COMP 370 Final Project — Breaking Down *Breaking Bad*: Analyzing Side Characters' Dialogue Topics and Engagement

Written by
Nicholas Belev, Priyanshu Bhandari, David Bosnjak
McGill School of Computer Science
3480 Rue University, Montréal, QC H3A 2A7
nicholas.belev@mail.mcgill.ca, priyanshu.bhandari@mail.mcgill.ca, david.bosnjak@mail.mcgill.ca

Abstract

Evaluating the dialogue patterns of side characters from the critically acclaimed television series "Breaking Bad" offers valuable insights into their roles and personalities. This study stems from the question of how dialogue reveals which topics and themes the show's characters are most significant to (involved in). To dissect this query, the dialogue of four prominent side characters—Saul Goodman, Skyler White, Gus Fring, and Mike Ehrmantraut—from Season 3—were analyzed. Ten episodes' worth of these characters' lines were collected and annotated across seven key thematic categories: Money, Business, Health, Law, Family, Daily Life, and Danger. Using Term Frequency - Inverse Document Frequency (TF-IDF), the most relevant words for each category (theme subject) were identified, providing characterization and context for these topics. The results highlight distinct patterns in dialogue, such as Saul's focus on legal issues and criminal business, and Skyler's emphasis on family and health concerns. Visualizations, including histograms and pie charts, aid in comprehension of the distribution of dialogue across topics and characters, providing a quantitative perspective on character's engagement with specific topics. This data-driven analysis determines characters' narrative significance and aids in audience understanding of the storytelling in "Breaking Bad".

Introduction

This study's objective is to assess the manner and subjects across which the supporting characters in a television communicate. A number of potential series were deliberated upon, ultimately leading to the selection of *Breaking Bad*, which had previously been viewed (with complex understanding) by all members of the team, thereby enhancing the interest and accuracy of the annotation process.

Breaking Bad is an AMC television series that premiered in the US in 2008. It tells the story of a high school chemistry teacher who "breaks bad", turning into a methamphetamine manufacturer after a terminal cancer diagnosis. The protagonist takes extreme and criminal measures to secure his family financially, and this pursuit leads to moral decay, personal tragedy, and rising violence. However, identifying the transcripts for this show was an immensely challenging task due to a lack of properly formatted, reputable sources—that were free of copyright restrictions. An appropriate resource was found after extensive search, covering the majority of episodes in *Season 3* of the show.

This project's structure is framed as a standard data science project. The initial step of this process is question formulation. In this case, the question began as: "What do the side characters of *Breaking Bad* talk about?". Upon refinement into a more quantifiable task, the resulting formulation is "What topics and themes are discussed by key side characters in *Breaking Bad*, *Season 3*?"

The supporting characters were chosen considerately in order to yield a diverse range of plot-significant dialogues; this approach also helps to reduce unnatural bias about certain subjects when discussing them. Among these characters is Saul Goodman, the protagonist's questionable and unethical criminal lawyer. Next chosen is Skyler White, the protagonist's wife, who has only recently discovered her husband, Walter White's, meth lab operations (and its high financial proceeds). Third is Gustavo Fring (Gus), a powerful drug lord, whose front is a fast food chain. Fring contracts Walter as his empire's meth chemist. The final character chosen is Mike Ehrmantraut, a private investigator and hitman, who has worked with both Gustavo Fring and Saul Goodman in the past, and currently serves as Gus's right hand man.

Upon conducting an open coding, the dialogue neatly split into 7 thematic categories or topics of conversations: *Money*, *Business*, *Health*, *Law*, *Family*, *Danger*, and *Daily Life*. These topics cover a vast range of conversations and have been presented and elaborated upon in detail in the *Results* section.

Applying a TF-IDF analysis along with visualization methods, this study outlines a number of (reasonably anticipated) findings and conclusions. Every character is strongly associated with some discussion topics that relate to their description, primary occupation, and typical setting in the series. Also considered are the 10 terms of each theme category with the highest TF-IDF scores, explored further under the *Results* section.

Data

Data Collection

Conducting this dialogue–character–theme analysis began with locating official scripts for the show. After extensive research, the only publicly released scripts from *Breaking Bad* director, Vince Gilligan, were those of the

episodes in *Season 3*, as permitted by copyright law. Fortunately, the third season of this show is plentiful in character engagement with central themes, along with near-climactic plot points for a number of side characters.

Subsequently, upon acquiring *Breaking Bad, Season 3*'s script scans, it remained to process them into computer-parsable text. This was achieved with Python's *PyPDF PDF Reader*. Next, a Python pattern matching program was applied to collect the dialogue lines from the side characters of focus: Saul Goodman, Skyler White, Gus Fring, and Mike Ehrmantraut. Compiling these into a Tab Separated Value (TSV) file, the rows were ordered by episode, in preparation for annotation.

The TSV has the following structure:

"*episode character dialogue_line"*

Efforts were made to ensure the selection of dialogue was unbiased, focusing on lines that reflected meaningful interactions or conveyed character intent. Trivial exchanges, such as filler phrases or repetitive banter, were manually excluded to maintain the relevance of the dataset. The result of this, given the copyright-limited quantity of film available (10 episodes, to be precise) was a range of between 100 - 300 lines per character.

Annotation Methodology

Using this collected and sorted data, the team conducted an open coding across two rounds (on 100 of each character's lines) to determine a minimum spanning typology that captures the categories related to all dialogue across *Season 3*. The resulting classification system contains the following seven categories: *Money* (m), *Business* (b), *Health* (h), *Law* (l), *Family* (f), *Daily Life* (dl), *Danger* (d).

With the categories established, the team proceeded to manually (and separately) annotate the collected dialogue lines. Each line was assigned to one of the seven categories based on its dominant thematic content, with particular attention paid to ambiguous "edge cases" that could lean towards multiple categories. In such cases, the context of the dialogue (captured via the embedded director's notes in each line) and the character's intent were used to determine the most appropriate classification.

Finalization of Annotations

This yields the following annotated TSV structure:

"episode character dialogue_line annot_1 annot_2 annot_3"

Note that "annot_[i]" refers to the *i*th team member's annotation for a given line. Furthermore "annot_[i]" $\in$ {"m", "b", "h", "l", "f", "dl", "d"} (thematic categories).

Following individual annotations, in which there were seldom instances of full disagreement between annotators, a majority-dominant approach was used to finalize the annotated dataset of side character dialogue:

"episode character dialogue_line final_annot"

Analysis Process

The team implemented a Term Frequency–Inverse Document Frequency (TF-IDF) word scoring to determine which keywords were most unique to each thematic category.

In order to obtain TF-IDF scores, Python was used to group all lines of dialogue—by the category they were annotated as—into text files. These text files were then cleaned and processed into their constituent words. TF-IDF values were calculated for each word by counting their occurrence and presence in each category of line.

The resulting values highlighted the words that were most distinctive to each thematic category. Words with high scores appeared frequently within a specific category but were relatively uncommon across the others, making them valuable indicators of the dialogue's focus and intent within that theme.

Methods

Data Collection

The decision to focus on *Breaking Bad, Season 3* was influenced by both availability and relevance. *Season 3* features notable developments and interactions for the selected side characters—Saul Goodman, Skyler White, Gus Fring, and Mike Ehrmantraut—making it an ideal focus for dialogue analysis. Python's PyPDF Reader and pattern matching were used to automate the extraction of dialogue lines. This approach was both efficient and convenient to prepare the lines for manual annotation. Trivial and filler lines were excluded to preserve the integrity of "meaningful" content, as such lines would dilute the thematic analysis and introduce noise to the analysis.

Annotation Methodology

An open coding method to derive the typology was the optimal choice due to its flexibility and learn-by-doing style. Two rounds of coding were conducted to cross-check that the resulting categories were comprehensive and representative of the dataset. The seven categories—Money, Business, Health, Law, Family, Daily Life, and Danger—were intentionally broad enough to encompass the variety of themes in the dialogue while being distinct and differentiable. This aided in asserting the handling of ambiguous lines—take, for instance Saul's line "It's the Tax Man…" (Gilligan, 2010). It is spoken in the context of legal loopholes, rather than pure business strategy or money, and so is resolved as "Law".

Hence, this typology maps directly onto the characters' roles and storylines, capturing the diversity of topics

central to the narrative, from financial and operational concerns to familial conflicts and personal vulnerabilities. They balance specificity and inclusivity, with *Daily Life* serving as a catch-all to appropriately categorize casual, non-trivial conversation.

Lastly, the majority-dominant method (in parallel with individual annotations) was used to finalize annotations, as it provided a simple yet effective way to consolidate team inputs and reduce the impact of individual bias and comprehension error. The robust typology combined with the familiarity of the team with the *Breaking Bad* storyline and characters largely minimized annotator disagreement.

Analysis Process

The decision to use TF-IDF for topic characterization was driven by its effectiveness in exposing terms that are both frequent and uniquely significant to specific categories. TF-IDF performed exceptionally well at burying noise terms, repeated across categories—a vast advantage over simple word frequency. This method prioritizes words that differentiate one category from others, making it well-suited to frame the spectrum of defining terms of the show's themes. The filtering of top 10 terms per category was a deliberate effort to distill the results into the most significant term–category pairs.

Challenges and Limitations

One note-worthy difficulty was collecting enough data given the limited amount of publicly available primary sources. Consequently, only *Season 3* scripts were accessible. This restriction reduced the potential for a broader longitudinal reach spanning the full show and plot. Additionally, some categories of the typology, namely *Health* and *Danger*, had fewer dialogue lines compared to more general topics like *Daily Life*. This imbalance likely impacted the diversity of terms identified for those categories post-analysis.

Ambiguity in some dialogue lines also posed a challenge, especially when lines could belong to multiple categories. Although context and majority-domination helped address this issue, these edge cases brought along a degree of human bias into the annotation process.

Results

Detailing the Typology

The following typology yields from the open coding:

• **Money (m)**: Dialogue whose main focus is on finance, drug money, and payments. A positive example would be: "So all this is to say... we have the money." It is labeled as money as evidently, the focus here is on finance and cash acquisition. A negative example: "Absolutely not, and I need to keep it that way. Hank's got enough on his mind, right now." Clearly there is no mention of money whatsoever; the dialogue primarily focuses on Hank's (another side character) *Health*. Edge case: The dialogue is "You wanna stay out of jail, dontcha? You wanna keep your money and your freedom? Because I've got three little letters for you: IRS." Even though the dialogue directly namedrops "money", the primary focus is on the law and staying within the legal bounds. In this case, it would lean more towards Law than Money, categorically.

• **Business (b)**: Dialogue which primarily involves information about business operations, investments, logistics or strategies. Positive example: "Final step: integration. The revenues from the salon go to the owner —that's you." Evidently this dialogue represents an operation step for a business, thus gets labeled as business. Negative example: "When Walt was diagnosed—it, um, it changed him." This has nothing to do with business, it discusses Walter's cancer and thus regards *Health*. Edge case: "Hey, what? Hey, listen to it! Come on! I'm talking about your future here!" The dialogue does not seem to mention anything related to our definition but given the context, it is actually talking about the operation of a beauty salon. Thus it gets labeled as business.

• **Health (h)**: Dialogue which mainly mentions concerns related to illness, recovery, or well-being. Positive example: "You know I haven't slept since Thursday? I need to sleep." Evidently the dialogue talks about recovery by sleeping thus it gets labeled as health. Negative example: "Hi, I hope we're not too late." This has no verbal connection to the health of any character, thus it is not categorized as health. Edge case: "Marie, you will take our money. Use it to take care of Hank." This dialogue might seem like a good fit to be labeled as *Money*, however, given the context of the show the focus of the scene is how much other characters Hank's health, due to which this line receives the label, *Health*.

• **Law (l)**: Dialogues which bring up legal issues, breaking laws, taxes, or implied references to criminal activity. Positive example: "It's the Tax Man. And he's looking at you. He sees a young fella with a big fancy house, unlimited cash supply and no job." This dialogue clearly centers around government officials and taxes so is labeled as *Law*. Negative example: "Then you have it. But what does that mean, exactly?". No mention of taxes, legal issues or criminal activity whatsoever so is not marked as *Law*. Edge case: "Contact Pinkman. Get him to drop these charges." Contextually, this dialogue fails to be labeled as law. This line is spoken in a dramatic argument between Skyler and Walter, where the charges refer to Hank's assault on Jesse Pinkman (Walter's partner in crime). The significance of this scene is geared towards the conflict of Skylar's family commitment and the world of drugs and crime she has been roped into by Walter. This line is a better candidate for *Family*, highlighting the difference that context makes for interpreting a read line.

• **Family (f)**: The lines which talk about interpersonal dynamics, responsibilities, and conflicts with family or

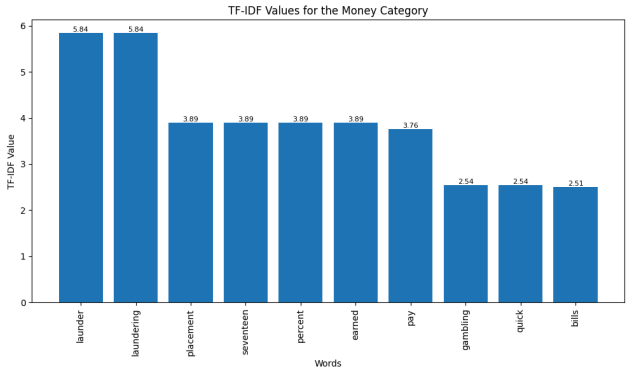
close relatives. Positive example: "Of course. My point being the divorce." The divorce here refers to an internal conflict between a couple and involves *Family*. Negative example: "Um, Marie, this is Ted. My boss." No evident mention of family matters. It rather focuses on Skyler's workplace dynamics, thus it cannot be labeled as *Family*. Edge case: "You really wanna do this now? Are you really gonna make me do this right now?". This may stand out as a candidate to be categorized as *Danger*, but given the context, this dialogue is said by Skyler, who is trying to separate from her husband, Walter. Thus this line is labelled as *Family*.

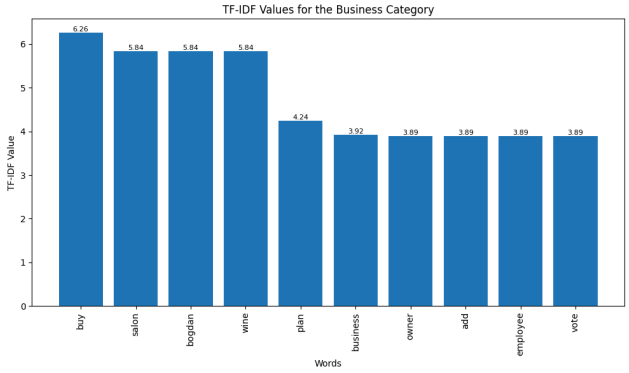
• **Daily Life (dl)**: Dialogues which involve routine or practical conversations, this is used primarily as a more informative "other" category that is still a valid classification. Positive Example: "You're so sure." This dialogue, even in its context of Saul's rant about criminal business strategy, does not offer a specific thematic subject, thus it can be labeled as *Daily Life*. Negative example: "Your medical condition. Has it grown worse?" Clearly the dialogue talks about someone's well-being thus would be labelled *Health*, and could not fall under *Daily Life*. There are no cases where a line might fall under a category apart from *Daily Life* but should still be classified as "dl", because the other categories are more plot-significant and unordinary.

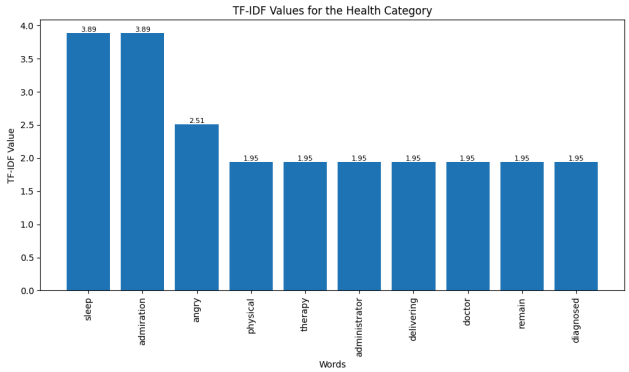
• **Danger (d)**: The dialogues which portray trust issues, imminent threats, or safety concerns. Positive example: "Are we safe?" This portrays a safety concern by the character, thus is a perfect candidate to be labeled as *Danger*. Negative example: "Dinner's almost ready, okay?" No implication of any sort of danger is mentioned in this dialogue, it is a good candidate to be labeled as *Daily Life* but definitely not as dangerous. Edge case: "I told you before. You will not kill Walter White. Not until my business with him has concluded." At first glance it appears that Walter's life is at threat, and the situation involves *Danger* (thanks to the word "kill"). However, the dialogue's primary focus is not on killing Walter White but rather executing a business strategy. Thus it is labeled as business and not danger.

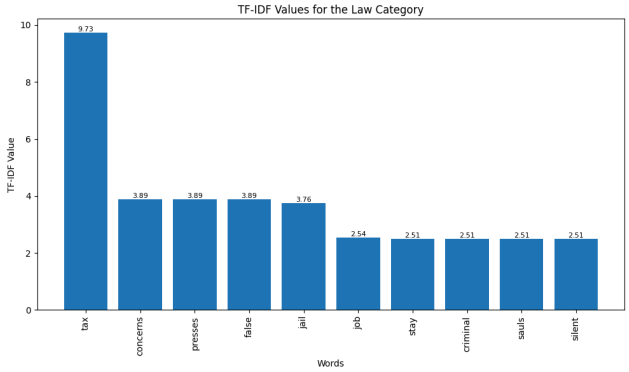
TF-IDF Results

The following histograms depict the top 10 TF-IDF scores for words in each of the 7 theme categories.









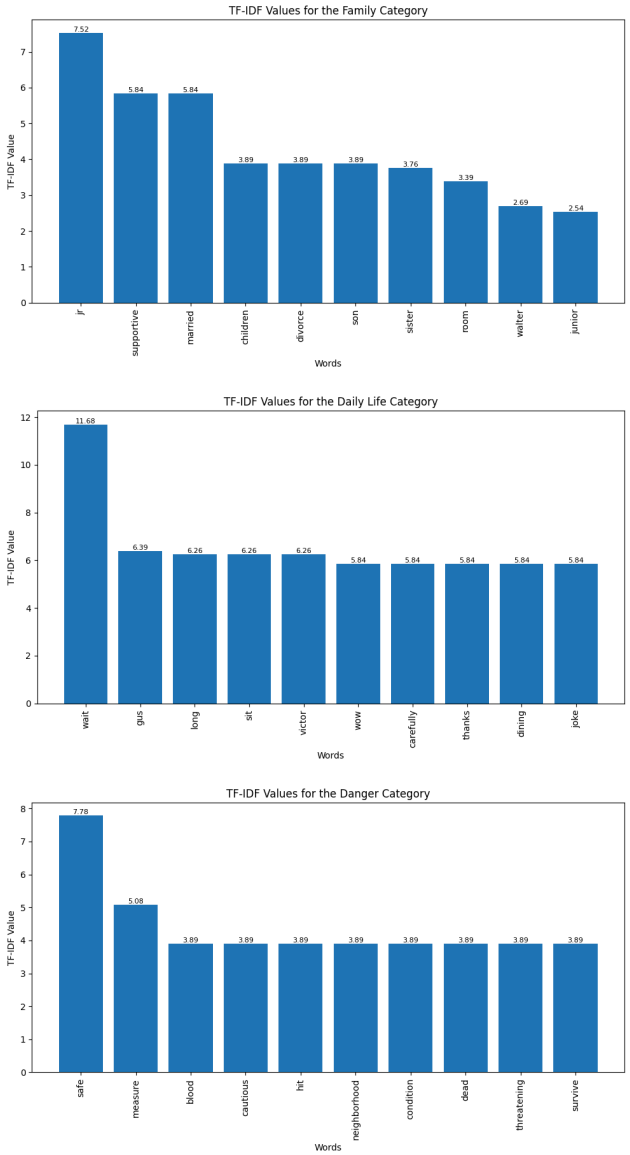


Figure 1: Top 10 words (x-axis) for each theme category sorted by highest to lowest TF-IDF scores (y-axis).

The top two categories in money are launder and laundering which indicates that there was a problem in our tf-idf since it couldn't segregate the noun and verb as the same word. Other than that we notice we have very reasonable top 10 words in each category. The most interesting observation is the words under business category. Words like “bogdan” and “salon” are at the top because Season 3 contains a large amount of content focused on the commencement of a beauty salon business and buying a carwash company from the owner named Bogdan. This indicates that the TF-IDF analysis has yielded sensible rankings, and has done well to catch these trends.

Annotation Results

The pie charts below demonstrate the distribution of categories discussed by each character. Having sufficient knowledge about the show, setting, and its characters, these figures depict a strongly expectable distribution of the categories per character.

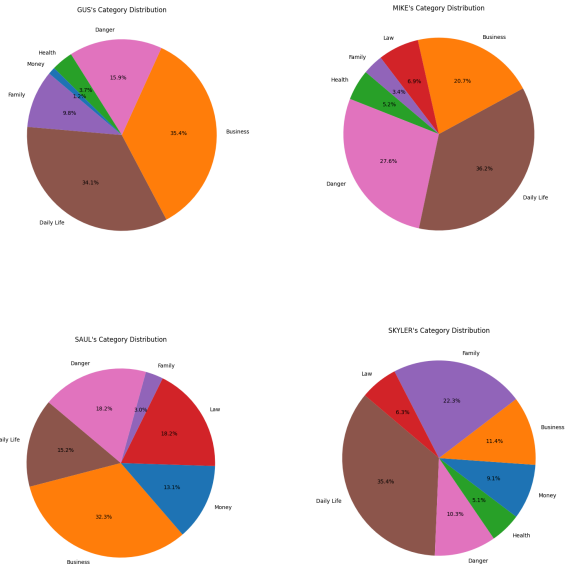


Figure 2: Pie chart breakdown of the percentage of dialogue spoken for each thematic subject, per side character.

Discussion

The analysis of *Breaking Bad* dialogue from *Season 3* reveals interesting patterns from key characters which reflect their roles and personalities. All the pie charts indicate very reasonable and logical insights about which topics were most frequently discussed by Saul, Mike, Skyler and Gus, respectively.

Saul Goodman’s pie chart is the most evenly spread across categories which may seem unusual at first seeing Saul is a practicing lawyer. After reviewing his dialog, the results are sensible as Saul’s expertise lies not only in law (18.2%), but in business (32.3%), money (13.1%), and he’s often associated with dangerous (18.2%) clients such as Mike. Most of Saul’s dialogue is centered around a combination of law, business, money, and danger which explains the nature of the chart. Particularly in the third season, Saul assists Walt in laundering his money using various schemes such as a car wash and a proposed laser tag business. This helps explain why 32.3% of Saul's dialog was business oriented.

Skyler White’s dialogue places a strong emphasis on family (22.3%) and business (11.4%), which reflects her focus on her familial duties in protecting her children and

aiding her husband Walter with his criminal enterprise. Her significant focus on money (9.1%) is mostly due to her helping run Walter's money laundering operation, disguised as a car wash. Skyler's growing concern about herself, Walter, and the rest of the family is shown through her focus on danger (10.3%) and law (6.3%). Overall, these themes paint a character torn between morals and survival as she evaluates turning against Walter as a major risk for the family.

Gus Fring's dialogue is largely dominated by business (35.4%) and danger (15.9%) which fits his role as a sophisticated drug lord disguised as a fast food businessman. His large focus on daily life (34.1%) also reinforces that notion that Gus is not your typical drug kingpin and he spends significant time helping out his employees at his chicken restaurants and doing other tasks such as charity work for a hospital seen in season 3.. Interestingly, Gus places a minimal focus on money (1.2%) which is usually unexpected for Gus's profession. As seen throughout season 3 and bolstered through the pie chart, Gus values power and relationships much more than he does money.

Mike Ehrmantraut's dialogue is heavily focused on danger (27.6%) and business (20.7%). This is logical as Mike is Gus's trusted enforcer and as a former police officer, he is attuned to risks. Something interesting about Mike's pie chart is despite the purpose of his criminal actions being providing for his family and especially his granddaughter, only 3.4% of Mike's dialogue was within the family category. This could be due to Mike rarely verbalizing his personal family struggles which he instead conveys through actions. Mike's dialogue chart reinforces his roles as a professional who prioritizes duty and survival over emotional expression.

Examining the daily life category in the different pie charts reveals an interesting insight. Mike, Gus, and Skyler's daily life dialogue ranged from 34.1-36.2% whereas Saul's was much lower at just 15.2%. After reflecting on Saul's role in season 3 and the rest of the show, we are rarely shown anything about his personal life or even anything about Saul outside of his morally dubious business and legal endeavors. It could be hypothesized that if the same project was performed for *Better Call Saul,* a spin off to *Breaking Bad*, Saul would likely spend more dialogue speaking about daily life and family.

Turning attention towards TF-IDF values, they reveal insights into the main themes per category and it can be inferred which characters contribute to which themes. In the danger category, prominent terms like "safe", "measure" and "threatening" reflect the high stakes environment usually navigated by Mike. The terms align well with his role as Gus's enforcer where he is frequently tasked with neutralizing threats and ensuring operational stability. Looking towards the health category, terms such as "sleep", "therapy" and "diagnosed" tie most directly to Skyler as throughout the season, she frequently reflects her concerns about both Walters physical health and the emotional toll their actions have on the family. The money category features terms like "laundering", and "placement" which can be attributed to Saul's involvement in Walters' money laundering schemes. Specifically, an example that strikes out is Saul walking Walter through the process during a famous scene that took place in a nail salon. Terms like "earned" and "percent" can be attributed to Skyler's in the car wash where she used her accounting expertise to aid Walters enterprise.

In conclusion, the analysis reveals that each character's dialogue aligns closely with their role and personality in *Breaking Bad* Season 3. Gus and Mike focus heavily on business and danger, reflecting their strategic and operational roles in Gus's empire. Saul balances law, business and money, highlighting his multifaceted expertise in navigating criminal and legal systems. Skyler's emphasis on family and money/business highlights her dual roles as a protective mother and reluctant criminal accomplice. The TF-IDF analysis serves as a great indication of key themes in each category. Based on these themes, inferences regarding which characters were most involved with each theme can be made.

Team Member Contributions

This project was completed with significant participation on the part of all team members. The final source scripts were discovered by David. Priyanshu developed the program to convert these scripts into parsable text, which Nicholas executed. Nicholas subsequently developed the program to extract the dialogue lines from target characters in the script and prepare the TSV for open coding and annotation. Priyanshu, David, and Nicholas each conducted an open coding and communicated results to decide on the final typology. After which, each team member conducted their own annotations in order to finalize this phase. For analysis, Nicholas developed the script to calculate and rank TF-IDF scores for words across all thematic categories. David used this data to create histogram visualization of highly ranked words for each theme. Priyanshu created pie chart visualizations for each character's dialogue distribution across topics. Regarding the final report, Nicholas wrote the Abstract, Data, and Methods sections. Priyanshu wrote the Introduction and Results sections. David wrote the Discussion section and compiled the References. All members peer-reviewed the paper together and approved this contributions section.

References

Published Scripts

Gilligan, V. 2010. Breaking Bad, Season 3 Scripts. Retrieved from https://bulletproofscreenwriting.tv/breaking-bad-tv-script-download/.