023. The Benefits of Label-Description Training for Zero-Shot Text Classification. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proc. of the 2023 Conf. on EMNLP*, 13823–13844. Singapore: ACL.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gera, A.; Halfon, A.; Shnarch, E.; Perlitz, Y.; Ein-Dor, L.; and Slonim, N. 2022. Zero-Shot Text Classification with Self-Training. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proc. of the Conf. on EMNLP*, 1107–1119. Abu Dhabi: ACL.

Hu, W.; Wang, Y.; Jia, Y.; Liao, Q.; and Zhou, B. 2024. A Multimodal Prompt Learning Framework for Early Detection of Fake News. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1): 651–662.

Kent, M. L.; and Taylor, M. 2016. From Homo Economicus to Homo dialogicus: Rethinking social media use in CSR communication. *Public relations review*, 42(1): 60–67.

Kim, S.; Joo, S.; Kim, D.; Jang, J.; Ye, S.; Shin, J.; and Seo, M. 2023. The CoT Collection: Improving Zero-shot and Few-shot Learning of Language Models via Chain-of-Thought Fine-Tuning. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proc. of the 2023 Conf. on EMNLP*, 12685–12708. Singapore: ACL.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In Moens, M.; Huang, X.; Specia, L.; and Yih, S. W., eds., *Proc. of the Conf. on EMNLP, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 3045–3059. ACL.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; and Tang, J. 2021. GPT Understands, Too. *CoRR*, abs/2103.10385.

Liu, Y.; Lu, Y.; Liu, H.; An, Y.; Xu, Z.; Yao, Z.; Zhang, B.; Xiong, Z.; and Gui, C. 2023b. Hierarchical prompt learning for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10888–10898.

Petrov, A.; Torr, P.; and Bibi, A. 2024. When Do Prompting and Prefix-Tuning Work? A Theory of Capabilities and Limitations. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21: 140:1–140:67.

Schick, T.; and Schütze, H. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, 255–269. Association for Computational Linguistics.

Shen, S.; Yang, S.; Zhang, T.; Zhai, B.; Gonzalez, J. E.; Keutzer, K.; and Darrell, T. 2024. Multitask vision-language prompt tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5656–5667.

Stieglitz, S.; Mirbabaie, M.; Ross, B.; and Neuberger, C. 2018. Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International journal of information management*, 39: 156–168.

Taylor, M.; and Kent, M. L. 2010. Anticipatory socialization in the use of social media in public relations: A content analysis of PRSA's Public Relations Tactics. *Public relations review*, 36(3): 207–214.

Valentini, C. 2015. Is using social media "good" for the public relations profession? A critical reflection. *Public relations review*, 41(2): 170–177.

Verhoeven, P.; Tench, R.; Zerfass, A.; Moreno, Á.; and Verčič, D. 2014. Crisis? What crisis?: How European professionals handle crises and crisis communication. *Public Relations Review*, 40(1): 107–109.

Wang, Y.; Cheng, Y.; and Sun, J. 2021. When public relations meets social media: A systematic review of social media related public relations research from 2006 to 2020. *Public Relations Review*, 47(4): 102081.

Wang, Z.; Panda, R.; Karlinsky, L.; Feris, R.; Sun, H.; and Kim, Y. 2023. Multitask Prompt Tuning Enables Parameter-Efficient Transfer Learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xin, Y.; Du, J.; Wang, Q.; Yan, K.; and Ding, S. 2024. Mmap: Multi-modal alignment prompt for cross-domain multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16076–16084.

Zhang, H.; Zhang, X.; Huang, H.; and Yu, L. 2022. Prompt-based meta-learning for few-shot text classification. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, 1342–1357.

Zhang, Y.; Wang, M.; Ren, C.; Li, Q.; Tiwari, P.; Wang, B.; and Qin, J. 2024. Pushing The Limit of LLM Capacity for Text Classification. *CoRR*, abs/2402.07470.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, we do not violate any privacy norms and actively seek to mitigate unfair profiling and disrespect to cultures through this work.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, see Tables 3 and 4, and the associated discussion in the Results & Discussion Section.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see the Material and Methods Section and the associated results in the Results & Discussion Section.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see Social Media Data for Public Relations Section.
   (e) Did you describe the limitations of your work? Yes.
   (f) Did you discuss any potential negative societal impacts of your work? Yes.
   (g) Did you discuss any potential misuse of your work? Yes.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, see the appendix for prompts, and the Experimental Evaluation Section for the code and experimental details.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see the Experimental Evaluation Section.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see the Experimental Evaluation Section.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see the Results & Discussion Section.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, see the Results & Discussion Section.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? NA.

(b) Did you mention the license of the assets? Yes, it is proprietary.

(c) Did you include any new assets in the supplemental material or as a URL? NA.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see Social Media Data for Public Relations Section.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA.

(d) Did you discuss how data is stored, shared, and deidentified? NA.

Appendix A – Manually translated prompts (Portuguese/BR to English)

Sentiment Analysis

Header Label the sentiment of the post in source. The post is enclosed in triple single quotes (''). The label should represent the post's sentiment polarity, which can be one of the following categories: ['POSITIVE', 'NEGATIVE']. Your task is to read the post from source and classify its sentiment polarity. The answer must be a single word, representing only one of the two categories, without any additional marks, bullets, or accents.

Specialist You are an expert in labeling the sentiment of social media posts.

Sentiment Strength

Header Label the intensity of sentiment within a social media post. The post is enclosed in triple single quotes (''). The sentiment of the post is known to be {post_sentiment}. Your task is to choose a label for the INTENSITY of the sentiment from the following categories: {classes}. To do this, read the social media post and classify the INTENSITY of its sentiment. Please provide a step-by-step explanation of your reasoning, and then select the appropriate intensity level from the available categories. What is the intensity of the sentiment for the post below?

Specialist You are an expert in labeling the sentiment of social media posts.

Topic Classification

Header You must act as a social media post analyst, as this is a multi-class task. The topic of the post belongs to one of the following categories: {classes}. Please follow these steps: Step 1 – Carefully read the post from {source}. Take as much time as needed to process the information, as this task is crucial for the entire company ecosystem. Step 2 – Classify the topic of the post by removing punctuation (', ", <, >, etc.) and diacritical marks (, ‘, ´, etc.). It is essential that you provide only a single topic name and nothing else. Keep your answer as concise as possible, and ensure it is within the provided list of categories. What is the topic of the post below?

Specialist You are an expert in labeling the sentiment of social media posts.

Multi-task Classification

Header You will act as a classifier for a social media post, performing three distinct and critical tasks. The post will be enclosed in triple single quotes (''). Carefully read the post and follow the instructions for each task. All answers must be as concise as possible.

1. **Sentiment polarity classification:**
   - Label the sentiment polarity of the post using only one of the following categories: ['POSITIVE', 'NEGATIVE'].
   - Provide only a single word for this task, without any marks or bullets.
2. **Sentiment intensity classification:**
   - Given that the polarity has already been determined, label the intensity of the sentiment's polarity as an integer value ranging from 1 to 10, where a higher number represents greater intensity.
   - Answer with a single number from 1 to 10, as concisely as possible for this task.
3. **Topic classification:**
   - You must choose only one of the following categories: {classes_topic} as the main topic of the post.
   - Be sure to remove punctuation (', ", <, >, etc.) and diacritical marks (, ‘, ´, etc.) before your analysis.
   - Provide only the name of the chosen topic, concisely.

Your answer must strictly follow this output format: Polarity: [YOUR POLARITY LABEL]. Intensity: [YOUR INTENSITY LABEL]. Topic: [YOUR TOPIC LABEL].

Specialist You are an expert in labeling the sentiment polarity, sentiment intensity, and topic of social media posts.