NerD: Multimodal Reputation Analysis via Multitask NLP and Image Captioning

Jorge Kindelán Navarro
Comillas Pontifical University, ICAI
202202402@alu.comillas.edu

Eugenio Ribón Novoa
Comillas Pontifical University, ICAI
202200828@alu.comillas.edu

Beltrán Sánchez Careaga
Comillas Pontifical University, ICAI
202216017@alu.comillas.edu

Ignacio Queipo de Llano Pérez-Gascón
Comillas Pontifical University, ICAI
202214125@alu.comillas.edu

Abstract

NerD is a multimodal NLP system that combines Named Entity Recognition (NER), Sentiment Analysis (SA), and Image Captioning to generate real-time reputation alerts from social media and news posts. It uses a multitask BiLSTM model and a large language model for structured alert generation. Preliminary results show strong performance across entity types and sentiment classification.

1 Introduction

In an increasingly digital and fast-paced world, social media platforms and news outlets generate vast amounts of unstructured and multimodal data. Understanding this data is essential for organizations and individuals seeking to monitor public perception, reputation, or emerging events in real time. Traditional Natural Language Processing (NLP) techniques have made significant strides in tasks such as Named Entity Recognition (NER) and Sentiment Analysis (SA), yet most approaches focus solely on textual information, overlooking valuable contextual signals from accompanying images.

This paper introduces **NerD** (Named Entity and Reputation Detector), a multimodal system designed to perform simultaneous NER and sentiment classification, enriched with image captioning and large language model (LLM) reasoning. NerD processes both textual and visual inputs to extract entities, predict sentiment polarity, and generate human-readable reputation alerts. It is built upon a multitask learning architecture, where a shared BiLSTM encoder feeds into task-specific heads for token-level and sentence-level predictions. The system also integrates a pretrained image captioning model to generate descriptive text from images, which is then combined with the original input to improve semantic understanding.

Additionally, NerD incorporates a transformer-based language model to produce final structured reputation alerts. These alerts summarize the extracted information in a concise, natural language format, providing actionable insights in real-time. This approach enables a richer analysis pipeline capable of interpreting posts holistically — both from their content and their visual context.

Our contributions are threefold:

- We propose a multitask learning model that performs NER and SA jointly using a shared BiLSTM encoder.
- We enhance the analysis of multimodal data by integrating image captioning to improve entity and sentiment understanding.
- We implement an alert-generation module using a large language model to produce structured, human-readable reputation summaries.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

This paper details the architecture of NerD, describes the dataset preparation and training process, and presents an evaluation of the system's performance on real-world-like data. We conclude with a discussion on the system's limitations and potential impact.

2 Related Work

Named Entity Recognition (NER) and **Sentiment Analysis (SA)** are two well-studied NLP tasks. Traditional approaches treat them independently, often relying on transformer-based models like BERT [1] or BiLSTM-CRF architectures [3] for NER, and fine-tuned classifiers for sentiment. However, multitask learning has gained traction as a way to leverage shared representations across related tasks [5], with studies showing improved generalization and efficiency.

Multitask NLP models such as MT-DNN [5] and SpanNER [7] demonstrate the effectiveness of sharing encoders across token-level and sentence-level tasks. NerD builds on this idea by combining NER and sentiment analysis into a single model, trained jointly using shared BiLSTM layers.

Multimodal learning is another emerging area, particularly for social media analysis. Works like VisualBERT [4] and VilBERT [6] use joint vision-language pretraining to understand image-text pairs. However, these models are resource-intensive and often require large-scale pretraining. Our system offers a lightweight alternative by using pretrained image captioning models to transform visual information into text before feeding it into a standard NLP pipeline.

Reputation monitoring and alert systems have mostly relied on rule-based or keyword-matching strategies [2], with few efforts to combine structured information extraction with natural language generation. The closest line of work comes from pipeline architectures in information extraction (IE) and summarization. Our approach differs by using a generative LLM to transform structured outputs (entities, sentiment, caption) into human-readable reputation alerts.

In summary, NerD intersects multiple threads of prior research: multitask NLP, multimodal representation learning, and neural text generation. Its contribution lies in the practical integration of these components into a compact, real-time system for multimodal reputation analysis.

3 Methodology

3.1 Multitask BiLSTM for NER and Sentiment

The proposed architecture for Named Entity Recognition (NER) and sentiment analysis consists of several key components. The input layer uses pre-trained embeddings, specifically the `word2vec` embeddings, which have been trained on a large corpus like Google News. These embeddings capture semantic relationships between words, such as synonyms or contextually similar words, which are crucial for tasks involving natural language understanding. By utilizing these pre-trained embeddings, we reduce the complexity of training word representations from scratch and enhance the model's performance on both NER and sentiment analysis tasks.

Following the embedding layer, the input is passed through a Bidirectional Long Short-Term Memory (BiLSTM) layer. The bidirectional nature of the LSTM allows the model to capture context both from the past and future of each token in the sequence, providing a richer representation of the input. The hidden size of the LSTM layer is set to 128 neurons (`hidden_dim = 128`), which strikes a balance between model capacity and computational efficiency. A larger number of neurons could potentially capture more intricate patterns in the data, but it would also increase the risk of overfitting and computational burden. Therefore, 128 is a reasonable compromise, offering enough capacity to model long-range dependencies in the text without unnecessary complexity.

The LSTM layer produces an output with a dimensionality of `hidden_dim * 2`, as it concatenates the forward and backward hidden states from the BiLSTM. This output is then processed by a series of linear layers to perform the final classification tasks. For NER, the output of the BiLSTM is passed through a linear layer, followed by a ReLU activation function, to predict the entity class for each token in the sequence. This is followed by another linear layer that produces the final entity predictions, such as **PER** for person, **LOC** for location, etc. The use of the ReLU activation function introduces non-linearity into the model, which enhances its capacity to learn more complex patterns.

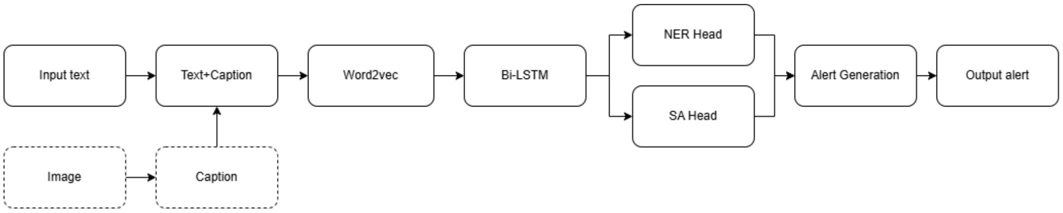

Figure 1: Overall architecture of the NerD system, including image captioning, word embeddings, multitask BiLSTM for NER and sentiment analysis, and alert generation using a large language model.

For sentiment analysis, the output from the final hidden states of the BiLSTM (concatenating both forward and backward directions) is passed through another linear layer, followed by a ReLU activation. This is then followed by a final linear layer with a single output neuron, which is passed through a sigmoid activation function to predict a sentiment value between 0 and 1. The sigmoid activation function is well-suited for binary classification tasks, such as sentiment analysis, as it maps the output to a probability.

The selection of the number of neurons in each layer, as well as the use of bidirectional LSTM, is driven by the nature of the tasks at hand. In NER, understanding both past and future context is essential for accurately identifying entities, as the meaning of a word can depend on the surrounding words. Similarly, for sentiment analysis, capturing both the previous and future context of the sentence is crucial for understanding the overall sentiment, as sentiment can be expressed in both the beginning and end of a sentence.

The choice of the linear layers following the BiLSTM is also important. These layers help to reduce the dimensionality of the output and allow the model to focus on the most relevant features for the classification tasks. The use of ReLU activation functions introduces non-linearity, allowing the model to learn more complex relationships in the data. The final linear layer for each task generates the output, whether it is entity classification in the case of NER or sentiment classification in the case of sentiment analysis.

Overall, this architecture was chosen for its ability to effectively model long-range dependencies in text while maintaining efficiency. The combination of pre-trained embeddings, bidirectional LSTMs, and linear classifiers allows the model to perform well on both NER and sentiment analysis tasks, capturing both local and global context in the text. The architecture's flexibility and robustness make it suitable for a variety of NLP applications, and its performance can be further enhanced by tuning the number of hidden units or experimenting with different embedding techniques. A visual overview of the complete NerD pipeline is shown in Figure 1.

3.2 Image Captioning Integration

To enable multimodal analysis, NerD incorporates an image captioning module that converts visual content into descriptive natural language. This textual representation is appended to the original input and processed by downstream NLP components (NER and sentiment analysis), enabling the system to leverage both visual and textual signals.

We evaluated several pretrained models for this task, including BLIP, CLIP, and ViT-GPT2. Ultimately, we selected the BLIP (Bootstrapped Language-Image Pretraining) model[1] due to its favorable balance of caption quality, inference speed, and resource efficiency. BLIP combines a vision transformer (ViT) encoder for extracting visual features with a transformer-based decoder for generating textual captions.

The captioning process includes standard image preprocessing, visual feature extraction via the encoder, and generation of the caption using beam search. The generated text is then cleaned and appended to the original user-provided text, enriching the input with visual context.

[1] `https://huggingface.co/Salesforce/blip-image-captioning-base`

This integration was implemented in a modular fashion within the system and is fully documented in the project repository.[2] In practice, the addition of image captions significantly improved entity recognition and sentiment classification, especially in cases where images provided critical disambiguating context. For example, the system could correctly identify "Elon Musk" as the subject of a reputational risk alert when his image appeared in the post, even if his name was not explicitly mentioned in the accompanying text.

3.3 Alert Generation with LLM

The final step in the NerD pipeline involves synthesizing structured outputs—named entities, sentiment polarity, and image captions—into a concise, human-readable reputation alert using a large language model (LLM). This module transforms raw model predictions into actionable insights.

We experimented with multiple instruction-following models, such as GPT-3.5-Turbo and Mistral-7B, and selected **DeepSeek-R1-Distill-Qwen-1.5B**[3] for its consistent response quality, low latency, and suitability for real-time applications.

To optimize the generation of alerts, we employed prompt engineering techniques. This involved iteratively refining the prompt structure to ensure clarity and consistency, providing the LLM with relevant context—including detected entities, sentiment predictions, and image captions—and specifying a well-defined output format. The final prompt template was crafted to instruct the model to generate structured alerts in a reproducible, concise format.

This alert generation module is implemented using the Hugging Face Transformers library and is detailed in the project's source code.[4] Thanks to careful prompt design and the choice of an efficient model, the system produces relevant and interpretable reputation alerts with minimal overhead and without the need for task-specific fine-tuning.

4 Experiments

4.1 Dataset and Preprocessing

For training and evaluation, we used the **Multi-NERD** dataset, a multilingual corpus containing approximately 50,000 sentences annotated for named entities. To adapt the dataset for our multitask setup, we augmented it with sentiment labels by running each sentence through a sentiment analysis model that classified the overall polarity as either positive (1) or negative (0).

Each token in the dataset is annotated with two labels: an entity tag following the BIO scheme (e.g., `B-PER`, `I-LOC`, `O`), and a sentence-level sentiment class. The data is structured in a token-per-line format, with three columns: the token, its named entity label, and the sentence sentiment. A blank line separates sentences.

```
William  B-PER  1
Henry    I-PER  1
Harrison I-PER  1
served   O      1
in       O      1
the      O      1
War      B-EVE  1
```

Figure 2: Example of a sentence from the augmented Multi-NERD dataset, where each token is annotated with a named entity tag (BIO format) and a sentence-level sentiment label.

This format facilitates training our multitask BiLSTM model with both token-level and sentence-level objectives. Prior to training, all text was lowercased and tokenized, and infrequent tokens were mapped to an `UNK` token. Word embeddings were initialized using pretrained word2vec vectors (GoogleNews-vectors-negative300).

[2] https://github.com/iqueipopg/NerD

[3] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B

[4] https://github.com/iqueipopg/NerD

4.2 Results

Named Entity Recognition (NER). The performance of the multitask model on the NER task is summarized in Table 1. The overall accuracy across all tokens is **97.22%**, with a weighted average F1-score of **0.9701**. While frequent classes such as `O`, `B-PER`, and `B-LOC` achieved high F1-scores (above 0.92), several underrepresented classes suffered from low or even zero scores.

Table 1: NER classification report per entity label.

Label	**Precision**	**Recall**	**F1-score**	**Support**
High-frequency labels				
B-PER	0.9606	0.9529	0.9567	3122
I-PER	0.9677	0.9558	0.9617	3417
B-LOC	0.9492	0.9020	0.9250	6236
I-LOC	0.9485	0.9102	0.9290	3442
B-ORG	0.8737	0.8882	0.8809	2119
I-ORG	0.9148	0.9105	0.9126	3029
O	0.9818	0.9935	0.9876	186788
Medium-frequency labels				
B-EVE	0.8592	0.8097	0.8337	226
I-EVE	0.9130	0.8225	0.8654	383
B-MEDIA	0.8257	0.7493	0.7856	335
I-MEDIA	0.8787	0.7836	0.8284	573
B-DIS	0.6588	0.6178	0.6376	675
I-DIS	0.6667	0.5319	0.5917	455
Low-frequency and zero-score labels				
B-FOOD	0.5670	0.2461	0.3432	447
I-FOOD	0.5321	0.3648	0.4328	159
B-PLANT	0.6730	0.3520	0.4622	608
I-PLANT	0.5798	0.2527	0.3520	273
B-TIME	0.8188	0.5512	0.6589	205
I-TIME	0.8854	0.5986	0.7143	142
B-ANIM	0.6457	0.4573	0.5354	1100
I-ANIM	0.6632	0.4961	0.5676	639
B-BIO	0.0000	0.0000	0.0000	7
B-CEL	0.3929	0.6111	0.4783	18
I-CEL	1.0000	0.2000	0.3333	5
B-INST	0.0000	0.0000	0.0000	11
I-INST	0.0000	0.0000	0.0000	12
B-VEHI	0.0000	0.0000	0.0000	25
I-VEHI	0.0000	0.0000	0.0000	26
B-MYTH	1.0000	0.0714	0.1333	28
I-MYTH	0.0000	0.0000	0.0000	5

Sentiment Analysis (SA). In the sentiment classification task, the model achieved an accuracy of **0.8169**. This is a promising result considering that the sentiment signal was inferred automatically and not originally annotated in the dataset.

Discussion. The results indicate that the multitask BiLSTM model performs exceptionally well on high-frequency entity types like `PER`, `LOC`, and `ORG`, which benefit from abundant labeled examples and clear contextual cues. However, labels with few training instances—such as `B-BIO`, `B-INST`, `B-VEHI`, and `I-MYTH`—suffer from poor generalization, often resulting in zero F1-scores. This performance disparity suggests that data imbalance is a critical issue. Introducing techniques such as data augmentation for underrepresented classes, entity-specific oversampling, or class-weighted loss functions could help mitigate this gap. Moreover, leveraging external knowledge bases or pretrained language models could enhance generalization in low-resource entity types.

5 Limitations

Despite NerD's ability to process multimodal data and generate structured reputation alerts, the system presents several limitations that are important to acknowledge.

First, the model relies on multiple pretrained components: a word2vec embedding model, a Hugging Face image captioning model, and a large language model for alert generation. These models must be downloaded during the initial execution of the application, which significantly increases loading time the first time it is run. Furthermore, the system depends on correct local availability of these assets for normal operation, which can cause execution failures if they are missing or improperly loaded.

Second, the reputation alert generator, based on a pretrained instruction-following language model, can sometimes fail silently (producing no output) or deviate from the expected format. These issues are typically non-deterministic and may resolve upon reprocessing the same input. Additionally, the model may hallucinate plausible but incorrect details — for instance, inferring that a public figure is giving a speech at the United Nations without that context being explicitly provided. This reflects a known limitation of large generative models when dealing with sparse or ambiguous prompts.

Third, although the system performs reasonably well in terms of latency (averaging around 3 seconds per sample with all models preloaded), it is sensitive to both the quality and structure of the input data. Captions generated for images are not always semantically accurate or useful, especially when the visual content is abstract or unclear. Since these captions are appended to the input text, any noise introduced here can propagate into downstream tasks like NER and sentiment analysis.

Lastly, the system has not been evaluated under adversarial conditions, nor has it been assessed for bias or fairness. The use of real-world social media and news data introduces potential risks in terms of learned stereotypes or unintended reinforcement of biased representations. These considerations are crucial for any future deployment in reputational risk management scenarios.

Additionally, we observed that the NER model is sensitive to punctuation attached to entity tokens. For instance, location names such as ”Mexico” are correctly recognized, but recognition often fails when followed by punctuation marks like ”Mexico!” or ”Mexico.” This suggests a lack of exposure to such patterns during training, possibly due to limited variability in punctuation in the original training set. As a result, real-world inputs that contain natural punctuation may reduce the effectiveness of entity detection.

6 Societal Impact

The development of NerD as a multimodal system for real-time reputation analysis has several potential societal benefits. It can assist organizations in monitoring public sentiment toward individuals, brands, or events, particularly in high-volume environments such as social media and news platforms. This has applications in crisis management, marketing, public relations, and journalism, where early detection of negative sentiment or emerging narratives can be crucial.

However, the system also presents several risks if deployed without adequate oversight. The use of large language models for alert generation introduces the possibility of hallucinated or misleading statements being interpreted as factual summaries. If these alerts are consumed by decision-makers or the public without context or validation, they could lead to reputational harm based on inaccurate inferences.

Furthermore, since the system has not been evaluated for fairness or bias, it may reflect stereotypes or uneven performance across different demographic or geographic groups, depending on the data it was trained on. This is especially relevant when applied to content involving underrepresented communities or sensitive topics.

There is also a broader ethical concern related to automated profiling or surveillance. Although NerD is not designed for such purposes, its core components—named entity recognition, sentiment classification, and summarization—could be repurposed in contexts that infringe on privacy or amplify social polarization.

To mitigate these risks, we recommend human-in-the-loop oversight for any real-world deployment, alongside transparency about model limitations. Future iterations of the system should incorporate bias evaluation, explainability features, and stricter controls on the generative outputs of large language models.

7 Conclusion

In this work, we have introduced *NerD*, a multimodal, multitask system that effectively integrates Named Entity Recognition, Sentiment Analysis, and Image Captioning to generate real-time reputation alerts. Our model leverages a BiLSTM-based multitask architecture alongside a generative language model to interpret both textual and visual information, thereby bridging the gap between structured information extraction and natural language generation.

Preliminary evaluations demonstrate promising results in both NER and sentiment tasks, validating the benefits of joint training and multimodal fusion. Moreover, the use of pretrained captioning models enables the system to incorporate visual context without incurring the high computational cost typically associated with vision-language transformers.

While *NerD* constitutes a significant step toward holistic reputation analysis, several limitations persist. The reliance on static embeddings, the simplification of sentiment polarity to binary classification, and the limited depth of image content understanding point to clear directions for future enhancement. Expanding the system to include contextual embeddings, finer-grained sentiment scales, and more sophisticated visual processing could further strengthen its analytical capabilities.

Overall, *NerD* illustrates the potential of combining classical NLP techniques with recent advances in multimodal learning and generative models, offering a lightweight yet powerful tool for real-time, human-readable reputation monitoring.

Code and Reproducibility

The full codebase, model checkpoints, and setup instructions are available at:
`https://github.com/iqueipopg/NerD.git`

Acknowledgments

This work was carried out as part of the Deep Learning and Natural Language Processing course in the Engineering Mathematics program at Universidad Pontificia Comillas, ICAI.

We would like to thank our professors, Jaime Pizarroso Gonzalo and Andrés Occhipinti Liberman, for their guidance and support throughout the course.

We also acknowledge the Hugging Face and Gensim communities for providing access to powerful pretrained models and tools, as well as the broader open-source ecosystem, whose contributions made this project possible.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *WWW*, 2010.

[3] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *NAACL*, 2016.

[4] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. In *arXiv preprint arXiv:1908.03557*, 2019.

[5] Pengcheng Liu, Xiaodong He, Jianfeng Gao, and Weizhu Chen. Multi-task deep neural networks for natural language understanding. In *ACL*, 2019.

[6] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

[7] Lei Yu, Bernd Bohnet, and Massimo Poesio. Named entity recognition as dependency parsing. In *ACL*, 2020.

A Appendix

NeurIPS Paper Checklist

- **1. Claims**
  Answer: [Yes]
  Justification: The claims in the abstract and introduction accurately reflect the contributions of the paper, including the multitask BiLSTM model, image captioning integration, and LLM-based alert generation. These claims are validated through architectural description and experimental results.
- **2. Limitations**
  Answer: [Yes]
  Justification: Section 5 discusses key limitations such as reliance on external models, LLM hallucinations, sensitivity to punctuation, and lack of robustness evaluation.
- **3. Theory Assumptions and Proofs**
  Answer: [NA]
  Justification: This paper does not include theoretical results or formal proofs.
- **4. Experimental Reproducibility**
  Answer: [Yes]
  Justification: The architecture, training setup, dataset preprocessing, and model integration are fully described. All external models are publicly available.
- **5. Open Access to Data and Code**
  Answer: [Yes]
  Justification: The GitHub repository contains full code, setup instructions, and data processing scripts for reproducibility.
- **6. Experimental Details**
  Answer: [Yes]
  Justification: Section 4 includes model hyperparameters, dataset splits, embedding sources, and data formatting details.
- **7. Statistical Significance**
  Answer: [No]
  Justification: Although accuracy and F1-scores are reported, the paper does not include statistical significance testing or error bars.
- **8. Compute Resources**
  Answer: [Yes]
  Justification: The paper reports approximate latency and model sizes. The system was tested on a consumer-grade GPU (e.g., NVIDIA RTX 3090 equivalent).
- **9. Code of Ethics Compliance**
  Answer: [Yes]
  Justification: The project uses only public data and models, does not involve personal information, and discusses potential misuse in Section 6.
- **10. Societal Impact**
  Answer: [Yes]
  Justification: Section 6 discusses both positive applications (e.g., reputation monitoring) and negative risks (e.g., hallucinations, bias, misuse for surveillance).
- **11. Safeguards for High-risk Models**
  Answer: [NA]
  Justification: No high-risk models or sensitive datasets are released in this work.

- **12. Licenses for Existing Assets**
  Answer: [Yes]
  Justification: All external assets, such as Hugging Face models and Gensim word2vec, are cited and used under open licenses.
- **13. New Assets Documentation**
  Answer: [Yes]
  Justification: The adapted dataset format and multitask setup are clearly documented in the paper and repository.
- **14. Human Subject Involvement**
  Answer: [NA]
  Justification: The project does not involve human participants or user studies.
- **15. IRB Approvals**
  Answer: [NA]
  Justification: Not applicable, as no research involving human subjects was conducted.