Prompt Engineering for Spanish Sexism Detection

Sexist content has become increasingly prevalent across the internet, including on platforms like X (formerly Twitter), which has 436 million monthly active users. Sexism is defined as discrimination or prejudice based on gender and disproportionately affects women. To combat this, using artificial intelligence (AI) to automatically identify and address sexism in tweets has been proposed as a way to combat this form of violence against women. With the rise of large language models (LLMs), prompt engineering has become crucial to take advantage of their capabilities. It involves designing prompts, that is, specific instructions, to guide models toward desired outputs. Effective prompts significantly impact response quality. In this study, we investigate the impact of using two LLMs with Spanish language support and different prompt formulations on the performance of sexism detection in tweets.

Our study used the EXIST 2024 dataset, which contains a variety of sexist and non-sexist tweets in both English and Spanish. We focused our analysis on the 1159 Spanish-language tweets classified as either "Sexist" and "Non-sexist" where all annotators were in agreement. (493 categorized as "Sexist" and 666 as "Non-sexist"). We selected two Spanish-compatible models supported by LangChain: GPT-4o and Mistral 7B. Both models were evaluated using four system prompts, which initialize the interaction by providing essential definitions and instructions, and two user prompts, specifically designed to identify sexism. These prompts are presented in Table 1.

Table 1: Prompts Design

System prompt	SP1.Consider the following definition of sexism: Sexism is the discrimination of people based on gender. SP2.You are a classification model with a very good performance in classifying whether a tweet is sexist or not. SP3.You are an expert in sexism and you know how to classify if a tweet is sexism or not. SP4.Take into consideration the following examples that are classified as sexist: {examples}
User prompt	UP1.Is this tweet sexism? Answer with Yes or No: {tweet} UP2.Does this tweet contain explicit or implicit sexist language? Answer with Yes o No: {tweet}

Performance metrics, including precision, recall, and f1-score, for each system and user prompt combination are presented in Table 2. Our results show that GPT-4o delivers more consistent outcomes across several combinations, while Mistral 7B demonstrates greater variability. Overall, GPT-4o achieved a higher f1-score, reaching a value of 0.83 for identifying sexist tweets. However, recall remained below 0.72, leading to missed sexist tweets (false negatives). Some Mistral 7B prompts achieved notably higher recall, detecting more sexist tweets, but also produced more false positives. Additionally, few-shot prompting had little effect on GPT-4o's performance, whereas Mistral 7B's performance declined when examples were included, suggesting a limitation in its ability to manage this prompt effectively.

Table 2: Precision, recall and f1-score for prompts combinations and LLMs

			GPT-4o						Mistral 7B					
			Sexist			Non-sexist			Sexist			Non-sexist		
	System prompt	User prompt	p	r	f1	p	r	f1	p	r	f1	p	r	f1
Zero-shot	SP1	UP1	0.99	0.64	0.78	0.78	0.99	0.88	0.70	0.91	**0.80**	0.91	0.70	0.80
	SP1	UP2	0.97	0.70	0.82	0.81	0.99	0.88	0.61	0.96	0.75	0.94	0.53	0.68
	SP2	UP1	0.98	0.63	0.77	0.78	0.99	0.87	0.83	0.54	0.66	0.72	0.91	0.81
	SP2	UP2	0.98	0.72	**0.83**	0.82	0.99	**0.90**	0.68	0.88	0.77	0.88	0.68	0.76
	SP3	UP1	0.98	0.66	0.79	0.79	0.99	0.88	0.79	0.75	0.77	0.82	0.85	**0.83**
	SP3	UP2	0.98	0.68	0.80	0.80	0.99	0.88	0.68	0.90	0.78	0.90	0.68	0.77
Few-shot	SP4	UP1	0.98	0.70	0.82	0.81	0.99	0.89	0.53	0.80	0.63	0.93	0.29	0.45
	SP4	UP2	0.98	0.72	**0.83**	0.82	0.99	**0.90**	0.58	0.57	0.57	0.90	0.30	0.44

Legend: p: precision, r: recall, f1: f1-score