Named Entity Recognition in U.S. Politics: a comparative analysis of Obama, Trump and Biden administration Press Releases

David Branes, Valeria Cerciello, Francesco Manzionna

November 15, 2024

Abstract

Named Entity Recognition (NER) is a powerful tool within Natural Language Processing (NLP) that allows for the extraction of valuable information from unstructured text.
This project utilizes NER to analyze U.S. government press releases from the Obama, Trump, and Biden administrations, aiming to uncover trends and shifts in political communication. By applying pre-trained NER models, such as DistilBERT and Flair, to press conference transcripts, we extracted key information regarding notable individuals, organizations, locations, and events central to each administration's domestic and foreign policies. The dataset, sourced from the Office of the Secretary of State, includes press releases structured to capture consistent metadata across different administrations. Our preprocessing involved tokenization, punctuation removal, and text structuring to optimize the data for NER tasks.
Evaluation metrics, including precision, recall, and F1-score, revealed that each model performed differently across the datasets, reflecting administration-specific focuses. Visualization of entity distributions highlighted the most frequently mentioned entities and their classification by type, offering insights into the political priorities emphasized by each administration. Future extensions of this work could include expanding the taxonomy of entities, comparative studies across international governmental datasets, and fine-tuning models for domain-specific language in political texts. This project contributes structured insights into U.S. political discourse, supporting more data-driven perspectives on recent historical events and figures.

The code relative to this project can be found in the following GitHub repository: https://github.com/valeriacerciello/Essentials_of_NLP.git

1 Introduction

In the vast landscape of political communication, textual data from government sources offers valuable insights into evolving global priorities and domestic policies. By analyzing these records, we can track historical trends and support leaders in making data-informed decisions that reflect shifts in governance. Our project applies Named Entity Recognition (NER), a key area in Natural Language Processing (NLP), to U.S. government press conference transcripts spanning three pivotal administrations: Obama, Trump, and Biden. Using advanced NER models like DistilBERT and Flair, this work extracts critical information about key figures, organizations, locations, and events that define U.S. domestic and foreign policies.
This project draws on multiple areas of linguistics, especially syntax, semantics, and pragmatics, to capture the meaning embedded in complex political discourse. Syntax is essential as the models parse formal sentence structures commonly found in press releases. Semantics, particularly the study of named entities and their categories (e.g., person, organization, location), underpins the NER task. Pragmatics also plays a role, as understanding the situational context (e.g., references to diplomatic or political entities) is vital

to accurate extraction.
Through NER, this project transforms unstructured press release texts into structured data, allowing us to observe patterns and uncover trends that map out each administration's priorities and policies. The structured data produced by NER enables further quantitative analysis, represented through tables and visualizations, to track key issues in areas such as foreign policy and domestic affairs.
Positioned at the intersection of NLP and Information Extraction (IE), this work contributes to the broader field by offering structured insights into unstructured political discourse. By focusing on linguistic and machine learning aspects, this project provides a unique perspective on the dynamics of U.S. politics, enriching our understanding of the global events and figures that have defined recent history.

2 Data set

We chose a collection of press releases from the Office of the Secretary of State spanning three U.S. administrations: Obama, Trump, and Biden. These datasets consist of statements and briefings delivered by the Secretary of State or other representatives, covering various topics of both domestic and international significance. The datasets were sourced from Kaggle [1] and provide an invaluable resource for tracking the evolution of political communication across administrations.
The dataset was constructed by scraping and collecting press releases from official U.S. government websites. Each dataset is structured with key metadata fields such as the title of the release, publication date, author, and the full text of the press release. Despite slight inconsistencies due to differing content schemas from archived websites, the datasets offer rich and structured textual data.
Each dataset contains several important fields, including:

- Text of the Press Release: The full body of the press release, split into paragraphs and lines.
- Title: The title of the press release.
- Publish Date: The date the press release was issued, formatted as MMMM dd, yyyy.
- Link: A URL directing to the original document.
- Document Type: The type of press release (e.g., media note, briefing, or statement).
- Document Author: The entity or person responsible for the release.
- Document Author Name: The name of the author(s) where available.
- Document Author Title: Honorifics or roles of the author (e.g., Secretary of State, Spokesperson).
- Tags: Although the Obama dataset lacks detailed tags, the Trump and Biden datasets include tags with both names and links, facilitating taxonomy-building and future analyses.

We selected this dataset because it offers a unique opportunity to apply Named Entity Recognition (NER) to analyze political language across multiple U.S. administrations. The formal nature of these press releases makes the dataset ideal for structured extraction tasks, such as identifying people, organizations, locations, and key events. Analyzing the

[1] https://www.kaggle.com/datasets/speckledpingu/secretary-of-state-press-releases-obama, https://www.kaggle.com/datasets/speckledpingu/secretary-of-state-press-releases-trump, https://www.kaggle.com/datasets/speckledpingu/secretary-of-state-press-releases-biden

language used by the Secretary of State allows for insights into shifting political priorities, governance strategies, and international relations over time. Furthermore, the consistent structure of the dataset across different administrations ensures that our analysis can focus on language patterns without the noise of unstructured or informal data formats.
While other datasets were considered—such as news media transcripts, congressional records, or social media posts—they were not as well-suited to our goals. News transcripts, for instance, may reflect biases or editorial decisions from media outlets, whereas social media posts, although more direct, tend to lack the formal structure required for large-scale entity extraction. The Secretary of State press releases, by contrast, provide formal, government-issued statements that are consistently structured, making them more reliable and conducive to tasks like NER.

3 Preprocessing

To prepare the dataset for NER, we applied several auxiliary NLP tasks, including text tokenization and punctuation removal. These preprocessing tasks were essential to ensure the data was properly structured and clean, optimizing it for the NER model.
We performed tokenization to split the text of each press release into individual words and phrases. Tokenization is critical because it allows the NER model to accurately identify entities like people, organizations, and locations while preserving the capitalization of named entities. Lowercasing non-entity words reduces variability in the dataset, helping to focus on the relevant terms for entity recognition. This task was implemented using the SpaCy library[2], which provides robust and flexible tokenization tools well-suited for NER tasks.
Additionally, we removed all punctuation marks, which are unnecessary for the entity recognition task. This step ensured that the model concentrated on the content of the text without being distracted by punctuation. For instance, in the raw text, sentences like "The U.S. Department of State, the Fulbright Foreign Scholarship Board, and the National Archives and Records Administration today formalized a new partnership to establish the first-ever Fulbright-National Archives Heritage Science Fellowship." were transformed into "The U.S. Department of State the Fulbright Foreign Scholarship Board and the National Archives and Records Administration today formalized a new partnership to establish the first ever Fulbright National Archives Heritage Science Fellowship" making the input cleaner for the NER model.
Finally, to handle the fragmented format of the raw dataset (where text was stored as a list of paragraphs), we combined all text for each press release into a single continuous string. This ensured that the press releases were processed as coherent documents. After applying these steps, the preprocessed text was saved into structured JSON files, a format compatible with various NER models due to its ability to handle structured data fields such as text and metadata.
These preprocessing choices were made to enhance the model's performance, as cleaner input data leads to more accurate and reliable entity recognition results. While other tools could have been used for preprocessing, SpaCy was selected for its efficient tokenization and its seamless integration with downstream tasks like NER in NLP models such as Flair and BERT.

4 Models

In our NER analysis, we used two pre-trained models to extract entities: DistilBERT, which is a transformer-based model, and Flair, which utilizes a different approach based on bidirectional LSTM networks combined with contextual embeddings. We selected these

[2] https://spacy.io/

models due to their complementary strengths in NER tasks and ability to handle large-scale textual data efficiently. DistilBERT, a distilled version of BERT, was chosen for its balance between accuracy and computational efficiency, making it suitable for high-throughput processing without sacrificing much performance Sanh et al. (2019). Flair was chosen for its contextual embeddings, which are effective in capturing nuanced entity relationships, especially in complex political texts like press releases.

To perform NER with Flair, several additional steps were necessary. Initially, we found that processing the entire dataset would require approximately 30,000 seconds per dataset, totaling around 25 hours. We attempted to use cloud-based computation tools, such as Colab, Kaggle Notebook, and Amazon SageMaker Studio Lab, but these significantly worsened performance compared to running NER with Flair locally.

To optimize the process, we performed sampling of the press releases, limiting the data to 100 releases per president, and set a seed value to ensure the reproducibility of the sample. After tokenizing, we applied NER with Flair to this sample, obtaining a list of classified entities for each speech.

For the distilBERT work we limited the dataset to a manageable sample size of 7,000 records per administration. This sampling allowed us to perform NER while keeping processing times reasonable.

Regarding hyperparameters, while the models we used were fine-tuned on the CoNLL-03 dataset and came with default settings, we didn't explicitly modify or tune additional hyperparameters within the model's internal architecture for this analysis. This decision helped maintain a streamlined approach, as the pre-trained models' hyperparameters were already optimized for entity recognition tasks.

This task does not involve sequence-to-sequence modeling. Unlike seq2seq models, which generate variable-length outputs, NER produces a set of labels that directly correspond to each input token. This characteristic is ideal for entity recognition, as the output structure and length directly match the input text, ensuring that each token receives exactly one label or none if it is not an entity.

Lastly, while these models utilize internal embeddings for contextual understanding, we did not apply external embeddings in this analysis. At a high level, the entity labels themselves act as abstractions or "embeddings" of the entities, capturing essential information about each recognized name, place, or organization.

5 Evaluation

To evaluate the performance of our Named Entity Recognition (NER) models, we employed both quantitative metrics and qualitative analysis. We used precision, recall, and F1-score to assess the effectiveness of entity recognition, calculating these metrics by comparing the models' output against manually labeled samples. These metrics were selected due to their ability to capture different aspects of the model's performance: precision reflects the accuracy of the entities identified, recall indicates the completeness of the entities extracted, and F1-score provides a balanced measure of both precision and recall. By choosing these metrics, we aimed to ensure a robust evaluation that accounts for both over- and under-identification of entities.

For Flair evaluation, we performed manual labeling on 10 randomly selected press releases. In this process, entities were manually annotated and then compared with the labels generated automatically by the Flair models. It's worth noting that manual labeling was conducted on press releases already processed with Flair to maintain consistency across the manually annotated data and the automatically identified entities. This approach allowed us to accurately assess model performance by providing a reliable, direct comparison between manual and automatically generated labels.

The main outcomes of this evaluation revealed that our models performed differently across

the datasets, with each president's press releases reflecting varying topical focuses and entity types. FLAIR achieved precision and recall scores that ranged around 56-70%, and Flair delivered comparable performance, with variations depending on the entity types and their frequency in each administration's texts. These results underscore the models' capacity to capture key entities in structured political text, although with some limitations in precision and recall due to complex language and domain-specific terms.

To gain a deeper understanding, we analyzed specific entity distributions using graphical visualizations. The Top 10 Specific Entities Mentioned by Obama, Trump, and Biden chart illustrates the most frequently recognized entities within each administration, offering insights into the primary topics and actors emphasized by each president. For instance, Obama's press releases frequently mention "Clinton" and "Kerry," reflecting key figures in his administration, while Trump's releases highlight figures like "Pompeo" and organizations such as "UN" and "ISIS," which align with his administration's focus on foreign policy and international security. Biden's top entities include "Blinken" and locations like "Ukraine" and "Russia," which correspond to ongoing geopolitical priorities during his administration.

The Entity Type Classification for Obama, Trump, and Biden graph provides a breakdown of the types of entities identified (Persons, Organizations, and Locations) for each administration. This distribution is indicative of the prominent role that locations and organizations play in political press releases, reflecting the focus on international relations and institutional actions in governmental communications. Each administration's specific focus on certain types of entities further highlights the political priorities and communication styles of their terms.

Overall, this evaluation reveals that our combined use of DistilBERT and Flair models captures critical entity information from U.S. government press releases, effectively highlighting political priorities and key figures across different administrations. The precision-recall performance confirms the models' utility for structured entity extraction, while the entity analysis charts provide valuable insight into the unique linguistic patterns and focal points in each administration's public discourse.

Dataset	**Model**	**Precision**	**Recall**	**F1 Score**
Obama	FLAIR	0.3412	0.5088	0.4085
Obama	distilBERT	0.3507	0.4123	0.3790
Trump	FLAIR	0.5733	0.6615	0.6143
Trump	distilBERT	0.4214	0.3026	0.3522
Biden	FLAIR	0.5618	0.6993	0.6231
Biden	distilBERT	0.4435	0.3566	0.3953

Table 1: Precision, Recall, and F1 Scores for Each Model and Dataset

6 Future work

In the future, this work could be extended by scaling the analysis across a broader range of governmental texts and applying NER models to other historical datasets beyond the U.S. Secretaries of State press releases. By incorporating data from different government bodies or international organizations, a comparative study could reveal global trends in political communication over time.

Another potential extension could involve expanding the taxonomy of entities to include more granular subcategories or new entity types relevant to political analysis, such as policies, treaties, or legislative actions.

Finally, evaluating the performance of NER models on multilingual datasets could be explored, allowing for a broader, more inclusive study of international diplomacy and

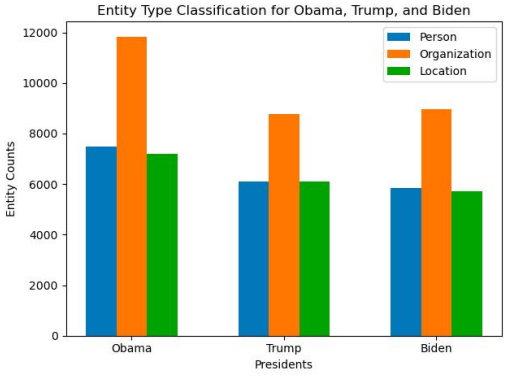

Figure 1: Entity type classification using DistilBERT

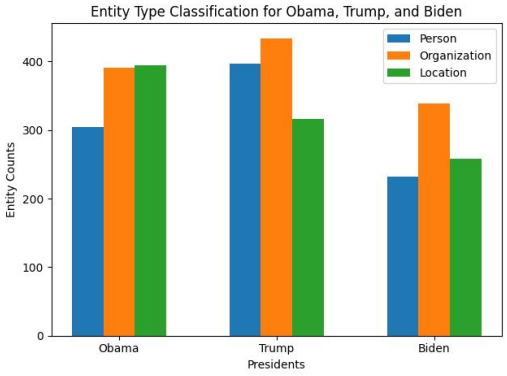

Figure 2: Entity type classification using Flair

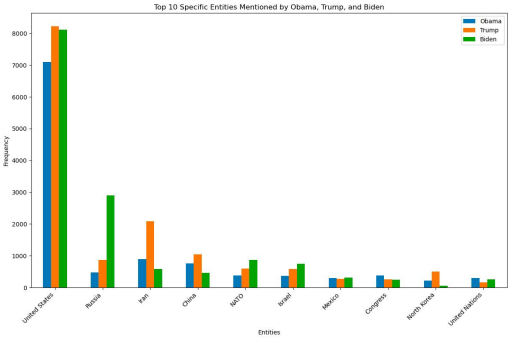

Figure 3: Top 10 entities identified by DistilBERT

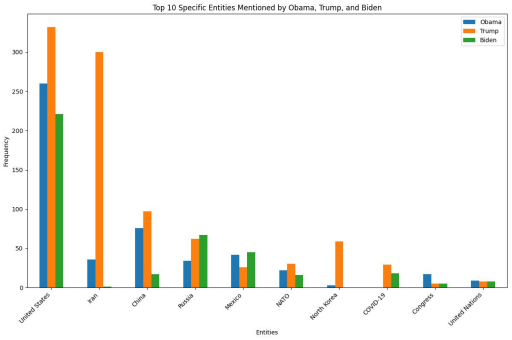

Figure 4: Top 10 entities identified by Flair

governance. Combining these efforts with deeper fine-tuning of models could also improve their ability to capture more complex or domain-specific named entities, leading to richer and more accurate analysis.

References

Abadeer, M. (2020). Assessment of distilbert performance on named entity recognition task for the detection of protected health information and medical concepts. In *Proceedings of the 3rd clinical natural language processing workshop*, pages 158–167.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59.

Elov, B. (2022). Making use of a 'spacy' module in the natural language processing. *Journal of Science and Innovative Development*, 5:41–54.

Greenwade, G. D. (1993). The Comprehensive Tex Archive Network (CTAN). *TUGBoat*, 14(3):342–351.

Magajna, T. (2022). *Natural Language Processing with Flair: a practical guide to understanding and solving NLP problems with Flair*. Packt Publishing Ltd.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Pakhale, K. (2023). Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges. *arXiv preprint arXiv:2309.14084*.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Vasiliev, Y. (2020). *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press.