Department of Statistics 2019

Personality Mining for Sentiment Analysis

28th August 2019

Candidate numbers

11851

17531

Submitted for the Master of Science, London School of Economics, University of London

Acknowledgements

We would like to thank and acknowledge the various people who were involved in helping us bring this dissertation together. Firstly, we would like to thank our supervisor Dr Kostas Kalogeropoulos, for his constant support and advice throughout the year. We really appreciate the amount of time you have generously contributed in discussing the contents of our project. We would also like to thank Aleksandar Matic and Joao Leitao Guerreiro from Alpha for providing us with interesting tasks and stimulating information about the applications of personality mining. Despite not being able to work on data from Alpha itself, we really appreciate the opportunity and are very thankful for their continuous support and guidance. We hope our research proves to be useful for their recommendation system and look forward to learning more about the work at Alpha from them.

Contents

List of Figures

List of Tables

Executive summary

Personality describes stable underlying attributes that shape peoples' behaviour and thinking. Knowing personality traits can help institutions or companies to support their users or customers in their needs, when used appropriately and ethically. At the same time, personality traits are rarely available. In this project, we use Amazon movie review data to explore the idea of mining personality traits from review texts and use them in sentiment analysis. We assume that the movie category is suited for this task because personality plays a role in movie taste and preference.

The task of mining personality from the dataset is challenging because review datasets do not contain labels for personality traits. To mine personality, we thus use a pre-trained dictionary developed as part of the Word Well-being Project (Schwartz et al. (2013), Kern et al. (2014)). This dictionary consists of words and n-grams and their correlation strength with the big five personality traits (agreeableness, openness, extraversion, neuroticism and conscientiousness). These correlations were trained on social media posts and not on review texts. Given that linguistic features are context specific and there is no pre-existing dictionary trained on review texts, this adds a further challenge to our task. Moreover, since we do not have personality labels in the amazon reviews dataset, evaluating our results is not straightforward.

We use several approaches to evaluate the success of our personality mining method. Firstly, we look at examples of reviews with extreme manifestations of each of the personality traits and compare the results with the underlying theoretical description of the big five personality traits. We find that the quality of this result is quite ambiguous. While some cases display correspondence with the psychological theory, others point out the problems rising from the shift of context from social media to movie reviews. For instance, the usage of swear words is indeed a theory-grounded indicator for low agreeableness. In contrast, the use of words such as "murder" does not express the same sentiment when describing a movie plot as compared to when it may be used to express feelings.

Openness plays a particularly interesting role in the context of movies. Since this trait is associated with, and partly even referred to as intellect, it should result in different movie tastes. Hence, we check the success of mining this particular personality trait by the following additional exploration: We look at two movies (out of a subsample) that display the largest discrepancy

in ratings from users with above median and below median openness values. We find that the movie with higher preference from customers that we categorized to high openness is a drama about ancient philosophy while the other movie is "Mortal Kombat II", a video game based thriller about fighting invaders. This is a very intriguing result and confirms the mined trait. Hence, we see that qualitative examination supports the idea that personality features can be mined in this context to some extent.

Next to these rather small-sample explorations, we also use our mined personality features in various ways to improve the sentiment analysis of the reviews. We build two sentiment prediction models that incorporate personality in different ways. The first model splits the dataset with regards to a personality feature into two parts and then builds separate bag-of-words (BoW) word embeddings on each of them. We use both shallow classification trees and the median for the split of the data. This model does not show an improvement over a single bag-of-words model without personality involved.

The second model, on the other hand, adds the personality features to the bag-of-word embedding features and uses these combined features for sentiment classification with a logistic regression. Adding the personality features increases the test accuracy from 82% to 86% on a balanced dataset which is a remarkable improvement. For comparison, we double the word embedding dimensions, implement neural networks, fastText word embeddings and a pre-trained sentiment analyser but do not reach accuracy values above 84%. This supports the hypothesis that the personality features add important and useful information to the word embeddings. Moreover, using only the personality features and additionally mined stylistic features such as punctuation and word type usage, we reach an accuracy of about 68%. This indicates that these variables contain relevant information.

To examine which of the many variables are relevant for the sentiment prediction, we also conduct a linear regression analysis on the binary rating (positive/negative) and look at the significance of each of the coefficients. To this regression model, we further add interaction terms between seven movie genres (e.g. drama, comedy) and the five personality variables. We reckon that if personality is mined appropriately, there would be an interplay with movie genres. A significant interaction term indicates a preference or aversion of a certain personality type for a genre. As mentioned above, based on the psychological theory, we expect users with high openness to prefer art house movies and indeed, the interaction term for this is significant and positive, confirming this expectation.

Several other interaction terms are significant as well, demonstrating the interplay of personality and movie tastes. Comparing the adjusted R-squared value for the regression both with and without the genre-personality interaction terms we find an increase from 0.125 to 0.133 when adding the interaction terms. In addition, grammatical features (word types and punctuation) are significant as well. We also find that for the pure personality features, only agreeableness and neuroticism are significant.

These combined results are evidence that it is possible to mine personality from review data to some extent and that these features are helpful in tasks such as sentiment analysis.

These findings open doors for a wide range of applications. For instance, self-optimizing applications that help their users to maintain and increase healthy or productive habits are becoming increasingly popular. These applications can provide more targeted advice on the basis of personality features. For example, a more conscientious user will more easily follow challenging plans than a less conscientious user. Or for instance, extroverted users can benefit from group-based interventions and activities.

Another promising field of application are recommendation systems. Instead of using only ratings for movies, music or products, the recommendation system can incorporate personality traits of its customers. This is promising because personality is an underlying attribute influencing various aspects of an individual such as taste. While there are many beneficial applications for such personality-based models, ethical considerations should not be neglected. Personality is a highly private and sensitive form of information. Therefore, users need to be made aware when their personality is being mined and any application should be based on the explicit, informed consent of the user.

1 | Introduction

1.1 Objectives and Research Questions

Personality describes stable underlying attributes that shape an individuals' behaviour and thinking. An understanding of an individual's personality traits can help institutions or companies to support their users or customers in their needs, when used appropriately and ethically. At the same time, personality traits are rarely available. In this project, we use Amazon movie review data to explore the idea of mining personality traits from review texts and use them in sentiment analysis. We assume that the movie category is suited for this task because personality plays a vital role in movie taste and preference.

The task of mining personality from the dataset is challenging because review datasets do not contain labels for personality traits. To mine personality, we thus use a pre-trained dictionary developed as part of the Word Well-being Project (Schwartz et al., 2013). Besides these dictionary based personalities, we also mine stylistic features such as the frequency of word types, word and comment lengths and punctuation as they might also contain information on personality.

The Word Well-being Project dictionary consists of words and n-grams and their correlation strength with the big five personality traits (agreeableness, openness, extraversion, neuroticism and conscientiousness). These correlations were trained on social media posts and not on review texts. Given that linguistic features are very context specific and there is no pre-existing dictionary trained on review texts, this adds a further challenge to our approach. Moreover, since we do not have personality labels in the amazon reviews dataset evaluating our results is not straightforward. Therefore non-standard methods for the evaluation of the success of our personality mining method have to be implemented.

One important pillar of the evaluation is to use the mined personality features for sentiment analysis. We incorporate them in different ways into sentiment classifications and compare the results with various other sentiment prediction methods such as predictions with just word embeddings, neural nets or a pre-trained sentiment analyser.

Another research question around personality traits is their relationship with movie genres. We

study whether individuals with certain personality traits prefer a specific genre of movies. One of the five personality traits is investigated more deeply. Since openness is associated with, and partly even referred as intellect, it should result in different movie tastes. Hence, we check the success of mining this particular personality trait by the following additional exploration: We look at two movies (out of a subsample) that display the largest discrepancy in ratings from users with above median and below median openness values.

Next to these rather small-sample explorations, we also use our mined personality features in various ways to improve the sentiment analysis of the reviews. We have built two sentiment prediction models that incorporate personality in different ways. The first model splits the dataset with regards to a personality feature into two parts and then builds separate bag-of-word word embeddings on each of them. We use both shallow classification trees and the median for the split of the data. The second model, on the other hand, adds personality features into different classifiers and compares the accuracy with the classification without these features.

Overall, the main objective of this report is not the general task of sentiment analysis. Our research focuses on investigating how the different features from text, especially personality, have an impact on polarity, rather than on prediction optimization. Hence, even though this paper aims to find the models that perform the best when it comes to classifying sentiment using different stylistic features, it places greater emphasis on exploring the impact of different features on sentiment prediction and the quality of mining them in the first place.

1.2 Overview of Sentiment Analysis and Personality

Sentiment Analysis, also known as Opinion Mining is a field of Natural Language Processing (NLP) that builds models by identifying features and extracting attributes from text to perform automated classification of text into positive or negative sentiment.

Moreover, sentiment analysis systems could be useful in transforming unstructured information into structured data of public opinions about products, services or any topics that people wish to express their opinions about. Hence, it can not only help process data at scale in an efficient way but can also allow us to establish a consistent criteria for evaluating the sentiment of a text. Since this report aims to investigate people's comments which is a very subjective concept of describing people's feelings and sentiments, this can prove to be quite a beneficial way for analyzing sentiment because it can ensure that companies use a consistent system to categorize all its data.

Sentiment Analysis can be useful both for business applications and for research. Businesses have several incentives to automatically scan vast amounts of user comments for their sentiment. For instance, they might want to select the most positive comments for display and advertisement. Moreover, since positive comments are indicative of happy customers, businesses can use positive sentiment to recognize customers who might be more receptive to spending more and can then use this as an opportunity to upsell. Therefore, sentiment analysis can allow businesses to provide a flexible and more efficient service by training its team to adapt its services to the mood of the customer. For instance, by identifying a disgruntled customer, agents can make sure they provide quick responses and helpful solutions to ultimately retain the customer.

Within research, on the other hand, automated sentiment analysis can be used to not only identify key emotional triggers but also investigate on societal trends. For instance, phrases like 'Please wait' may often trigger people's annoyance which implies that sentiment analysis may be used to identify what messages or words might act as emotive triggers that might change people's moods. Moreover, performing sentiment Analysis of "tweets" during elections can act as a great indicator of the overall opinion of society.

With the rise of online services, there has been an increase in comments and ratings from customers. Review data with a text review and a corresponding rating of a product or service can function as a labelled dataset for sentiment analysis. Due to the vast amount of these comments, review systems have become increasingly important for customer decisions. Google maps reviews assess any facility (including shops, universities, restaurants, beauty salons and medical facilities), Facebook reviews rate their pages (including business, cultural or political pages), Glassdoor lets employees review their employers and Amazon provides reviews for their products. Rising software platforms such as Airbnb, Treatwell or Uber maintain a reviewing system as well. In all these cases the sentiment of the review text is known by the rating assigned to it.

Therefore, this report aims to further investigate the different features that can be extracted from review text which can in turn help determine the polarity of the review. These include stylistic features such as the type of punctuation used, the use of different word types (verbs, adjectives, nouns etc) and words indicative of an individual's personality.

Most previous works of sentiment analysis have used writing style of tweets by different users as a whole to extract sentiment without actually considering the diverse word use of people (Maeda et al., 2012). Thus, some sentiment words may be neglected in the process of analysis because they are only used by people of specific groups. For instance, people who identify themselves

as extroverts are more likely to talk about friends and family and use words like 'dancing' and 'drinks' because people matching that personality type are expected to enjoy socializing more. Therefore, inspired by such psychological findings and perceptions that personality influences the way people write and talk, this report proposes an in-depth exploration into personality based sentiment classification on review data.

Personality can be defined as as individuals unique patterns of thought, emotion, and behavior. It is often seen as a product of an individual's social interaction. This is because individuals have different habits, attitudes as well as physical traits of a person which are developed during the process of socialization in a culture of a specific group or society. However, personality is not a characteristic that an individual can be measured on a scale of having more or less. Therefore, it is often observed as a part of a myriad of traits.

The most widely accepted of these traits are the Big Five which were developed in the 1970s by two research teams that were led by Paul Costa and Robert R. McCrae of the National Institutes of Health and Warren Norman and Lewis Goldberg of the University of Michigan at Ann Arbor and the University of Oregon. These are:

- Openness - This is a characteristic representative of an individual that is 'open' to experience. Hence, people with high openness scores tend to be curious, creative, adventurous and more appreciative of art, imagination and innovation. Whereas, people low in openness tend to be characterized as more conventional such that they prefer sticking to their habits and routine and shy away from new experiences.
- Conscientiousness - This dimension of personality measure an individual's degree of organisation. People who are conscientious are thought to be disciplined, motivated and achievement-focused with a strong sense of duty while irresponsibility and spontaneity are indicative of an individual who is found at low end of the conscientious scale.
- Extroversion - This is a very broadly defined feature which is a measure of an individual's communication and social skills. Extroverts are individuals that tend to be cheerful, sociable and assertive in social interactions, while introverts are more reserved and shy and often prefer solo or small-group activities.
- Agreeableness - This is a trait that measures the extent of a person's warmth and kindness. Hence, an individual with high agreeableness is seen as one who is friendly, helpful and compassionate, while a shy, suspicious and a low cooperative person scores low on the spectrum.

- Neuroticism - This attribute of personality is a measure of an individual's emotional stability. People with high scores are anxious, inhibited, moody and less self-assured. Whereas, those on the lower end are calm, confident and contented.

Hence, the Big Five can described as the building blocks that make up each individual's personality. For instance, an individual could be disagreeable, neurotic, introverted but not at all conscientious or open.

1.3 Summary of Main Results

Looking at the review examples of extreme values for personality and comparing the results with the underlying theoretical description of the big five personality traits, we find the quality of the results quite ambiguous. While some cases display correspondence with the psychological theory, others point out the problems rising from the shift of context from social media to movie reviews. For instance, the usage of swear words is indeed a theory-grounded indicator for low agreeableness. In contrast, the use of words such as "murder" does not express the same sentiment when describing a movie plot as compared to when it may be used to express feelings. Overall, we notice that most of the examples show meaningful instances.

For openness, we find particular evidence that the personality mining process is fruitful. We discover that the movie with higher preference from customers that we categorized to high openness is a drama about ancient philosophy while the other movie is "Mortal Kombat II", a video game based thriller about fighting invaders. This is a very intriguing result and confirms the mined trait. Hence, we see that qualitative examination supports the idea that personality features can be mined in the context to some extent.

We observe that several personality-genre-interaction terms turn out to be significant and the overall model fit in a linear regression predicting sentiment increases with these features. This can be seen as proof that these features contain relevant information not already covered in the other features.

Adding personality features to bag-of-word embeddings in a logistic regression model increases the test accuracy from 82% to 86% on a balanced dataset which is a remarkable improvement. For comparison, we used neural networks, fastText word embeddings and a pre-trained sentiment analyser but did not reach accuracy values above 84%. This supports the hypothesis that the personality features add important and useful information to the word embeddings for sentiment

analysis. Moreover, using only the personality features and additionally mined stylistic features such as punctuation and word type usage, we could reach an accuracy of about 68%. This indicates that these variables contain relevant information as well.

1.4 Project Outline

By developing an understanding of the different personality features, through our report, we aim to further investigate the link between an individual's personality features and its effect on the overall sentiment of one's views about a certain product or service. The structure of the paper is as follows: Section 2 consists of the literature review of the various methods used in this research. Section 3 will provide an overview of the datasets used. Section 4 describes the methods used to extract the different features from text. After establishing the most relevant variables with particular emphasis on personality traits, we will investigate and test out the different sentiment analysis models for the task of sentiment classification. Section 5 provides the experimental results of the study. Section 6 provides a summary of the limitations in this research and the paper concludes with Section 7.

2 | Literature Review

2.1 Sentiment Analysis

Sentiment Analysis is a problem that has become very popular within text processing. Tan et al. used the amazon review dataset to perform sentiment analysis in their goal of multiclass classification, where each rating category from 1 to 5 stars is a class. They tried out a range of different methods and achieved the best test accuracy results when using LSTM of approximately 71%. Our report is partly comparable with this paper, as we similarly build and compare different models to perform binary classification of sentiment of Amazon movie reviews.

In our research of studying the correlation between movie reviews and the rating assigned to the movie by the customers, feature mining is a very essential process. The Twitter Sentiment Analysis Based on Writing Style paper (Maeda et al., 2012) focused on the usage of expressions such as emoticons and punctuations to determine that there exists a link between these expressions and a user's identity which could be useful for sentiment analysis. Following up on this, our report mines features from text such as grammar, use of punctuation and length to study whether there exists any connection between writing style and a review's sentiment.

Furthermore, in our research we include the use of a pre-existing sentiment tool. "VADER" (for Valence Aware Dictionary for sentiment Reasoning) is a rule-based model for sentiment analysis, introduced by Hutto Gilbert (2014). It uses both qualitative and quantitative methods. Empirical tests on Twitter data have shown superior classification accuracy compared to human classification (Hutto and Gilbert, 2014). It was also validated on movie and product reviews. Since it is pre-trained, no training and learning is necessary.

To create the VADER Analyser, additional grammatical and syntax laws were incorporated that increase the quality of pure bag-of-words (BoW) tools. These qualitative analysis features include (1) punctuation, (2) capitalization, (3) intensifiers, (4) consequences of the word "but" and (5) tri-gram predecessor of a sentiment-word. Examples of this tool can be seen in Table 2.1.

In addition to using a pre-existing sentiment tool, we also use an open-source library fastText to create word embeddings for the reviews in our dataset. FastText was developed by the Facebook

Table 2.1: Vader

Example Text	Vader Score
"Greatest and best and most wonderful thing in the world!"	0.93
"Nice."	0.42
"Not Nice."	-0.32
"Nice but I hated it"	-0.71

AI Research (FAIR) team and plays a very significant role in the task of text classification (Bojanowski et al., 2017). In our research, we use the fastText library to learn word representations of words and sentences. Our main motivation behind using fastText word embeddings instead of word-to-vec is its capability of learning character level representations, which in turn achieves good performance for our models by learning word embeddings for rare words as well. This means that each word is represented as a bag of character n-grams in addition to the word itself. For example, the word 'carrot' with n = 3 has fastText representations for the character n-grams which are <ca, car, arr, rro, rot, le>. The boundary symbols (<, >) are added to distinguish the ngram from a word itself.

Other pre-trained sentiment models use lexicon-based approaches (Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010). Words are classified as positive or negative and then counted in the respective text corpus. Some lexicons (such as the Affective Norms for English Words (ANEW) lexicon) provide not only the sentiment category but also the strength of the sentiment (Bradley and Lang, 1999). Thus, "good" and "excellent" have different impacts on a sentiment score and more differentiable sentiment scores can be obtained. Nevertheless, these forms of sentiment analysis tools are not context-sensitive. Different meanings of words cannot be understood based on the context of their usage.

2.2 Personality

One focus of our research is the correlation between an individual's personality traits and how it might affect the sentiment of their reviews. Similar tasks are often carried out by the use of dictionaries which are then used as a starting point for an unsupervised personality recognition tool (Celli and Rossi, 2012). As Celli and Polonio (2013) describe: "Given a set of correlations between personality traits and some linguistic or extra linguistic features, we are able to develop a system that builds models of personality for each user in a social network site whose data are publicly available". This idea of the usage of dictionaries was further supported by Golbeck et al.

(2011) in his paper as he predicted personality features from Twitter, partly by using the LIWC dictionary.

Therefore, to establish a link between personality and linguistic features, a training dataset with both these features is needed. In the context of the World well being project (WWBP), a dataset based on 69,729 individuals was created (Schwartz et al., 2013). The personality scores were obtained through the facebook application MyPersonality (Quercia et al., 2011) which lets users fill out a personality questionnaire. This questionnaire is based on the International Personality Item Pool (IPIP). About 40% of the users of this application agreed to not only their results being used for research, but also their facebook status updates being collected and analysed. This way, a large training dataset on personality and written text was built. Only data from English-Speaking users was used. The average participant was around 24 years old and mostly from the United States or Canada.

Regarding the text, single words, 2-grams and 3-grams are used that occur at least in 1% of the samples. To obtain the relation between these n-grams and the personality, separate ordinary least squares linear regressions are fit for each n-gram and personality feature combination. All regressions control for gender and age. The 100 n-grams with the strongest relation and a significant p-level are selected for each personality feature and included in a dictionary. Every word is listed with the beta-value with which it is associated with the feature. Positive values indicate a positive relationship and vice versa.

Previous methods to link linguistic features and personality were based on indirect correlations. A popular dictionary for this is the LIWC (Linguistic Inquiry and Word Count) (Pennebaker and Francis, 1999). These closed-word dictionaries provide categories (such as positive emotion, pronouns, friends etc.) for frequently used words and were created manually. Some of these categories can be linked to personality features. For example, high neuroticism correlates with more negative emotions. The open-ended vocabulary dictionaries provided by the WWBP demonstrate several advantages. Not only are they directly linking linguistical and personality features, but they also include misspelling and abbreviations (“cant” instead of “can’t’, “ur” instead of you’re). This way, the open-ended approach can detect connections in naturally occurring text that the LIWC dictionaries cannot capture. In our case of Amazon reviews, this is also of relevance since these natural text usages occur frequently.

This report further looks into the relationship between personality and a preference for a specific genre of movie. Previous research has suggested that often an individual’s preference for a certain

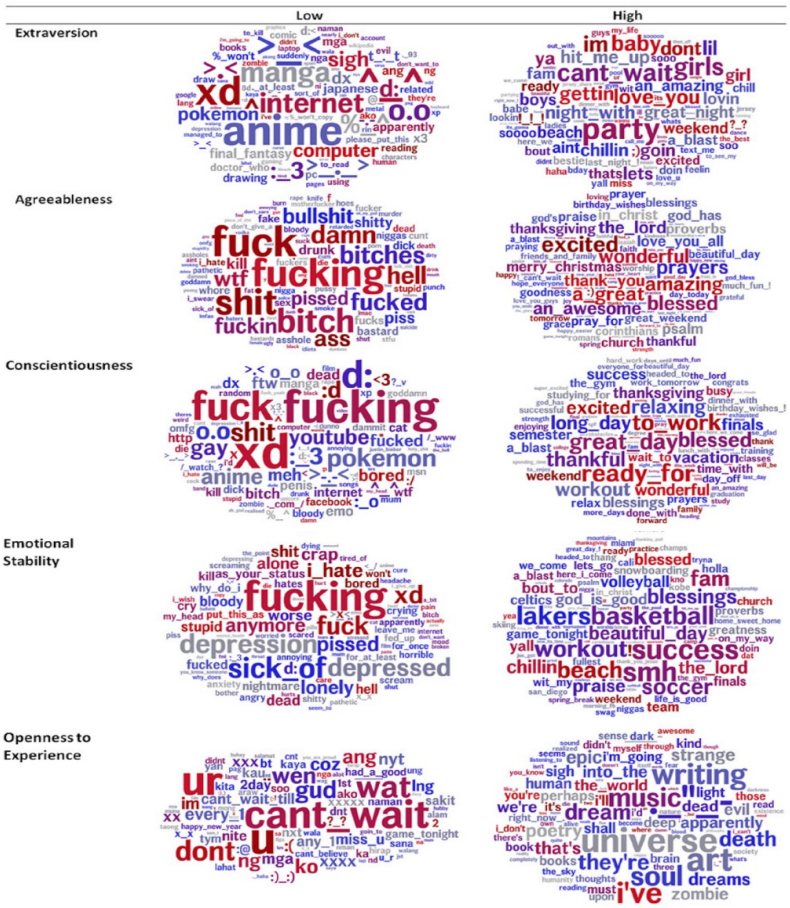

Figure 2.1: Word Cloud (from Kern et al. 2014) is a visualization that illustrates the different indicators for each of the five personality features. The correlation strength is symbolized by the size of the word and the relative frequency is coded through the color. Red words occur often, while blue words are rare.

type of movie might be a result of the features of their personality. For instance, Duuren (2008) described how there is a significant correlation between a preference for arthouse movies and the openness personality trait. Hence, to expand on this, this report includes movie covariates so

as to explore their interactions with personality factors such as finding out whether people that score high on openness would give more positive movie reviews overall, or just in the case for arthouse movies.

3 | Data Overview

To perform the task of sentiment analysis and personality mining, we use two different datasets. The first one provides us with the reviews and ratings and the second one with genres.

1. Amazon Movie Reviews Dataset: This is a publicly available dataset of 1,697,533 reviews from the movies and TV category of the Amazon website (He and McAuley (2016), McAuley et al. (2015)). For each review, the dataset provides information about reviewer's ID, reviewer's name, movie ID, review summary and text as well as the review rating. The reviews vary in length with the shortest review containing only 4 words and the longest containing 32766 words. The original ratings for the reviews on the dataset are 1-5 star ratings, which will then be classified as either being 'positive' (greater than 3) or 'negative' (less than 3). An example of a positive review would be

"I was very happy with the service and with my purchase. The series is wonderful– I was a big fan of the show when it was on TV"

Whereas a bad review would be

"Hope I can find twenty words to say how bad this film is. Started watching this movie, and after ten minutes turned it off and threw it in the trash. Save your money on this one, and use it to buy something good. One star is too high a score for this one, it should be -5."

2. Movie Category Dataset: This is a dataset of 253,059 movies from the Amazon movie reviews dataset with their respective categories that is available on Kaggle (Ground truth labels - Amzn movie reviews dataset). There are a total of 1292 unique genres in this dataset but we only consider movies with the following seven genres:

- Comedy
- Art House & International
- Drama
- Action & Adventure
- Horror

- Science Fiction
- Animation.

The main incentive for using this dataset is to investigate whether there exists a link between personality-movie interactions and sentiment. For instance, people with high openness scores might rate Art & International movies higher.

For the analysis, the two datasets are combined using Amazon Standard Identification Number (ASIN), which is a 10-character alphanumeric unique identifier assigned by Amazon.com as it is a common variable found in both datasets. Neutral reviews are ignored and reviews with ratings 4 and 5 are labeled as good whereas reviews with ratings of 1 or 2 are labeled as not good.

Controlling for all these factors, the dataset consists of mostly unique users. This means most users in the dataset only contributed one review. Thus the features are review-based and no user-specific features were designed. The dataset is then further balanced regarding the target using under sampling. The resulting data set had 18,810 observations and it will be used to carry out the feature-mining process.

4 | Methodology

4.1 Feature Mining

Following is an overview of the different methods that we are using to extract features from the review text as part of this research:

- **Word Embeddings**

  Classical machine learning applications work on data that comes in the form of observations and numerical features. In contrast, textual data does not come in this quantified form. Being able to use text as input is important to make use of the large amounts of data available whether it be in social media, reviews or literature. Therefore, it is crucial to find ways to convert natural language into numerical features.

  Bag-of-Words

  A bag-of-words model (BoW), as mentioned above, is a common feature extraction procedure that extracts features from text for use in modeling, such as with machine learning algorithms. It can be seen as a representation of text that describes the occurrence of words within a document. The complexity exists in deciding how to design the vocabulary of known words (or tokens) and how to score the presence of known words. Additionally, it is called a "bag" of words, because any information about the order or structure of words in the document is discarded. Hence, the model is only concerned with the presence of known words in the document, not their position.

  In this approach, we use the tokenized words for each observation and find out the frequency of each token. Let's take an example to understand this concept in depth. We have the following sentences: "It was a great movie", "It was a bad movie", "I liked the movie", "I hated the movie".

  We treat each sentence as a separate document and we make a list of all words from all the four documents excluding the punctuation. We get: 'It', 'was', 'a', 'great', 'movie','bad', 'liked',

'hated', 'the', 'I'. The next step is the create vectors which will convert text that can be used by the machine learning algorithm. We take the first document — "It was a great movie" and we check the frequency of words from the 10 unique words:

"it" = 1

"was" = 1

"a" = 1

"great" = 1

"movie" = 1

"bad" = 1

"liked" = 0

"hated" = 0

"the" = 0

"I" = 0

Rest of the documents will be:

"It was a great movie" = [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]

"It was a bad movie" = [1, 1, 1, 0, 1, 1, 0, 0, 0, 0]

"I liked the move" = [0, 0, 0, 0, 1, 0, 1, 0, 1, 1]

"I hated the movie" = [0, 0, 0, 0, 1, 0, 0, 1, 1, 1]

In machine learning, there are two different ways to convert text into vectors. These are:

- Counting the number of times each word appears in a document.
- Calculating the frequency that each word appears in a document out of all the words in the document.

In our model, to convert the NLP text into numbers, we will be using the Count TF-IDF Vectorizer. This is similar to the count vectorizer, in the manner that it counts the number of times a word appears in the document. However, additionally it normalizes the count and

adds a weight depending on how important a word is such that it increases proportionally to the number of times but is offset by the frequency of the word in the corpus.

Managing the vocabulary is a very important part of building a bag-of-words model because increasing the vocabulary size in turn increases the vector representation of documents. Hence, in the case of a very large corpus, the length of the vector might be thousands or millions of positions. Furthermore, each document may contain very few of the known words in the vocabulary resulting in a very sparse vector representation. This will not only make the model more computationally heavy but also make the modeling process very challenging. The approach we will be using to decrease the size of the vocabulary when using a bag-of-words model is creating a vocabulary of grouped words where each word or token is called a "gram". Hence, we will be using bi-grams to build our word embeddings in this report.

Finally, we are going to investigate whether bag-of-words models would perform differently with different sized vocabularies. Hence, in order to test that, we are going to build models using vocabulary size 500 and 1000 so as to be able to compare its effect on sentiment.

Text Pre-processing

The bag-of-words features were built on pre-processed texts. Using the NLTK Library (Bird et al., 2009), the following parts were included in the pre-processing:

- Delete punctuation.
- Transform to lowercase.
- Delete stop words. Stop words are considered words that are not useful for the natural language processing task because they do not contain information, such as "the" or "a".
- Stem the remaining words. Stemming defines the process in which words are reduced to their stem.

For instance, the words "know, say, yes, say, batteries, charged" are relevant in the first sentence as they are not seen as stop words. The remaining words are ignored. Finally, using top 500 words and up to 2 grams, a vector of counts for each of these 2-grams for each review is created.

Table 4.1: Text processing: stemming

	Before Stemming	After Stemming
1	"Now I don't know what to say other than yes they were batteries and they were charged when I got them."	"know say ye batteri charg got"
2	"they arrived in different packages but the batteries are still the same. the best batteries on the market. good price too!"	"arriv differ packag batteri still best batteri market good price"

FastText

We implement the open word-embedding approach when using fastText which creates the word-embeddings by using the movie reviews dataset rather than a pre-trained model. As a result, these word-embeddings are domain-specific and much more relevant and useful for our context. However, a drawback of training our own word embeddings is the increased running time. When training the word-embeddings for the movie reviews dataset using fastText, we use a size parameter of 500, which determines the dimension of the word vector. For the rest of the parameters, we use the default values.

We hypothesize that a model built using fastText word embeddings would perform well on its own because of fastText's ability of accounting for character level n-grams. However, it would be interesting to find out whether the addition of other variables such as grammar or personality could result in further improvement.

- **Grammatical Features**

  Personality can not only manifest itself in the form of vocabulary but also in various stylistic features. These should also be captured. Therefore, we additionally extract the following grammatical features from the full text reviews.

  - overall length
  - reverse average word length described by white spaces divided through overall length
  - punctuation marks "!","?",",",","."
  - relative amount of word types like nouns, adjectives etc "noun", "adj", "verb (simple)","verb (simple past)", "verb (past participle)", "verb (gerund)", "adverb", "preposition"

All these features are counted and divided through the overall length of a review. Afterwards they are normalized by subtracting the mean and dividing through the standard deviation.

- **Vader**

  For comparison, the VADER analyser, as a pre-trained sentiment analyser, is used on the original review texts and its overall ratings are used as features. Examples of this tool can be seen in Part 2 in Table 2.1.

- **Personality**

  To obtain personality features, the original review text without punctuation is searched for matches with the World well being project (WWBP) dictionary vocabulary and the corresponding correlation strength is added to a counter. Afterwards, the obtained counter sum is divided through the overall length of the comment. Finally, the personality columns are normalized. As an example, Table 4.2 shows the top and lowest five examples for openness.

  The five most positively correlated indicators are highlighted in red. The five most negatively correlated indicators are not highlighted and have a negative sign.

Table 4.2: Example for openness indicators

1 to 3-Grams	Strength of Relation to Openness
u	-0.090361
ur	-0.085626
cant wait	-0.083149
cant	-0.081017
wat	-0.073425
the universe	0.107483
writing	0.109566
i've	0.111732
art	0.115162
universe	0.117249

We base the personality on text without punctuation. For better interpretability, we have punctuation marks as separate features. Moreover, punctuation marks can create a bias in the results since their effect is not linear. Using one point at the end of a sentence can be a

sign of higher openness whereas using ten points is a sign for lower openness. In our additional model, this biases the result. Furthermore, the personality features become highly correlated as many reviews use dozens of explanation marks, whose effect superimpose the vocabulary effects and cause correlations in the personality features that are driven by punctuation. Extreme value examples of personality features including punctuation marks is included in the appendix.

Note that neuroticism is coded inversely in the WWBP dictionary used. This means that high values indicate high emotional stability and low values indicate neuroticism.

To qualitatively examine the personality trait feature mining process, the openness trait with regards to movie preference is analyzed. Openness is known to be associated with intellect and creativity. For this, a subset of movies is randomly selected and within the movies with the highest discrepancy between rating from user with above-median openness and below median-openness are chosen and examined.

- **Genre**

  The Genre dataset contains a list of Amazon movie IDs and the respective categories they belong to. Each movie belongs to a range of categories with at most five categories at one time. For instance, the movie Avengers has three categories: Action, Adventure and Sci-Fi. Hence, we create dummy variables for each of the considered genres. In our research, we focus on these seven popular genre categories: Comedy, Art House & International, Drama, Action & Adventure, Horror, Science Fiction and Animation. We then fill the dummy variables in our dataset by assigning binary values for each genre, where 1 is assigned when the movie belongs to the specific genre and 0 otherwise.

- **Interaction**

  Since some words of the personality dictionaries like “anime”, “murder”, etc can be used in the context of a movie plot without any relation to the user’s personality, interaction terms with the genre are added to control for these effects.

  All combinations of the seven genre dummy variables and the five personality scores are added as features, creating 35 interaction terms.

4.2 Prediction Models

To be able to use all mined features for prediction purposes, several models are implemented. Firstly, a linear regression on mined features was conducted (see part 4.2.1). Moreover, we have built multiple classification models using different machine learning techniques (Logistic Regression, Naive Bayes and Support Vector Machines) on different combinations of features shown in part 4.2.2. The sklearn library for python (Pedregosa et al., 2011) is used for all classifiers. We additionally create a model where the dataset is split by personality and seperate BoW models are trained on each of the sets (see part 4.2.3). Lastly, we also build neural nets (see part 4.2.4). In all cases, the quality of the prediction is measured by the test accuracy.

4.2.1 Linear Regression

We run a linear regression on all mined features to get a better understanding of which features are significant predictors for sentiment and thus relevant to our problem.

The python library statsmodels.api (Seabold and Perktold, 2010) is used for this.

4.2.2 Classification on Feature Combinations

Furthermore, we use logistic regression, support vector machines and naive bayes classifiers on 16 different combinations of features. These combinations are the following.

- Bag-of-Words with 500 words
- Bag-of-Words with 1000 words
- fastText features
- Personality features
- Grammatical features
- Sentiment (VADER)
- Interaction (between Genre and Personality)

- Personality+Grammar+Interaction
- BOW+Personality
- BOW+Personality+Interactions
- BOW+Personality+Interactions+Grammar
- BOW+Personality+Grammar
- fastText+Personality
- fastText+Personality+Interactions+Grammar
- fastText+Personality+Interactions
- fastText+Personality+Grammar

In the last eight models, we create bag-of-words or fastText vector representations for all the reviews in the dataset and then combine it with the different features that we have mined from the text.

We use the gensim library for python to obtain the fastText models.

The obtained test accuracy and its variance is obtained over running the models on 20 random train-test-splits of the data.

4.2.3 Separate Bag-of-Words Models

Besides the inclusion of the personality features as additional independent variables in a classification model, a second approach of incorporating personality for sentiment analysis was implemented. Personality features are used to separate the data into subsets on which different bag-of-word models are run. The theoretical consideration behind this is the idea that people with different personalities use language differently. These differences can be captured with separate bag-of-word models.

Two different ways of splitting the data are implemented. Firstly, a shallow classification tree is built and its splitting points are used to split the data into two. Alternatively, the medians of

the personality values are used for splitting the data. Using the median ensures to have more balanced splits.

From the decision tree classification of the effect of the Big 5 Personality traits in 4.1, we discover that agreeableness and neuroticism are the two most relevant features for classifying the sentiment of a review. Therefore we build separate bag-of-word models for agreeableness and neuroticism using the splitting point the tree suggests.

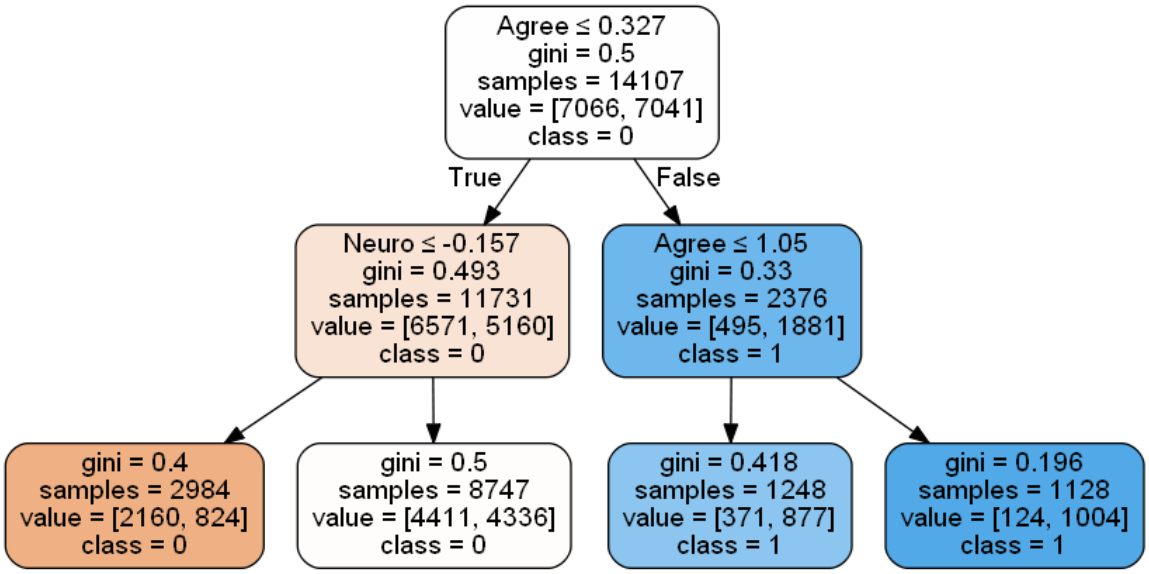

Figure 4.1: Results of a shallow decision tree classification

Hence, overall we build four separate bag-of-word models for high agreeableness, low agreeableness, high neuroticism and low neuroticism respectively. For a better comparison, the overall data set is also split randomly into two equally large subsets.

The obtained average test accuracy is obtained by running the models over 30 random train-test-splits of the data.

4.2.4 Neural Networks

We build a neural network for the purpose of looking into whether it might detect underlying stylistic features related to personality. Other than that, this model can also prove to be very useful in terms of comparison. Hence, we explore how Recurrent Neural Networks (RNNs) can be used for this task because deep learning methods have been gradually showing good performance in data classification as they learn data representation using a deeper hierarchy of structures in neural networks.

RNN has an internal memory that allows it to process a sequence of inputs such as text by performing the same task for every element of a sequence with the output being dependent on the previous computations, which is in a way similar to remembering information that has already been processed. Moreover, this built-in feedback loop also allows it to act as a good forecasting engine. Theoretically, RNNs can make use of the information in arbitrarily long sequences, but in practice, the standard RNN is limited to looking back only a few steps due to the vanishing gradient or exploding gradient problem. This is where the Long Short-term memory might come in handy. Long Short-Term memory network is a special type of RNN, which solves the problem of vanishing gradient by allowing the network to learn long-term dependencies.

Before we can use the reviews as inputs for the recurrent neural network, we clean the data by lowercasing, removing special characters and removing stop-words for the purpose of shrinking the observation space. We then create a Word-to-Index map, which assigns a unique integer value to each word in the dataset. Thus the word-to-index map has the same number of entries. This step of assigning a unique integer to words in the dataset is crucial because we cannot feed in string data into a neural network. Instead, word-to-index allows us to use integers to represent whole sentences and reviews.

For instance, this review

'A truly "kids" movie. Kids are the stars, and the aliens. What a terriffic twist to the end. Great Sci-fi in outer space adventure and fun from start to finish.'

transforms to

[33756, 42265, 23705, 42265, 16739, 73634, 54517, 39717, 6902, 25509, 1617, 79958, 73515, 57105, 58590, 76714, 30374, 6391]

Finally, since the reviews differ in length, we trim each review so that it is of equal length by padding these sequences. For instance, in our model we pad the reviews by a value of 250, which means that any reviews which have a length smaller than that will have 0's appended to the start of its sequence. In contrast, a review with a length longer than 250 will simply be cut off after 250 words. After finishing the data pre-processing, we split the data into training and testing so that 80% of the reviews are trained as part of the training dataset. We construct a neural network that uses Keras with an Embedding, LSTM and Sigmoid Activation layer. The embedding layer learns a word embedding for each word in the dataset and the addition of the LSTM layer will help overcome the problem of the vanishing gradient in an RNN. Finally, the Sigmoid Activation Layer is a dense layer which turns all output values to a value between 0 and 1. Hence, it predicts a probability of a review being positive or negative.

5 | Results

5.1 Results for Feature Mining

This section provides the results related to the feature mining process. It shows visualizations of the grammar and personality variables and investigates on relations around personality and genre interaction. In the following, extroversion, agreeableness, neuroticism, openness and conscientiousness are partly abbreviated with "Extro", "Agree", "Neuro", "Open", "Consc".

Histogram of Stylistic Features

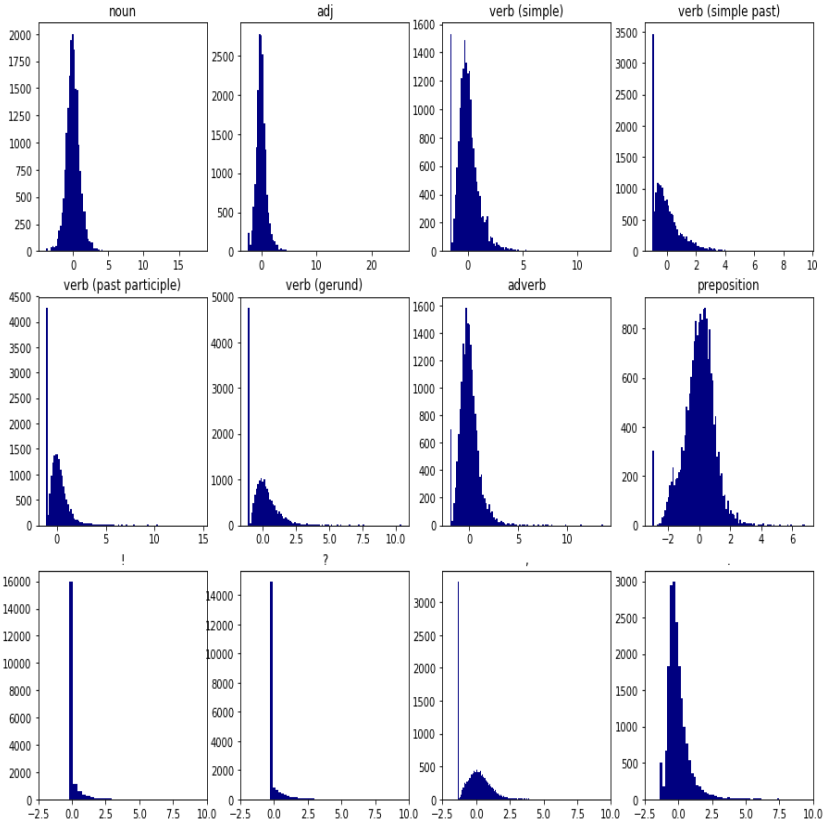

Figure 5.1: Histograms of grammatical features. The spikes at the left represent reviews without any usage of this feature.

The distribution of most grammatical (stylistic) features is approximately bell-shaped (compare with Figure 5.1). One can detect a significant spike at the left-hand side of most of the distributions that corresponds to the absence of such a feature in reviews. After normalization, this value is different from zero. Default features of writing such as nouns, verbs and prepositions seem to be distributed normally. This is expected, as they are used more necessarily and frequently. Optional features of writing such as exclamation marks or special verb forms are closer to a Poisson distribution. This can be considered as a sanity check for these features.

Extreme-Personality Review Examples

Table 5.1: Examples of most extreme instances of personality features

Personality Trait	Lowest Value Review	Highest Value Review
Agreeableness	Same story line as murder she wrote, with sex.Stupid writer and very stupid police-woman manage crime.,Not very good really.tlt-	great
Openness	for its time it was really good twisted and u culd feel the emotion of t he actors and it took u where they were and made u think wat you wuld do in a situation like that	AWESOME!!! MUST SEE!!
Extroversion	This is a well done anime. I highly recommend this show to an anime lover.	if you want to see him act its ok but its not real his movie
Neuroticism	CRAP CRAP CRAP CRAP CRAP CRAP CRAP CRAP CRAP worst made movie of ALL TIME!!!!!!!!!!!!!!!!!!!!!!!!1 Very bad don t buy it. Don t take it for FREE!!!!!!!!!!!!!!!!	This is my favorite of the 3. You can really feel your core workout and I walk away feeling accomplished and like I actually got a workout done.
Conscientiousness	This is a well done anime. I highly recommend this show to an anime lover.	great

Table 5.1 shows the ten reviews with the most extreme instances of personality attributes. As

one can see, low openness is related to misspelling and low emotional stability displays the usage of swear words. These are appropriate findings. On the other hand, low extroversion and conscientiousness display the word "anime" while low agreeableness writes about "murder". The last two are examples of context shift. Using "anime" or "murder" has a different meaning when used in social media than when it describes movie plots. For instance, the use of words such as “murder” does not express the same sentiment when describing a movie plot as compared to when it may be used to express feelings. We find that the quality of this result is ambiguous. While some cases display correspondence with the psychological theory, others point out the problems rising from the shift of context from social media to movie reviews.

Histogram of Personality Traits

In Figure 5.2, we observe that except for neuroticism, the other four features display a rather symmetric Gaussian-shaped distribution. Also, openness is uniquely bimodally distributed whereas the other are unimodal. These patterns do not indicate irregularities.

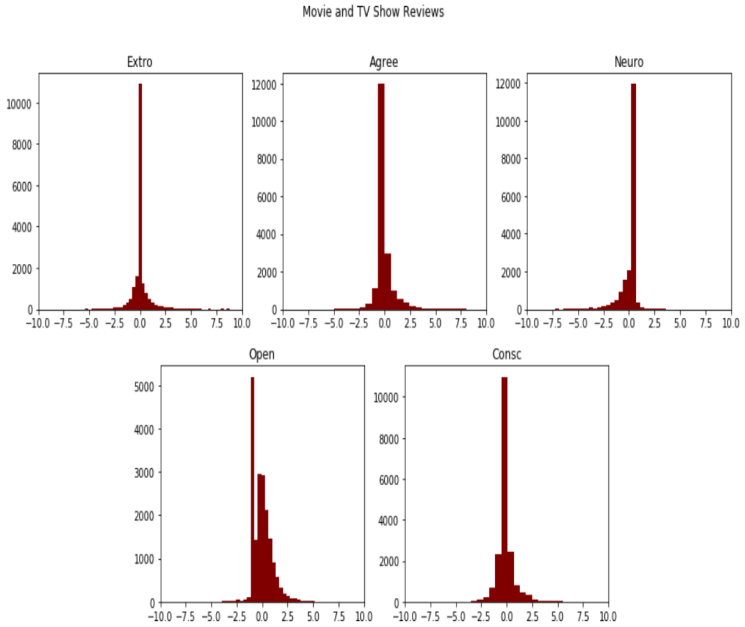

Figure 5.2: Histograms of personality traits

Scatter Matrix of Personality Traits

The scatter matrix 5.3 shows the correlation between each of the personality features. Moreover, one can see how they are related to positive and negative reviews. Positive reviews are marked green, the negative reviews are marked red.

As one can visually notice, agreeableness most clearly distinguishes the ratings. Openness and extroversion do not seem to separate ratings. Conscientiousness and agreeableness are positively correlated with each other which can be caused by the usage of swear words in case of low values for both traits and rather motivational vocabulary for high values.

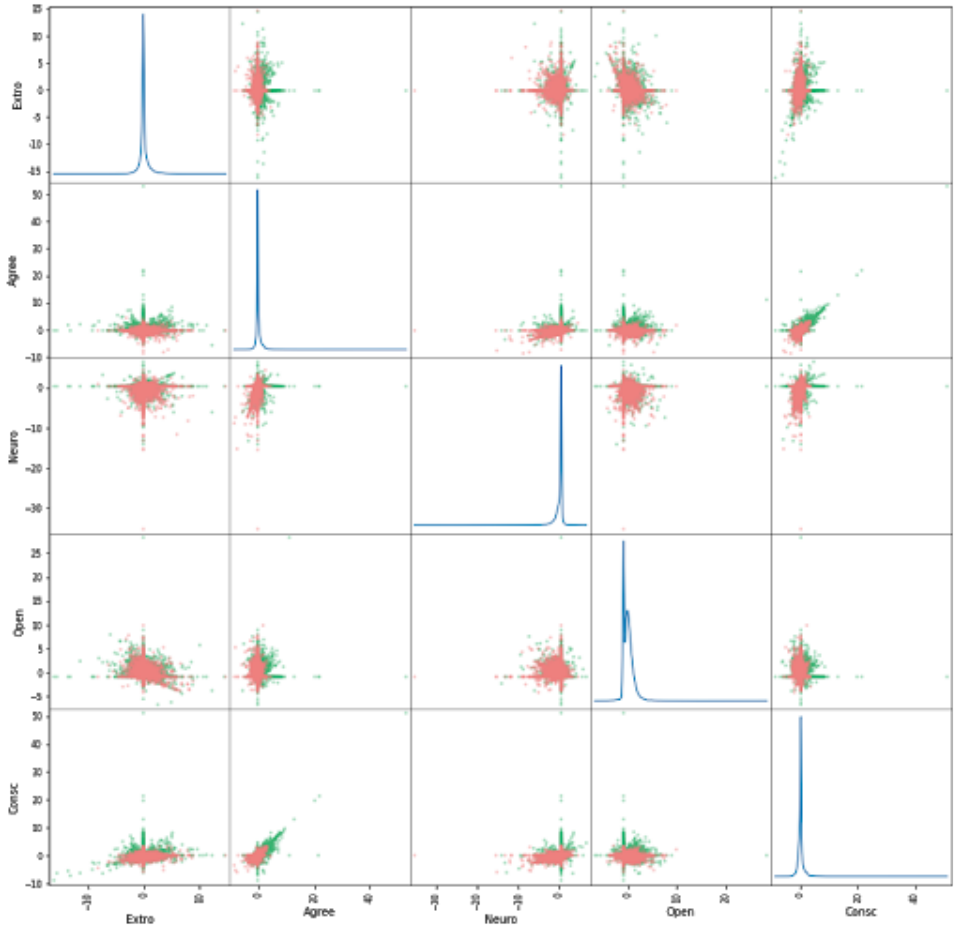

Figure 5.3: Scatter Matrix of five personality traits. Red points represent negative reviews, green points positive reviews. The diagonal fields show kernel plots of the feature.

Polarising Movies with Regard to Openness

We examine the theory whether higher openness manifests in an increased interest and preference for rather intellectual art. We look at the two movies (out of a subsample) that display the largest discrepancy in ratings from users with above median and below median openness values.

The movie with a higher preference from people with more openness is "A Funny Thing Happened on the Way to the Forum" (see Figure 5.4) which is, as Wikipedia describes, "inspired by the farces of the ancient Roman playwright Plautus (251-183 B.C.)".

Figure 5.4: Movie poster of "A Funny Thing Happened on the Way to the Forum". The Greek scenery is recognizable. This movie is preferred by reviewers with high openness.

The movie with a higher preference from people with low openness is "Mortal Kombat II" (see Figure 5.5): a video game based thriller about martial arts warriors who save the earth from an extra-dimensional invasion.

Figure 5.5: Movie poster of "Mortal Kombat II - Destroy all Expectation". The action-thriller theme is noticeable. This movie has more positive reviews by users with lower openness values.

The movies confirm the hypothesis that high-openness users prefer intellectual movies while low-

openness users prefer less intellectual movies. This is an indicator that the word-based feature mining is capturing important characteristics and that our "openness" indicator works.

Further results regarding the features can be obtained from the following results as well, e.g. the regression output.

5.2 Results for Prediction

5.2.1 Linear Regression

Using the adjusted R-squared criterion, we can see that adding the personality-genre-interaction terms increases the fit of the model (see Table 5.2) from 0.125 to 0.133. This confirms their explanatory power and information content.

Table 5.2: Results for linear regression

Independent Variables	Grammar & Personality	Grammar & Personality & Interaction terms
Adjusted R-squared	0.125	0.133

Table 5.3 demonstrates the different variables in the linear regression model, their coefficients and their significance for predicting the binary outcome variable of rating sentiment. We expect this table to show that if personality is mined correctly, there would be a possible correlation between personality and movie types. A significant interaction term is an indicator of a certain personality type's preference or aversion for a certain type of movie. For instance, we expect users with high openness scores to like more artsy movies (Duuren, 2008).

Table 5.3: Results for a linear regression: beta-values and significance levels for all features.

Variable	beta	p-value
const	0.498	0.000
length	-0.006	0.120
!	0.011	0.002
?	-0.051	0.000
,	-0.022	0.000
.	-0.011	0.002

Inverse Word Length	0.019	0.000
Extroversion	0.012	0.277
Agreeableness	0.108	0.000
Neuroticism	0.089	0.000
Open	-0.017	0.049
Conscientiousness	0.007	0.667
Extro_Comedy	-0.018	0.220
Extro_Art House & International	0.007	0.523
Extro_Drama	0.007	0.659
Extro_Action & Adventure	-0.017	0.256
Extro_Horror	-0.019	0.224
Extro_Science Fiction	-0.020	0.215
Extro_Animation	-0.002	0.907
Agree_Comedy	-0.047	0.020
Agree_Art House & International	0.037	0.021
Agree_Drama	-0.045	0.032
Agree_Action & Adventure	-0.031	0.170
Agree_Horror	0.019	0.391
Agree_Science Fiction	0.001	0.980
Agree_Animation	0.023	0.277
Neuro_Comedy	-0.039	0.014
Neuro_Art House & International	0.018	0.100
Neuro_Drama	-0.039	0.012
Neuro_Action & Adventure	-0.040	0.019
Neuro_Horror	-0.067	0.000
Neuro_Science Fiction	-0.031	0.078
Neuro_Animation	0.007	0.652
Open_Comedy	0.015	0.249
Open_Art House & International	0.022	0.031
Open_Drama	0.014	0.274
Open_Action & Adventure	0.015	0.277
Open_Horror	0.042	0.002

Open_Science Fiction	0.032	0.026
Open_Animation	0.002	0.901
Consc_Comedy	-0.011	0.592
Consc_Art House & International	-0.018	0.248
Consc_Drama	0.045	0.032
Consc_Action & Adventure	0.039	0.082
Consc_Horror	0.038	0.085
Consc_Science Fiction	-0.009	0.683
Consc_Animation	-0.008	0.699
noun	-0.015	0.000
adj	-0.017	0.000
verb (simple)	-0.059	0.000
verb (simple past)	-0.054	0.000
verb (past participle)	-0.011	0.002
verb (gerund)	-0.030	0.000
adverb	-0.064	0.000
preposition	-0.025	0.000

For a significance level of 0.05, we find out the following about the significance of the different features:

- Grammatical features (punctuation and word types) are significant, except for the overall length of the review. The average inverse word length is significant. Also, the longer the used words the more negative the review.
- Agreeableness and neuroticism are highly significant with large coefficients, confirming the results of the classification tree.
- Openness is narrowly a significant personality feature at a significance level of 0.05. Extroversion and conscientiousness are not significant.
- There is some evidence of significance of interaction terms. Neuroticism displays the most significant interaction terms with a lead of 4 terms. (e.g. Neuroticism-Horror, Neuroticism-Action-Adventure and Neuroticism-Drama and Neuroticism-Comedy). This is followed by

openness and agreeableness each with three significant interaction terms. Conscientiousness has only one significant interaction term (Conscientiousness-Drama) whereas extroversion has none. It is interesting to notice how the three most significant personality traits are the ones with the majority of the relevant interaction terms.

- Finally, we observe that there exists a significance for the openness and Art House movie interaction term which confirms our expectation that individuals with high openness are likely to prefer Art House & International movies.

5.2.2 Classification on Feature Combinations

Table 5.4: Logistic regression: Prediction accuracy results for multiple combinations of features.

Logit	Mean	SD
Bag-of-Words (500 dimensions)	0.818208	0.004017
Bag-of-Words (1000 dimensions)	0.838703	0.003986
fastText	0.833918	0.004688
Personality	0.625758	0.007798
Grammar	0.617597	0.00787
Sentiment	0.707071	0.008115
Interaction	0.624216	0.005333
Personality+Grammar+Interaction	0.674561	0.009727
BOW+Personality	0.856725	0.005142
BOW+Personality+Interactions	0.856619	0.004307
BOW+Personality+Interactions+Grammar	0.855582	0.003645
BOW+Personality+Grammar	0.857044	0.00388
fastText+Personality	0.83453	0.00455
fastText+Personality+Interactions+Grammar	0.837241	0.003128
fastText+Personality+Interactions	0.832164	0.004246
fastText+Personality+Grammar	0.837932	0.004317

The leading most optimal models in our research are built by using logistic regression. The Table 5.4 gives an overview of the performance of the different models when using logit. The results for support vector machine classification and naive Bayes classification can be found in the appendix.

The accuracy for a bag-of-words model with 500 dimensions is around 82%. Adding the mined five personality features increases the test accuracy to 86%. For comparison: doubling the feature space from 500 to 1000 words in the bag-of-words model only increases the accuracy to around 84%. This value is similar to a fastText model with 500 dimensions. This supports the hypothesis that the personality features add important and useful information to the word embeddings and contain useful information. Once the personality features are added to the model, further stylistic or interaction variables do not show an improvement.

Another interesting comparison is between the pre-trained VADER sentiment and the combination of all mined features, such as personality, grammar and genre interaction. The pre-trained sentiment analyser reaches an accuracy of around 71%. The mined features 67%. Although this is still 4 percentage points lower, the result is quite impressive. The VADER sentiment analyser was built highly sophisticated using human judgment and multiple rules for understanding language. By only using personality, grammar and genres, one can get results at a comparable level.

It is furthermore interesting that personality, grammar and genre interaction terms alone only get accuracies of up to 62%. It is their combined predictive power that increases the accuracy to 67%.

The following interesting observations from the above table are also noticeable:

- FastText

  FastText tends to provide useful representations for word embeddings. We notice, however, that the fastText model performs well on its own with the model predicting the correct sentiment of a review about 83% of the time. Additional personality or stylistic features show no improvement when added. This can be seen in the results in the table which show that the addition of other features to the fastText model does not result in an improvement. Hence, we could assume that fastText word embeddings hold sufficient useful information to perform the task of sentiment analysis well.

- Interactions

  The main idea behind building the movie-personality interaction model was to investigate if

individuals with certain personality features prefer a certain genre of movies. Hence, with an accuracy of about 62%, there is some evidence to support this hypothesis.

5.2.3 Separate Bag-of-Words Model

As we can see in Figure 4.1, the classification tree sets its first split for the agreeableness variable at a value of 0.327. Since this variable is normalized, a majority of data (around 12,000) belongs to the group with rather positive reviews and a minority (around 2,400) belong to a group with rather negative reviews. In the second layer of the tree neuroticism and agreeableness are used for splitting.

Table 5.5: Results for separate BoW models (decision tree thresholds)

	Size of Subset	Accuracy
Random Split 1	9405	0.82
Random Split 2	9405	0.81
High Agreeableness	2734	0.85
Low Agreeableness	16076	0.81
High Neuroticism	5091	0.82
Low Neuroticism	13719	0.82

The first split separates the data into a branch of high agreeableness with 81% positive reviews and another branch with lower agreeableness with only 45 % positive reviews.

Table 5.6: Results for separate BoW models (median thresholds)

	Size of Subset	Accuracy
Random Split 1	9405	0.82
Random Split 2	9405	0.81
High Agreeableness	14964	0.82
Low Agreeableness	3846	0.80
High Neuroticism	12248	0.82
Low Neuroticism	6562	0.82
High Openness	9405	0.81
Low Openness	9405	0.82

Using the tree-based splitting points, the following results were achieved (see Table 5.5). There is no improvement to the full bag-of-words model. The 85% accuracy have to be considered with the weight of their subset that is quite low. The higher accuracy here is probably due to the

higher null rate in the small subset.

The medians for agreeableness, neuroticism and openness are -0.22, 0.37 and -0.13, respectively. Using the median value for the splits, the following accuracies in Table 5.6 were reached. As can be seen, there is no improvement when splitting.

5.2.4 Neural Networks

In our neural network which is created using the Keras sequential model, we are able to achieve an accuracy of about 82%. It is interesting to notice that the neural networks model performs very similarly to the bag-of-words model, with both yielding similar accuracy levels for predicting sentiment correctly. Nevertheless our combined models with personality features show superior results.

6 | Limitations

The used pre-trained personality dictionaries are pre-trained on another context. Social media posts and movie reviews are different in many regards. Context matters highly regarding the interaction of personality and language. Therefore the transfer to amazon reviews from social network text has to be taken with caution.

Moreover, generalizations about personality and preferences might be biased due to the selection both in the WWBP dictionary generating process and in the decision to write a review. It is likely that rather extroverted consumers write reviews and that people who care about their data privacy opt out of the access to their social media posts.

Moreover, while the WWBP controls for age and gender, these attributes are missing in this study. It is likely that these factors are confounders for personality and genre preference.

The bag-of-words model is a very essential aspect of our personality model. However, it suffers from some shortcomings. For instance, it discards word order, ignores the context, and in turn meaning of words in the document (semantics). Context and meaning can offer a lot to the model, that if modeled could tell the difference between the same words differently arranged (“this is a nice movie” vs “is this a nice movie”), synonyms (“old bike” vs “used bike”), and much more. Furthermore, in terms of vocabulary, it might produce very spare vector representations for large vocabulary size because of the challenge of harnessing little information in a large representational space.

7 | Conclusion

7.1 Outlook

On a technical level, it would be of interest to investigate a multi-class classification and include neutral reviews as well to compare the performance of our models when neutral reviews play a role themselves as well. As shown in the appendix, other areas of reviews such as apps or health products are similarly interesting but show different behaviour when it comes to personality and sentiment prediction. Further investigations can be helpful here to determine the generalisability of the personality models.

One option, to improve the usage of personality dictionaries for different contexts, would be to manually delete all n-grams that will bias the results due to an assumed context shift. Although this would add a subjective component, it seems justifiable in many cases that have clearly different semantic meanings such as "murder" in the context of describing a thriller, compared to expressing one's feelings.

7.2 Guidelines

In this report, we show that personality features can, to some extent, be mined from text using dictionaries that are trained on other contexts. In our case, we use a dictionary that is trained on social media posts and extract personality features from Amazon movie reviews. Openness particularly seems to work well which might be due to its high correlation with misspellings which is less context-sensitive.

All five personality features, and openness in particular, can be used for recommendation systems. Instead of using only ratings for movies, music or products, the recommendation system can incorporate personality traits of its customers. This is promising because personality is an underlying attribute influencing various aspects of an individual such as taste. While there are many beneficial applications for such personality-based models, ethical considerations should not be neglected. Personality is a highly private and sensitive form of information. Therefore, users need to be made aware when their personality is being mined and any application should be based on the explicit, informed consent of the user.

Another interesting area of application is self-optimizing tools that help their users to maintain and increase healthy or productive habits. These applications can provide more targeted advice on the basis of personality features. For example, a more conscientious user will more easily follow challenging plans than a less conscientious user. Or for instance, extroverted users can benefit from group-based interventions and activities.

As part of this research, we communicated and exchanged ideas with two members of the data science team within Alpha. Alpha is a moonshot innovation start-up that was launched by a huge Spanish telecommunications firm whose aim is to try to tackle the social issue of health. They believe that one of the leading causes of chronic disease is an individual's everyday behaviour. Hence, their aim is to build an application that would prompt individuals to change their usual behaviour and recommend user-specific changes. This is where our research comes into play. We believe that our research would be very useful to Alpha as they could use it to build a recommendation system incorporating personality aspects.

However, the application of our personality model is definitely not restricted to the health industry. Both scientific research and companies have an interest in precise sentiment analysis. We have shown that the mined personality features can remarkably increase sentiment prediction accuracy. When working with sentiment analysis, be it in the context of scientific research or for corporations, this tool can improve the analysis. The closer the area of analysis is to the area on which the dictionary is trained on, the better the results will probably be.

7.3 Ethical Considerations

While knowing personality traits of people can support helping them, it also opens doors for manipulative applications. Personality is something very private and the usage of sensitive insights like personality can easily cross ethical limits.

We suggest to follow a code of conduct when mining personality and using these kind of information for any application.

Transparency

Most social media users or customers are not aware of the potential revelation of their personality traits when writing texts. Moreover, they do not expect their texts to be mined for this kind of insights. When collecting, mining and using these texts and personality features, affected individuals should be informed about the analysis. The analysis itself should only be conducted

when there is an explicit consent.

Ethical Areas of Application

We believe ethical applications are those that help customers. Potential use cases can be health or productivity applications where users try to increase their healthy habits and chose to obtain personality-targeted advice.

Any kind of manipulation, e.g. targeted advertisement or political influencing or the usage of mined personality for credit score ratings etc. can be harmful both for the individual and on a societal level and thus should be taken with caution.

8 | Bibliography

References

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5: 135–146, 2017.

Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer, 1999.

Fabio Celli and Luca Polonio. Relationships between personality and interactions in facebook. *Social networking: Recent trends, emerging issues and future outlook*, pages 41–54, 2013.

Fabio Celli and Luca Rossi. The role of emotional stability in twitter conversations. In *Proceedings of the workshop on semantic analysis in social media*, pages 10–17. Association for Computational Linguistics, 2012.

David Duuren. The relationship between personality and preference for either arthouse or mainstream movies. B.S. thesis, University of Twente, 2008.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 149–156. IEEE, 2011.

Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517. International World Wide Web Conferences Steering Committee, 2016.

Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014.

Margaret L Kern, Johannes C Eichstaedt, H Andrew Schwartz, Lukasz Dziurzynski, Lyle H Ungar, David J Stillwell, Michal Kosinski, Stephanie M Ramones, and Martin EP Seligman. The online social self: An open vocabulary approach to personality. *Assessment*, 21(2):158–169, 2014.

Hiroshi Maeda, Kazutaka Shimada, and Tsutomu Endo. Twitter sentiment analysis based on writing style. In *International Conference on NLP*, pages 278–288. Springer, 2012.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM, 2015.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

James Pennebaker and M Francis. Linguistic inquiry and word count: Liwc, 1999, 1999.

Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 180–185. IEEE, 2011.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61. Scipy, 2010.

Wanliang Tan, Xinyu Wang, and Xinyu Xu. Sentiment analysis for amazon reviews.

Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

9 | Appendix

Project Delays

We would like to thank our advisor and Alpha's representatives for their constant support as they tried to provide guidance in the best way possible despite some difficult circumstances. Within our research, we faced a lot of hurdles in terms of our Capstone partnership with Alpha mainly due to legal complications. We had to wait till July in hopes of acquiring the data our project was going to be based on but the issue unfortunately never got resolved. Moreover, as regards to the research question, due to the lack of data, it was a challenge to define one task that the report could focus on resulting in us simultaneously working on different research questions. However, what made the task even more challenging was the final research question which was agreed on was quite different to what we set out to do initially and it was decided on at the beginning of July with not a lot of time to spare. However, we are very thankful to our advisor and Alpha's representatives for their help and positive feedback. We tried our best to adjust and make the most out of a difficult position and we hope our hard work comes through in this research.

Further Results

Effect of Including Punctuation Marks

The following table shows the effect the inclusion of punctuation marks has on personality mining.

Table 9.1: Example of the effect of punctuation marks on subset of data

	Lowest Value	Highest Value
Agreeableness (incl. PM)	Best of the murder mystery spoofs.,If you like this one, see also "Murder by Death" and "Haunted Honeymoon".	great show
Agreeableness (excl. PM)	I liked the first one better. This had too many of the cliche moments I have come to hate.Especially the obvious fate of the fresh faced rookie. damn that pissed me off.	great show
Openness (incl. PM)	All I have to say about this movie is BUY IT!!!!!!!!!!!!!!!! Morgan Freeman, is acting skills, I can't describe!!!! The man is that good!!!!!!	Great Movie. I really enjoyed it. . . . .
Openness (excl. PM)	Best show ever cant wait to watch	If you like music like the old classic stuff then you will like this music.
Extroversion (incl. PM)	Great Movie. I really enjoyed it.	amazing movie!!!
Extroversion (excl. PM)	A mind-bending anime. Enjoyed very much!	Its silly.
Neuroticism (incl. PM)	All I have to say about this movie is BUY IT!!!!!!!!!!!!!!!! Morgan Freeman, is acting skills, I can't describe!!!! The man is that good!!!!!!	Bought this on a whim. I was bored so I turned it on. I was bored so I turned it off.
Neuroticism (excl. PM)	these two should stay as a team, they work off each other perfectly	Bought this on a whim. I was bored so I turned it on. I was bored so I turned it off.
Conscientiousness (incl. PM)	A mind-bending anime. Enjoyed very much!	great show
Conscientiousness (excl. PM)	A mind-bending anime. Enjoyed very much!	great show

Including the punctuation makes explanation marks and periods the most relevant factor for many features, as seen in Table 9.1.

Personality Feature Mining for Other Review Areas

The following two scatter matrices show personality features for other Amazon reviews. One can see that their predictive power for sentiment fluctuates. While Apps reviews are visible separated by personality, no such pattern is noticeable for health and personal care reviews. Purple points are negative reviews and yellow points are positive ones.

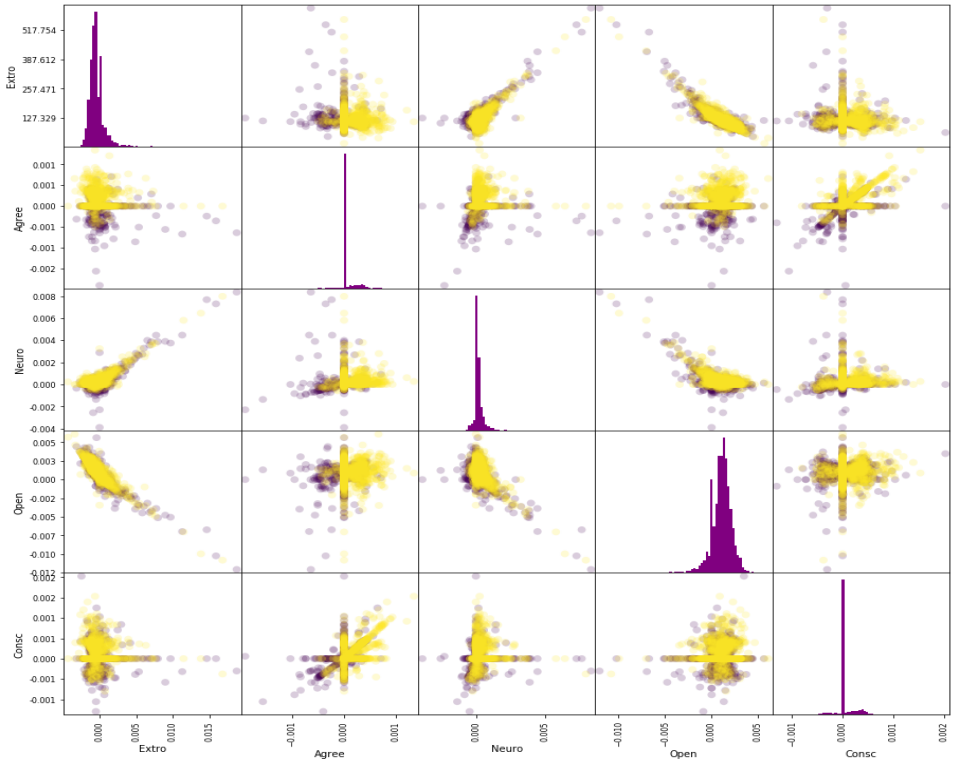

Figure 9.1: Apps reviews scatter matrix. Personality features seem to distinguish the rating as good as for movie reviews. Classification results confirm this.

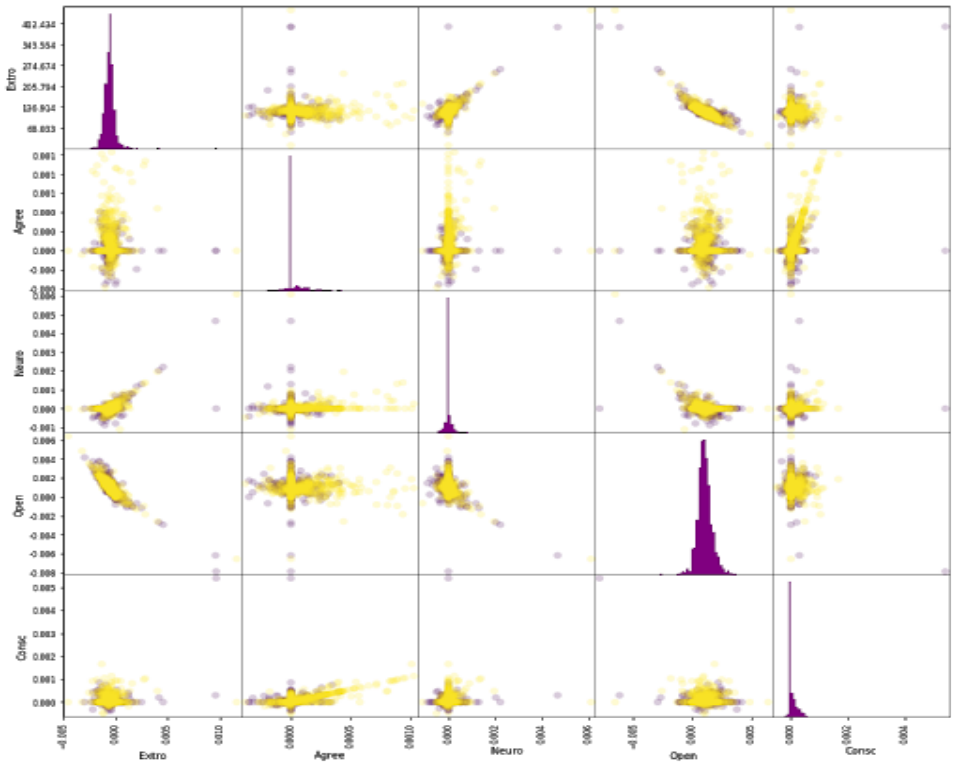

Figure 9.2: Health and personal care data scatter matrix. No visible classification of reviews by personality can be found. Classification results confirm this.