ElectionBot: A Bilingual RAG-Based Question Answering System

Stanford CS224V Final Project

Anusha Kuppahally, Malavi Ravindran
Department of Statistics
Stanford University
akupp@stanford.edu, mr328@stanford.edu

Abstract

Our paper presents ElectionBot, a retrieval-augmented generation (RAG) question-answering (QA) system to analyze and compare English and Spanish language perspectives on the 2020 U.S. Presidential Election, using a corpus of 237,185 relevant articles. Leveraging Sentence Transformers for bilingual text embedding and Facebook AI Similarity Search (FAISS) for efficient retrieval, our system provides structured and concise responses to user queries regarding the election. Both LLM-based and human evaluation were performed to assess response quality across the following metrics: helpfulness, relevance, accuracy, depth, creativity, and level of detail. While the system concisely and precisely summarizes and compares viewpoints present in the articles, we note several areas for future improvement, such as including more diverse articles and conducting more rigorous human evaluation. Future work could extend the system to the 2024 U.S. Presidential Election, incorporate additional languages, and analyze results across sources to consider bias and reliability.

1 Introduction

As events occur around the world, it can be difficult to grasp what is being reported and understand the nuances and biases present. Especially when considering news in multiple languages, there can be diverse perspectives that enhance understanding of events. This type of news analysis has many potential use cases, including predicting financial markets, learning about the political climate of different countries, and monitoring global conflicts.

This project aims to focus on a political use case and determine if large language models (LLMs) can analyze news articles efficiently and output concise, thoughtful analysis about varying viewpoints. This can help with information flow and encourage users to learn more about reporting around the world. We will focus on news articles about the 2020 U.S. Presidential Election in both English and Spanish to determine if it is possible to obtain cross-lingual insights and understand the similarities and differences in how the election was reported in different languages.

To accomplish this, we implemented a question-answer (QA) system where the user is able to ask for any information related to the election, and the system can answer from the news corpus. This includes a natural language interface so the user can ask a question and the system analyzes the data to obtain an answer.

2 Related Work

Tarekegn uses LLMs and clustering analysis to perform event detection within news sources [4]. The author uses LLMs for pre-event detection, which includes keyword extraction and text embedding, and post-event detection, which includes summarization, cluster labeling, and topic identification.

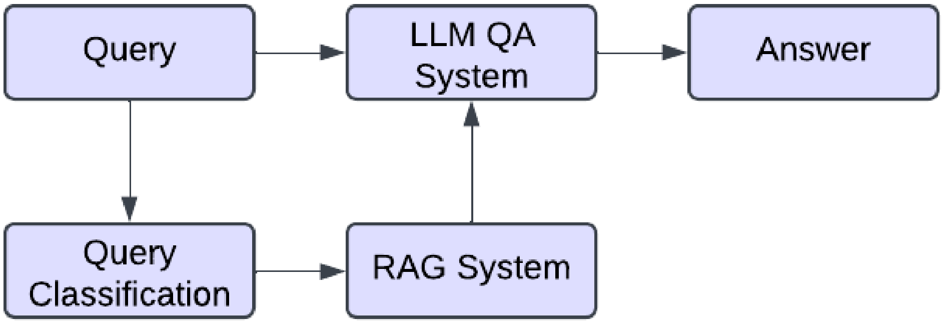


Figure 1: Project Architecture

To evaluate results, the paper utilizes stability-based evaluation, which assesses the clustering on training and test data. Stability is measured as the distance between features of the training and test data within each cluster. This was a useful paper to read to understand how LLMs can extract and identify important information in news sources.

Sainz et. al provide guidelines for LLMs to be able to perform zero-shot information extraction [3]. This paper highlights guidelines that GoLLIE follows, including an input-output representation, guidelines enhanced representation, and training regularization. The model can perform the following 5 tasks: named entity recognition, relation extraction, event extraction, event argument extraction, and slot filling. Results indicate higher performance for zero-shot evaluation with these annotation guidelines on almost all datasets. The techniques used in this paper were useful to understand how guidelines for the LLM can influence performance.

Jain et. al fine tune an off-the-shelf LLM with QA pairs from broadcast television news, which is similar to this project's goal of tuning an LLM for news-related questions and answers **?**. This paper also uses multilingual data to fine tune and uses a RAG pipeline to give contextualized answers. While fine-tuning can improve LLM performance, it may not fix issues related to reliability and accuracy. Our project utilizes some of the techniques in this paper, including using a RAG system, using cosine similarity for the retrieved context, and analyzing results by topics. Comparing broadcast news against web articles may be an interesting extension to this project.

3 Approach

The methodology of this project utilizes RAG-based question-answering and incorporates some of the tools and techniques discovered in the related works. 1 details our system architecture. Below we discuss the key components of our approach.

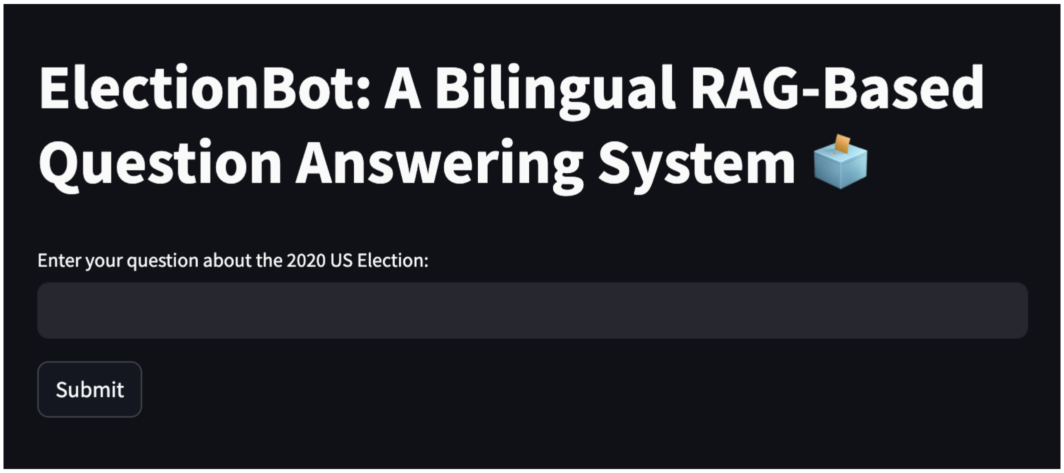
Article Embedding and Indexing The first step in our RAG system was to embed various news articles related to the 2020 election using the Sentence Transformers *paraphrase-multilingual-mpnet-base-v2*, which has multilingual capabilities appropriate for embedding both English and Spanish articles. We only generated embeddings for articles that fit the following restrictions:

1. The article was published between 11/03/2020 and 11/10/2020, which was the week following the 2020 U.S. Presidential election.

2. The article was written in either English or Spanish.

3. The article contains one or more election-specific keywords. English keywords include (but are not limited to) "election", "vote", and "presidential", and their Spanish counterparts were "elección", "voto", and "presidencial".

Embeddings were indexed using Facebook AI Similarity Search (FAISS), for efficient retrieval in high dimensional spaces.


ElectionBot: A Bilingual RAG-Based Question Answering System 🗳

Enter your question about the 2020 US Election:

Figure 2: Streamlit UI

Query Classification After the articles were embedded, the next step was to take in a query as input from the user. As ElectionBot is meant solely to answer questions about the 2020 U.S. Presidential Election, we employed a classification step to determine whether the input was in fact relevant to the election. For follow-up queries that required context to determine relevancy, the classification system considered the chat history, specifically the last two exchanges, in order to make a more appropriate determination. Classification was performed with Together AI's *Llama-3.1-Menotron-70B-Instruct-HF*. Queries that were flagged as irrelevant were not answered by the system, and the user would instead see a message indicating that the question could not be answered.

Article Retrieval After determining if the input query could indeed be answered, the query was embedded using the same Sentence Transformers textifparaphrase-multilingual-mpnet-base-v2 as the news articles. This query embedding was then compared to the FAISS-indexed article embeddings using cosine similarity. The two most relevant articles per language were retrieved.

Answer Generation Once the appropriate articles were retrieved for each language, the queries, along with the retrieved context and prompt, were inputted to Together AI's *Llama-3.1-Menotron-70B-Instruct-HF* in order to generate structured responses that followed the requirements detailed in the prompt. The architecture also supports follow-up questions, and uses at most two previous exchanges in the context window to maintain fluidity in the conversation.

User Interface The user is able to interact with the system using the Streamlit UI, where they can submit questions, view responses, and ask follow ups 2.

4 Experiments

4.1 Data

We used a dataset of news articles collected from 11/3/20 - 11/10/20. This date range was used to capture news during the aftermath of the 2020 Presidential Election. Before filtering, there were 851,503 articles. After applying filtering related to date, language, and election-related keywords, there were 237,185 articles, with 173,963 in English and 63,222 in Spanish.

4.2 Evaluation method

This project employs both human and LLM based evaluation. 15 question-answer pairs were rated on a scale of 1-5 on the following metrics: helpfulness, relevance, accuracy, depth, creativity, and level of detail. Questions were grouped into the following categories:

Election Candidates and Campaigns

1. Who were the candidates in the 2020 US election?

2. What were the major campaign promises of Joe Biden?
3. How was climate change addressed in the campaigns?
4. What stances did the candidates take on immigration?
5. How did the election discussions address systemic racism?

Election Outcomes and Their Implications

1. What were the key critiques of Donald Trump's campaign in English and Spanish articles?
2. What was the coverage of voter fraud or irregularities in the 2020 election?
3. What were the reactions to Biden's victory in English vs. Spanish media?
4. What was the reaction to Donald Trump's refusal to concede?
5. How does Biden's victory in the 2020 presidential election influence global trade?

Voter Engagement and Demographics

1. How did the candidates engage with Latino voters?
2. How was the response to the pandemic discussed during the election?
3. What were the pivotal swing states in the 2020 election?
4. How was voter turnout during the 2020 election described in English vs. Spanish media?
5. What were the main concerns of Black voters in the 2020 election?

4.3 Experimental Details

Sentence Transformers *Paraphrase-Multilingual-Mpnet-Base-v2* was used to embed the articles and queries with batch size = 8 [2]. This is a semantic similarity model that can find similar sentences across languages, including English and Spanish.

Llama-3.1-Nemotron-70B-Instruct-HF was used for the QA agent, query classifier, and LLM-based evaluation. For the QA agent, we used max tokens = 512, temperature = 0.7, top p = 0.7, frequency penalty = 0.1, and presence penalty = 0.1. These parameters were chosen to slightly discourage repetition, avoid completely deterministic responses, and have relatively creative responses. The prompt was as follows:

You are an expert analyst of the 2020 US Presidential Election. Based on these sources: {context_text}. Answer this question: {query}. Important:

- *Use both English and Spanish sources to provide a complete perspective.*
- *Keep Spanish article titles in their original language.*
- *Only include information from the provided sources.*
- *Include in-text citations when relevant.*
- *Keep responses less than the max number of tokens (512).*
- *If you cannot answer based on these sources, say so clearly.*

For the classifier that was used to determine query relevance, max tokens = 10 and temperature = 0.1 were selected to encourage the model response to be more deterministic. Here, the prompt was:

Given the following conversation about the 2020 US Presidential Election, determine if the new query is a relevant follow-up question. Previous Conversation: {context_text}. New Query: {query}. Is this query relevant to the ongoing conversation about the 2020 US Presidential Election? Answer with 'YES' or 'NO'. Consider:

- *Is it a follow-up to the previous discussion?*
- *Does it ask for clarification about previously discussed election topics?*
- *Is it related to any aspect mentioned in the previous answers?". If the query is deemed to not be relevant, the following is outputted: "I apologize, but I can only answer questions related to the 2020 US Presidential Election".*

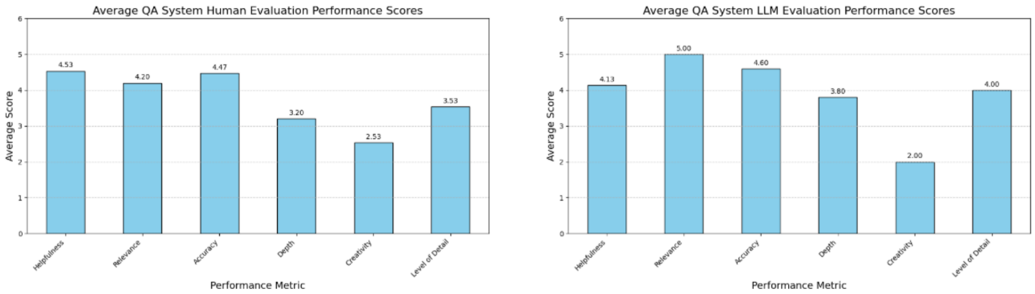


Figure 3: Overall Human and LLM Evaluation

The same parameters were used when using the model for evaluation. The prompt for evaluation was: *Rate the following Q&A interaction about the 2020 US Presidential Election on a scale of 1-5: Question: {query}. Answer: {answer}. Please rate each category and provide a brief explanation:*

- *Helpfulness: [1-5]*
- *Relevance: [1-5]*
- *Accuracy: [1-5]*
- *Depth: [1-5]*
- *Creativity: [1-5]*
- *Level of Detail: [1-5]*

4.4 Results

Within the RAG system, we found that switching from Euclidean distance to cosine similarity yielded answers that referenced articles more relevant to the query. This is supported by the work done by Jain et. al [1].

We performed prompt engineering to ensure that the output was formatted as desired. At first, we had added more instructions to the QA system prompt, but noticed that answers contained the instructions and often contained information unrelated to the question. After making the prompt more simple and tweaking the parameter values, we noticed answers were more concise, grounded, and did not contain irrelevant information.

To evaluate the model, we used 15 questions about the 2020 Presidential Election. The LLM evaluated each QA pair according to the aforementioned metrics, and we performed human evaluation by manually going through the results.

5 Analysis

Overall, the LLM based evaluation and human evaluation had similar results 3. The figures indicate that the LLM performed worse on average for questions about Election Candidates and Campaigns, and best for questions about Voter Engagement and Demographics 4 5. The responses had low creativity on average, but this is preferable as our QA system is meant to be concise and factual.

In addition to understanding how our system performed across the 6 aforementioned metrics, we also analyzed the similarities and difference in how our system represented the English and Spanish perspectives on the election. Broadly, we note that the English perspective is slightly more comprehensive, and cites empirical data more often than the Spanish. Spanish perspectives appear to be rooted more largely in cultural themes and narrative. While both provide insights into the topics that were queried, the English sources focused on policy, historical context, and politics, while Spanish sources provide more of a cultural and global lens. Queries specifically addressing the Latino population and voter turnout resulted in more specific responses from the Spanish sources than other topics, which is understandable as these topics will be more prevalent in Spanish media. Overall, there was not a huge discrepancy between the content of English and Spanish perspectives, although tonal and stylistic

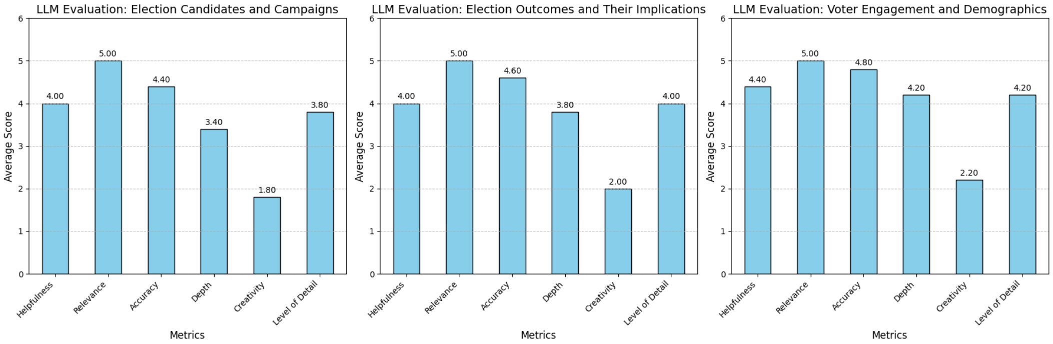


Figure 4: LLM Evaluation by Group

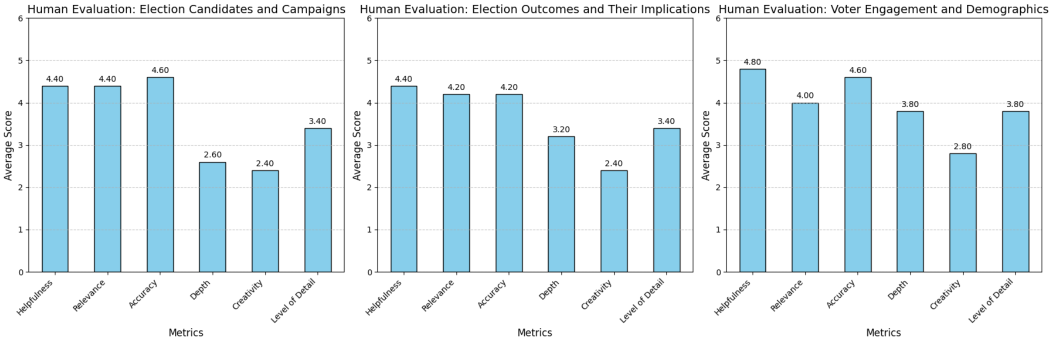


Figure 5: Human Evaluation by Group

deviations were present. This may suggest a need for more comprehensive Spanish-language articles to offer a more distinct and contrasting perspective compared to the English ones.

6 Conclusion

Through the implementation of a RAG-based QA system leveraging Sentence Transformers for embedding and Facebook AI Similarity Search (FAISS) for retrieval, ElectionBot was able to concisely and precisely summarize viewpoints present in the articles and represent both English and Spanish perspectives on the U.S. 2020 Presidential Election. Responses were relevant, accurate, and detailed, as measured through both human and LLM based evaluation.

We note certain limitations of our current system. Firstly, as certain questions are more factual and straight-forward (i.e. "Who were the candidates"?), we may consider including an additional classification layer to determine whether both language perspectives are necessary for a response. Furthermore, we would likely benefit from including more articles, especially in Spanish, to capture more diversity in perspective. We also conducted human evaluation with only two evaluators (ourselves), and would benefit from more evaluators as well as additional questions and topics for testing. Future work on this project could focus on extending the system's capabilities to include the 2024 U.S. Presidential Election. We can further incorporate additional languages, transforming a bilingual system into something that could reach global audiences. Other avenues for extensions include analyzing results across different news sources, considering both bias and reliability, and more robust fact checking and bias assessment.

References

[1] Tarun Jain et al. *News Reporter: A Multi-lingual LLM Framework for Broadcast T.V News*. 2024. arXiv: `2410.07520 [cs.CL]`. URL: `https://arxiv.org/abs/2410.07520`.

[2] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. URL: `http://arxiv.org/abs/1908.10084`.

[3] Oscar Sainz et al. *GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction*. 2024. arXiv: `2310.03668 [cs.CL]`. URL: `https://arxiv.org/abs/2310.03668`.

[4] Adane Nega Tarekegn. *Large Language Model Enhanced Clustering for News Event Detection*. 2024. arXiv: `2406.10552 [cs.CL]`. URL: `https://arxiv.org/abs/2406.10552`.