Funny, not Funny

Walter Squires, Kyungchan Lim, and Autumn Henderson

I. INTRODUCTION

Joke theft, or joke poaching, has been prevalent in society as early as the end of the 19th century, where vaudeville relied on stock materials[1] and became more prominent in the 1970s, where more prominent figures such as Robin Williams have been singled out for stealing certain jokes or whole acts[2]. In today's society, joke theft is still rampant, but with the emergence of social media, recording, and information being at our fingertips, there is more visibility on a comedian and their acts, leaving them vulnerable to someone studying their act in great detail with the information readily available. With that being said, determining the originator of the joke can be difficult and can lead to financial loss for the originator as other comedians make money off of the originator's creations.

Written works such as books and scholarly articles have streamlined processes in the legal world when someone utilizes their productions without proper citations. Prosecuting thieves of stand-up works is more challenging. Copyright laws defend the expression of an idea, but not the idea itself, leading to cases where a comedian can tell a joke about a topic with slightly different wording and it is hard to prove that it was stolen. Likewise, there are "informal, network-based institutions" that, in the place of formal legal interventions, use social norms to "define exclusivity and guide adjudication of rights violations." Even with these institutions in place, they serve more as a deterrent and are "ineffective regarding the sanctioning of those that transgress property rights norms."[3]

II. OBJECTIVE

The goal of this study is to establish a system that, when given a joke and a comedian, can determine the likelihood that the comedian was the originator of that joke.

Originality can be defined in two ways: first, as "literary originality or originality of content", and second as "performance originality or originality of style"[4]. For the purpose of this study, we will apply both definitions when compiling and analyzing the data.

A joke can be broken down into three essential parts, each typically occurring in a systematic way: setup, punchline, and taglines. The setup usually poses a question or observation that provides an opportunity for the punchline to be given in response. Following that is the punchline itself, which is the climactic conclusion to the setup that is designed to make the audience laugh. Finally trails the tagline(s); they are optional and are any laugh lines after the original punchline relating back to that first set up, even if other setups and punchlines have come since.

This study will develop a baseline tool that can be used to detect joke theft. There are "several online tools to check whether someone is trying to take any undue credit by using plagiarized content," but no equally plausible methods for stand-up comedy[5]. In addition, it will provide a data collection mechanism for compiling works of a comedian and the likelihood that the comedian is the originator of those works.

The motivation for such a study is multi-layered: Can a tool be created that can determine the originality of stand-ups with accuracy? How prevalent is joke-poaching? And who are the bad actors?

III. METHODS

A. Data Collection

Before any analysis can be completed or models can be formed, there must be data. Transcripts of many comedian's stand-up routines can be found online on "Scraps from the Loft", and by using a web-crawler to collect raw data we will be able to extract components in order to have a meaningful data set.

B. Data Analysis

Routines typically follow a specific flow where the setup occurs first, then the punchline, and then finally the tagline. The model we build will utilize those specifications to determine how many jokes occur in a stand-up act as well as what types of jokes are delivered and the manner in which they are executed; in order to do so though, the raw data must be analyzed.

By going over the initial raw data, following components will be extracted:

- Laughs in the script
- Words that are used by performer

Due to the fact that both audience and performer can both laugh during the performance and both of them can be written in the script(raw data), differentiating between laughs from audience and performer is necessary;only laughs from the audience can be used to determine the presence of a punchline or tagline.

Extracting meaningful data from words that are used by performer will allow for a model of that performer to be made; after collecting and analyzing data is complete for one comedian, this process will be repeated several times in order to collect several different data sets for different performers.

C. Comedian Modeling and Testing

Different performers will have different types of speaking habits and word choices. By analyzing collected and extracted data, our model will reflect how each performer has different habits and uses different word choices. The validity of that model will then be tested against the jokes that makeup its data set; the model should be able to identify a joke that was used in the construction of the model as belonging to it's respective comedian with high probability.

Once confidence in the model is established, it will be tested against jokes from routines known not belong to the comedian on which the model was based. The goal in this case is for the model to be able to have high certainty that the given joke was not written by the modeled comedian.

The ultimate goal then, once these methods have been proven to be successful, is to model a comedian who has a reputation for joke poaching. In creating such a model, the hope is to be able to determine what jokes, if any which are attributed to them, were in fact poached from another non-credited comic.

D. Assumptions

A setup can have one or multiple lines, whereas punchlines and taglines are one-liners. All three elements always occur in the same order: setup, punchline, tagline. The punchline and the tagline always make the audience laugh, and differentiating between the punchline and tagline involves evaluating when the most recent setup occurred and wherein the transcript the laughter prompt followed.

E. Roles

- Walter Squires
  - Develop methodology and milestones that the project should accomplish.
  - Ensure that other team members are on track to accomplish milestones as described in the timeline
- Kyungchan Lim
  - Develop scraper for transcripts.
  - Implement conversion of routines into aspects of the comedian that can be modeled
- Autumn Henderson
  - Draft proposal and convert into IEEE format.
  - Draw graphs.

F. Timelines

- October 9th: Scrap stand-up site for transcripts
- October 23rd: Break routines up into components
- November 6th: Build comedian model (test accuracy against itself)
- November 20th: Test model against jokes known not to belong to that comedian
- November 27th: Build a model for questionable comedian

IV. POSSIBLE IMPLICATIONS

A. Robin Williams and Others

Robin Williams has been targeted for being a notorious joke-stealer by other comedians. Once this model is built, an end-goal would be to utilize the model to determine, with a specific accuracy, the extent with which Robin Williams, and other popular thieves, stole jokes.

B. Legal Ramifications

As stated in the Introduction, lawsuits brought against bad actors for joke-poaching are complicated. This model has the potential to provide a systematic way to prosecute thieves. For example, say the model determines with statistically significant accuracy that a comedian replicated 85 percent of an originator's work, where a threshold of 80 percent is considered prosecutable; then litigation becomes simplified in a way unseen in history.

C. Future Research

The first version of this model is going to rely on transcripts that are structured in such a way that a web crawler can easily parse them to determine where the setup, punchline, and taglines occur. However, not all transcripts are formatted the same. Therefore, there may be the potential to enable the model to process differently formatted transcripts.

Databases could be constructed utilizing the program that stores an originator and their transcripts as well as specific data about those transcripts, such as how many jokes it consisted of, how many lines of setups and taglines there were for stand-up acts, how much those works resemble other works, and other data as needed. Those databases can then have various other applications unrelated to this study.

REFERENCES

[1] Library of Congress, "Bob Hope and American Variety: Bits 'I&' Sketches," Library of Congress. [Online]. Available: https://www.loc.gov/exhibits/bobhope/bits.html. [Accessed Sep. 26th, 2021].

[2] R. Zoglin, Comedy at the Edge: How Stand-up in the 1970s Changed America, Bloomsbury USA, 2003.

[3] P. Reilly, "The Weakness of Sanctioning in Norms-Based Property Systems: An Investigation of Joke Theft," Academy of Management Proceedings, vol. 2015, no. 1, November, 2017, [Online]. Available: Academy of Management, https://journals.aom.org/doi/abs/10.5465/ambpp.2015.15659abstract. [Accessed Sep. 26th, 2021]

[4] G. Pate, "Whose Joke Is It Anyway? Originality and Theft in the World of Standup Comedy," Theatre Journal, vol. 66, no. 1, March, 2014, [Online]. Available: The Johns Hopkins University Press, https://www.press.jhu.edu/. [Accessed Sep. 25, 2021].

[5] RK Dewan 'I&' Co, "'Joke Theft': Not a joke," Lexology. [Online]. Available: https://www.lexology.com/library/detail.aspx?g=5d3d7764-098f-45b1-83ef-f3d2e2e5abe0. [Accessed Sep. 26th, 2021]