Translation of Code-Mixed and Code-Switched Tweets Using LLMs

Guru Gobind Singh Indraprastha University

Durga Sharma, Rahul Johari

durgasharma5899@gmail.com

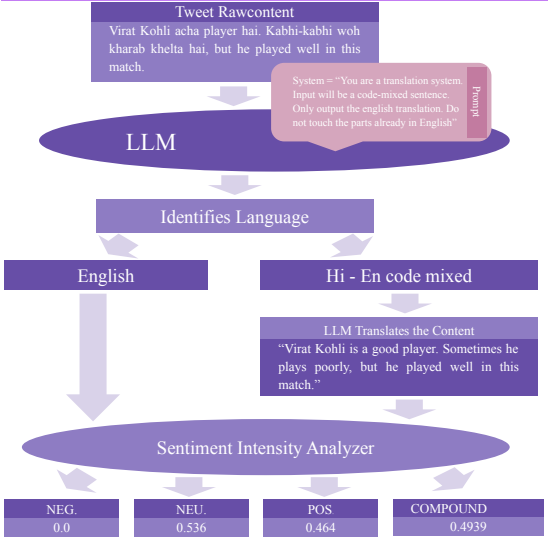
What is Code-Mixing and Code-Switching?

Code-Mixed Hi-En: "Yaar, kal ka match bohot intense tha, but Virat ne amazing performance di!"
Code-Switched Hi-En: "I can't believe we lost the game, lekin Virat ne bohot achha khela."

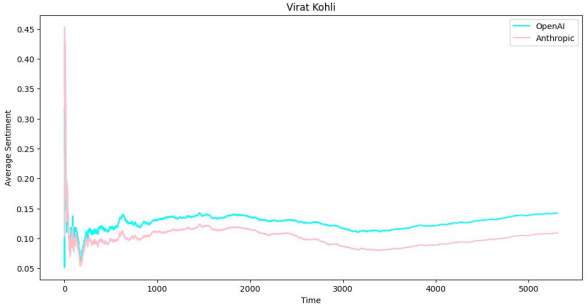
Background

Sentiment analysis of such data often lacks accuracy. For example, sentiment analyzers classify the English word **"Nice"** as **positive**, but interpret the Hindi-English code-mixed word **"Badiya"** as **neutral**, as they fail to recognize the language.

Methodology



Results



This is an **Average Sentiment V/S Time** graph using LLMs **OpenAI** and **Anthropic**. This helps in analysing the changing public perception over time. In this graph, the x-axis represents the number of tweets, used as a proxy for time, since they are in chronological order, showing the progression of sentiment. The y-axis displays sentiment values, ranging from -1 to +1.

Name	Gender	Sport	Total Tweets	Positive	Neutral	Negative
Virat Kohli	M	Cricket	5313	2349	1521	1443
Harmanpreet Kaur	F	Cricket	5100	2603	1930	567
Vijender Singh	M	Boxing	4280	1129	1628	1523
Sarita Devi	F	Boxing	5003	1790	838	2375
Sushil Kumar	M	Wrestling	4445	1061	570	2814
Sakshi Malik	F	Wrestling	5305	1722	1587	1996

Table 1: *Distribution of positive, neutral, and negative tweets for each of the six sports personalities using* ***Anthropic***

Name	Gender	Sport	Total Tweets	Positive	Neutral	Negative
Virat Kohli	M	Cricket	5313	2520	1594	1199
Harmanpreet Kaur	F	Cricket	5100	2673	2007	420
Vijender Singh	M	Boxing	4280	1157	1674	1449
Sarita Devi	F	Boxing	5003	1804	862	2337
Sushil Kumar	M	Wrestling	4445	1093	602	2750
Sakshi Malik	F	Wrestling	5305	1779	1631	1895

Table 2: *Distribution of positive, neutral, and negative tweets for each of the six sports personalities using* ***OpenAI***

Discussion

	Positive	Neutral	Negative
Google Translate	36.19%	29.41%	33%
Anthropic	36.18%	27.42%	36.40%
OpenAI	37.44%	28.42%	34.13%

Table 3: *Percentage distribution for positive, neutral, and negative tweets for Google Translate, Anthropic, and OpenAI*

This study employed LLMs, specifically **Anthropic** and **OpenAI**, to translate **Hindi-English (Hi-En) code-mixed** tweets for enhanced **sentiment analysis** of cyberbullying experienced by Indian athletes on Twitter. A notable shift from neutral to positive and neutral to negative was observed, improving the accuracy of previously misclassified neutral tweets. Anthropic identified 27.42% and OpenAI 28.42% as neutral, compared to 29.41% using Google Translate (8,661 neutral tweets out of 29,446). For positive tweets, Anthropic found 36.18% and OpenAI 37.44%, while for negative tweets, Anthropic detected 36.40% and OpenAI 34.13%, against the original 33%.

Conclusion

This study demonstrates the effectiveness of LLMs in improving sentiment analysis of Hindi-English code-mixed and code-switched tweets on social media. The findings highlight the importance of developing NLP models that are capable of handling the nuances of code-mixed and code-switched language.

References

Mark Sebba, Shahrzad Mahootian, and Carla Jonsson. *Language mixing and code-switching in writing: Approaches to mixed-language written discourse*. Routledge, 2012. **M**ichael Groves and Klaus Mundt. Friend or foe? google translate in language for academic purposes. *English for Specific Purposes*, 37:112–121, 2015. **O**penAI. Openai gpt-4. https://cdn.openai.com/papers/gpt-4.pdf. **A**nthropic. Anthropic's claude. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. **C**layton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.