# Machine Learning Regression Assignment

Huang HuiYi
Meneghini Jade

# Predicting Popularity of eCommerce Reviews

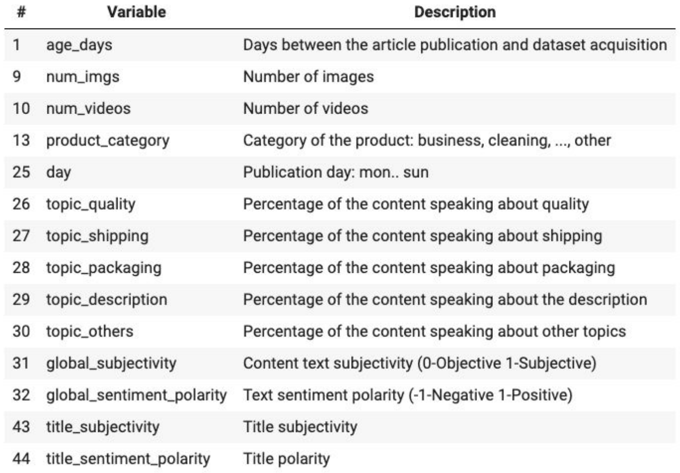
*Objective:*

Predict number of shares on product comments with the end goal of managing potential negative feedback.

# Variables of interest based on: Context

From our understanding of the variables:

- The number of **days elapsed** from publication might impact the number of shares.
- Including **images** and **videos** may attract more attention from viewers.
- The relatable **topics** might push people to draw parallels with their lives.
- **Subjective** and **sentiment-driven titles** might enhance empathy among readers.



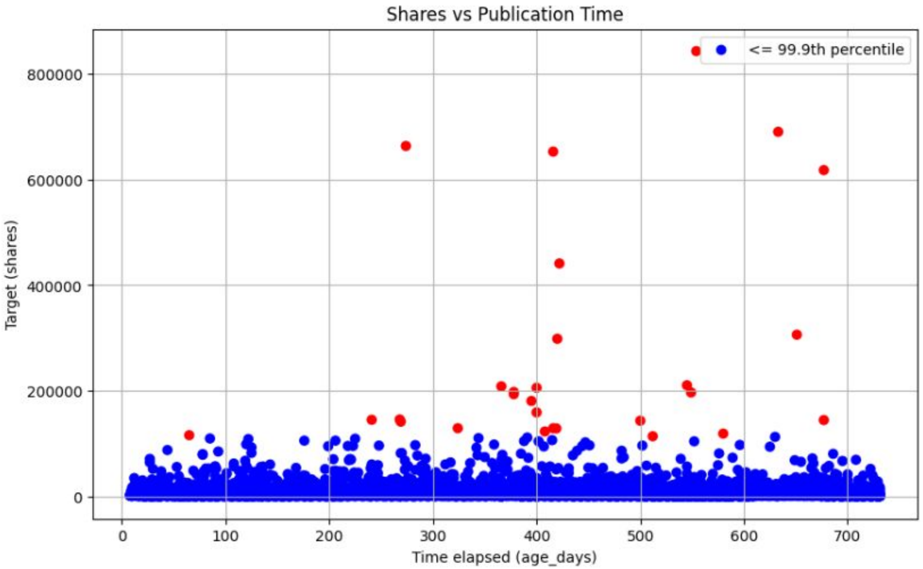
| # | Variable | Description |
|---|---|---|
| 1 | age_days | Days between the article publication and dataset acquisition |
| 9 | num_imgs | Number of images |
| 10 | num_videos | Number of videos |
| 13 | product_category | Category of the product: business, cleaning, ..., other |
| 25 | day | Publication day: mon.. sun |
| 26 | topic_quality | Percentage of the content speaking about quality |
| 27 | topic_shipping | Percentage of the content speaking about shipping |
| 28 | topic_packaging | Percentage of the content speaking about packaging |
| 29 | topic_description | Percentage of the content speaking about the description |
| 30 | topic_others | Percentage of the content speaking about other topics |
| 31 | global_subjectivity | Content text subjectivity (0-Objective 1-Subjective) |
| 32 | global_sentiment_polarity | Text sentiment polarity (-1-Negative 1-Positive) |
| 43 | title_subjectivity | Title subjectivity |
| 44 | title_sentiment_polarity | Title polarity |

# Deep dive into days elapsed

Typically, there is a correlation between the **days since publishing** and the **number of shares**.

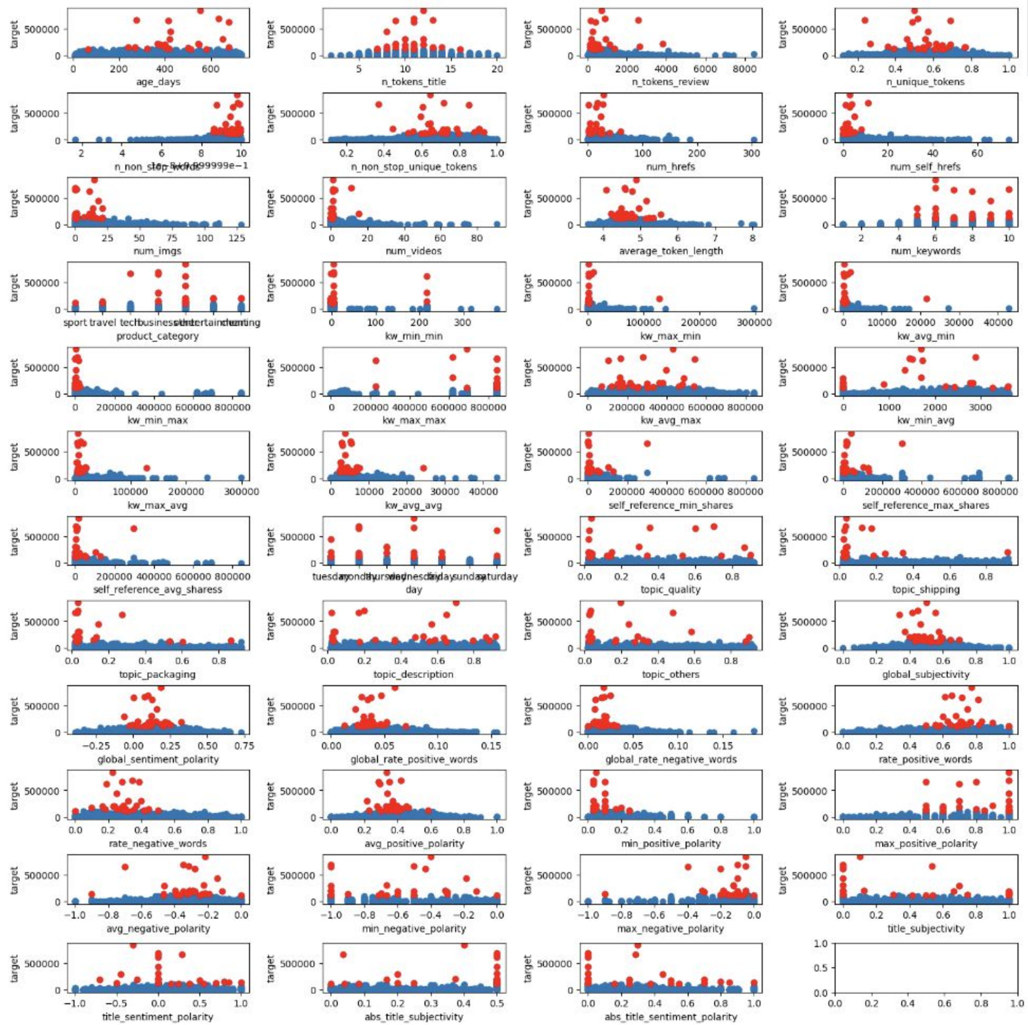
In this case, that correlation is not present.

Therefore, we will focus on the **outliers** in the data, which might be **viral comments,** and investigating the underlying reasons for this phenomenon.

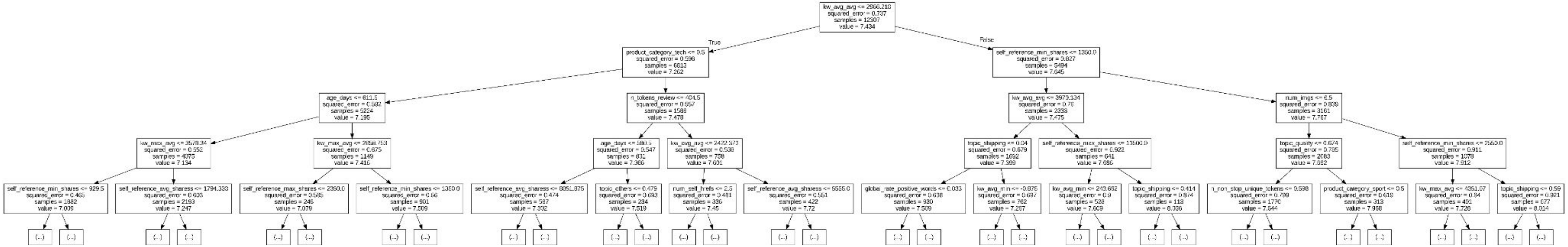


# How to craft a viral comment?

From all the numerical variables, there's no obvious pattern for viral comments, as they are **highly connected with the content of the comments.**



# RANDOM FOREST MODEL | *gs_99_2*

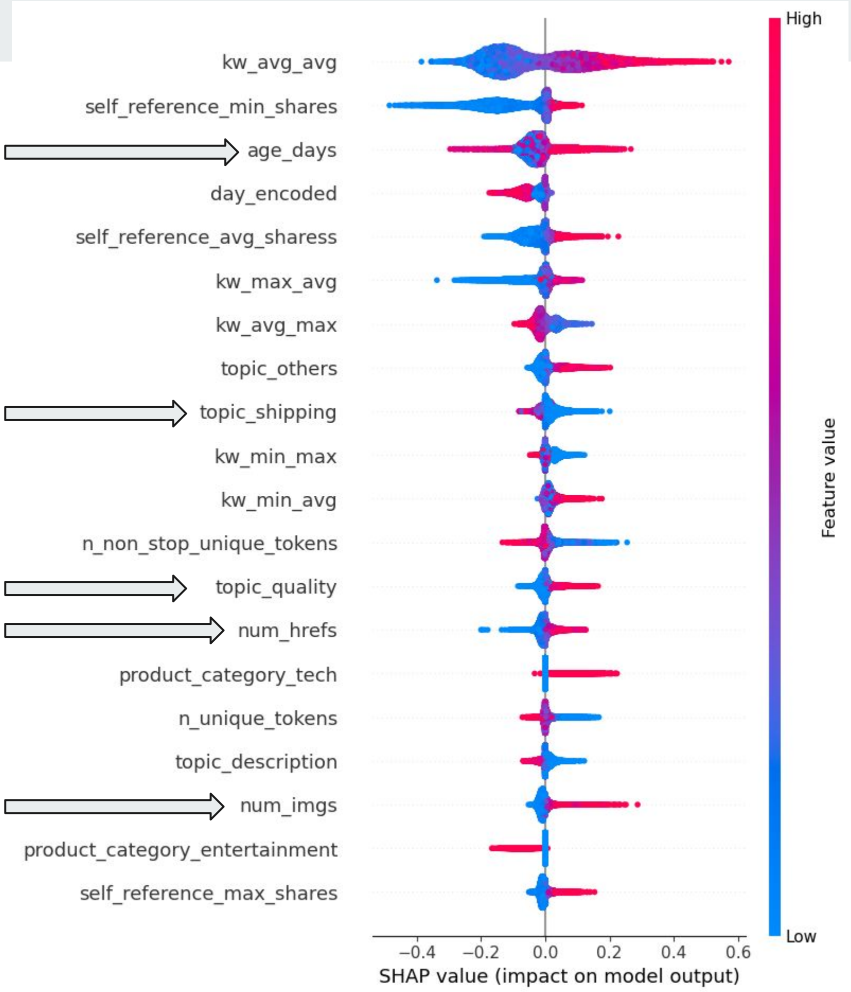


By filtering out **viral** comments (1% of comments), we found the patterns that **popular** comments follow.

# What makes comments *popular*?

Some variables impact  shares more than others.

- Days after publication
- Percentage of the content about shipping
- Percentage of the content about quality
- Number of links
- Number of images



# Project Limitations

- Excluding two categorical columns, all the variables are **numerical variables**.
- The **data lacks detailed context knowledge** regarding the variables, which may impact the depth of the analysis.
- It's unclear whether the **data has undergone pre-processing** that could affect the model.

# Can we know for sure how many shares a comment will gain?

***Predictive model gs_99_2***

While the model provides some predictions of how many shares a comment will receive, there is space for improvement.

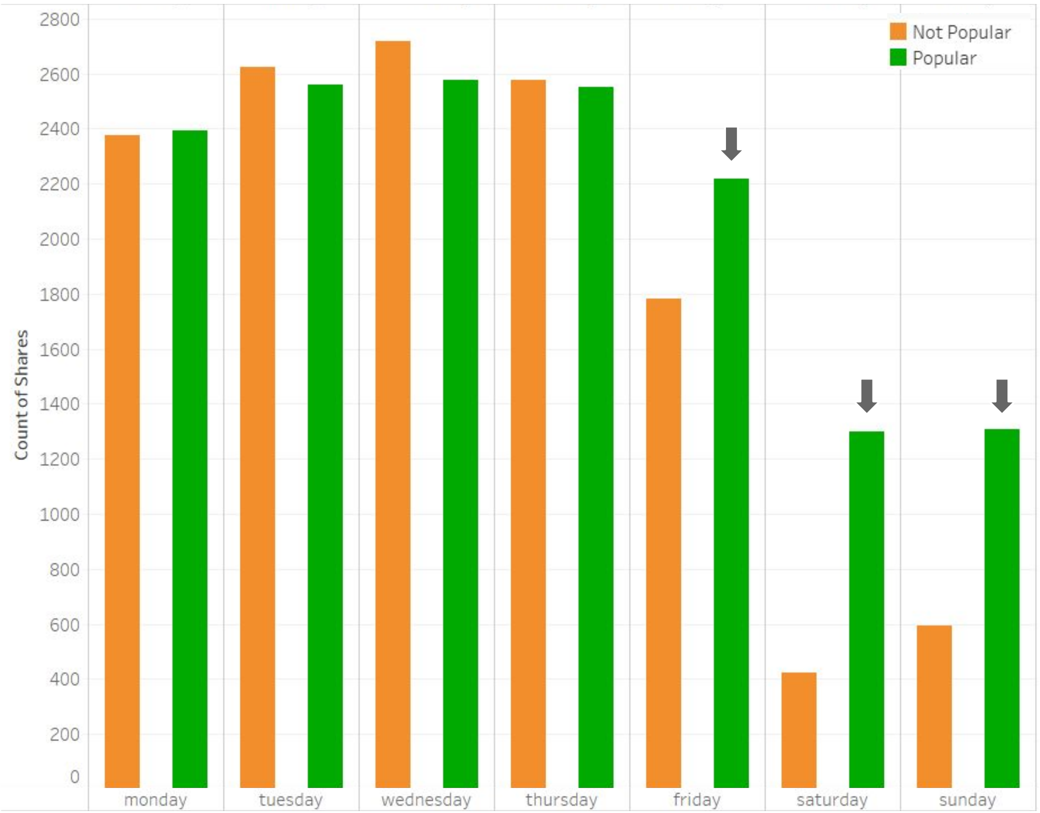
Areas to further investigate:

- data imbalance
- user behaviour

# “Popularity” what is it?

# +1400 shares

This is the median of the shares gained by comments in the past 2 years



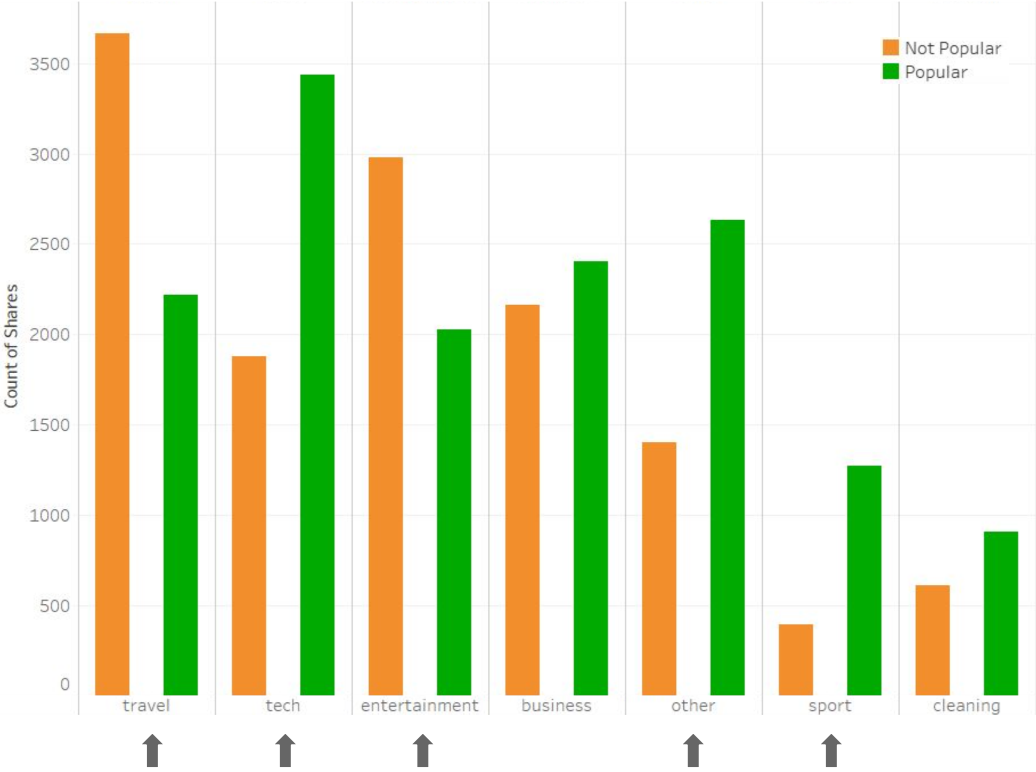
# Popular comments VS Day of the Week

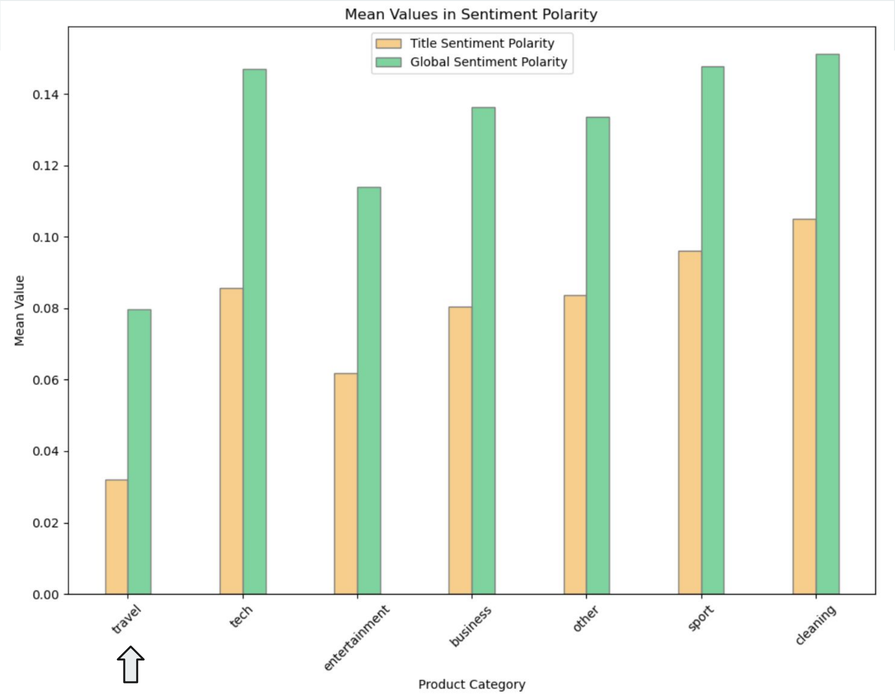
- Users are less active on the weekend
- Comments posted on the weekend are more likely to gain popularity
- Most comments are posted week days

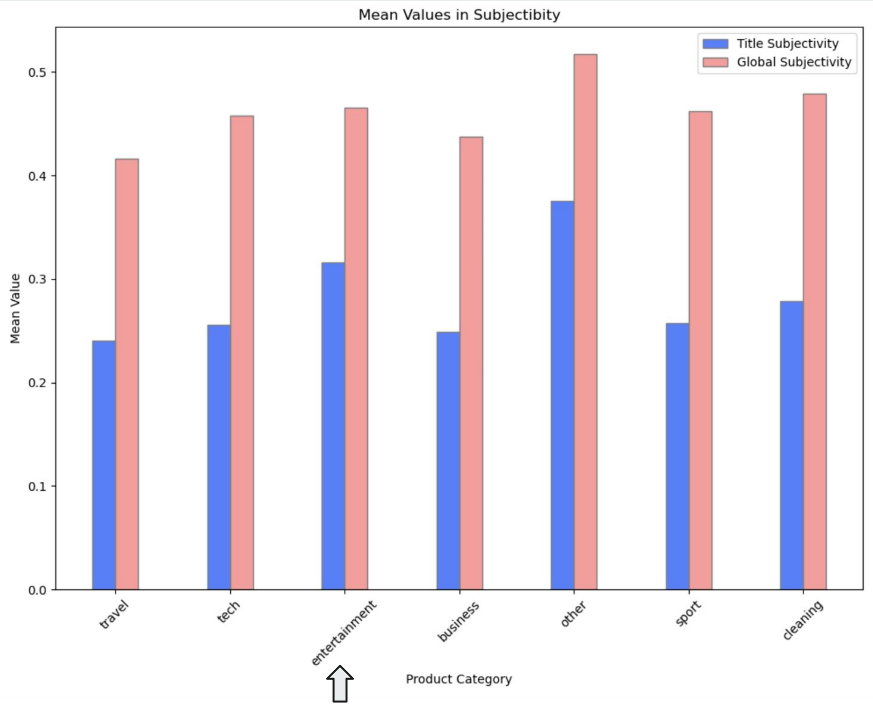
# Popular comments VS Product categories

- **Travel** and **entratainment** comments are  less likely to become  popular
- **Tech, sport** and **other** comments are likely to become popular
- Category representation is uneven

*Note: this is based on the assumption that all these services started to be offered at the same time.*







- **Travel** comments have lower avg. sentiment polarity, which means that they have more **negative words**

- **Entertainment** comments have higher possibility to have **subjective titles**

The **subjectivity and sentiment polarity** in each category are similar.

**As we mentioned before, viral comments are highly influenced by the content of the comments.**

so…

do we need to know the **exact numbers of shares** a comment will gain?

# Not really.

**Objective**: [ … ] **managing potential negative feedback.**

To do this, it is enough to create a model able to predict if a comment will become infamous or not instead of the exact numbers of shares.

# Thank you!