Text mining approach to social media posts and MBTI
STA2101 Final Project

Yuxuan Wang

December 13, 2022

1 Part I - Non technical

1.1 Scientific problem of interest

MBTI [1] stands for Myers Briggs Type Indicator, which is a personality type system that divides everyone into 16 distinct personality types across 4 axis:

- Energy axis: Introversion (I) – Extroversion (E)
- Information axis: Intuition (N) – Sensing (S)
- Decision axis: Thinking (T) – Feeling (F)
- Lifestyle axis: Judging (J) – Perceiving (P)

MBTI is an increasingly popular tool which is frequently used to help individuals understand their own communication preference and how they interact with the world around them. Having an awareness of what MBTI is can help people adapt their interpersonal approach to different situations and audiences.

With different MBTI personal type, people tends to behave, think and express themselves in different ways. In this project, we are particularly focusing on the specific aspect of how they express themselves on internet. The paper will investigate this particular scientific problem of interest, trying to build a model to evaluate their MBTI type by their posts on social media. Exploring this question can help people have a better understanding of their personality type in a new way, which will help them adjust their relationships with others in their daily lives.

1.2 How and why the data was collected

This data was collected through the PersonalityCafe [2] forum, as it provides a large selection of people and their MBTI personality type, as well as what they have written. Since MBTI is used widely in businesses, for research, for fun and lots more, numeric MBTI personality test has came out, but the reliability of them need to be questioned. A simple search on google will show people lots of unreliable tests, which might lead to different unfaithful results. This dataset was aimed to improve the uncertainty of the MBTI test by achieve a better connection between MBTI and the people by data mining. The dataset provides a chance to implement a interdisciplinary approach to MBTI. Psychological analysis, statistical model and evaluation, machine learning based natural language computing can all contribute to this specific scientific problem. Since I do not have a background of psychology, this paper will focus on the language processing and statistical modelling.

1.3 Data and preprocessing

The dataset contains 8675 observations of data, for each observation is a person's MBTI personality type and a section of each of the 50 things they have posted on social media, which are seperated by '|||'. A brief presentation of an observation is showed in Table 1.

id	MBTI_type	posts
24	INFP	45016 urh sorry uh. couldn't resist.||| all of you enfjs, please collectively marry me.||| oh my goodness... these are great.||| **(other 46 posts)**||| Consider how stressed you are about this as well. Sometimes the 'obivious' thing is the right thing. It's easy to over- analyze.'

Table 1: Brief presentation of data

The distribution of the MBTI type for the dataset is described in Figure 1.

Since the raw data is a mixture of 50 posts with redundant and meaningless components, we need to do the preprocessing. We first constructed a filter list, which contains stop words,

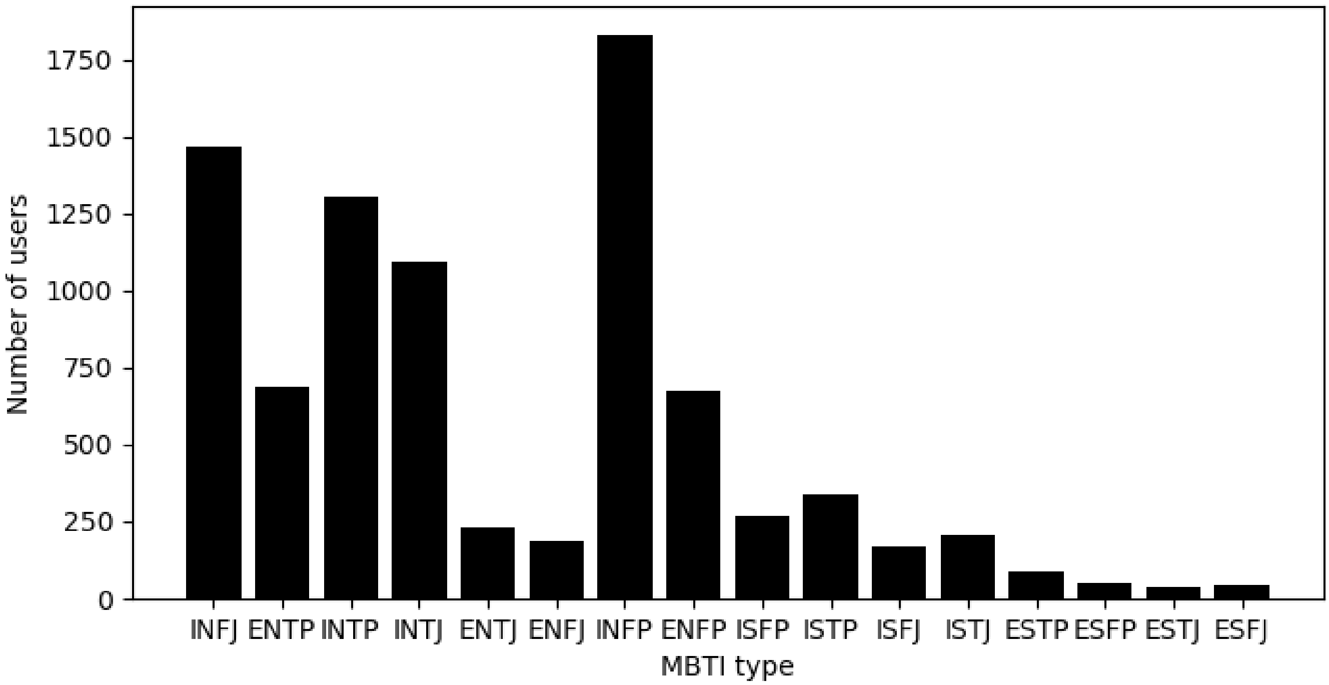

Figure 1: Distribution of MBTI

punctuation and numbers. We then split the 50 posts of each user. For each post, we apply the filter list to remove the redundant characters. In addition, we removed the posts contains website url link since those links are hard to analysis. Finally, we removed the case sensitivity and achieved the word lemmatizer. After all, we concatenate the posts for each person together. We will implement further feature extraction to these posts in the following sections. An example of how the post is cleaned is shown in Table 2.

Original post	Cleaned post
Steve Job's was recognized for his striving for efficiency and practicality. His genius is in his systemization of inventions, less so than in invention. This is where claims of Se and Te come from.	steve job recognized striving efficiency practicality genius systemization invention le invention claim se te come

Table 2: Example of preprocessing

1.4 Non-technical summary

In order to solve this scientific problem, we need to build a model [3] which can estimate the MBTI personal type by the text they posted. There are various methods to build the relationship between them. If we want to explore them in a direct way, which means an end to end model from the text to the MBTI prediction, that work will contain deep learning based natural language model like LSTM. But that's much about machine learning and it's not the content of this course. To build a more statistical model, we need to first extract several features from the post texts. In this paper, we particularly chose 9 topics generated by Latent Direchlet Allocation [4] and 2 sentiment scores produced by NLTK [5] based on Opinion Lexicon [6]. Our goal is to fit a model which is complex enough to fit the data well and simple enough to interpret. We first use all of the features to construct the model, then remove the features which do not have significant contribution to the model until the rest explanatory variables can provide a adequate fit. Finally we will do several test to check whether the model explain the dependent variable well and have anything unusual or unreasonable about the model.

Since MBTI have 16 personal types, we separately build four models for each axis of it to have better approach toward the analysis. For each model, the selection of explanatory variables are not same, but there are some global cases: the score of positive sentiment plays a more significant role in the model than the negative sentiment. Despite it is rather hard to make the prediction only by the text, our work still produce a significant progress toward the problem. The data preprocessing is done by python in jupyter notebook, the modeling and analysis are done by R in R markdown.

2 Part II - Technical

2.1 Model and analysis

2.1.1 Latent direchlet allocation

Latent direchlet allocation is a generative probabilistic model for collections of discrete data. It explains a set of observations through unobserved groups, and each group explains the similarity of the particular parts of data. In the context of text mining in this paper, it provides a topic distribution of each post(document) in the whole dataset(corpus). LDA have the assumption of the generative process of the document is in the following way:

1. Choose N from Poisson(ϵ).
2. Choose θ from Dir(α).
3. Choose each of the N words w_n:
    (a) Choose a topic z_n from Multinomial(θ).
    (b) Choose a word w_n from p($w_n|z_n, \beta$), which is a multinomial probability conditioned on topic z_n.

Where α and β are corpus level parameters for the generation from all posts to topic distribution and the generation from a topic to each word occurred in the corpus. θ is the document level parameter representing the topic distribution for the specific post, which is used in our further logistic regression [7] model. z_n and w_n are word level parameters showing the topic and the word itself for the specific word in the document.

After acknowledging the assumption, we can infer that the joint distribution of topic distribution θ, topics z_n and word w_n can be expressed in the following probability distribution in condition of α and β:

$$p(\theta, z_n, w_n|\alpha, \beta) = p(\theta|\alpha)p(z_n|\theta)p(w_n|z_n, \beta)$$

After taking the product of each word in the document, integrating over θ, summing over z, and taking the product of the marginal probabilities of single document, we obtain the probability of the whole corpus:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d$$

We can use expectation-maximization or Gibbs sampling to estimate the parameters in the equation. Finally, we get the topic distribution for each post θ. The only hyperparameter for this task is the numbers of topics k. Our setting is k = 10 in order to meet the requirements of this project. The output will give us the probability distribution for every post toward 10 topics.

2.1.2 Sentiment analysis

For sentiment analysis, we implemented python NLTK package based on Opinion Lexicon. Opinion Lexicon is a widely used model in the area of sentiment analysis and opinion mining from social media. [8] The lexicon contains a list of around 6800 words which express positive and negative opinions. The article's sentiment is evaluated based on the number of positive words, negative word and neutral words in the main part of the sentence, which are counted based on the lexicon. The output of the opinion mining is a feature based summary of the document input. The result will give us three scores of positive, negative and neutral for every post.

2.1.3 Model selection

To begin with the model building, my first idea was to build a multiclass logistic regression to directly predict the 16 MBTI type by the dependent variables. However, the result turned out to be not so good. Then, we decided to build four logistic regression models to separately predict four axis of MBTI, which means each logistic regression is a binary classifier. The assignment from MBTI axis toward target value of independent variable is shown in Table 3. We will select parameters and perform checking for each model separately.

Axis	type assigned to 1	type assigned to 0
Energy	Extraversion(E)	Introversion(I)
Information	Sensing(S)	Intuition(N)
Decision	Thinking(T)	Feeling(F)
Lifestyle	Judgement(J)	Perception(P)

Table 3: Assignment of target values

We totally have 13 dependent variables, 10 of them are topic distributions from LDA and 3 of them are sentiment scores from sentiment analysis. Since the topic distributions and sentiment scores all have the property that they sum to 1, we choose 9 of the 10 topics distributions and positive and negative scores as our original dependent variables in order to prevent the multicollinearity.

For each of the four models, we first use all of the 11 original dependent variables. For model selection, we consider the p-value of the dependent variables. For each iteration, we check the dependent variable with the highest p-value. If removing it can gives the model a lower BIC [9], we will remove it and check the next dependent variable. If removing it make the BIC of the model become higher, we will not remove it and stop the process. After the process stopped, we will double check whether removing other parameter will make BIC lower, the result is they will not.

The final model of energy axis contains topics 1, 3, 5, 7, 8, 9 and positive score. The final model of information axis contains all topics from 1 to 9 and positive score. The final model of decision axis contains all topics from 1 to 9 and positive score. The final model of lifestyle contains topics 2, 3, 5, 6, 8, 9. Negative sentiment score are not included in all models due to low significance. The BIC of the models are shown in Table 4.

Axis	BIC of original model	BIC of revised model
Energy	8817.78	8788.18
Information	6461.67	6457.42
Decision	6371.30	6367.28
Lifestyle	11411.13	11378.75

Table 4: BIC metrics of models

Finally, it is crucial to do the data and model check after constructing our model. For the data check [10], we used cook's distance plot and jackknife residuals, which are all widely preferred regression model diagnostics. Due to the space limitation, we only use plots of energy axis as example. Their plots are shown in Figure 2 and Figure 3 in the appendix.

From the two plots, we can clearly see that there do have some outliers that will affect the fit of the model. This is in the range of tolerance because they are not strong enough to significantly influence the model, and this kind of condition is ubiquitous in the field of text mining on account of the uncertainty of the text volatility.

For the model check [11], besides the BIC check we already done, we also implemented anova to the original model and the model after revised to check whether the simplified models are still adequate. The result are presented in Table 5.

Axis	Deviance	Degree of freedom
Energy	6.6713	4
Information	4.8216	1
Decision	1.3538	1
Lifestyle	12.962	5

Table 5: Anova test

From the table, we can see that the deviance is acceptable, while simplifying the model, the underfit situation did not occur.

2.2 Technical summary

For the LDA part, we extracted 10 topics from the posts. For each topic, we calculated several relevant words with the highest probabilities, which is helpful for us to explore its practical implications. Some examples are shown in Table 6.

topic	key words
2	love, relationship, friend, like , guy
7	type, think, infj, intj, mbti
9	music, like, book, movie, song
4	would, people, think, point, say

Table 6: Example of topics

From the table, we can see that most key words of the topics can build an understandable theme, topic 2 is about relationship with others, topic 7 is about MBTI personality type, topic 9 is about hobbies. But there are some topics like topic 4, which do not have a specific theme. As a result, topic 4 is used less in the models than other topics, which meets our anticipation and perception. So, BIC and anova indexes are necessary to help us adjust which of the topics are be used as explanatory variables.

For the sentiment analysis part, we can see that positive attitude score is used in three models and the negative attitude score is not used in all models. This infers that people's positive attitudes on social media is more related to their personal type than their negative attitudes. This might partly reached beyond our common sense, but it is supported by significance indicators.

For the modeling part, the summary of regression are shown in Table 7,8,9,10 in the appendix. The P-value of all explanatory variables are all low, which shows that all of the terms are significant. Since there are four models, we cannot concisely give a summary for all models in a few sentences. But we can make a conclusion that the positive value of topics' coefficient estimation shows that the people who posts more about this topic have a larger probability of belonging to the personality type assigned to value 1, and vice versa. The positive value of the positive attitude score's coefficient estimation shows that the people who posts more positive text have a larger probability of belonging to the personality type assigned to value 1, and vice versa. In addition, topic 3, 5, 8, 9 are important topics since they contribute to all of the four personality axis.

2.3 Future Work

Due to time limitation, this paper only concentrates on two type of features extracted from social media posts. We can implement some further methods to construct more explanatory variables. This will absolutely improve not only the fit and integrity of the model, but also the interpretability and practical meaning. Moreover, we might try to build the model on top of the feature embedding generated by deep learning based transformer [12], which might significantly improve the quality of the extracted features. But there will be a trade off of reducing interpretability and practical meaning. Finally we might combine our work with some psychological models to achieve some improvement. But I do not have an exact idea about it since I do not have any psychology background knowledge.

3 Appendix

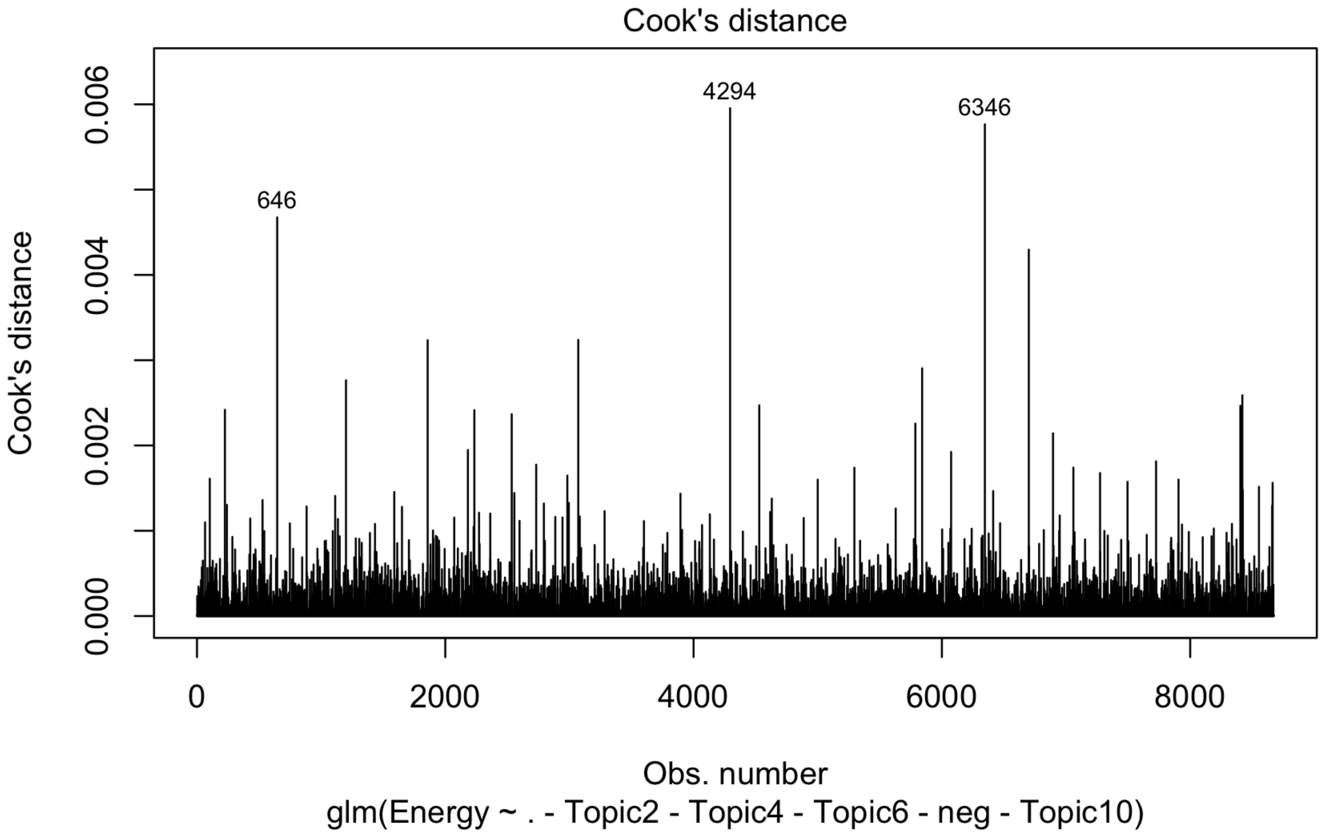

Figure 2: Cook's distance

Term for Energy	P-value	Estimate	2.5%	97.5%
Topic1	0.00367	-1.58	-2.65	-0.52
Topic3	1.85e-09	-2.68	-3.56	-1.81
Topic5	0.00066	-1.79	-2.82	-0.76
Topic7	<2e-16	3.37	2.72	4.02
Topic8	<2e-16	4.58	3.71	5.45
Topic9	<2e-16	-5.58	-6.71	-4.48
positive	<2e-16	5.61	4.61	6.61

Table 7: Summary for Energy axis

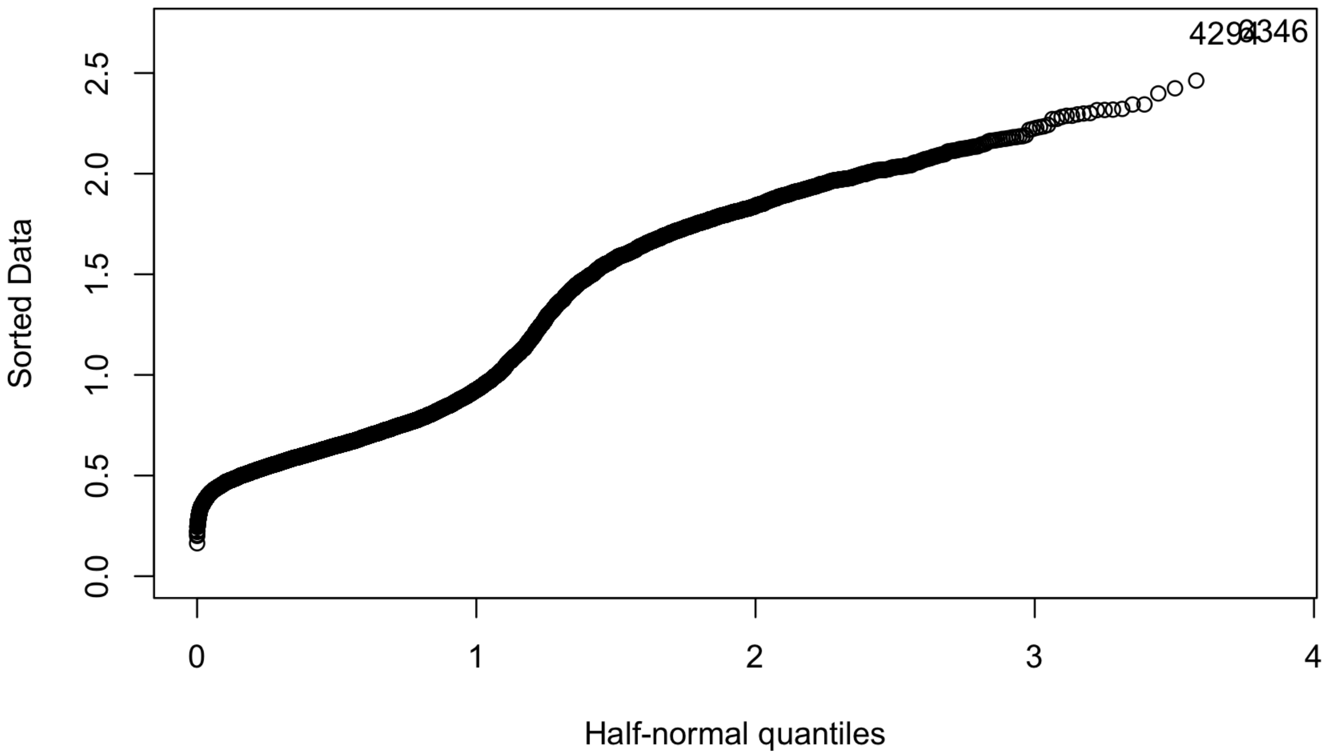

Figure 3: Jackknife residuals

Term for Information	P-value	Estimate	2.50%	97.50%
Topic1	<2e-16	-7.16	-8.67	-5.66
Topic2	<2e-16	-11.81	-13.30	-10.34
Topic3	7.25E-15	-5.24	-6.56	-3.92
Topic4	<2e-16	-15.65	-17.39	-13.95
Topic5	<2e-16	-7.47	-8.97	-5.99
Topic6	<2e-16	-6.88	-8.37	-5.40
Topic7	4.66E-10	-3.10	-4.08	-2.12
Topic8	5.40E-09	-3.80	-5.08	-2.53
Topic9	<2e-16	-8.32	-9.84	-6.83
positive	0.0199	1.85	0.29	3.41

Table 8: Summary for Information axis

Term for Decision	P-value	Estimate	2.50%	97.50%
Topic1	<2e-16	-7.49	4.97	6.54
Topic2	<2e-16	-14.45	-8.73	-6.27
Topic3	<2e-16	-5.33	-15.71	-13.21
Topic4	<2e-16	7.80	-6.45	-4.22
Topic5	<2e-16	-8.15	6.56	9.06
Topic6	<2e-16	-7.07	-9.34	-6.96
Topic7	<2e-16	-4.48	-8.31	-5.84
Topic8	2.45E-12	-3.98	-5.37	-3.60
Topic9	<2e-16	-10.46	-5.09	-2.87
pos	0.0054	-1.75	-11.72	-9.21

Table 9: Summary for Decision axis

Term for Lifestyle	P-value	Estimate	2.50%	97.50%
Topic2	9.18E-08	1.73	1.09	2.36
Topic3	5.12E-12	-2.42	-3.10	-1.73
Topic5	5.06E-07	2.15	1.31	2.99
Topic6	3.40E-04	1.38	0.63	2.14
Topic8	<2e-16	-3.99	-4.83	-3.17
Topic9	6.86E-09	-2.55	-3.41	-1.69

Table 10: Summary for Lifestyle axis

References

[1] Isabel Briggs Myers. The myers-briggs type indicator: Manual (1962). 1962.

[2] Jeff Ihaza. Personality cafe. *The New York Times Magazine*, pages 22–L, 2019.

[3] Julian J Faraway. *Linear models with R*. Chapman and Hall/CRC, 2004.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[5] Amrita Shelar and Ching-Yu Huang. Sentiment analysis of twitter data. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1301–1302. IEEE, 2018.

[6] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[7] Julian J Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC, 2016.

[8] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

[9] Kenneth P Burnham and David R Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[10] Azzedine Boukerche, Lining Zheng, and Omar Alfandi. Outlier detection: Methods, models, and classification. *ACM Computing Surveys (CSUR)*, 53(3):1–37, 2020.

[11] Edmund M Clarke. Model checking. In *International Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 54–56. Springer, 1997.

[12] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.