



CS-401

Applied Data Analysis

Themes and semantics

Authors

András Horkay

Introduction

Humour provides valuable insight into how societies interpret the world around them; what they find acceptable, absurd, or taboo. The *New Yorker Caption Contest* offers a distinctive opportunity to examine these dynamics. Since 2005, readers have been invited each week to submit captions for a single cartoon, competing to amuse both editors and other participants. With thousands of entries submitted and ranked, the contest constitutes a rich dataset that captures how humour functions in everyday language and how perceptions of what is “funny” have evolved over time.

This exploratory data analysis investigates how thematic and lexical patterns influence humour perception within the contest's captions and how these patterns have changed across two decades. The study focuses on four core dimensions: differences in word usage between winning and losing captions; the frequency and average score of specific entities such as people, places, and objects; the distribution of thematic categories including politics, religion, and relationships; and temporal trends in the representation of taboo topics such as violence, sex, and death.

Together, these perspectives provide a multidimensional understanding of how humour both reflects and responds to cultural transformation. By combining linguistic, thematic, and temporal analyses, the project seeks to establish how the captions of the *New Yorker Caption Contest* mirror prevailing social values, tensions, and sensitivities. It therefore asks how thematic and lexical choices have shaped humour perception over time, and what role taboo or socially sensitive themes play in determining caption success.

By linking word use, thematic framing, and temporal evolution, this study aims to demonstrate how humour functions as a social barometer which reveals what audiences find amusing, what they reject, and how these boundaries shift across two decades of cultural life.

Methodology

The dataset for this analysis contains the images from said captions, with no temporal data. Therefore, the first step was to scrape the contest's website to obtain the dates associated with each caption: most dates could be found online, and others were inferred based on the sequence of contests. This allowed for a temporal analysis of the captions, but must be treated with caution due to potential inaccuracies in the inferred dates. The dataset also contains captions submitted to each contest, along with their rankings and total received votes. The votes are either funny, somewhat funny or not funny. Finally, the access to the metadata of each contest is also available: data such as a short description of the cartoon, the location of the cartoon, the number of votes and the number of captions submitted. This file must be treated with caution as data is only available for the first 230 contests, while for the final 150 contests, the data is empty.

Results

Wordcloud of the most frequent words in winning and losing captions

The aim of this topic was to find the most frequent words in the most and least funny captions. To judge how funny a caption is, a new metric was introduced:

$$\text{Funniness_score} == \frac{\omega_1 \cdot \text{funny} + \omega_2 \cdot \text{somewhat_funny} + \omega_3 \cdot \text{not_funny}}{\text{number_of_votes}}$$

where the weights

$$\omega_1 = 1.863 \quad \omega_2 = 0.308 \quad \omega_3 = -1.171$$

were found using a linear regression. Using this, a funniness score is attached to each caption, which will be used to order comments on how funny they are.

After identifying common stopwords in english, the captions are preprocessed. This means a new column is created in each dataframe called 'tokens'. Fig. 1 shows the twenty most common words that occur in all captions. Similarly, Fig. 2 displays the words which occur at least three times in the top 100 comments but not in the worst 1000 comments.



Figure 1: The most frequent words in all captions.

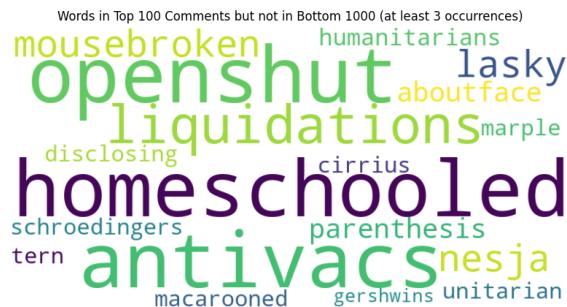


Figure 2: Words that occur in the 100 top performing comments at least 3 times, but not in the least performing 1000 comments.

Frequency of entities in the captions and their average score

Discussion

Conclusion

References

Appendix