



ASSIGNMENT COVER SHEET

**THIS FORM MUST BE AT THE FRONT OF EACH ASSIGNMENT
CANDIDATES MUST KEEP A COPY OF THEIR ASSIGNMENT**

Candidate to complete the following section (and update details in header and footer):

| | |
|--|------------------------------|
| Member ID: | COURSE: DSA |
| DATE DUE: Monday, 19 April 2024 at 12:00pm (AEST) | |

PLAGIARISM

By submitting your assignment, you are implicitly stating that the work is your own.

Remember that an important aspect of being a professional actuary is to always act with integrity. Committing plagiarism by copying another person's work or not properly referencing other sources used in your assignment is a breach of the Integrity principle under the Actuaries Institute's Code of Conduct.

Any suspected plagiarism will be referred to the Institute's Executive General Manager, Education for review. Depending on findings, a complaint regarding the member may be made to the Institute's Conduct Committee. Subject marks may not be released until the matter is resolved.

Be aware that your assignment may be vetted using Turnitin.



Question 4 – Executive Summary

4.1 – Summary

Movie scripts containing offensive language or taboo topics present a potential risk to the company due to stipulations in some actor contracts. Unknowingly including problematic lines in scripts may result in significant legal costs, expensive recasts and reshoots, as well as strained actor relationships. Therefore, a statistical model was built to assist with identifying problematic movie quotes. Ultimately, it was found that the model was *too risky* to implement as it was not able to identify problematic quotes to an acceptable standard.

A model that accurately identifies problematic quotes could significantly reduce legal and production risks, however the current model's performance falls short. The model missed a material proportion of problematic quotes (50%), leaving them undetected and at risk of breaching actor contracts. Relying solely on this model to review movie scripts would expose the company to significant legal risk. Possible consequences of contract breaches include unforeseen legal costs, delays due to recasts, rewrites and reshoots, or actors boycotting our movies. Therefore, naively using this model without safeguards is likely to have a negative impact on our bottom line and operational efficiency.

Most of the quotes flagged as problematic by the model had no issue and were clean (86%). While mislabelled quotes can be easily verified by humans, having to manually check model outputs at scale is likely to be arduous and resource intensive. This is costly and may contribute to delays in the scriptwriting process. Overall, the level of manual verification the model requires undermines the goal of automation and suggests the model is impractical from a cost-benefit standpoint.

Our statistical model performed similarly to a simpler approach and was limited in detecting more nuanced problems in quotes. As a benchmark, we compared our model to a keyword-based classifier that searched for the presence of certain words such as expletives. The benchmark model achieved similar results to the statistical model, with both identifying 50% of problematic quotes. Interestingly, the models were better at identifying different types of issues: the keyword approach caught expletives, while our model captured rude tone. This suggests that the models could be combined to improve coverage. However, both approaches struggled to accurately identify nuanced topics like religion and politics, a key component of actor stipulations. Relying solely on our model exposes us to a higher risk of contract breaches where actors have conditions around these taboo topics. Moreover, addressing these sensitive issues in scripts likely requires more complex and time-consuming rewrites than simply replacing explicit terms. Therefore, accurately flagging such quotes early in the script development process holds far greater value than identifying rude tone and expletives. This highlights that our current model has limited usefulness for the most critical aspect of the business objective.



In conclusion, deploying the current model poses an unacceptably high risk of breaching actor contracts and is not justified from a cost-benefit perspective. Further exploration is needed to develop a reliable system for identifying problematic quotes, potentially by combining approaches and focusing on detection of sensitive topics.

Word Count: 493

4.2 – Limitations and Response

Limitations during the modelling phase contributed to the unsuitability of the final model. We identify these limitations as well as actions that could be taken to overcome them now and in future modelling projects.

Limitation 1 – Data Quantity

To effectively identify problematic quotes, models require massive amounts of data to understand the complexity of language and diverse topics. Our model was trained on a small dataset, hindering its ability to grasp these complexities, and contributing to its poor performance. There are several aspects that can make a quote problematic, but the dataset contained only a few quotes for each category, further exacerbating this data scarcity issue. The model may perform well at identifying problematic quotes similar to those it was trained on, but struggle with the new and original quotes created by our scriptwriters. This will have a direct impact on our business goals as inaccurate identification of problematic quotes may result in significant legal penalties from contract breaches, eroded trust with actors and wasted resources from investigating quotes with no contractual risk.

There are several mitigation strategies to reduce the impact of this limitation in future projects. While acquiring a massive amount of text data could address the data quantity issue, data is expensive to store and time-consuming to clean and label. A more cost-effective approach might be to leverage pre-trained language models from companies like OpenAI or Google. These models are trained on vast amounts of data and can be fine-tuned for our specific need. This approach may also achieve superior accuracy compared to a model built entirely in-house. The data gap is a critical limitation of the model and future projects should focus on addressing it to improve modelling outcomes.

Limitation 2 – Interpretability

Currently the model simply outputs whether it believes a quote is problematic to flag to scriptwriters that there may be an issue. However, as the flag is binary it does not offer any detail on what the nature of the problem is (e.g. politics, religion) and so does not provide clear direction to scriptwriters. While some flags may be obvious, others arising from nuanced topics like politics may generate ambiguity. This may lead to wasted time and frustration from attempting to decipher the output as well as possibly mistaken revisions. Moreover, actors may have stipulations concerning only certain topics and so potentially



problematic quotes may not affect all actors. Therefore, more specific output labels pertaining to the nature of the quote's problem would facilitate better informed decision-making.

This limitation could be overcome by building a model with more granular output labels. Alternatively, it may be more informative to build models for each problem label as these categories may not be mutually exclusive with quotes having multiple problematic elements. It has already been demonstrated that a keyword-based classifier is effective for identifying quotes with explicit terms. As an interim solution, a human expert could provide a more nuanced classification for flagged quotes that can be passed on to the relevant parties.

Limitation 3 – Data Quality and Relevance

Our model training relied on unreliable data labels generated by third-party software. While efficient, this method lacked control and oversight, leading to labels that were inconsistent with the stipulations in our actual actor contracts. Essentially, many quotes externally labelled as problematic were extremely unlikely to constitute a breach of any contracts, and vice versa. This biased and flawed data resulted in similarly unreliable model outputs. Consequently, the model was likely to miss quotes that could lead to contract breaches, which holds significant risk of financial repercussions.

To address this issue in the future, the legal team should set out clear criteria for identifying quotes that could lead to violations across different stipulation types. Humans should review quote labels in the dataset based on these criteria. This approach will strengthen model governance and ensures the model outputs align with our contracts, reducing the risk of legal breaches. While manual labelling incurs costs, these are dwarfed by the potential reputational and legal costs associated with a contract breach, so we believe these measures are justified.

Limitation 4 – Model Complexity

Our model was relatively simplistic compared to the complexity of different problem areas in quotes. The simple model structure was a result of limited data and computing power. This inconsistency between the problem and model complexity means that the model performs poorly at recognising subtle references to taboo topics. While the model may be able to identify blatant or simple issues with quotes, these obvious problems are easy for scriptwriters to identify or disclose themselves. Alternatively, there would be more benefit to the business if the model could automatically review quotes and pick up subtle issues that may have been overlooked by scriptwriters. Therefore, the simplicity of our model limits its usefulness, and it cannot be relied upon to thoroughly screen movie scripts for possible contract breaches.

While our model is not recommended for deployment, any model built for this task is likely to simplify reality and fail to pick up on all nuances in quotes. This highlights the need to have additional checks on scripts to mitigate the risk of contract breaches. To manage model



risk, it is recommended that scriptwriters self-disclose any potential problems since they have intimate knowledge of the content of their scripts. Clear criteria of what may result in contract breaches is needed to ensure this measure is effective. This process will assist with guarding against quotes missed by a model, as well as help monitor model performance and inform future improvements.

In the future, a more complex model could be built to reduce the impact of this limitation. However, this will likely require the company to invest significant funds and resources, such as for cloud computing services. The board may want to explore interim solutions like those discussed above before committing to such a project.

Word Count: 970

4.3 – Video Summary Link

<https://youtu.be/bkVVrFYG5dc>