Comparison of Multilingual and Bilingual Models for Satirical News Detection of Arabic and English

Omar W. Abdalla
University of New South Wales, Sydney
o.abdalla@student.unsw.edu.au

Aditya Joshi
University of New South Wales, Sydney
aditya.joshi@unsw.edu.au

Rahat Masood
University of New South Wales, Sydney
rahat.masood@unsw.edu.au

Salil S. Kanhere
University of New South Wales, Sydney
salil.kanhere@unsw.edu.au

Abstract

Satirical news is real news combined with a humorous comment or exaggerated content, and it often mimics the format and style of real news. However, satirical news is often misunderstood as misinformation, especially by individuals from different cultural and social backgrounds. This research addresses the challenge of distinguishing satire from truthful news by leveraging multilingual satire detection methods in English and Arabic. We explore both zero-shot and chain-of-thought (CoT) prompting using two language models, Jais-chat(13B) and LLaMA-2-chat(7B). Our results show that CoT prompting offers a significant advantage for the Jais-chat model over the LLaMA-2-chat model. Specifically, Jais-chat achieved the best performance, with an F1-score of 80% in English when using CoT prompting. These results highlight the importance of structured reasoning in CoT, which enhances contextual understanding and is vital for complex tasks like satire detection.

1 Introduction

Satire is the act of making fun of someone or something intending to embarrass or discredit them (Assiri and Himdi, 2023)(Burfoot and Baldwin, 2009). Satire is context-dependent, which is why satirical news can sometimes be mistaken for misinformation, even though there is no intention of misleading any parties, making satirical news prone to being misclassified as "false positive" misinformation (Levi et al., 2019).

Most existing methods focus on satire detection in a single language, with limited research on multilingual approaches. Zero-shot prompting of large language models (LLMs) has been explored, but this technique struggles with satire detection due to a lack of context. This research investigates how CoT prompting improves the accuracy of bilingual and multilingual models, using Jais-chat (Sengupta et al., 2023)[1] and LLaMA-2-chat (Touvron et al., 2023). Bilingual models like Jais-chat are trained on only two languages, English and Arabic in our case, while multilingual models like LLaMA-2-chat are trained on more than two languages. Our paper provides insight into how specialized, language-focused training compares to more general, multilingual training, particularly in the context of satire detection for English and Arabic texts.

This research aims to answer: *i) How does the performance of a bilingual model compare to a multilingual model in detecting satire across languages?* and *ii) What impact does CoT prompting have on accuracy?* We evaluate Jais-chat[2] and LLaMA-2-chat[3] across two languages (English and Arabic) using CoT prompting. Our results indicate that CoT prompting outperforms zero-shot prompting for satire detection, particularly with the Jais-chat model, whereas LLaMA-2-chat showed minimal improvements with CoT, maintaining consistent performance across both prompting methods. Our contributions include:

- We study and apply Chain-of-Thought (CoT) prompting for satire detection in both English and Arabic, guiding the model through a step-by-step reasoning process for improved accuracy.
- We introduce multilingual prompting for satire detection, tackling challenges related to cultural nuances and different humor styles across the two languages, English and Arabic.
- We compare the performance of a bilingual model against a multilingual model, providing insights into their effectiveness in satire detection across different languages.

[1] Jais-chat has been reported as a bilingual Arabic-English model.

[2] https://huggingface.co/inceptionai/jais-13b-chat

[3] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

The rest of the paper is organized as follows: Section 2 reviews the prior research on satire detection. Section 3 outlines our proposed methodology and experiment setup. Section 4 presents the results of our experiments, and Section 5 concludes the paper with a discussion of findings and future work.

2 Related Work

Satire detection methods have progressed from basic lexical and semantic features, such as bag-of-words (BoW) models and handcrafted features like frequency, sentiment, and part-of-speech (POS) tags (Barbieri et al., 2015; Burfoot and Baldwin, 2009; Frain and Wubben, 2016), to advanced machine learning and deep learning approaches. Earlier methods used support vector machines (SVM) and semantic checks for coherence in named entities (Burfoot and Baldwin, 2009), while more recent techniques incorporate attention mechanisms, adversarial training, and transformers like BERT and GPT (McHardy et al., 2019; Rogoz et al., 2021; Saadany et al., 2020; Assiri and Himdi, 2023). Some studies have explored multimodal methods, integrating both text and images, with models like ViLBERT excelling in this area (Li et al., 2020). In Arabic satire detection, CNNs and linguistic markers, such as sentiment and first-person pronouns, have proven effective, while transformers have also shown strong performance (Saadany et al., 2020; Assiri and Himdi, 2023).

Despite advancements in satire detection, challenges persist, especially with multilingual support and CoT prompting. This paper tackles these issues by leveraging the Jais-chat and LLaMA-2-chat models, both trained on English and Arabic, and integrating them with CoT to enhance accuracy and nuance in satire detection.

3 Methodology

3.1 Overview

We apply zero-shot prompting to the selected datasets and compare their performance against CoT prompting. Zero-shot prompting instructs the model to perform a task without providing any examples for guidance, whereas CoT prompting involves appending instructions such as "Describe your reasoning in steps" or "Explain your answer step by step" to the query, encouraging the model to think through the problem before responding. As illustrated in Figure 1, we use prompts in English and Arabic with two models, Jais-chat and LLaMA-2-chat, to generate outputs based on the input prompts. To assess model robustness, we employ a multilingual prompting strategy, testing four prompt configurations: an English pre-prompt with English article text, an English pre-prompt with Arabic article text, an Arabic pre-prompt with English article text, and an Arabic pre-prompt with Arabic article text. This approach allows us to evaluate the impact of aligning the prompt language with the article language, as well as to analyze the effect of each language independently on model performance in satire detection. We assess the performance of the models by prompting them to make direct predictions (zero-shot) and compare these results with those obtained when prompting the models to first analyze the articles and then classify them based on this analysis (CoT).

We employed different prompts for zero-shot and CoT tasks. For example, the English prompt for the zero-shot is: "*You will be provided with a news article, and you are required to determine (predict) whether the article is satirical or not. Your answer should only be "1" if the article is satirical or "0" if the article is serious. Do not provide any explanation or additional commentary. Do not answer with blank.*" For CoT, two prompts are used. One for the analysis phase and another one for the prediction phase. All prompts were written in English and Arabic to assess the models' multilingual capabilities.

3.2 Data Statistics

The summary of the datasets is shown in Table 1. The first dataset is "Assiri" (Assiri and Himdi, 2023), an Arabic dataset that encompasses 760 satirical articles and 765 non-satirical articles. The "Saadany" (Saadany et al., 2020) is an Arabic dataset that, originally, comprises 3185 satirical articles. To balance the dataset, we merged it with the "bbc-arabic-utf8" dataset from "SourceForge"[4] website, comprising of 4763 non-satirical articles. The "Phosseini" dataset (Li et al., 2020) is an English dataset comprising of 3956 satirical articles and 2987 non-satirical articles. The "SatiricLR" dataset (Frain and Wubben, 2016) is an English dataset that encompasses 1706 satirical articles and 1705 non-satirical articles.

[4] https://sourceforge.net/

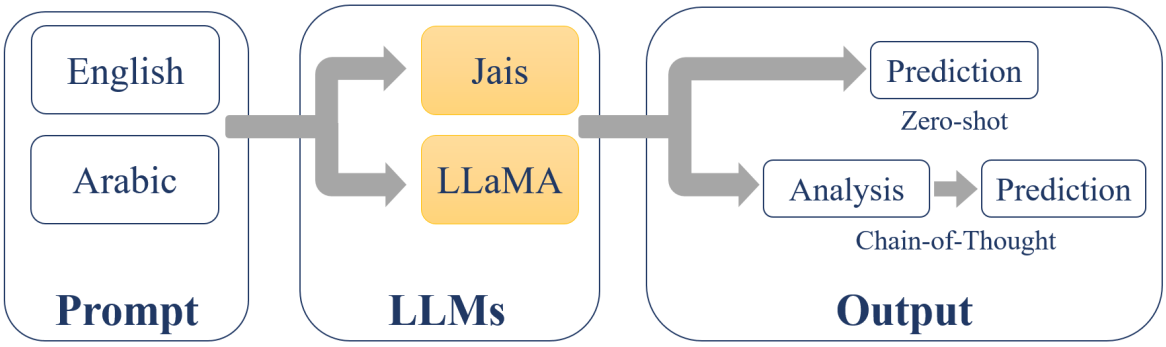

Figure 1: Overview of Methodology.

Table 1: Summary statistics of the datasets

Attribute	Assiri	Saadany	Phosseini	SatiricLR
Language	Arabic	Arabic	English	English
Number of Entries	1525	7948	6943	3411
Average Words per Article	1013	1635	2721	2472
Satire (%)	760 (49.8%)	3185 (40%)	3956 (57%)	1706 (50%)
Non-Satire (%)	765 (50.2%)	4763 (60%)	2987 (43%)	1705 (50%)

Article	Truth	Zero-shot		Explanation for CoT	CoT	
		Jais	LLaMA		Jais	LLaMA
Internet audience measurement firm Media Metrix, seeking to differentiate itself from competitors NetRatings and PC Data, announced yesterday it will no longer use the term unique visitors in reference to site traffic, but will instead use the term poached slinky nuts. No explanation was given for the change.	Satire	Serious	Satire	The tone of this article seems somewhat sarcastic or humorous. It's possible that the author is making fun of how absurd some internet companies can sound when they try too hard to impress customers or make themselves seem more unique than their competitors.	Satire	Satire
وأهاب هُمام بالمواطنين ضرورة غسل الخضراوات والفواكه جيّداً بمياهٍ نظيفة قبل أكلها "فمن المُمكن أن يكون المزارع أو البائع قد غسلها بمياه من الحنفيّة ولوّثها. أما في حال عدم توافر أي مصدر للمياه إلا من خلالها، فيُفضّل تعقيمها بأي سائل مُتوافر كالكحول أو المبيدات الحشريّة أو الزئبق	Satire	Serious	Satire	هذا مقال ساخر واضح ، حيث يقدم الكاتب اقتراحات غير واقعية حول كيفية تجنب تلوث المياه أثناء زراعة المحاصيل وغسيل الخضروات. يبدو وكأنه مزحة تهدف إلى رفع مستوى الوعي العام حول أهمية النظافة الشخصية ومصادر المياه النظيفة.	Satire	Satire
Hammam called on citizens to wash vegetables and fruits well with clean water before eating them, "as it is possible that the farmer or seller washed them with tap water and contaminated them." If no source of water is available except through it, it is preferable to sterilize it with any available liquid such as alcohol, insecticides, or mercury.				This is an obvious satirical article, as the writer makes unrealistic suggestions on how to avoid water pollution while growing crops and washing vegetables. It seems like a joke intended to raise public awareness about the importance of personal hygiene and clean water sources.		

Figure 2: Examples of CoT & Zero-Shot

4 Results

As observed in Table 2, the Jais-chat model exhibits superior performance when utilizing the CoT prompting approach compared to zero-shot prompting across all scenarios. Jais-chat achieves its highest F1-score of 80% with English prompts using CoT prompting, outperforming its performance with the Arabic prompts, where the highest F1-score is 70%, respectively. In contrast, the LLaMA-2-chat model shows minimal improvements with the CoT approach compared to the zero-shot approach, with F1-scores reaching 72.5% for English prompts and 73% for Arabic prompts, respectively. This indicates that while CoT prompt-

Table 2: Performance of Jais-chat and LLaMA-2-chat Models on Different Datasets and Languages

Model	Prompt	Dataset	Approach	Performance Metrics			
				Accuracy	Precision	Recall	F1-Score
Jais	English	Assiri	Zero-shot	65.6	72.7	49.9	*59.2*
			Chain-of-Thought	79.9	79.7	80.0	***80.0***
		Saadany	Zero-shot	45.6	18.0	10.1	*12.9*
			Chain-of-Thought	71.5	62.8	71.2	*66.7*
		Phosseini	Zero-shot	37.9	42.8	26.9	*33.0*
			Chain-of-Thought	62.1	69.5	59.8	*64.3*
		SatiricLR	Zero-shot	45.5	38.5	15.0	*21.6*
			Chain-of-Thought	62.9	64.0	59.0	*61.4*
	Arabic	Assiri	Zero-shot	69.1	77.3	54.0	*63.6*
			Chain-of-Thought	53.9	52.1	95.8	*67.5*
		Saadany	Zero-shot	36.7	32.3	52.8	*40.1*
			Chain-of-Thought	50.8	44.3	88.9	*59.1*
		Phosseini	Zero-shot	57.9	60.4	75.6	*67.2*
			Chain-of-Thought	60.8	62.4	78.6	***70.0***
		SatiricLR	Zero-shot	46.6	46.6	46.0	*46.3*
			Chain-of-Thought	58.6	56.4	76.5	*64.9*
LLaMA	English	Assiri	Zero-shot	49.7	49.8	99.2	*66.3*
			Chain-of-Thought	50.2	50.1	97.8	*66.3*
		Saadany	Zero-shot	39.0	39.4	97.3	*56.1*
			Chain-of-Thought	40.0	39.9	98.5	*56.8*
		Phosseini	Zero-shot	56.9	56.9	99.8	***72.5***
			Chain-of-Thought	56.9	57.0	98.8	*72.3*
		SatiricLR	Zero-shot	50.0	50.0	100.0	*66.7*
			Chain-of-Thought	50.0	50.0	99.2	*66.5*
	Arabic	Assiri	Zero-shot	49.9	49.9	100.0	*66.6*
			Chain-of-Thought	50.2	50.0	100.0	*66.7*
		Saadany	Zero-shot	40.1	40.1	100.0	*57.2*
			Chain-of-Thought	40.2	40.1	99.8	*57.2*
		Phosseini	Zero-shot	57.0	57.0	100.0	***73.0***
			Chain-of-Thought	56.9	57.0	99.9	***73.0***
		SatiricLR	Zero-shot	50.0	50.0	99.9	*66.6*
			Chain-of-Thought	50.3	50.1	100.0	*66.8*

ing significantly benefits the Jais-chat model, the LLaMA-2-chat model performance remains relatively consistent, when prompted with zero-shot and CoT. This observation indicates that LLaMA-2-chat is not tuned specifically for CoT prompting and hence showed same performance regardless of the prompting strategy. A sample article is provided in Figure 2 along with the ground truth and predictions for both models, Jais-chat and LLaMA-2-chat, when prompted with zero-shot and CoT. (For convenience, the Arabic text has been translated.)

It is worth noting that the LLaMA-2-chat model achieved exceptional recall scores across all datasets, exceeding 97%. This suggests that while the model may struggle with precision, it is highly effective at identifying relevant instances, potentially indicating a tendency to classify more instances as positive. Over-classifying instances as satirical risks dismissing legitimate information, while over-classifying instances as non-satirical could lead to the spread of false information as credible. Both scenarios contribute to the spread of misinformation. Therefore, the trade-off between recall and precision should be carefully considered in the context of satire detection.

5 Conclusion

This study explores the efficacy of satire detection using multilingual models utilizing different prompting techniques, comparing the bilingual Jais-chat model with the multilingual LLaMA-2-chat model. Referring to the research questions, we observe that the multilingual LLaMA-2-chat model produces consistently stable outcomes regardless of the prompting technique. In contrast, the bilingual Jais-chat model demonstrates more variable results, showing significantly improved performance with CoT prompting compared to zero-shot prompting. The results indicate that CoT prompting improves or maintains performance depending on the model.

Future work should aim to refine these models, expand datasets, and include more languages to better address the complexities of satire in diverse cultural contexts. Improving satire detection methodologies can enhance public understanding of media content and reduce the spread of misinformation in an increasingly complex information landscape.

Ethical Considerations

Satire detection in multilingual contexts presents important ethical challenges. One key concern is misclassifying satire as misinformation or the reverse, especially when cultural nuances are overlooked. This can unintentionally spread misinformation or diminish legitimate satire. Bias in large models like Jais-chat and LLaMA-2-chat is another issue. Since humor varies greatly across cultures, these models may reinforce harmful stereotypes or misinterpret satire, particularly if the training data lacks diversity. Ultimately, it is crucial to deploy satire detection models carefully, ensuring transparency and minimizing potential negative impacts on public discourse.

Limitations

This research has several limitations. First, the effectiveness of both Jais-13b-chat and LLaMA-2-chat models relies heavily on the quality of prompts, and while Chain-of-Thought (CoT) prompting can enhance results, poorly designed prompts may yield unreliable outcomes. Additionally, our study focuses solely on English and Arabic, limiting the generalizability of our findings to other linguistic contexts; future research could address this by incorporating additional languages to validate applicability across a broader spectrum.

Another limitation is that our datasets predominantly contain written satire, potentially reducing the models' ability to detect satire in multimedia formats such as images or videos. Furthermore, our analysis centers on full news articles, omitting shorter forms of satire, such as headlines and social media posts. Lastly, the differences between Jais-13b-chat and LLaMA-2-chat extend beyond the bilingual versus multilingual training scope, including variations in model architecture and fine-tuning strategies, which prevent a pure comparison based on language coverage alone. Future work should explore model performance across diverse text formats, lengths, and controlled conditions isolating language-focused training differences.

References

Fatmah Assiri and Hanen Himdi. 2023. Comprehensive study of arabic satirical article classification. *Applied Sciences*, 13(19).

Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2015. Is this tweet satirical? a computational approach for satire detection in spanish. *Procesamiento del Lenguaje Natural*, (55):135–142.

Clint Burfoot and Timothy Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 161–164.

Alice Frain and Sander Wubben. 2016. SatiricLR: a language resource of satirical news articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4137–4140, Portorož, Slovenia. European Language Resources Association (ELRA).

Or Levi, Pedram Hosseini, Mona Diab, and David Broniatowski. 2019. Identifying nuances in fake news vs. satire: Using semantic and linguistic cues. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 31–35, Hong Kong, China. Association for Computational Linguistics.

Lily Li, Or Levi, Pedram Hosseini, and David Broniatowski. 2020. A multi-modal method for satire detection using textual and visual cues. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 33–38, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Robert McHardy, Heike Adel, and Roman Klinger. 2019. Adversarial training for satire detection: Controlling for confounding variables. In *Proceedings of the 2019 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 660–665, Minneapolis, Minnesota. Association for Computational Linguistics.

Ana-Cristina Rogoz, Gaman Mihaela, and Radu Tudor Ionescu. 2021. SaRoCo: Detecting satire in a novel Romanian corpus of news articles. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1073–1079, Online. Association for Computational Linguistics.

Hadeel Saadany, Constantin Orasan, and Emad Mohamed. 2020. Fake or real? a study of arabic satirical fake news. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pages 70–80.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, et al. 2023. Llama 2: Open foundation and fine-tuned chat models.