Hook Document – Can a Machine Learn Comedy?

GitHub Repository: https://github.com/Ariana-Elahi/DS4002-Final-Project-CS3

Comedy is messy. Characters interrupt, yell, mumble, argue, contradict themselves, and somehow it all still works. Every line is fast, chaotic, and rarely follows textbook grammar. But what happens when you hand thousands of these lines to a machine and ask it a simple question:

“Who said this?”

In this case study, you are stepping into the role of an NLP researcher trying to teach a machine how to interpret comedic dialogue. You’ll be working with a curated dataset of lines from South Park, focused on ten of the show’s most frequent characters. Each has a distinctive voice — or at least, humans think they do. Your mission is to find out whether a computer can detect those differences too.

You’ll rebuild a text-classification pipeline that transforms raw dialogue into numerical features, trains a machine learning model, and evaluates how well the system can identify speakers. Along the way, you’ll see where the computer excels, where it fails spectacularly, and what linguistic quirks it learns to rely on. Do certain characters use unique catchphrases? Does tone translate into text? Do characters blend together when the script gets chaotic?

Your goal is not just to run the model, it’s to interpret it like a scientist. You’ll examine the model’s predictions, dig into misclassifications, and decide whether there are discoverable patterns in comedy… or whether human humor is still too unpredictable for algorithms.

By the end, you’ll have recreated the analysis, produced your own evaluation results, and reflected on what happens when machine learning meets one of television’s most chaotic shows.

All materials, including data, scripts, and instructions, can be found in the GitHub repository linked above.

Image Source: https://www.bbc.com/news/business-58109993