ASSIGNMENT 2: REPLICATION STUDY by DONATO SCARANO

In this assignment we are replicating the study conducted by Muller at al 16: https://canvas.ltu.se/courses/21054/files/3591424?wrap=1 to predict the helpfulness of online customer reviews. We are using just as in the original study the reviews for the video games category.

1.1 RESEARCH QUESTION

What we want to address is the question of 'What makes a helpful online review?' (Mudambi &Schuff, 2010). We are building a predictive model for reviewing helpful news that can be valuable in many practical and theoretical contexts from proper sorting to filtering to understanding how to write effective reviews and discovering hidden relationships between features.

1.2 DATA COLLECTION

We have pre-processed and transformed the data to restrict the focus on valuable information, remove duplicates and reduce the dataset.

Preprocessing steps include:

Importing the relevant libraries: pandas, matplotlib, seaborn, gzip and json

Importing the json review dataset

Create a data frame from the list of dictionaries.

Check for duplicate reviews based on a subset of columns.

Count the number of unique reviews.

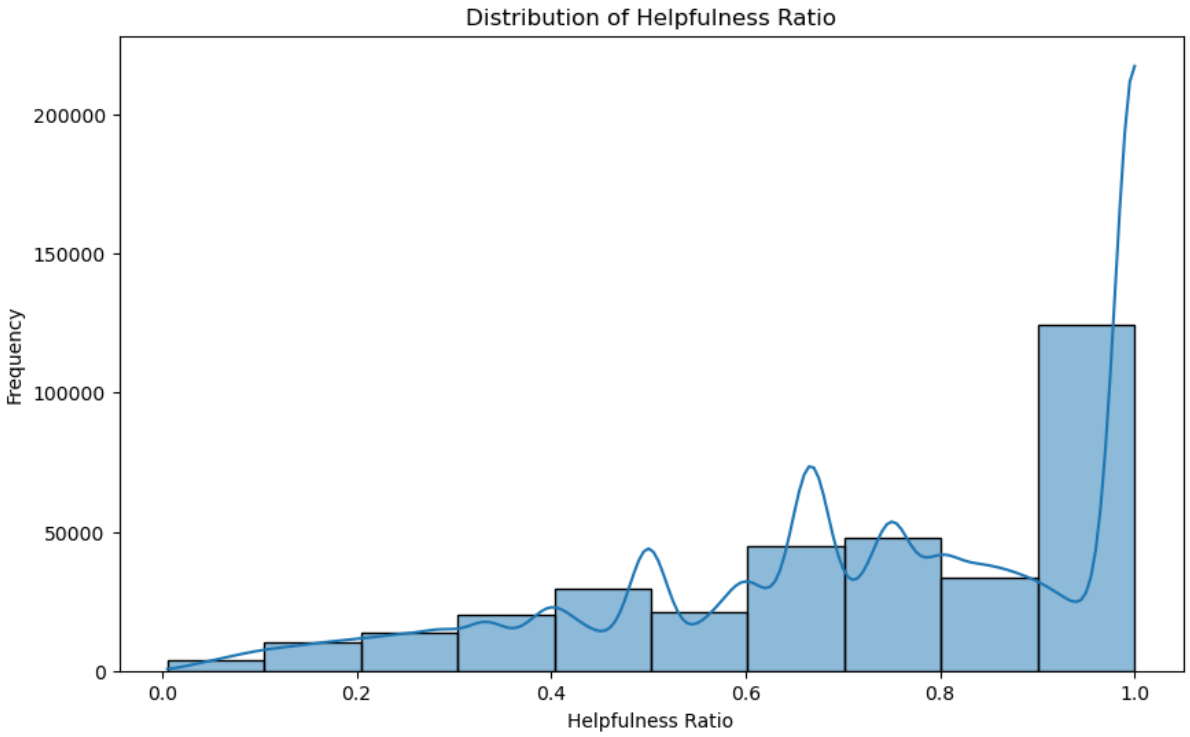
Count the frequency of each helpfulness ratio value to get an overview of helpfulness rating.

Preprocessing shows that over half of the reviews do not have any helpfulness rating (neither positive nor negative).

We have therefore created a function "filter_helpful" to exclude reviews with less than two helpfulness ratings to increase the reliability of the analysis.

We have also created a function called "calculate_helpfulness" to calculate the helpfulness ratio by dividing the number of helpful_votes by the number of total votes.

By plotting the distribution of helpfulness ratio values, we noticed some extreme distribution happening and to avoid statistical concerns arising from this we have dichotomised the review helpfulness variable (i.e., reviews with a helpfulness ratio of > 0.5 were recorded as helpful, and reviews less than 0.5 recorded as not helpful).



1.3 DATA ANALYSIS

Early research focused mostly on the overall rating or the star rating although recently many studies have started to analyze the text of the reviews (e.g., Mudambi & Schuff, 2010;Cao et al, 2011; Ghose & Ipeirotis, 2011; Pan & Zhang, 2011; Korfiatisa et al, 2012).

To capture the review content and its impact on the helpfulness of the review we have used probabilistic topic modelling using LDA (Latent Dirichlet Allocation) algorithm.

Probabilistic topic models are unsupervised algorithms that annotate the documents with topic labels (Blei et al, 2003; Blei, 2012).

The foundational idea is the distributional hypothesis of statistical semantics; words that occur together in similar contexts tend to have similar meanings (Turney & Pantel, 2010).

The step undertaken to achieve the above are the following:

Import relevant libraries (nltk, gensim, string, panda).

Preprocess the text using nltk ‘stopwords’ ad ‘punkt’.

Create a dictionary representation of the document.

Convert the collection of texts to a bag of worlds.

Training the LDA model.

Annotating each review with a vector of topic probabilities.

Create and format a new dataframe to store the separated values.

Remove NaN values by replacing them with an empty string.

To train the predictive model we used random forests.

Random forests are an ensemble supervised-learning technique that can process high dimensional data sets and is robust against data anomalies.

A random forest model is constructed by generating a multitude of decision trees based on boot-strapped sub-samples such that only a random sample of the available variables at each split of the tree is considered a potential split candidate (Breiman, 2001a).

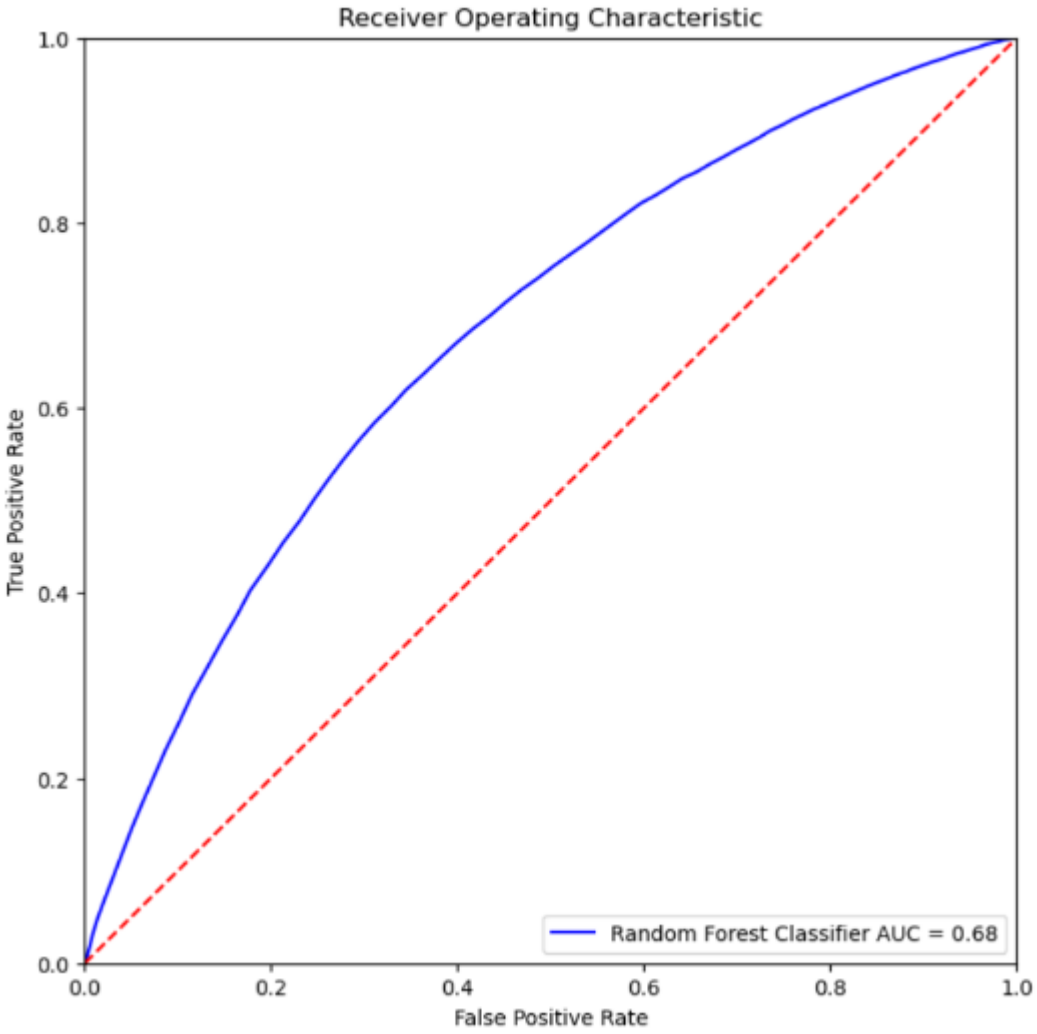
We use the implementation provided by the scikit-learn Python package and set the number of trees to 128.

The model accuracy achieved: Forest Score: 0.7731915930027651

Receiver Operating Characteristic (ROC) Curve show the predictive performance of our classification on the holdout test set (20% of the overall dataset).

It plots the true positive rate against the false positive rate.

The area under the ROC Curve amounts to 0.68 which means that the model has an accuracy of 68% in distinguishing between a randomly drawn helpful review and a non-helpful one.



The only way to interpret a random forest model are the variable importance measures.

We know the most influential variables for predicting the helpfulness of a review.

The most important variables are ranked below:

Feature ranking:

1. Feature 1 (unixReviewTime): 0.219470

2. Feature 2 (0): 0.201327

3. Feature 3 (1): 0.187303

4. Feature 4 (2): 0.142838

5. Feature 5 (3): 0.092845

6. Feature 0 (overall): 0.058151

7. Feature 6 (4): 0.052152

8. Feature 7 (5): 0.024311

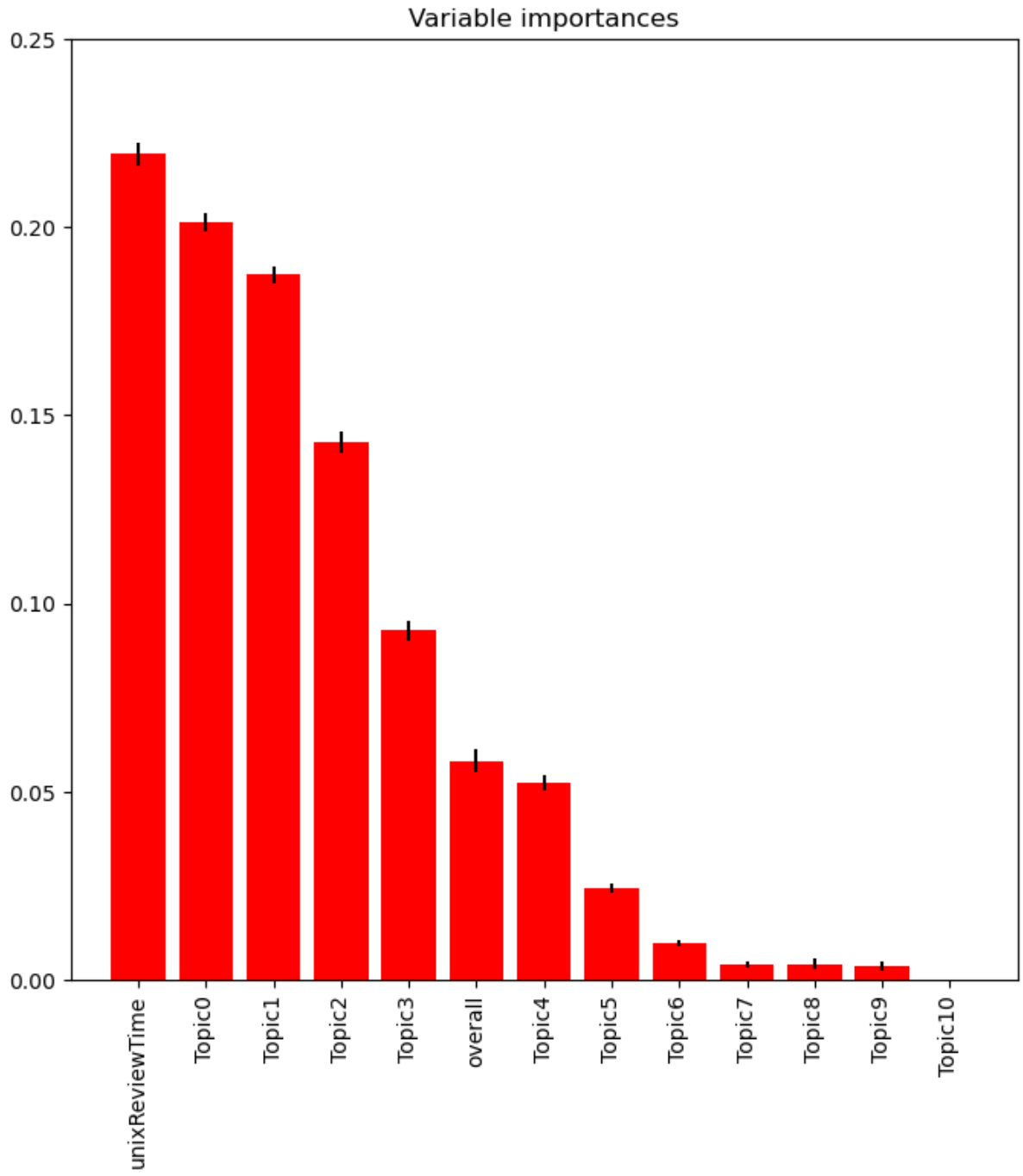
9. Feature 8 (6): 0.009646

10. Feature 9 (7): 0.004126

11. Feature 11 (9): 0.004082

12. Feature 10 (8): 0.003748

13. Feature 12 (10): 0.000000



We must empirically triangulate the LDA results, that is, the per-topic word distributions and the per-document topic distributions.

In a first step, we employed a word intrusion task to measure the semantic coherence of topics.

Since topics are represented by words that co-occur with high probability, the idea behind the word intrusion task is to insert a randomly chosen word (intruder) into a set of words representative of a topic and ask human judges to identify the intruder.

For each topic, we generated five randomly ordered sets of six words: the five most probable words for the given topic plus one randomly chosen word with low probability for the respective topic.

Word Set 1: ["'ask", 'one', 'xbox', "n't", "'s", 'ps4']
Word Set 2: ['one', "n't", 'xbox', 'ps4', "'s", 'moombas']
Word Set 3: ["n't", 'xbox', 'right.c', 'ps4', 'one', "'s"]
Word Set 4: ['xbox', "'s", 'one', 'capacity.graphically', 'ps4', "n't"]
Word Set 5: ['one', "'s", 'xbox', "n't", 'immuersion', 'ps4']
Word Set 6: ['game', "n't", 'play', 'get', 'like', 'valgas.my']
Word Set 7: ['turd.but', 'game', "n't", 'play', 'get', 'like']
Word Set 8: ['get', 'game', 'play', 'like', "n't", 'stealth.']
Word Set 9: ['came.so', 'get', 'play', "n't", 'like', 'game']
Word Set 10: ['keuth', 'like', 'play', 'get', 'game', "n't"]
Word Set 11: ['34', 'pokemon', 'unit.now', '8220', '8221', 'simcity']
Word Set 12: ['8220', 'pokemon', '34', '8221', 'ventrillo.needs', 'simcity']
Word Set 13: ['34', 'simcity', 'bobbled', '8220', 'pokemon', '8221']
Word Set 14: ['8221', 'pokemon', 'simcity', 'however.before', '34', '8220']
Word Set 15: ['8220', '8221', 'simcity', '34', 'pokemon', 'orange.please']
Word Set 16: ['cars', 'mode', 'like', 'disney.com', 'game', "'s"]
Word Set 17: ['cars', 'like', 'game', 'headcams', "'s", 'mode']
Word Set 18: ['game', "'s", 'mode', 'cars', 'update.in', 'like']
Word Set 19: ['like', 'mode', "'s", 'alivei', 'game', 'cars']
Word Set 20: ["'s", 'mode', 'like', 'cheerleader-routines', 'cars', 'game']
Word Set 21: ["n't", '--', 'play', 'weapons+detailed', 'game', '``']
Word Set 22: ['--', 'only-multiplayer', "n't", 'play', '``', 'game']
Word Set 23: ['game', 'purchase.simply', 'play', '``', '--', "n't"]
Word Set 24: ['drawback.recommended', "n't", '``', 'play', '--', 'game']
Word Set 25: ['--', '``', 'paced.you', 'game', 'play', "n't"]
Word Set 26: ["'s", 'game', "n't", '``', 'games', 'forwardy']
Word Set 27: ["'s", 'far.fans', '``', 'game', "n't", 'games']
Word Set 28: ["'s", 'games', '``', 'game', "n't", 'hacking/database']
Word Set 29: ["'s", 'speculated', 'game', "n't", '``', 'games']
Word Set 30: ["n't", 'game', 'games', '``', "'s", 'dumb-heads']
Word Set 31: ['mario', 'itdoes', 'wii', 'u', 'nintendo', 'games']
Word Set 32: ['u', 'mario', 'wii', 'nintendo', 'fun.contra', 'games']
Word Set 33: ['wii', 'u', 'mario', 'touch.really', 'nintendo', 'games']
Word Set 34: ['nintendo', 'mario', 'games', 'box-', 'u', 'wii']
Word Set 35: ['wii', 'mario', 'nintendo', 'enourmes', 'games', 'u']
Word Set 36: ['much.enemys', 'multiplayer', 'cod', 'battlefield', 'campaign', 'headset']
Word Set 37: ['simluations', 'campaign', 'cod', 'multiplayer', 'battlefield', 'headset']
Word Set 38: ['multiplayer', 'headset', 'shot.thank', 'cod', 'campaign', 'battlefield']
Word Set 39: ['-rolls', 'battlefield', 'campaign', 'headset', 'cod', 'multiplayer']
Word Set 40: ['battlefield', 'campaign', 'multiplayer', 'headset', 'marine/alien', 'cod']
Word Set 41: ["'s", 'like', 'cord-nest', 'get', "n't", 'game']
Word Set 42: ['game', "n't", "'s", 'like', 'playeranyway', 'get']
Word Set 43: ["'s", 'get', 'like', 'game', 'booma', "n't"]
Word Set 44: ['response.buyer', "n't", 'game', 'get', "'s", 'like']
Word Set 45: ['get', 'game', 'like', "n't", "'s", 'screen.beware']
Word Set 46: ['62', 'tales', 'card', 'cards', 'memory', 'premote']
Word Set 47: ['hughely', '62', 'tales', 'cards', 'memory', 'card']
Word Set 48: ['card', '62', 'tales', 'memory', 'biessman', 'cards']
Word Set 49: ['grab-itas', 'card', 'tales', 'memory', 'cards', '62']
Word Set 50: ['cards', 'memory', '62', 'tales', 'card', "'upgraded"]

We will present these sets to three independent human coders via the crowdsourcing platform Amazon Mechanical Turk and prompt them to identify the intruder.

In a second step, we conduct a best topic task to validate the topic assignments for each review. (The task is a variation of the topic intrusion task developed by Chang et al (2009).

Instead of identifying an intruder among a set of highly probable topics, we chose to identify the best match of a topic.

Topic Set for Review 1: [4, 1, 5, 3] ... (truncated)
Topic Set for Review 2: [0, 1] ... (truncated)
Topic Set for Review 3: [3, 8, 1, 5] ... (truncated)
Topic Set for Review 4: [1, 3, 4, 5] ... (truncated)
Topic Set for Review 5: [1, 4, 0] ... (truncated)

1.4 RESULT INTERPRETATION

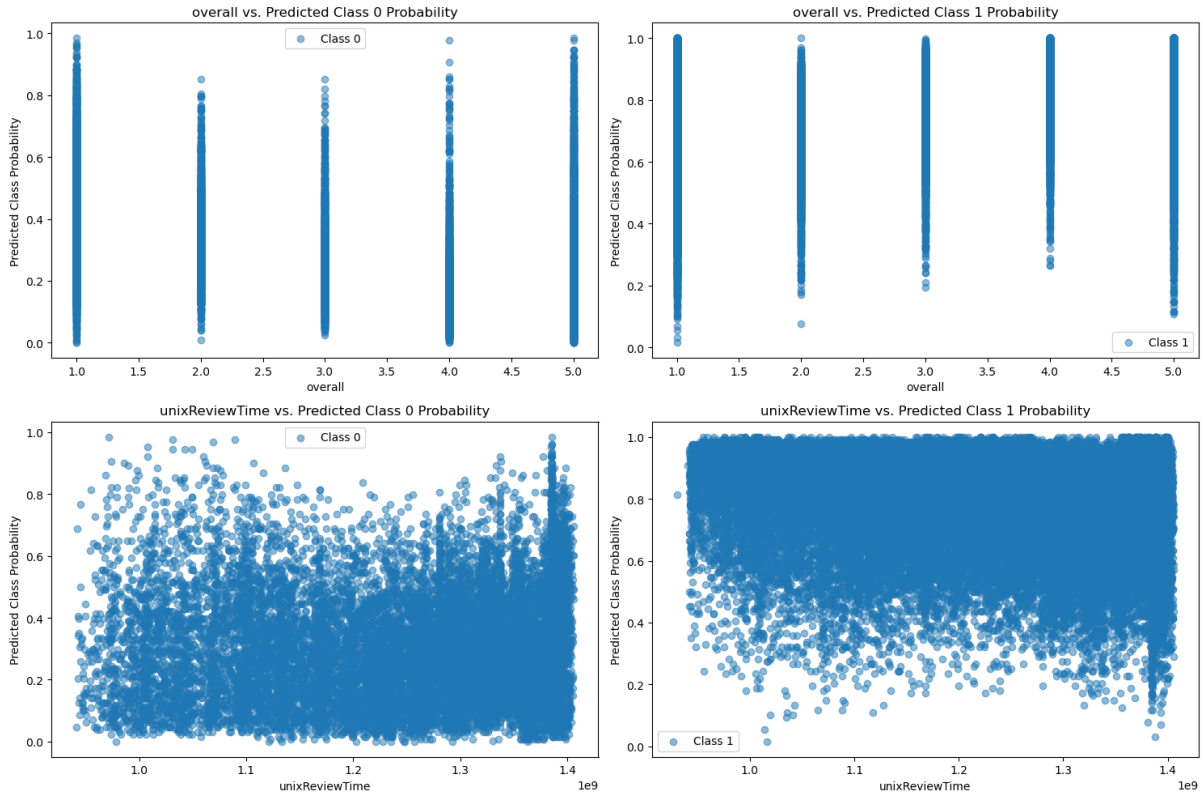
Interpreting the results of a black-box algorithm like random forests can be challenging.

One way to shed more light on a random forest model is to plot the values of a selected independent variable against the class probabilities predicted by the model (i.e., predictions of the dependent variable) (Friedman et al, 2013).

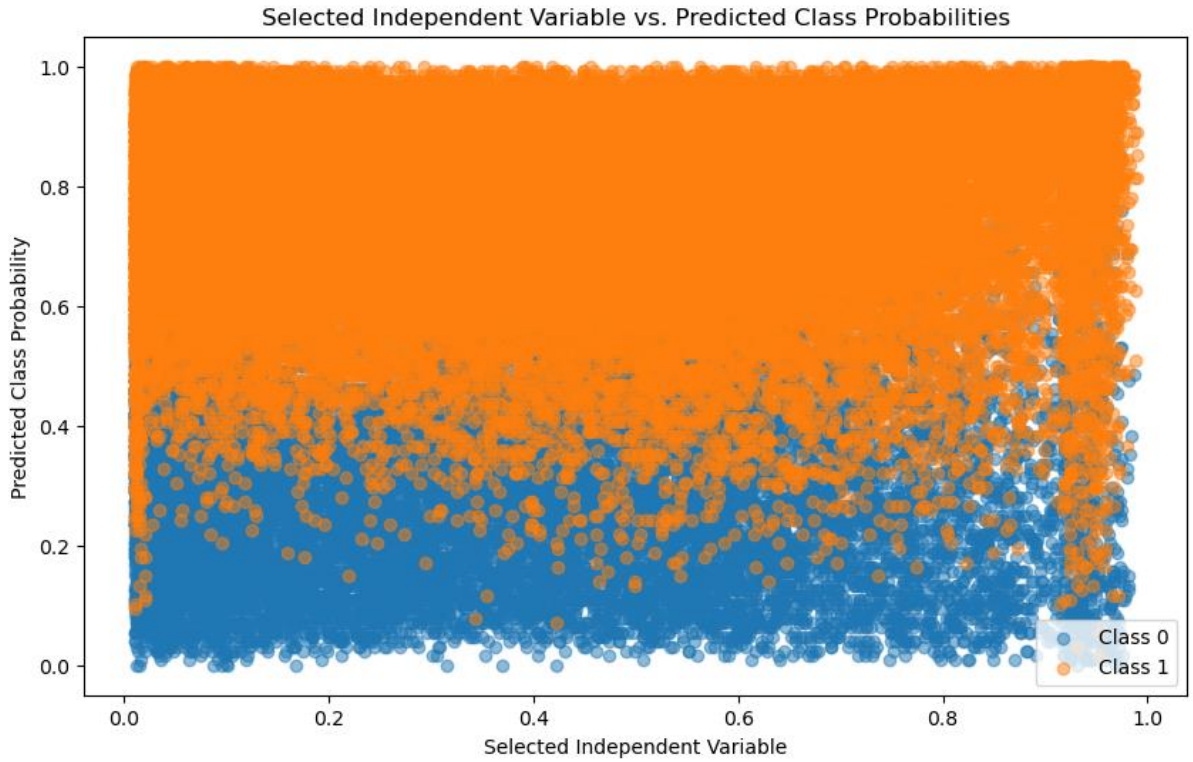
We achieve the above by:

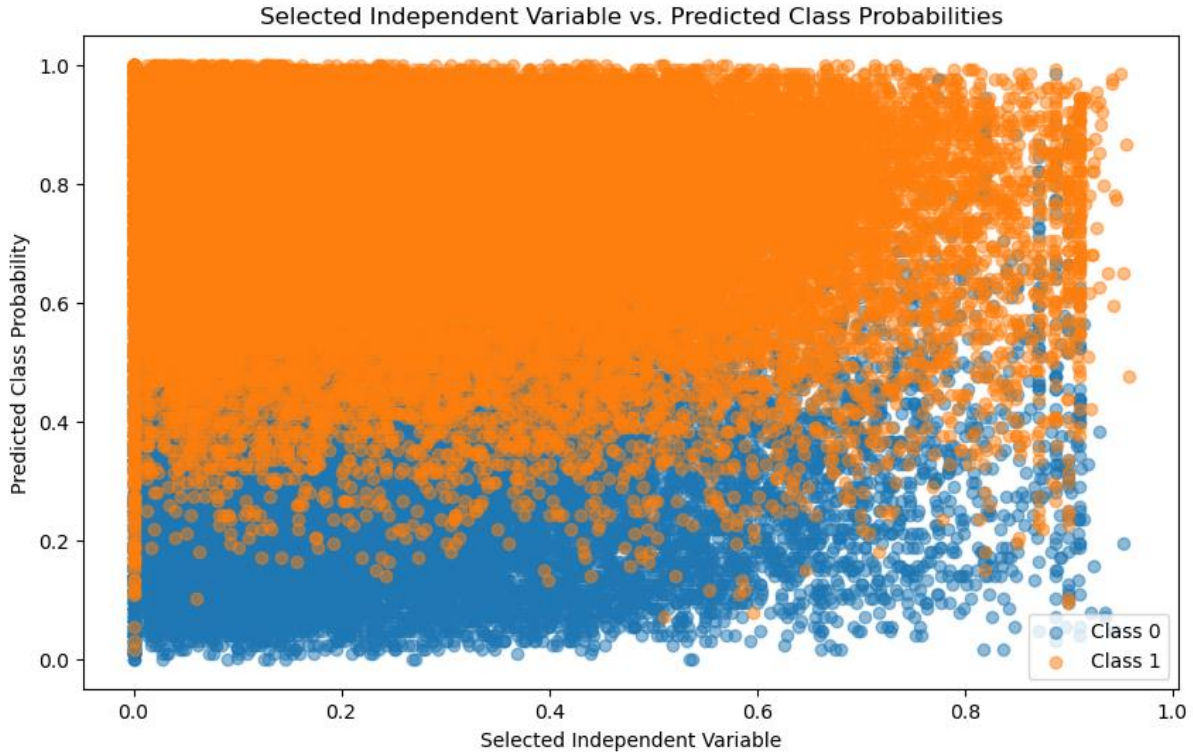
Making predictions on the testing set

Creating a multi-dimensional scatter plot for each pair of selected independent variables and class probabilities.



We analyze selected variables and find out whether a variable has a positive or negative influence on the probability of belonging to a certain class (i.e., helpful or unhelpful).





Final step of the result interpretation is then to compare and contrast the discoveries with theory and literature.

1.5 SUMMARY AND COMPARISON

The two studies, both the original and the replication have highlighted the potential of Machine learning and NLP to discover patterns and relationships and use human verification to interpret those results in a human context.

Visualizations helped to discover those trends and highlight patterns that would be difficult to understand by mere number analysis.

Big data analytics have the potential to provide a new and innovative approach to using large datasets to increase scientific knowledge and the comprehension of our world and its patterns.

From both my study and its original source I have had the chance to understand better the process to adopt and the relative weights assigned to each phase. I found out it is extremely important to set initial objectives (asking the questions you want to answer) and exploring the way of achieving them.

This is the priority, and it is extremely important in the first phase even before data collection.

The initial questions and its framing have guided the rest of the process.

Understanding and preparing the data is also vital to set the ground for analysis.

I have followed the original research phases and its guidelines and even if lack of computing resources did not allow me to reach the same depth of analysis the results are quite similar and showing the value of the guidelines for IS researchers in applying BDA.

These are an excellent starting point for further iterative testing and researching.

1.6 REFLECTION

1) 3 THINGS I HAVE LEARNED

1) Human feedback to detect the intruder in Amazon Mechanical Turk, I used Mechanical Turk for other purposes, and I was intrigued by the possibility to bring in the human equation in the study and enrich the research with those contributions. The role of human verification in interpreting the results of machine learning and NLP is vital to discover patterns and relationships in large datasets.

2) The empirical triangulation of the LDA results using a word intrusion task to measure the semantic coherence of topics by inserting a random word into a set of words and overall the importance of setting initial objectives and framing the research question before data collection.

3) The shift of focus from a star rating system in the early researches to a text analysis approach in the reviews (e.g.,Mudambi & Schuff, 2010;Cao et al, 2011; Ghose & Ipeirotis, 2011; Pan & Zhang, 2011; Korfiatisa et al, 2012) and the importance to visualize our analysis to understand trends and patterns.

2) 2 QUESTIONS STILL OPEN

1) We are facing an explosion of data, and it is still open the question if and how difficult will be for BDA to keep the pace and overcome the difficulties to measure and theorize. How can the process be optimized to handle even larger datasets, especially considering the limitations of computing resources?

2) Will guidelines remain applicable to future research and how we will have to adapt and modify them if necessary. How can the accuracy of the model be improved further? There will be other algorithms that could yield better results?

3) 1 THING I HAVE ENJOYED

I enjoyed learning about the practical application of machine learning and NLP in a real-world context, specifically in predicting the helpfulness of online reviews. It was fascinating to see how these advanced techniques can be used to extract valuable insights from seemingly simple data.

The challenge to overcome the roadblocks that I have faced due to my computer limited computing capabilities.

More than that the way to find a solution to problems I had never faced and that I did not know how to address but after deep research was able to understand and overcome.