5-2020

Predicting Consumers' Brand Sentiment Using Text Analysis on Reddit

Puti Cen

Predicting Consumers' Brand Sentiment Using Text Analysis on Reddit

Abstract

With the emergence of data privacy regulations around the world (e.g. GDPR, CCPA), practitioners of Internet marketing, the largest digital marketing channel, face the trade-off between user data protection and advertisement targeting accuracy due to their current reliance on PII-related social media analytics. To address this challenge, this research proposes a predictive model for consumers' brand sentiment based entirely on textual data from Reddit, i.e. fully compliant with current data privacy regulations. This author uses natural language processing techniques to process all post and comment data from the r/gadgets subreddit community in 2018 – extracting frequently-discussed brands and products through named entity recognition, as well as generating brand sentiment labels for active users in r/gadgets through sentiment analysis. This research then uses four supervised learning classifiers to predict brand sentiments for four brand clusters (Apple, Samsung, Microsoft and Google) based on the self-identified characteristics of Reddit users. Across all four brand clusters, the predictive model proposed by this research achieved a ROC AUC score above 0.7 (three out of the four above 0.8). This research thus shows the predictive power of self-identified user characteristics on brand sentiments and offers a non-PII-required consumer targeting model for digital marketing practitioners.

Keywords

brand sentiment; Reddit; natural language processing; consumer insight mining

Disciplines

Business

Predicting Consumers' Brand Sentiment

Using Text Analysis on Reddit

By

Puti Cen

The Wharton School, School of Arts & Sciences

An Undergraduate Thesis submitted in partial fulfillment of the requirements for the

JOSEPH WHARTON SCHOLARS

Faculty Advisor:

Professor Chris Callison-Burch

Associate Professor, Computer and Information Science

THE WHARTON SCHOOL, UNIVERSITY OF PENNSYLVANIA

MAY, 2020

ACKNOWLEDGEMENT

First and Foremost, I would like to thank my thesis advisor, Professor Chris Callison-Burch. Thank you for introducing me to the field of computational linguistics and the various wonderful applications of natural language processing (e.g. detecting depression and monitoring gun violence). This interdisciplinary project would not be possible without your wise guidance and kind support.

Thank you to Professor Daniel Goldstein, who showed me the power of digital marketing and inspired me with his research on social targeting and consumer behavior prediction. Thank you to Professor Catherine Schrand and Dr. Utsav Schurmans for guiding me through this year-long thesis-writing process.

Thank you to Aditya M. Kashyap for sharing the Reddit dataset and offering all the technical help whenever I am stuck. Thank you to my WH399 group for all the great advice on my thesis proposal.

Thank you to Wu, Heng, Helen and Jacqueline for offering me the best research environment I could hope for during the COVID-19 pandemic. Thank you to Nova for all the mental support and idea-bouncing – this project gets better with every conversation we have. Thank you to all my friends, near and far, for giving me the strength to keep working on this project even in times of uncertainty due to the pandemic.

Thank you to my parents – thank you for all the love and unwavering support, without which I would not have had the courage to pursue this interdisciplinary path both for my thesis and for my college career.

ABSTRACT

With the emergence of data privacy regulations around the world (e.g. GDPR, CCPA), practitioners of Internet marketing, the largest digital marketing channel, face the trade-off between user data protection and advertisement targeting accuracy due to their current reliance on PII-related social media analytics. To address this challenge, this research proposes a predictive model for consumers' brand sentiment based entirely on textual data from Reddit, i.e. fully compliant with current data privacy regulations.

This author uses natural language processing techniques to process all post and comment data from the r/gadgets subreddit community in 2018 – extracting frequently-discussed brands and products through named entity recognition, as well as generating brand sentiment labels for active users in r/gadgets through sentiment analysis. This research then uses four supervised learning classifiers to predict brand sentiments for four brand clusters (Apple, Samsung, Microsoft and Google) based on the self-identified characteristics of Reddit users. Across all four brand clusters, the predictive model proposed by this research achieved a ROC AUC score above 0.7 (three out of the four above 0.8). This research thus shows the predictive power of self-identified user characteristics on brand sentiments and offers a non-PII-required consumer targeting model for digital marketing practitioners.

Keywords: brand sentiment; Reddit; natural language processing; consumer insight mining

1. INTRODUCTION

Since the advent of Internet, this highly-individualized channel for information foraging has been the top of mind for marketers interested in reaching potential customers with personalized messages and offers. However, with the rapid evolution of technology, the enormous amount of personal data available online and the lack of agency over one's own data have triggered concerns from regulators, scholars and average consumers (Charters, 2002). For digital marketers, this rising trend leads to the fundamental question in Internet advertising practices: Is personally identifiable information (PII) necessary to serve accurate and effective advertisement to Internet users? In other words, with data privacy regulations tightening worldwide, is it possible for marketers to target consumers based solely on insights generated from their words, without any other personal information? Interested in this topic as a student in marketing and computer science, this author would love to explore the potential of alternative, non-PII-required techniques, i.e. natural language processing, in business settings such as consumer insight mining.

This research is important for researchers and industry practitioners as it addresses the potential threat to the long-thriving digital marketing landscape, i.e. user data privacy. Currently, the use of private user data is essential to online advertising providers (e.g. Google Ads, Amazon Advertising) as these data on users' search and browsing history often reveal users' personal demographic or psychographic characteristics, shopping preferences and topics of interests, which in turn can be used by digital marketers optimize the targeting of their advertising messages (Rossi, McCulloch and Allenby, 1996). In the broader digital marketing landscape, Internet advertising surpassed TV advertising to become the largest digital marketing channel in 2016, placing data-powered Internet marketing at the core of digital marketing industry. Therefore, the increasing regulatory focus on data privacy should

be regarded not only as a threat to data-powered Internet advertising, but also the broader digital marketing industry.

Since European Union's General Data Protection Regulation (GDPR) coming into effect in May 2018, multiple governments and organizations around the world have enacted similar regulatory processes to offer their citizens full control over their own digital data, including browsing history, cookies and even profiles (EU General Data Protection Regulation, 2016). Therefore, anticipating a decline in user data access in the coming years, it's important for digital marketers to find an alternative way to profile their potential customers online for personalized communications, without breaching users' data privacy.

This research thus explores a consumer insight mining model that uses natural language processing techniques to extract and predict consumers' brand sentiments based on users' publically shared content online, fully in compliance to data privacy regulations around the world. To further ensure the consistent availability of user data, this author uses textual data from Reddit, where all posts and comments under subreddits are always visible to everybody – different from other social media platforms such as Facebook or Twitter.

Based on the homophily theory in marketing (Shalizi and Thomas, 2011) and existing research on the effectiveness of social targeting (Goel and Goldstein, 2014), this author further advances the argument by proposing that people with similar self-identified characteristics share similar brand sentiments. Motivated by this hypothesis, the model uses both demographic and psychographic characteristics – self-identified and shared by Reddit users – to predict their sentiments towards specified brands and products.

This author uses natural language processing techniques to process Reddit textual data: named entity recognition for the extraction of frequently-discussed brands, and sentiment analysis for consumers' sentiments towards those brands and their products. This author then uses four different machine learning algorithms (Logistic Regression,

Complement Naïve Bayes, Bernoulli Naïve Bayes and Random Forest) to develop the model and compare the skills of sentiment analyzers and machine learning algorithms with Receiver Operating Characteristic (ROC) curve.

Across all four brand clusters, the predictive model proposed by this research achieved a ROC AUC score above 0.7 (three out of the four above 0.8). This research thus shows the predictive power of self-identified user characteristics on brand sentiments and offers a non-PII-required consumer targeting model for digital marketing practitioners.

2. EXISTING RESEARCH

(1) Digital Marketing Powered by Social Media Analytics

With the explosive growth of Internet technology and customers' adaptation to the Web 2.0 era, social media, also known as customer-generated media, has become one of the most important communication channels for marketers in 21th century. In contrast to traditional marketing communication channels such as print advertising and personal selling, social media marketing offers a cost-effective approach to reach more audience as the impact of customer-to-customer communications is magnified in the digital marketplace (Mangold and Faulds, 2008).

While the "free" word of mouth marketers get on social media comes at marketer's less of control over customer-to-customer messages, it also provides massive behavioral data for marketers to analyze for better segmentation and targeting. This remedy, known as social marketing analytics, has emerged in recent years as "a study of social media metrics that help drive business strategy (Sterne 2010)." Specifically, for digital marketers, social media analytics enables them to create customer-centric marketing (e.g. behavioral targeting, personalized messaging, social customer relationship management) and improve the effectiveness of their marketing communications (Misirlis and Vlachopoulou 2018). Across

age groups and business verticals, data-powered, personalized marketing has been proved to be more effective than generic, default marketing communications within same digital channels (Sunikka, Bragge and Kallio, 2011; Taken, 2012).

(2) GDPR and Its Impact on Data-Powered Marketing

However, with the increasing sophistication of data-powered digital marketing including "personalized product offerings and recommendations, price discounts, free services, and more relevant marketing communications and media content (Martin and Murphy, 2017)," there has been increasing privacy concerns due to potential consumer data abuses, which translate to unwanted marketing communications as well as highly targeted, obtrusive marketing communications in the digital marketing world. In addition to extensive academic research striving to define and raise concerns for these privacy issues (Nissenbaum, 2009), policy makers have reacted to this privacy trend as well, with the most representative being European Union's General Data Protection Regulation (GDPR) that came into effect in May 2018.

While the exact financial impact of GDPR on targeted digital marketing is unclear, a one-year-in pulse check on European Union's web traffic and e-commerce sales shows a ten percent decrease in recorded pageviews and recorded revenues while controlling for a user's average time on site and average page views per visit (Goldberg, Johnson and Shriver, 2019). Qualitatively speaking, while not all consumers will exercise their GDPR rights to have their data removed for targeted marketing purposes, the decision architectural change from opt-in to opt-in would anticipate a significant decrease in data-sharing for digital marketing, as illustrated by Johnson, Bellman and Lohse in their research on default framing and privacy (2002). What's more, the data privacy regulation trend started by GDPR has just started for digital marketers around the world (Breitbarth, 2019). From America's California Consumer

Privacy Act (CCPA) to Brazil's *Lei Geral de Proteçao de Dados* (LGPD) which will both come into effect in 2020, digital marketers need an alternative way to profile their potential customers online for personalized communications, without breaching users' data privacy.

(3) Text Mining on Social Media

One alternative of non-intrusive online profiling is text analysis on social media posts. For instance, sentiment analysis, a natural language processing technique that identifies attitudes embedded in speech, can help determine a user's opinions on brands and services without accessing private data such as personal profiles or purchase history. In addition, sentiment analysis on consumer-to-consumer communications also offers insights on how consumers' preferences are influenced by others' opinions and thus help marketers understand how to best control and utilize these eWOM (Neri, Aliprandi, Capeci, Cuadros and By, 2012).

Existing research on social media text analysis have largely focused on microblog datasets like Twitter, Facebook and Sina-Weibo given their huge user base and readily available Application Programming Interface (API) for collecting public data from these sites (Ravi and Ravi, 2015). For applications most relevant to this research, i.e. profiling users based on the ways they speak, researchers have used sentiment analysis to develop better recommender systems where like-minded people (clustered by their speech style) receive similar interest recommendations on micro-blogging sites (García-Cumbreras, Montejo-Ráez and Díaz-Galiano, 2013; Bao, Li, Liao, Song and Gao, 2013).

(4) Social Targeting & Homophily in Digital Marketing

While this research is the first to apply text analysis on social media data to consumer preference prediction, existing marketing research have established connections between

consumer homophily and behavioral similarity both offline and online. A term most used by sociologists in the study of social networks, homophily describes the phenomenon where people with similar demographics and psychographics tend to come together and consequently influence each other behaviorally due to physical or psychological proximity (McPherson, Smith-Lovin and Cook, 2001; Shalizi and Thomas, 2011). Examining this behavioral influence in digital marketing contexts, researchers find homophily a robust predictor of consumer preference and recommend social targeting for online advertisers when transactional data is unavailable for accurate behavioral targeting (Goel and Goldstein, 2014).

Building upon existing research, this author applies text analysis techniques to identify similar individuals online and examines if self-identified characteristics online can effectively predict consumer preferences – as suggested by existing social targeting research.

3. METHODOLOGY

(1) Data

(1) Reddit

This research uses Reddit as the textual data source for two reasons. First, with 330 million monthly active users in 2018 (Reddit, 2018) and 1.7 billion unique web visits in July 2019 (SimilarWeb, 2019), Reddit is one of the most popular online communities nowadays with a demographically and psychographically representative user base. Thus, using Reddit data offers greater external validity in comparison to niche web forms where selection bias may undermine the generalizability of proposed targeting methodology. Second, compared to other popular social sites such as Twitter and Facebook, Reddit has maximized content availability for researchers as it has no character limit and little privacy settings for its threads. In other words, Reddit is a larger gold mine compared to tweets and a more public one compared to Facebook.

While Reddit offers its own API for post scraping and data collection, this author uses the Reddit dataset collected and generously shared by Professor Callison-Burch and his PhD student Aditya M. Kashyap. Located in NLP Grid, the computing cluster shared among natural language processing researchers at University of Pennsylvania, the dataset contains all post-related information (38 properties per post, see Appendix A for sample data format) for all subreddits and all Reddit users between January 2006 and December 2018. For the purpose of this research, only three properties are considered: author (i.e. user name on Reddit), body (i.e. the content of the post or comment) and subreddit (i.e. the within-Reddit community where the post is submitted).

(2) r/gadgets in 2018

To limit the scope of data processing for this pilot research in consumer sentiment mining, this author uses 2018 Reddit data for its recency and its representativeness in terms of potential seasonality effects. Furthermore, to reveal consumer sentiment towards certain brands and products, this research chooses r/gadgets as the target subreddit for text analysis as most of the comments and posts within r/gadgets are product-focused and opinionated towards different electronic brands. Created in 2007 with a member base of 16.7 million Reddit users as of May 2020, r/gadgets (https://www.reddit.com/r/gadgets/) is a subreddit community "all about discussing, reviewing, and enjoying gadgets." In 2018, out of the total 1,133,665,361 posts on Reddit, r/gadgets has a total of 340,824 posts and comments, with a monthly average of 28,402 posts and comments (see Table 1 for detailed statistics on number of posts in r/gadgets versus the entire Reddit platform).

(3) Characteristics of All Reddit Users

In addition to post data on Reddit, this research uses the Reddit user characteristics dataset for user profiling. Generously shared by Professor Callison-Burch and Aditya M. Kashyap, the characteristics dataset contains all self-identified characteristics of Reddit users, collected from "I am (characteristic)" expressions in posts and comments on Reddit. For instance, User A's comment "I am a college student" will result in an entry associating User A with the characteristic "college student" in the dataset. The dataset contains 2,981,575 characteristics extracted from Reddit, with each characteristic associated with multiple Reddit users. For the online profiling part of this research, this author transformed the dataset to a "User → Characteristics Set" format, i.e. associating each Reddit user with multiple self-identified characteristics (see sample data in Appendix A).

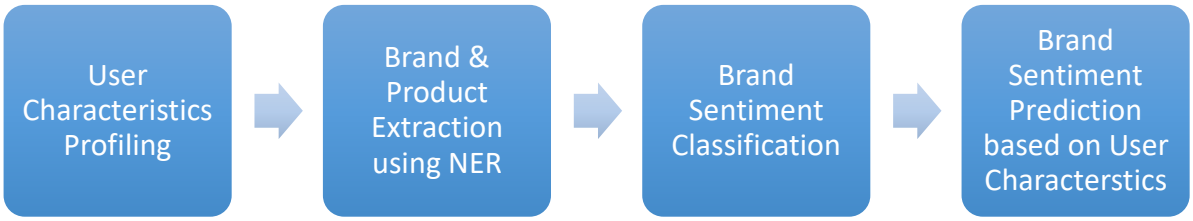
(4) Active r/gadgets Users in 2018

For maximized text analysis accuracy and best marketing application of predictive model, this research only analyzes the behavior of r/gadgets users who are active (i.e. posted more than 30 times) in the subreddit community in 2018. From a technical perspective, the sentiment label generated for each user comes from a stable average, less likely to be influenced by special contexts. From the business side, targeted advertising is more effective when displayed on websites a user frequents to, so predicting the sentiment of users who barely visit a particular website adds little value to the digital marketers' decision process.

(2) Method

The main analysis in this research follows a similar structure as Chamlertwat, Bhattarakosol, Rungkasiri and Haruechaiyasak proposed in their consumer insights study using Twitter data for sentiment analysis (Chamlertwat, Bhattarakosol, Rungkasiri and

Haruechaiyasak, 2011). Specifically, with readily available characteristics data on Reddit users, the analysis can be divided into three main stages: topic extraction, sentiment classification and predictive model evaluation.



(1) Brand & Product Extraction with Named Entity Recognition (NER)

In the first stage of main analysis, this author identifies the electronic products and their brands frequently discussed by Reddit users in the r/gadgets subreddit. To ensure both the comprehensiveness and the accuracy of this topic extraction process, this author first uses pre-trained named entity recognition models from spaCy to generate a list of brands and products, and then manually sifts the list to identify the most frequently-discussed brands (and their products) in r/gadgets.

For NER, this author uses the en_core_web_md model from spaCy as it performs significantly better than en_core_web_sm while insignificantly worse than en_core_web_lg with a F-score of 86.25 for NER tasks. The en_core_web_md model is an English multi-task CNN trained on OntoNotes, with GloVe vectors trained on Common Crawl. For NER tasks, it assigns various labels to identified entities, such as “CARDINAL,” “ORG,” “LANGUAGE,” “PERSON” and “PRODUCT.” For this research, only entities with “ORG” and “PRODUCT” labels are considered.

When sifting through the entity list, this author groups all entities by electronic products and their brands, while discards all “ORG” and “PRODUCT” entities unrelated to the gadgets product space. Recognizing that there are certain limitations to the NER topic extraction approach, such as failure to identify non-capitalized brands and products that share

spelling with normal entities (e.g. “Apple” vs. “apple”, Nintendo “Switch” vs. “switch), this author regards the processed brand and product list (see Appendix B for the list generated from a random sample of 50 users with >30 r/gadgets posts in 2018) quite reflective of electronic brands currently popular in society.

Upon combining products with their respective brands (e.g. “iPhone” “Mac” for Apple, “Galaxy” for Samsung, “Pixel” “Android” for Google), this author concludes that four consumer electronics brands are most frequently discussed by Reddit users in r/gadgets in 2018: Apple, Samsung, Microsoft and Google.

(2) Brand & Product Sentiment Classification with Sentiment Analyzers

Upon identifying brands and products frequently-discussed by active users in r/gadgets, this author uses three different sentiment analyzers to extract and classify brand sentiments from posts and comments containing brand or product names.

To prepare textual data for sentiment analysis, this author extracted sentences containing the four brands and their frequently-discussed products from all posts and comments made by active r/gadgets users in 2018. Specifically, lower-case transformed sentences containing the following entities are extracted into clusters under each brand:

Apple: “apple” “iphone” “macbook” “mac” “ios”

Samsung: “samsung” “galaxy”

Microsoft: “microsoft” “windows” “xbox”

Google: “android” “pixel” (“google” not included to avoid ambiguity with the verb)

For each brand cluster of sentences under each active r/gadgets user, this author uses three sentiment analyzers discussed below to first extract sentiment from each sentence, and then average the sentiment scores across the number of sentences as the final sentiment score for each brand cluster. For sentiment classification based on sentiment score, this author

assigns “1” for non-negative scores and “0” for negative sentiment scores. In marketing applications, positive and neutral sentiment towards a brand (score ≥ 0, classification = 1) implies the user is a potential consumer and thus a good candidate for advertising influence. In comparison, negative sentiment (score < 0, classification = 0) implies the user dislikes the brand and thus not an ideal candidate for digital marketing campaigns.

$$SentimentClassification = \begin{cases} 1, if \ \frac{\sum SentenceSentimentScore}{\#sentence} \geq 0 \\ 0, if \ \frac{\sum SentenceSentimentScore}{\#sentence} < 0 \end{cases}$$

(a) TextBlob sentiment analyzer

TextBlob is a Python library inspired by NLTK and Pattern. When passing in a sentence for text analysis, TextBlob outputs a sentiment polarity score ([-1.0, 1.0], 1.0=extremely positive, -1.0=extremely negative) and a subjectivity score ([0.0, 1.0], 1.0=extremely subjective, 0.0=extremely objective).

(b) Liu & Hu opinion lexicon on NLTK

The Liu & Hu opinion lexicon on NLTK is an opinion lexicon curated by Minqing Hu and Bing Liu over the years, ever since their first paper on customer review mining (Hu and Liu, 2004). Built upon years of sentiment analysis and insights mining of customer reviews, the Liu & Hu opinion lexicon contains over 6800 words with “positive,” “negative,” or “neutral” sentiment labels. To classify sentence sentiment using Liu & Hu opinion lexicon, this research first tokenizes and lemmatizes the sentence into word stems before assigning each word with a label from the opinion lexicon (words not present in the lexicon are considered “neutral”). Upon comparing the number of positive, negative and neutral labelled

words in the sentence, this research uses the majority label as the sentiment score for the entire sentence (1.0 = positive; 0.0 = neutral; -1.0 = negative).

(c) VADER sentiment analyzer on NLTK

The Valence Aware Dictionary and Sentiment Reasoner (VADER, available through NLTK) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media (Hutto and Gilbert, 2014). Specifically, VADER takes into account word-order sensitive relationships in addition to the traditional bag-of-words model such as Liu & Hu opinion lexicon approach mentioned above. Moreover, VADER considers features like capitalization and the use of punctuation or emoticons in a sentence, to improve its sentiment analysis beyond lexicon and syntax. This research uses the compound score from VADER analyzer as the sentence sentiment score, which is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1.0 (extremely negative) and +1.0 (extremely positive).

(3) Predicting Brand Sentiment Using User Characteristics

Upon processing the input variables (i.e. self-identified characteristics of active r/gadgets users) and output variable (i.e. binary sentiment towards the four electronic brands), this research uses four supervised machine learning models for the ultimate sentiment prediction task: In other words, this author aims to develop a robust classifier that predicts consumers' binary brand sentiment based on multiple self-identified characteristics. User characteristics, the input variables for training and testing samples, are represented as binary-valued feature vectors in the model (see Table 2 for sample input variable data). Specifically, a (1xN) array is created for each user, where N equals the number of unique characteristics of all users in a particular brand cluster. Then, for each vector in the (1xN) array, a binary value

(1 or 0) signals if the user self-identified as having a particular characteristic. The consumer brand sentiment, the output variable in the model, is also binary as mentioned in the previous section (see Table 2 for sample output variable data).

For all four supervised learning classifiers discussed below, this author randomly splits the processed data into training and testing sets with the train_test_split function in scikit learn, where test_size = 0.5 and random_state = 42.

(a) Bernoulli Naïve Bayes

Bernoulli Naïve Bayes model is an implementation of the generic Naïve Bayes classification algorithm, tailored to binary-valued data distributed according to multivariate Bernoulli distributions (Manning, Raghavan and Schütze, 2008). Compared to other Naïve Bayes models, Bernoulli Naïve Bayes model explicitly penalizes the non-occurrence of a feature *i* that is an indicator for output variable *y*. The decision rule for Bernoulli Naïve Bayes model is as follows:

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i)$$

(b) Complement Naïve Bayes

Complement Naïve Bayes (CNB) algorithm is an adaptation of the standard multinomial naive Bayes (MNB) algorithm to improve its performance on imbalanced data sets. Illustrated by inventors to significantly outperform MNB on text classification tasks, CNB uses statistics from the complement of each class to compute the model's weights (Rennie, Shih, Teevan and Karger, 2003). The weight calculation is as follows:

$$\widehat{\theta_{ci}} = \frac{\alpha_i + \sum_{j:y_j \neq c} d_{ij}}{\alpha + \sum_{j:y_j \neq c} \sum_k dkj} \ (\alpha = smoothing\ hyperparameter)$$

$$w_{ci} = \log \hat{\theta}_{ci} \qquad\qquad w_{ci} = \frac{w_{ci}}{\sum_j |w_{cj}|}$$

The classification rule for Complement Naïve Bayes model is:

$$\hat{c} = \arg min_c \sum_i t_i w_{ci}$$

(c) Logistic Regression

Logistic regression is a linear model for classification, using a logistic function to describe the probabilities of potential outcomes of a single trial. This research uses the "lbfgs" optimization algorithm as solver which approximates the Broyden–Fletcher–Goldfarb–Shanno algorithm belonging to quasi-Newton methods (Fletcher, 1987). For penalization scheme, this research users *l2* which minimizes the following cost function:

$$min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp\left(-y_i(X_i^T w + c)\right) + 1)$$

(d) Random Forest

The random forest classifier is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It uses the perturb-and-combine technique on various decision tree classifiers to introduce randomness in classifier construction. The introduced randomness helps cancel out prediction errors from individual decision tree models that typically exhibit high variance and a tendency to overfit.

(4) Predictive Model Evaluation

In order to tune the machine learning models described above and evaluate the effectiveness of consumer sentiment prediction, this author uses Receiver Operating Characteristic (ROC) curve as the metric for model evaluation. ROC curve plots the false

positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0.0 and 1.0.

$$False\ Positive\ Rate = Sensitivity = \frac{False\ Positive}{False\ Positive + True\ Negative}$$

$$True\ Positive\ Rate = 1 - Specificity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

The area under the ROC curve, also known as AUC, is an indicator of the given model's skill, with a no-skill model having a baseline AUC of 0.5 (i.e. prediction labels are randomly assigned). The better the model is, the higher the AUC and the more bowed to the upper left the ROC curve is in the plot.

For this research, this author uses ROC curves to: 1. compare and select the most effective sentiment analyzer given same model and brand sentiment data; 2. tune all four supervised learning algorithms to the best performance for each user-brand cluster; 3. evaluate the overall performance of the proposed consumer prediction approach using consumer characteristics.

4. RESULT

(1) Summary of Brand Posts and Consumers' Brand Sentiments in r/gadgets

Upon defining the criteria of user data this research will use for predictive modelling, this author constructs four brand clusters (Apple, Samsung, Microsoft & Google) from all r/gadgets posts and comments submitted by active users of the subreddit community in 2018. In each brand cluster with dictionary format, each active r/gadgets user has a list of posts and comments containing the name of either the brand or the products of the brand (see Methodology for the list of product names). The statistical details of each brand cluster are as follows (see post distribution graphs in Figure 1-4):

Brand Clusters	NumUser	NumUser with Characteristic Data	Average NumPost per user	Median NumPost per user
Apple	789	756	9.64	6.0
Samsung	528	507	3.09	2.0
Microsoft	481	464	2.65	2.0
Google	616	591	4.16	3.0

From the table above and the post distribution graphs, this author notes a consistent brand-related post frequency distribution pattern across the four brand clusters. Specifically, despite the scale difference in number of users and post frequencies between Apple and the other three brands, all four brand clusters have the same left-skewed distribution pattern, i.e. a small portion of users submit significantly more brand-related comments than the majority of users.

This pattern further adds value to the characteristic-based consumer targeting approach proposed by this research: with training data from a small portion of users and their substantial amount of sentiment-embedded posts, the proposed model can predict the brand sentiments of the vast majority of users without explicit brand-related data, as long as these users share similar characteristics with the hyper-active user group. In other words, this model is cost-effective for digital marketers, as sentiment mining on a small set of Reddit users can yield predictions for a much larger set of consumers.

(2) Comparison of Sentiment Analyzers: TextBlob, Liu & Hu Lexicon and VADER

As discussed in the Methodology section, this research uses three sentiment analyzers – TextBlob, Liu & Hu opinion lexicon and VADER (NLTK ver.) – to extract brand sentiments from Reddit posts and comments. With Liu & Hu using a simple lexicon-based approach, VADER and TextBlob both a combination of lexicon-based and rule-based models, the three sentiment analyzers produced slightly different distribution in sentiment labels. The sentiment label distribution among all users with characteristic data is summarized below in tuples containing the number of positive or neutral labels and the

number of negative labels (see Table 3-6 for detailed breakdowns of sentiment label distribution in all users by brand):

	TextBlob	**Liu & Hu Lexicon**	**VADER**
Apple	(614, 142)	(567, 189)	(549, 207)
Samsung	(398, 109)	(395, 112)	(377, 130)
Microsoft	(360, 104)	(363, 101)	(330, 134)
Google	(509, 82)	(470, 121)	(430, 161)

For brand clusters Samsung and Microsoft, TextBlob and Liu & Hu Lexicon yield similar positive/negative label ratio while VADER labels more users as having negative sentiment towards the brand. For brand clusters Apple and Google, the between-model difference is higher, with TextBlob yields the most positive labels and VADER yields the least. Overall, VADER model has a tendency to produce more negative labels than the other two, potentially due to more stringent criteria for positive valence when imposing normative rules upon opinion lexicon patterns.

When using the three different set of sentiment labels as output variables in supervised learning models (see Figure 5-8 for ROC curves of different supervised learning models using different label sets), this author observes a small but consistent superiority in model performance for sentiment labels produced by TextBlob analyzer. Due to the similarity in sentiment label ratio, the models using Liu & Hu labels have a similar performance with TextBlob whereas models using VADER labels perform worse than the other two. However, this author does not attribute the performance gap to the inferiority of VADER as a sentiment analyzer – with a significantly higher percentage of positive labels across all three analyzers, the supervised learning models could bias towards outputting positive labels, and thus result in a higher discrepancy between predictions and a truth label set with more negative labels.

As sentiment labels produced by TextBlob analyzer yield best model performance, this research proceeds with only TextBlob-generated sentiment labels in the tuning and evaluation of supervised learning models. While it is of the author's interest to further

explore the difference between sentiment analyzers in Reddit sentiment mining, this research only requires sentiment labels from the best-performing analyzer to illustrate the validity of proposed predictive model.

(3) Supervised Learning Model Performance

As discussed in the Methodology section, this research develops the optimal model by fine-tuning four supervised learning algorithms available through scikit-learn (Pedregosa, 2011) – Bernoulli Naïve Bayes, Complement Naïve Bayes, Logistic Regression and Random Forest. To avoid overfitting given the enormous number of features (i.e. user characteristics) in training set input data, this author applies features selection to the feature vectors and selects K features with highest chi-square scores using the scikit-learn function feature_selection.SelectKBest(chi2, K). In other words, only the K most class-dependent user characteristics are passed into the predictive model for training and testing.

(a) Apple

Among all 756 r/gadgets active users with characteristic data and have submitted Apple-related posts and comments in 2018, there are 10916 unique self-identified characteristics. Therefore, the pre-feature-selection input vector space has a size of 756x10916. Upon tuning the number of selected features and hyper-parameters for each classifier, the optimized models for Apple cluster and their performance are showed below with the best classifier achieving a 0.836 AUC score (see Figure 9 for ROC curve graphs):

Classifier	**Number of Input Feature Selected (K)**	**Classifier Parameters**	**ROC AUC (no skill: 0.5)**
Bernoulli Naïve Bayes	4500	alpha = 0.02	0.836
Complement Naïve Bayes	4500	alpha = 0.1	0.799
Logistic Regression	8000	C = 0.001 class_weight = “balanced”	0.633

Random Forest	8000	max_depth = 35 class_weight = “balanced”	0.613

(b) Samsung

Among all 507 r/gadgets active users with characteristic data and have submitted Samsung-related posts and comments in 2018, there are 6870 unique self-identified characteristics. Therefore, the pre-feature-selection input vector space has a size of 507x6870. Upon tuning the number of selected features and hyper-parameters for each classifier, the optimized models for Samsung cluster and their performance are showed below with the best classifier achieving a 0.807 AUC score (see Figure 10 for ROC curve graphs):

Classifier	**Number of Input Feature Selected (K)**	**Classifier Parameters**	**ROC AUC (no skill: 0.5)**
Bernoulli Naïve Bayes	3200	alpha = 0.005	0.807
Complement Naïve Bayes	3200	alpha = 0.8	0.765
Logistic Regression	6000	C = 0.7 class_weight = “balanced”	0.581
Random Forest	6000	max_depth = 53 class_weight = “balanced”	0.611

(c) Microsoft

Among all 464 r/gadgets active users with characteristic data and have submitted Microsoft-related posts and comments in 2018, there are 7811 unique self-identified characteristics. Therefore, the pre-feature-selection input vector space has a size of 464x7811. Upon tuning the number of selected features and hyper-parameters for each classifier, the optimized models for Microsoft cluster and their performance are showed below with the best classifier achieving a 0.732 AUC score (see Figure 11 for ROC graphs):

Classifier	**Number of Input Feature Selected (K)**	**Classifier Parameters**	**ROC AUC (no skill: 0.5)**
Bernoulli Naïve Bayes	3600	alpha = 0.0000005	0.732

Complement Naïve Bayes	3600	alpha = 2.0	0.723
Logistic Regression	7000	C = 0.4 class_weight = "balanced"	0.693
Random Forest	7000	max_depth = 80 class_weight = "balanced"	0.628

(d) Google

Among all 591 r/gadgets active users with characteristic data and have submitted Google-related posts and comments in 2018, there are 8527 unique self-identified characteristics. Therefore, the pre-feature-selection input vector space has a size of 591x8527. Upon tuning the number of selected features and hyper-parameters for each classifier, the optimized models for Google cluster and their performance are showed below with the best classifier achieving a 0.829 AUC score (see Figure 12 for ROC curve graphs):

Classifier	**Number of Input Feature Selected (K)**	**Classifier Parameters**	**ROC AUC (no skill: 0.5)**
Bernoulli Naïve Bayes	3300	alpha = 0.0001	0.829
Complement Naïve Bayes	3300	alpha = 1.0	0.791
Logistic Regression	7000	C = 1.5 class_weight = "balanced"	0.602
Random Forest	7000	max_depth = 80 class_weight = "balanced"	0.645

(e) Summary

Across all four brand cluster data, the Bernoulli Naïve Bayes model yields best predictive performance with input feature size half of that required by logistic regression and random forest models. With Bernoulli Naïve Bayes model's ROC AUC scores consistently above 0.7 (max at 0.836), this author concludes that the predictive model proposed by this research has not ideal yet notable skill. In other words, the brand sentiments of users similar

to target consumers (in self-identified characteristics) have predictive power over the brand sentiments of target consumers on Reddit.

5. CONCLUSION

While sentiment mining and behavioral predictions on social media have gained popularity in marketing research in recent years, this research is among the first to examine brand sentiments on Reddit using natural language processing techniques. This research contributes to the digital marketing field in three ways: First, it tests the validity of homophily-only social targeting in a largely anonymized social media platform; Second, it explores the potential of computational linguistic techniques in business applications, such as brand insights mining and consumer behavior predictions; Last but not least, it proposes an alternative to current digital advertising targeting practices with full compliance to Internet user privacy regulations.

First, this research contributes to the study of social targeting as it tests an expansion of the range of influential social contact online – from previously established direct and indirect contacts to total strangers who post and comment in the same social media platform. Different from other social media platforms popular in social targeting research, Reddit requires zero personal information (e.g. email address, name) during user registration and thus maintains a higher degree of anonymity compared to Twitter or Facebook (Overdorf and Greenstadt, 2016). Therefore, this research is able to filter out more social influence by contact (i.e. users know each other offline and thus may influence each other in ways not discoverable from online interactions) and offers some preliminary evidence on how homophily alone can have predictive power over the brand sentiments of anonymous users.

Second, this research explores the business application of natural language processing techniques, such as named entity recognition and sentiment classification. Admittedly, there

are many limitations to the text analysis models used in this research. For instance, as most currently available named entity recognition models are developed and used by researchers in computer science, recognition of business entities (e.g. brands, products) is not optimized in algorithms and thus require manual validation. Similarly, due to the lack of interdisciplinary research on brand sentiments in social media, the most relevant opinion lexicon one can use for classifying brand sentiments is trained on customer reviews on Amazon, which may not best reflect the lexical valence people observe on social media. In addition, general NLP problems also pose challenges to the accuracy of text analysis in marketing applications, such sarcasm detection and context understanding.

Nevertheless, this research hopes to show the potential of computational linguistics in digital marketing to researchers working in both fields. Despite the limitations mentioned above, this research is able to produce a working model for consumers' brand sentiment prediction with a natural language processing pipeline involving minimal human intervention (i.e. the validation and grouping of brands and products). With potential future research on tasks such as brand-product relationship classification and sentiment mining in brand or product recommendations, this author is optimistic about the transformative power of natural language processing in both academic marketing research and real-life digital marketing applications.

Last but not least, this research explores an alternative for digital marketing practitioners currently using private user data for online advertising targeting. Showing that user characteristics extracted from public posts and comments also have predictive power on consumers' brand sentiments, this research hopes to offer marketers a consumer targeting alternative that's less reliant upon personally identifiable information – so that the digital marketing industry can keep striving under data protection regulations and the trade-off between effective targeting and user privacy is no longer necessary.

TABLES & FIGURES

Month	Reddit post	r/gadgets post
18/01	91,558,594	34,287
18/02	86,467,179	36,625
18/03	96,490,262	34,877
18/04	98,101,232	18,249
18/05	100,109,100	31,493
18/06	100,009,462	27,838
18/07	108,151,359	31,397
18/08	107,330,940	31,309
18/09	104,473,929	30,074
18/10	112,346,556	32,278
18/11	112,573,001	29,988
18/12	16,053,747	2,409
Annual Total	1,133,665,361	340,824
Monthly Average	94,472,113	28,402

Table 1. Number of Posts and Comments in r/gadgets and Entire Reddit (2018)

	german	big fan	longtime fan	retard	western er	straight male	...	pretty big mustang fan	SamsungSentimentLabel
defrgthzjukiloaqsw	1	0	0	0	0	0	...	0	1
Exist50	0	1	0	0	0	0	...	0	1
wookiebath	0	1	0	0	0	0	...	0	1
AkirIkasu	0	1	0	0	0	0	...	0	1
WaidWilson	0	1	0	0	0	0	...	0	1

Table 2. Sample Input & Output Variable for Supervised Machine Learning Models
("SamsungSentimentLabel": binary output variable)

Sentiment Label	TextBlob	Liu & Hu Lexicon	VADER
Positive/Neutral	643 (596+47)	593 (420+173)	577 (529+48)
Negative	146	196	212

Table 3. Sentiment Label Distribution for All Users in Apple Brand Cluster

Sentiment Label	TextBlob	Liu & Hu Lexicon	VADER
Positive/Neutral	416 (318+98)	410 (220+190)	389 (268+121)
Negative	112	118	139

Table 4. Sentiment Label Distribution for All Users in Samsung Brand Cluster

Sentiment Label	TextBlob	Liu & Hu Lexicon	VADER
Positive/Neutral	373 (291+82)	374 (222+152)	344 (274+70)
Negative	108	107	137

Table 5. Sentiment Label Distribution for All Users in Microsoft Brand Cluster

Sentiment Label	TextBlob	Liu & Hu Lexicon	VADER
Positive/Neutral	530 (457+73)	491 (309+182)	450 (381, 69)
Negative	86	125	166

Table 6. Sentiment Label Distribution for All Users in Google Brand Cluster

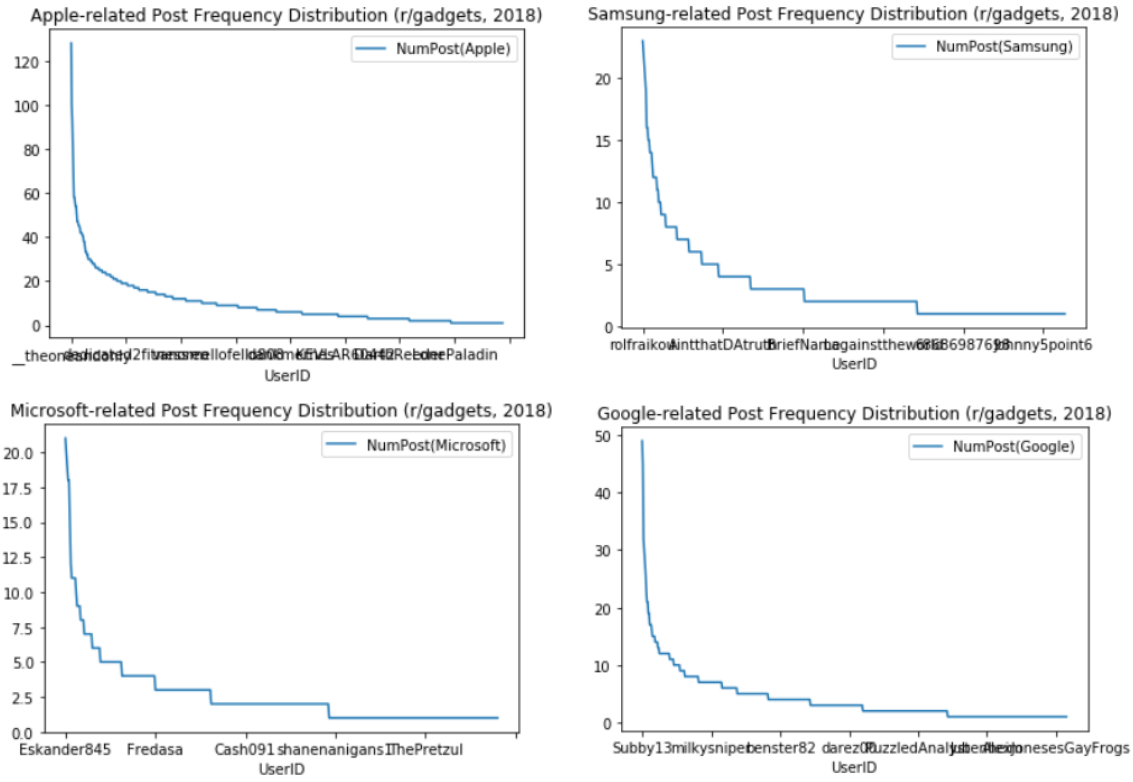


Figure 1-4. Distribution of Posts and Comments in Brand Clusters
(Upper-left to Lower-right: Apple, Samsung, Microsoft, Google)

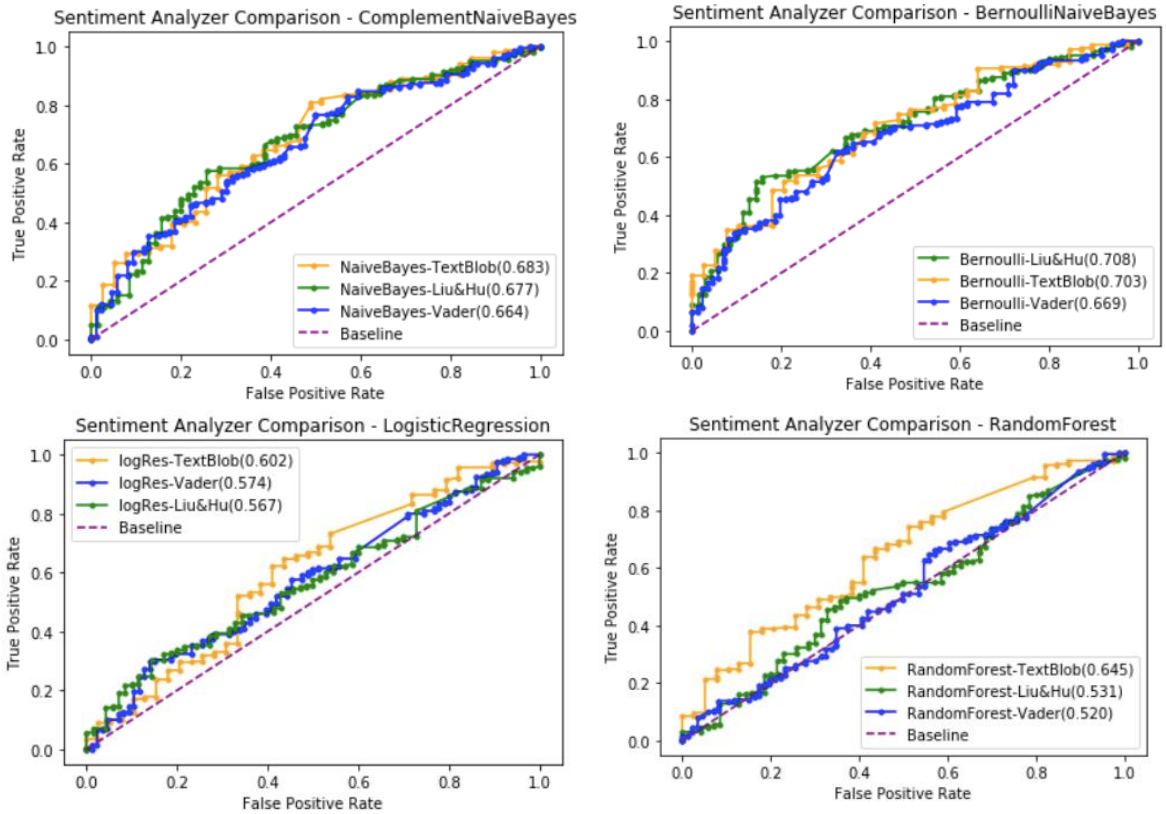


Figure 5-8. ROC Curves of Models using Different Sentiment Label Sets
(Complement Naïve Bayes, Bernoulli Naïve Bayes, Logistic Regression, Random Forest)

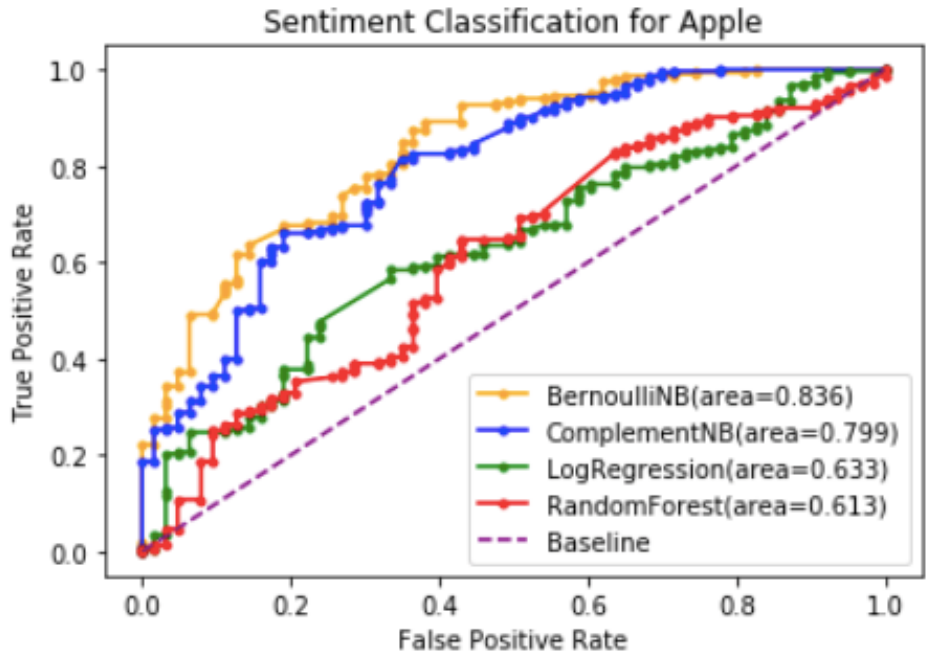


Figure 9. ROC Curves for Optimized Classifiers on Apple Brand Cluster Data

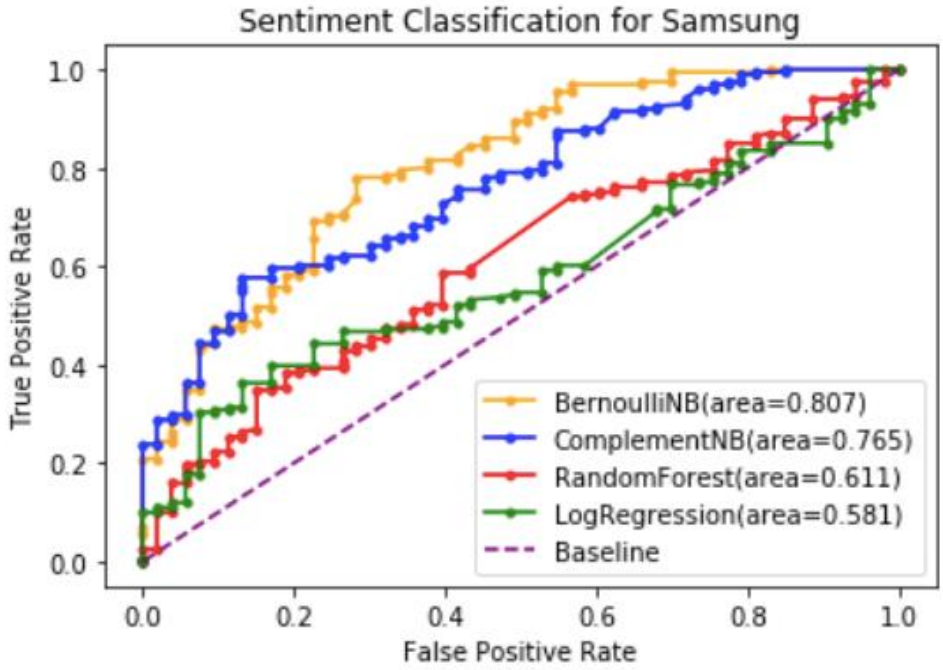


Figure 10. ROC Curves for Optimized Classifiers on Samsung Brand Cluster Data

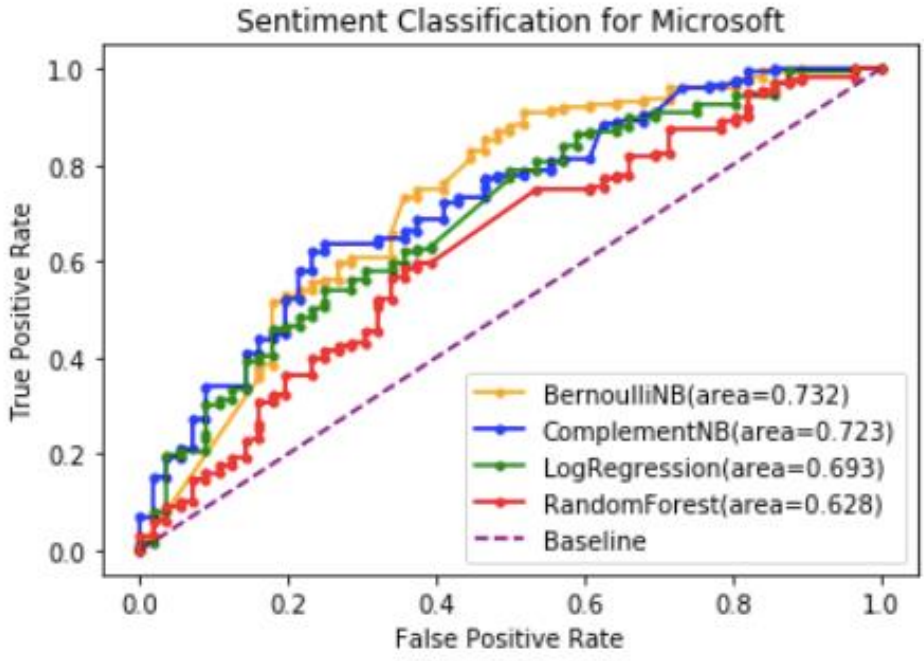


Figure 11. ROC Curves for Optimized Classifiers on Microsoft Brand Cluster Data

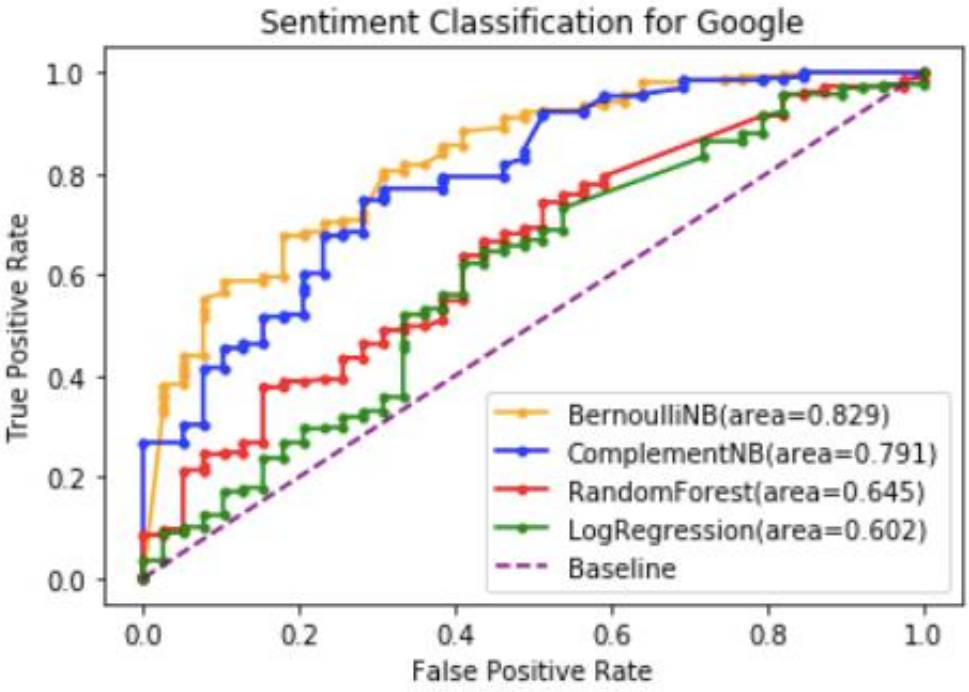


Figure 12. ROC Curves for Optimized Classifiers on Google Brand Cluster Data

APPENDIX A – Sample Data

1. Sample data from original Reddit dataset (data used in this research highlighted)

Property	Data	Property	Data
archived'	FALSE	edited'	FALSE
author'	PM_Me_Whatever_lol'	gilded'	0
author_created_utc'	1465214592	gildings'	{'gid_1':0, 'gid_2':0, 'gid_3':0}
author_flair_background_color'	None	id'	eatm5fl'
author_flair_css_class'	None	is_submitter'	FALSE
author_flair_richtext'	[]	link_id'	t3_a1u640'
author_flair_template_id'	None	no_follow'	TRUE
author_flair_text'	None	parent_id'	t1_eatl1xh'
body'	You said gaming rigs may soon be replaced by cloud gaming. '	permalink'	/r/gadgets/comments/a1u640/microsoft_may_actually_release_its_foldable/eatm5fl/'
author_flair_type'	text'	removal_reason'	None
author_fullname'	t2_yi3e1'	retrieved_on'	1546258301
author_patreon_flair'	FALSE	score'	2
author_flair_text_color'	None	send_replies'	TRUE
can_gild'	TRUE	stickied'	FALSE
can_mod_post'	FALSE	subreddit'	gadgets'
collapsed'	FALSE	subreddit_id'	t5_2qgzt'
collapsed_reason'	None	subreddit_name_prefixed'	r/gadgets'
controversiality'	0	subreddit_type'	public'
created_utc'	1543622453	distinguished'	None

2. Sample data from processed Reddit user characteristics dataset

User Name	Self-identified Characteristics
Exist50	'big fan', 'fan', 'male', 'bit', 'huge fan', 'idiot', 'american', 'part', 'native speaker', 'strong proponent', 'big proponent', 'troll', 'shill', 'twin', 'cheap bastard', 'liar', 'fast typer', 'conservative', 'little skeptical', 'softy', 'beekeeper', 'hardware guy', 'apple hater', '" shill', 'trump supporter', " hater', '" creep', 'little more pessimistic / skeptical', 'astroturfer', '" apple hater', 'huge oled fan', 'good enough bot', 'massive irrational , fact - free apple hater'

APPENDIX B – List of Electronic Products & Brands in r/gadgets

The following list is manually processed after extraction using NER model from a random sample of 50 Reddit users with >30 r/gadgets posts in 2018 (abbr. for occurrence >=5)

Brand/Product	Occurrence	Brand/Product	Occurrence	Brand/Product	Occurrence
Apple	167	Gamecube	9	thunderbolt	5
Samsung	36	Ray Tracing	9	x5690 xeon	5
iPhone	36	the USB 3	5	the usb medium	5
Xeon	23	LG/Sammy	5	usb 2.0	5
EGR	18	Vega M	5	Sandy Bridge	5
Google	18	vega APU	5	the Kessel Run	5
LG	18	Vega M GH	5	Mario Odyssey	5
GPU	18	Meizu	5	Netflix	5
Corsair	18	nokia	5	GPS	5
iOS	18	XPS	5	AMD Wraith	5
Ryzen	14	the XPS 13	5	the Pro M.	5
intel	14	IIRC	5	samsung	5
HP	14	xiaomi robotrock	5	Pixel 1	5
Zenbook	9	Xiaomi	5	Pixel	5
linux	9	elantra	5	MacBook	5
IBM	9	6GB RAM	5	linux mint	5
Nokia	9	Mi Home	5	ubuntu	5
HEDT	9	bixby	5	Windows	5
Infinity Fabric	9	Samsung Account	5	Mac	5
Vega	9	Xeon 1366	5	GPM	5
NUC	9	NVIDIA	5	Denon AVR	5
Logitech	9	Tensorflow	5	Chromecast	5
ABS	9	Linux / OSX	5	Galaxy	5
Cherry MX	9	Linux	5	CES	5
Android	9	Sony	5	Alexa	5
dell	9	Super Smash Bros	5	Google Home	5
MBP	9	Nintendo	5	Siri	5

Bibliography

Bao, Hongyun, Qiudan Li, Stephen Shaoyi Liao, Shuangyong Song, and Heng Gao. 2013. *A New Temporal and Social PMF-Based Method to Predict Users' Interests in Micro-Blogging*. Vol. 55. doi://doi.org/10.1016/j.dss.2013.02.007. http://www.sciencedirect.com/science/article/pii/S01 67923613000663.

Breitbarth, Paul. 2019. *The Impact of GDPR One Year On*. Vol. 2019. doi://doi.org/10.1016/S1353-4858(19)30084- 4. http://www.sciencedirect.com/science/article/pii/S1353485819300844.

Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T., & Haruechaiyasak, C. , 2012. Discovering Consumer Insight from Twitter via Sentiment Analysis. *J. UCS*, *18*(8), 973-992.

Charters, 2002. Electronic Monitoring and Privacy Issues in Business-Marketing: The Ethics of the DoubleClick Experience. *Journal of Business Ethics* **35,** 243–254 (2002). https://doi.org/10.1023/A:1013824909970

EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

Fletcher, Roger, 1987, *Practical methods of optimization* (2nd ed.), New York: John Wiley & Sons, ISBN 978-0-471-91547-8

García-Cumbreras, Miguel Á, Arturo Montejo-Ráez, and Manuel C. Díaz-Galiano. 2013. *Pessimists and Optimists: Improving Collaborative Filtering through Sentiment Analysis*. Vol. 40. doi://doi.org/10.1016/j.eswa.2013.06.049.

Goel, Sharad and Goldstein, Daniel G., Predicting Individual Behavior with Social Networks (2014). Marketing Science, Vol. 33, No. 1, 2014; pp. 82-93; DOI: 10.1287/mksc.2013.0817. Available at SSRN: https://ssrn.com/abstract=2397052

Goldberg, Samuel, Garrett Johnson, and Scott Shriver. 2019. *Regulating Privacy Online: The Early Impact of the GDPR on European Web Traffic & E-Commerce Outcomes*. Rochester, NY. doi:10.2139/ssrn.3421731. https://papers.ssrn.com/abstract=3421731.

Hu, Minqing and Bing Liu. "Mining and summarizing customer reviews." *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper),* Seattle, Washington, USA, Aug 22-25, 2004.

Hutto, C.J. & Gilbert, E.E. , 2014. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text". *Eighth International Conference on Weblogs and Social Media (ICWSM-14).* Ann Arbor, MI, June 2014.

Mangold, W. Glynn and David J. Faulds. 2009. *Social Media: The New Hybrid Element of the Promotion Mix*. Vol. 52. doi://doi.org/10.1016/j.bushor.2009.03.002. http://www.sciencedirect.com/science/article/pii/ S0007681309000329.

Manning, C.D., P. Raghavan and H. Schütze, 2008. Introduction to Information Retrieval. Cambridge University Press, pp. 234-265.

Martin, Kelly D. and Patrick E. Murphy. 2017. "The Role of Data Privacy in Marketing." *Journal of the Academy of Marketing Science* 45 (2): 135-155. doi:10.1007/s11747-016-0495-4. https://doi.org/10.1007/s11747-016-0495-4.

McPherson, Miller, Lynn Smith-Lovin and James M Cook, "Birds of a Feather: Homophily in Social Networks", *Annual Review of Sociology* 2001 27:1, 415-444 https://doi.org/10.1146/annurev.soc.27.1.415

Misirlis, Nikolaos and Maro Vlachopoulou. 2018. *Social Media Metrics and Analytics in Marketing – S3M: A Mapping Literature Review*. Vol. 38. doi://doi.org/10.1016/j.ijinfomgt.2017.10.005. http://www.sciencedirect.com/science/article/pii/S0268401216305291.

Nissenbaum, Helen. 2009. *Privacy in Context: Technology, Policy, and the Integrity of Social Life* Stanford University Press.

Overdorf, R., & Greenstadt, R. 2016. Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution. *Proceedings on Privacy Enhancing Technologies*, *2016*(3), 155-171.

Pedregosa *et al.*, Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.

Ravi, Kumar and Vadlamani Ravi. 2015. *A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications*. Vol. 89. doi://doi.org/10.1016/j.knosys.2015.06.015. http://www.sciencedirect.com/science/article/pii/S0950705115002336.

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. , 2003. Tackling the poor assumptions of naive bayes text classifiers. In ICML (Vol. 3, pp. 616-623).

Rossi, Robert E. McCulloch, and Greg M. Allenby, *The Value of Purchase History Data in Target Marketing*, Marketing Science 1996 15:4, 321-340, https://doi.org/10.1287/mksc.15.4.321

Shalizi, C. R., & Thomas, A. C., 2011. Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociological Methods & Research*, *40*(2), 211–239. https://doi.org/10.1177/0049124111404820

Sterne, Jim. 2010. *Social Media Metrics: How to Measure and Optimize Your Marketing Investment* John Wiley & Sons.

Sunikka, Anne, Johanna Bragge, and Henrik Kallio. 2011. "The Effectiveness of Personalized Marketing in Online Banking: A Comparison between Search and Experience Offerings." *Journal of Financial Services Marketing; London* 16 3-4 (Special Issue: Advertising Financial Services: Challenges): 183-194. https://search-proquest-com.proxy.library.upenn.edu/docview/907073411/abstract/DB87B17DDA3D492BPQ/1?accountid=14707.

Taken, Smith Katherine. 2012. "Longitudinal Study of Digital Marketing Strategies Targeting Millennials." *Journal of Consumer Marketing* 29 (2): 86-92. doi:10.1108/07363761211206339. https://doi.org/10.1108/07363761211206339.

Tucker, Catherine E. 2014. "Social Networks, Personalized Advertising, and Privacy Controls." *Journal of Marketing Research* 51 (5): 546-562. doi:10.1509/jmr.10.0355.

https://doi.org/10.1509/jmr.10.0355.