Topic Identification and Classification of GooglePlay Store Reviews[1]

Daniel David M. Pamplona[2]
University of the Philippines Visayas

ABSTRACT

Digital distribution platforms, such as Google®[3] Play Store, contain an enormous quantity of information related to app data and user reviews. A particularly challenging task is to classify a large unstructured dataset into smaller clusters or topics. With this, data from 19,886 user reviews was extracted from Google Play Store. The main task is to determine app characteristics, though common themes, that are commonly mentioned in positive and negative reviews. Text data was preprocessed and then common topics were identified using LDA for positive reviews and negative reviews. The accuracy of topics was assessed using perplexity-based approach and human interpretation. To further validate the topic model, the topic assignment was used as the outcome variable in Naive Bayes model with reviews as input. Empirical results show that the extracted topics can be predicted well using text reviews. Finally, the distribution of topics was calculated according to different app categories.

Keywords: *Topic Modeling, Latent Dirichlet Allocation, Naive Bayes Classifier, Perplexity*

I. INTRODUCTION

Google Play, also known as Google Play Store, is a digital distribution service developed and maintained by Google. It is an online store where Android users can find and download Apps for a wide variety of use on their mobile phones or tablets. As of September 2022, Statista reports that there are 2,683,925 available Apps on the platform. Gaming Apps were the most popular app category, accounting for 13.8% of available Apps worldwide. Education Apps ranked second with 10.47% and business Apps at third with 7.11% percent share. With the vast collection of Apps, Google Play has provided app information that provides consumers' data such as the number of downloads, number of reviews, average star ratings, and the text reviews of users. This data set is important for two reasons: (1) for consumers, it provides indicators of quality and popularity and may influence their decision in downloading the app, and (2) for developers, it returns valuable feedback for evaluation and future improvements.

One, however, cannot go through the sheer volume of online reviews and is forced to arbitrarily select a small sample as the basis. This practice ignores the large majority of available data that may provide valuable insights. Thus, an automated and intelligent system, capable of organizing and classifying key insights (topics) from these reviews, will be of vital importance for both digital consumers and the developers. This system may enable the consumers to quickly extract the key topics covered by the reviews without having to go through all of them, and help the developers get consumer feedback in the form of topics (extracted from the consumer reviews).

The general task of this paper is to utilize available review data from Google Play and extract valuable insights in the form of topics. Specifically, this paper aims to find app characteristics that are frequently mentioned in good reviews and bad reviews, and then find the distribution of these topics across different app categories. The results of this paper may

[1] Presented at the 15th National Convention on Statistics, 03-5 October, 2022, Crowne Plaza Manila Galleria, Ortigas Center, Quezon City and published in this issue as a *Technical Paper*.
[2] Corresponding Author: *dmpamplona@up.edu.ph*
[3] Note that all 'Google®' will be simply referred to as 'Google' for the rest of the paper.

provide an organizational framework for classifying user reviews and determining the possible reasons for user ratings.

II. RELATED WORK

Topic modeling is a popular tool for understanding text-based data. Numerous methods have been developed to perform topic modeling but the most frequently utilized method is the Latent Dirichlet Allocation (LDA), which was presented by Blei et al. in 2003. The LDA model is a Bayesian mixture model for discrete data where topics are assumed to be uncorrelated. This approach also assumes that there are three levels of data structure – word, topic, and document and that each document may be composed of multiple topics.

The hyperparameters α and β in LDA play a crucial role in this approximation. The hyperparameter, α is defined as the prior count of times an individual topic is observed in the document while β refers to the prior knowledge of the occurrence of words generated from a topic. Both hyperparameters depend on the size of the text data and number of topics k. In this paper, the R package "topicmodels" by Grun and Hornik (2011) was used to implement LDA. The hyperparameters were initially set as $\alpha = 50/k$ and $\beta = 0.1$ as suggested by Steyvers and Griffith (2004). Gibbs sampling was used maximize the marginal log-likelihood of the data in estimating the hyperparameters.

There is often no predefined number of topics that best fit the data before conducting topic modeling. In order to determine this value, researchers often create numerous topic models for the same data using different k values and then analyze certain metrics, such as perplexity, to find the better model. Perplexity or predictive likelihood is the way of measuring in what way the model is able to predict based on a sample, and it can be of great help in determining an optimal number of topics (Prichard et. al, 2000). A single perplexity score however does not provide helpful information since results can fluctuate due to the random nature of LDA approximation algorithms and internal weight sampling (Zhao et al., 2015). A common approach is to divide the data randomly into m subsets or folds $(S_1, S_2, \dots, S_m)$ and models were constructed utilizing the m-fold cross-validation for a topic range $k = k_1$ to $k = k_m$. The perplexity is calculated for every held-out subset with D documents by

$$perplexity(S_{test}) = \exp\left\{\frac{\sum_{d=1}^{D} \log p(s_d)}{\sum_{d=1}^{M} N_d}\right\}. \quad (1)$$

The average perplexities from m testing sets for each topic value k are then calculated. A satisfactory model will have high likelihood and consequently low perplexity.

After the topics were formed using perplexity measures, it may also be of interest to further test the accuracy of the topic assignments by considering classification models with topics as outcome variable and text data as features. In this paper, Naive Bayes method was used as a classifier. Wang (2015) showed that the Naive Bayes method can be used to create classification models using 10,000 reviews from Yelp as feature space. Although it performed weaker when compared to other iterative algorithms like Support Vector Machines (SVM) and Random Forest (RF), it still provides relatively good predictions with less time and computational power. Using this method, the posterior probability of a review being drawn from a certain topic given the text feature x_i is computed using the Bayes' rule

$$p(y = 1|x) = \frac{\left(\prod_{i=1}^{n} p(x_i|y = 1)\right) p(y = 1)}{\left(\prod_{i=1}^{n} p(x_i|y = 1)\right)p(y = 1) + \left(\prod_{i=1}^{n} p(x_i|y = 0)\right) p(y = 0)}. \quad (2)$$

To evaluate prediction performance, recall and precision can be used as error metric. These are defined as

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP}, \tag{4}$$

where TP, FP, and FN stand for number of true positives, false positives, and false negatives.

III. DATA COLLECTION AND PREPROCESSING

Web scraping software, Octoparse 8.5.8, was used to extract data from Google Play. A sample of 19,886 user reviews from 497 different Apps across 12 app categories was collected. The app categories included were Game, Education, Business, Tools, Entertainment, Music & Songs, Food, Shopping, Lifestyle, Productivity, Social, and Dating. On average, 40 sample user reviews were selected from each sample app. The data collected includes the star rating, date of review, the text review, and the number of votes that found the review helpful. Only reviews in the English language were included. This study did not collect any information that may personally identify the user. The data from a sample user as seen on the Google Play Store app is shown in Figure 1.

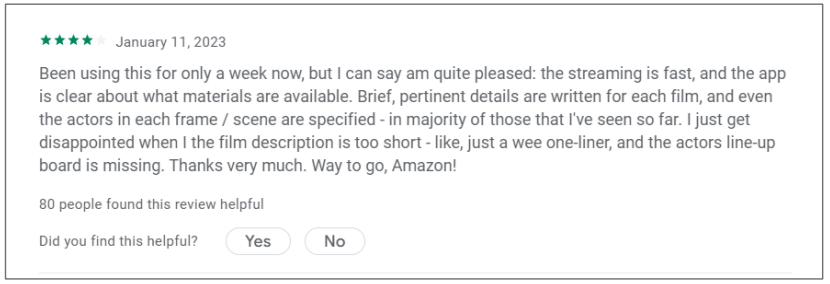

Figure 1. Sample Google Play App review data.

Review text data were cleaned by removing the format, punctuation, and extra whitespace. All letters were converted to lowercase. Stopwords, that is, words that hold little to no information value in the context of text mining (and, or, is, that, by) were also removed. Stemming was also performed to transform words into their root or stem form, for instance, the words (running, run, runner) were all converted to the word run. This process minimizes the word complexity and also reduces the size of the document-term matrix.

The reviews were divided into two groups. The “positive” reviews are those with associated star ratings of 4 or 5 while the “negative” reviews are those with star ratings 3 or below. Text analysis was then performed for each group as discussed in the next sections.

IV. DATA DESCRIPTION

IV.1 App Star Rating by Category

The data extracted from Google Play were a mixture of positive and negative reviews. It can be seen in Figure 2 that the median star ratings for each category varies between two to four stars. Game Apps are the highest app category amounting to 21.7% of the sample reviews extracted. This is followed by Social Apps with 15.9% share and Music & Songs with 13.5% share. The lowest app category collected are Tools (3.8%), Food (2.6%), and Lifestyle (2.2%). A total of 8,900 positive reviews and 10,986 negative reviews were extracted.

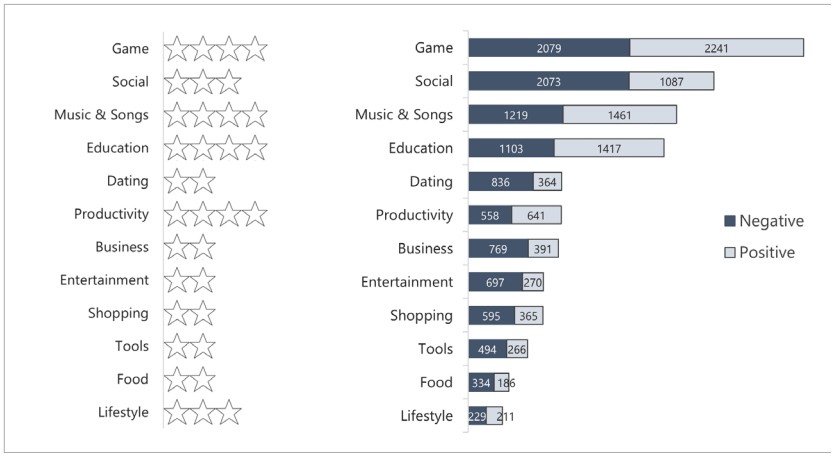

Figure 2. Google Play App ratings from 19,886 sample seviews: (Left) Median Star Rating by category and (Right) number of positive and negative reviews by category.

IV.2 Text Characteristics by Review Type

Traditional topic modeling approaches for online classification and trend analysis struggle when it comes to documents with significantly short lengths (Hong and Davison, 2010). These are often observed when extracting data from websites with character limits on posts like Twitter. Fortunately, Google Play does not impose the same restrictions and are open to long and articulate reviews from users. Table 1 shows the text characteristics of the reviews. From the data extracted, the median length for each review is 73 words. The average length of review was also found to be the same between positive and negative reviews. The median number of votes as "helpful" for each review is 27 and the median number of punctuations used per review is seven.

Table 1. Average word count by review type.

Review Type	n (%)	Median of Counts (P_{50})		
		Words Per Review	Punctuations Per Review	People Who Rated the Review as Helpful
Positive (4-5 stars)	8900 (44.8%)	71	7	27
Negative (1-3 stars)	10986 (55.2%)	74	8	26

V. TOPIC MODELING OF REVIEWS

The Latent Dirichlet Allocation (LDA) method was used in this paper to extract topics from positive and negative reviews. Unlike *k*-means clustering, where each word can only belong to one cluster (hard-clustering), LDA allows for "fuzzy" allocations (soft-clustering) where a word or document can belong to one or more clusters (Silge and Robinson, 2017). Currently, there are different methods in topic modeling such as Correlated Topic Model (CTM) and Dynamic Topic Models (DTM), but only LDA will be used in this study since the topic correlations and time of reviews will not be incorporated in the analyses. LDA is already a flexible and adaptive method in topic modeling when complex topic relationships are not of interest (Vayansky and Kumar, 2020).

V.1 Number of Topics

The ideal number of topics for unsupervised topic clustering remains a challenge to this day. After all, there is no unified idea of what a topic is. In LDA, a trial-and-error evaluation method may be used to evaluate how well the LDA model with k topics fits the dataset. This is commonly known as the perplexity-based approach. In this study, 5-fold cross-validation was implemented and the perplexity score was calculated on each held-out group. The perplexity was then averaged over all folds for every iteration. The algorithm implemented in R software took 2.3 hours for around 10,000 documents and implemented for positive and negative reviews. The lower perplexity denotes the better probabilistic model. The perplexity score and the quantiles of document-topic probabilities for different values of the number of topics (k) is shown in Table 2.

Table 2. Perplexity and quantiles of document-topic probabilities of different *k*-topic LDA models.

Number of Topics (k)	Positive Reviews (n=8900)		Negative Reviews (n=10986)	
	Perplexity	Document-Topic Probabilities (P_{50}, P_{75})	Perplexity	Document-Topic Probabilities (P_{50}, P_{75})
2	1422.2	0.500, 0.548	1308.7	0.498, 0.549
3	1299.7	0.323, 0.369	1193.5	0.289, 0.327
4	1214.4	0.239, 0.277	1135.9	0.214, 0.267
5	1177.2	0.189, 0.221	1089.2	0.188, 0.222
8	1073.1	0.125, 0.139	999.8	0.114, 0.139
10	1030.8	0.091, 0.111	959.2	0.089, 0.121
15	965.6	0.066, 0.074	891.7	0.059, 0.076
25	910.5	0.034, 0.044	826.7	0.031, 0.042
64	900.1	0.011, 0.014	797.2	0.009, 0.012
120	921.8	0.006, 0.007	815.2	0.004, 0.006

It can be observed from Table 2 that the perplexity decreases as the number of topics increases. This is expected since choosing a higher k value can provide more granular sub-topics. Ideally, the best number of topics represents the point where perplexity no longer improves with increasing values of k. This characteristic can be seen around 15 to 25 topics. This might seem to be the ideal range of the number of k, however, there is another consideration to follow. The document-topic probability (γ), which denotes the proportion of words in a document that belongs to a specific topic, needs to be a substantial amount to be able to classify later on whether a review relates to a certain topic. Values close to one indicate that a document was drawn from the topic while values close to zero say otherwise.

The 5-topic LDA model was chosen as the final topic model for the positive reviews and negative reviews. Although the decrease in perplexity is not the lowest at $k = 5$, it is considerably lower than the LDA models with four or less topics. It can also be seen that the 75th percentile of the distribution of document-topic probabilities is still around 22% which indicates that the probability of documents being drawn from a certain topic is not close to zero – unlike other models with very high k values.

The final justification for choosing the 5-topic LDA model was interpretability. Using the top words that describe each topic and top reviews that fall under specific topics, the 5-topic LDA model coincides the most with human interpretation of coherent topics. In other words, the 5-topic model was the model that "made sense". Similarly, a paper by Chang et al. (2009) argues that human evaluation of topic coherence based on top words is usually not related to predictive perplexity.

V.2 Topic Definition

The next task is to properly assign labels or interpretations per topic. This can be done in two ways: first, to look at the estimated word-topic probability; and second, to read the reviews that have high estimates of document-topic probabilities for each topic. The five topics extracted from the set of 8,900 positive reviews along with the top words and documents from each topic are shown in Table 3. The topics were named *Easy and Convenient*, *Worth Paying*, *Fun and Enjoyable, A Place to Learn*, and *Media-Related Quality*.

The first topic, *Easy and Convenient*, was named after examining the top words and top documents based on the word-topic probabilities and document-topic probabilities of this topic. The word *easy* had the highest word-topic probability of 0.033 followed by the terms *people*, *feature*, *nice, user, experience* with word-topic probabilities between 0.010 to 0.024. This indicates that the words related to this topic could be about the convenience experienced of the app users. The top documents further provided clarity, with statements such as "I like the app because it's quick and convenient." and "The app is easy to use.", which indicates that the topic relates to positive user experience.

The second topic was named *Worth Paying* since the terms and documents under this topic are mostly about account upgrades. The words *love, free, version, pay,* and *complete* were computed by the model to have high word-topic probabilities (0.021 to 0.062) of being drawn from this topic. Some parts of documents with high probabilities of being drawn from this topic include "I paid for the premium after the month of usage" and "Upgraded to paid version to make more than 2 routines. So worth it!"

The third topic was labelled *Fun and Enjoyable* since it was observed that most terms and documents under this topic were from good game reviews. The words *play*, *fun*, *level*, and *game* have high probabilities (0.025 to 0.065) of being generated from this topic. Furthermore, the model computed that positive game reviews have high probabilities of being generated from this topic as show in Table 3.

In the fourth topic, the model computes that the terms *learn*, *help*, *improve,* and *language* have high probabilities of being drawn from this topic. Also, the documents with high probabilities of being generated from this topic were found to be closely related to good reviews from Apps that teach a new language, hence, this topic was named *A Place to Learn*.

The last topic was named *Media-Related Quality*. This comes after observing that the terms *music*, *song*, *video,* and *sound* have high word-topic probabilities (0.021, 0.038). This may indicate that the topic closely relates to media characteristics of the app. Similarly, most

documents that have high document-topic probabilities under this topic also talks about video, sound, or file-handling quality of the app.

Table 3. Top words and reviews under 5-topic LDA model of **Positive** Google Play user reviews.

Topic Name	Top Words Based on Word-Topic Probabilities	Top Documents Based on Estimated Document-Topic Probabilities
Topic 1 *Easy and Convenient*	**easy**, people, **feature**, **nice**, friend, option, **user**, custom, **experience**, message, suggest	"I like <app name> because it's quick and convenient. Setting up my business account was hassle free, within minutes, I am able to send professional invoices to my customers…" "The best online shopping platform out there! The app is easy to use. The important features are easy to find. All the best deals are easily accessible..." "I've had terrific user and customer experience with <app name>, first app I've made an international purchase on and never had a problem, very easy and very good deals on items…"
Topic 2 *Worth Paying*	**love**, time, **free**, lot, day, app, version, **add**, option, **pay**, **complete**, set, task	"I found this app, this amazing app! I paid for the premium after the first month of usage; yes, I feel that it is worth it for all that it does actually do…" "This app is wonderful. I did splurge and pay the $2.99 instead using the free version but either way works fine. I like that I can set the time and date of my reminders, have multiple reminders, different task with subtasks, repeat tasks, and different alarms…" "Upgraded to paid version to make more than 2 routines. So worth it! Helps get me ticking through tasks that are boring and/or overwhelming to begin. So much more helpful than pomodoros for me…"
Topic 3 *Fun and Enjoyable*	**play**, **fun**, **level**, **game**, **enjoy**, pretty, watch, hard, character, star, buy	"This game is the greatest!!! You can play it offline, and while you're away, you'll still earn a lot of profit! I just installed it today and already earned a lot in food and coins…" "The best game ever! We can feed Pou, wash him/her, collect coins and play games! Fortunately, its ad free!!" "This is a really fun game, but upgrades cost a lot and income per level is low, so it takes too long to gain enough cash to buy the next upgrade…"
Topic 4 *A Place to Learn*	**learn**, recommend, **help**, amaze, make, **improve**, **read**, **language**, **word**	"This app is my go-to app for learning languages so far. I usually read the grammar notes before each lesson and take down the vocab before getting started…" "Works great, is intuitive and well designed and has good options to best help you learn, and I've used ALOT of Apps, and this one has to be in the top three…" "This app is really thorough. There is _SO_ much here (hear). I love it. I really do. I know a lot about reading music, but the ear training is something new to me. Thus far, it is all I have done in the app, and there are literally hundreds of exercises…"

Topic Name	Top Words Based on Word-Topic Probabilities	Top Documents Based on Estimated Document-Topic Probabilities
Topic 5 *Media-Related Quality*	**music**, **song,** phone, update, fix, issue, , **video**, **sound**, change, edit, star, save	"After the developer informed me of the issue with Android 10 concerning music players and music files stored on SD cards, I've updated my rating from 3 to 5 stars..." "I just downloaded the app so I haven't yet to subscribed for a PRO Version but I used it offline and IT'S GREAT. It has everything you just need: Edit music cover, Equalizer (SOUNDS QUALITY IS ALREADY GREAT)…" "This app is amazing for what I use it for. The only thing I would recommend for the developers, are to add a stop button. I connect to an external DAC, that usually turns itself off after playback is stopped…"

The five topics extracted from the set of 10,986 negative reviews along with the top words and documents from each topic are shown in Table 4. The topics were named *Issues after Update*, *Ads*, *Features Not Working*, *Not Worth Paying*, and *Subscription / Account Payment Issues*.

Using the same approach as the interpretation of the five topics from the positive reviews, the common topics using the negative reviews were also interpreted based on the top words and top documents under each topic. The following interpretations seek to understand the common complaints by app users in giving negative feedback towards an app. The first topic, *Issues after Update*, was named straight from the top three words generated from this topic. These words are *update*, *fix*, and *issue* with word-topic probabilities between 0.031 to 0.045. Reading the top documents further supports the interpretation of this topic. Some parts of the reviews with high document-topic probabilities are "since I installed the latest update my internet has run unbelievably slow…" and "I'm rating a 1 star as after the new update, <app name> does not work at all".

The second topic was simply named *Ads*, which is not exactly a surprise since most app users, if not all, dislike interruptions from advertisements. In the same way, the terms and documents under this topic are mostly about ad-related complaints. The word *ads* has word-topic probability of 0.065 which is relatively higher than the following top words such as *play, start, annoy,* and *stop* with word-topic probabilities between 0.020 to 0.026 of being drawn from the same topic. Some parts of documents with high probabilities of being drawn from this topic were "there's other ads that are so loud my ears start to hurt" and "ads began to appear in the streams, not at the beginning or end but randomly to the point where it became unusable".

The third topic was interpreted as *Features Not Working* or *Bad Design* since it was observed that most terms and documents under this topic were from the interface and functionality of the Apps. It may also relate to poor design choices by the developers. At first, reading the top unigram words did not provide a coherent interpretation of the topic but when the top documents were examined the interpretation became clearer. As shown also in Table 4, reviews related to this topic indicate complaints with app functionality such as "The UI is miserable, it's so taxing to navigate anywhere within the settings" and "Amazing in principle, awful in practice. The limited functionality has caused me hours of work".

The fourth extracted topic was named *Not Worth Paying / Buying* since the context of words and reviews under this topic was mostly about an app or game purchase that they regret. The top words include *play*, *money*, *time*, and *buy* with word-topic probabilities between 0.015 and 0.035. Some complaints cited in the top documents are "Obviously it's done that way to

encourage you to buy upgrades, but you can only do so through buying random loot boxes" and ". I spend money to pass levels, or beat hard levels and only get 1 star lol NOT FAIR…".

The last topic was named *Subscription / Account Payment Issues*. This comes after the terms *account, pay, subscription,* and *delete* were found to have high word-topic probabilities (0.011 to 0.025) of being generated from this topic. Furthermore, the top documents on this topic speak against the app's deceiving payment schemes.

Table 4. Top words and reviews under 5-topic LDA model of **Negative** Google Play user reviews

Topic Name	Top Words Based on Word-Topic Probabilities	Sample Parts from Top Documents Based on Document-Topic Probabilities
Topic 1 *Issues After Update*	**update**, **fix**, **issue**, phone, screen, time, load, **uninstall**, **bug**, **crash**, close, connect, error, log	"since i installed the latest update my internet has run unbelievably slow, were websites dont even work 95% of the time, not sure if its just mine but nothings working…" "New update and I can no longer upload files. I get a notice, will upload when internet connection is back online. My interent is working fine. I uninstall/reinstall the app, restart, shut down any and all other Apps, programs and connections so that only dropbox is using the internet, and still nothing will upload…" "I'm rating a 1 star as after the new update, <app name> does not work at all. I have tried using WIFI like normal, accessing through another app…"
Topic 2 *Ads*	**ads**, **play**, video, **start**, time, **watch**, download, **annoy**, music, **stop**, minute	"Why are the ad volumes so inconsistent? Some ads are way to quiet you have to turn your volume all the way up just to hear anything. Then there's other ads that are so loud my ears start to hurt.." "It really annoys me that I'm listening to music and I'll pause and try to play an hour+ later and half the time it doesn't know what the last song I was listening to was. Restarts my playlist a lot! It's unfortunate that I found this out after I payed to remove ads…" "Recently and without notice audio ads began to appear in the streams, not at the beginning or end but randomly to the point where it became unusable…" "I've been using this app for years but now I'm on the verge of deleting it because they are now periodically bursting into the streams to run 2-3 ads, which completely messes it up…"
Topic 3 *Features Not Working (Bad Design)*	option, **feature**, **change**, **version**, set, notif, add, click, button, edit, Google®, save, page, list	"Lacks Basic Functionality. App came standard on Pixel 6. I like that it syncs with Gmail because it's easy to transfer large amounts of writing/notetaking I do on my phone into Googledocs.." "The UI is miserable, it's so taxing to navigate anywhere within the settings. Options like Print page and Search through Bookmarks are gone.." "It isn't bad, but it lacks a lot of key visual elements and organizatioal tools for tablet interfaces, making this browser absolutely useless on 7+ inch screens. The main issue is not having an easy to read list of tabs at a quick glance.." "Amazing in principle, awful in practice. The limited functionality has caused me hours of work. Highlight coloring is limited to 9…"

Topic Name	Top Words Based on Word-Topic Probabilities	Sample Parts from Top Documents Based on Document-Topic Probabilities
Topic 4 *Not Worth Paying / Buying*	**play**, level, **money**, **time**, star, fun, lot, **game**, complete, upgrade, hard, **buy**, player	"You don't have to spend money, but you have to repeat difficult missions over and over again, in order to have enough money to get to the next level. Or keep repeating contracts…" "Obviously it's done that way to encourage you to buy upgrades, but you can only do so through buying random loot boxes. As higher level require ever increasing costs.... Don't spend any money here, this kind of behavior shouldn't be encouraged…" "Welp it's my time to move on, I loved to play but I came to realize it gets harder but the rewards either remain the same or it's not worth it. I spend money to pass levels, or beat hard levels an only get 1 star lol NOT FAIR…"
Topic 5 *Subscription / Account Payment Issues*	message**,** people, day, **account**, free, **pay**, month, **subscription**, **service**, experience, item, custom, **delete**, profile, user	"I used the free trial for a couple days and didn't like it too much so I decided to cancel. My trial ends TOMORROW 11/18 so I canceled today, 11/17 and just got charged a subscription fee AFTER I canceled…" "I updated to the <app name> Premium with the understanding there was a 30-day free trial. I was charged the full price / rate through my mobile carrier. I immediately cancelled the subscription only to see that (1) I would not receive a refund and the subscription would cancel one year from today…" "The worst ever refuse to sell after two good years for no reason. Will not. Ship sometimes accepts my visa then no then freezes account treats me terrible this advertising is full of lies…"

V.3 Topic Evaluation

To evaluate the extracted themes, the topics that resulted from the LDA models were used as outcome variable and classification models were constructed using 70% of the reviews as training set and 30% as validation set. A binary variable for each topic was created, where 1 – presence of the topic in the review and 0 – otherwise. Then the preprocessed text data was used as feature space in creating classification models. If the extracted topics are adequate, the presence of the topic can be predicted using the reviews.

The Naive Bayes Classifier (NBC) was used to perform the prediction. Based on the summary of results in Table 5, both the training and test errors were observed to be around 20%. This indicates that there is no overfitting in the training set and the model performed well with about 80% accuracy. The classification model was better at controlling the false negatives than the false positives as shown in higher values of recall as compared to precision. In general, the topics extracted can be adequately predicted using text data as predictors based on the overall metrics.

Table 5. Error metrics using NBC in predicting topics per review.

Extracted Topics	Train Error	Test Error	Recall	Precision
Positive Review				
1 Easy and Convenient	0.21	0.22	0.82	0.56
2 Worth Paying	0.22	0.27	0.67	0.43
3 Fun and Enjoyable	0.24	0.23	0.90	0.55
4 A Place to Learn	0.23	0.24	0.83	0.49
5 Media-Related Quality	0.17	0.19	0.82	0.58
Negative Review				
1 Issues After Update	0.25	0.27	0.85	0.49
2 Ads	0.22	0.24	0.82	0.50
3 Features Not Working (Bad Design)	0.18	0.22	0.83	0.55
4 Not Worth Paying / Buying	0.21	0.23	0.85	0.50
5 Subscription / Account Payment Issues	0.18	0.20	0.84	0.58

V.3 Distribution of Topics

The distribution of topics for positive reviews for each app category is shown in Figure 3. Based on the sample of reviews from Google Play, it is shown that the majority (77%) of the positive reviews of Game Apps characterize the game as fun and enjoyable. For Social Apps, almost half (53%) of positive reviews relate to the Apps being easy and convenient, and the same topic was also found to be frequently mentioned in Dating Apps (76.6%), Business Apps (70.6%), and Shopping Apps (68.8%).

For Apps that offer premium subscription to access all features, results show that most positive reviews relate to Apps being worthy of the payment or the upgrade. This was observed in 66.8% of positive reviews in Productivity Apps and 50.2% of positive reviews in Lifestyle Apps. Media-related quality appears to be the most common topic among positive reviews in Music & Song Apps (66.5%) and Entertainment Apps (56.3%), this is expected since sound and video qualities are the main features of these app categories.

Lastly, for Education Apps, 73.4% of the positive reviews in this category speak about the app being a good place to learn. The same can also be seen in 36.5% of lifestyle Apps, although the share is not as large as in the education category.

For negative reviews, the distribution of topics by App category is illustrated in Figure 4. From the sample of negative reviews in Google Play, there is a notable frequency in having issues after installation of updates among negative reviews in Tool Apps (48.6%), Social Apps (37.9%), and Business Apps (30.4%).

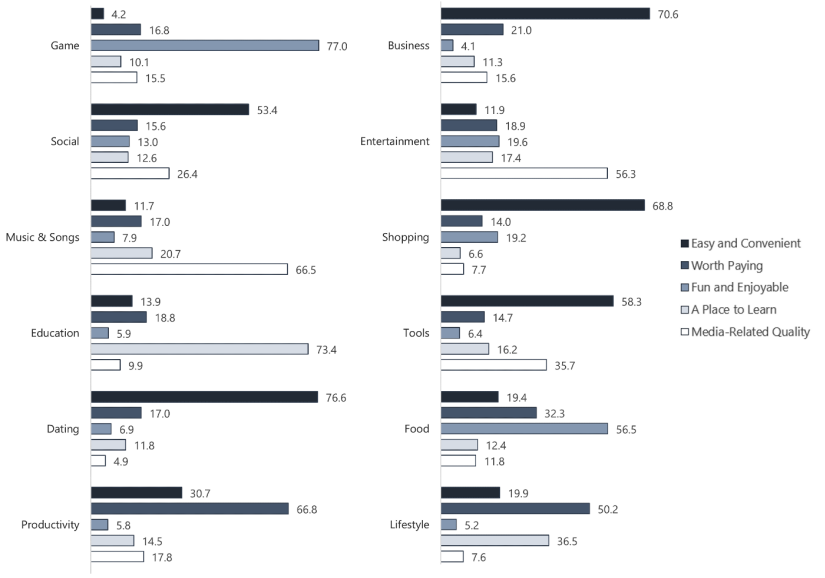

Figure 3. Percent distribution of topics by category in **Positive** App reviews.

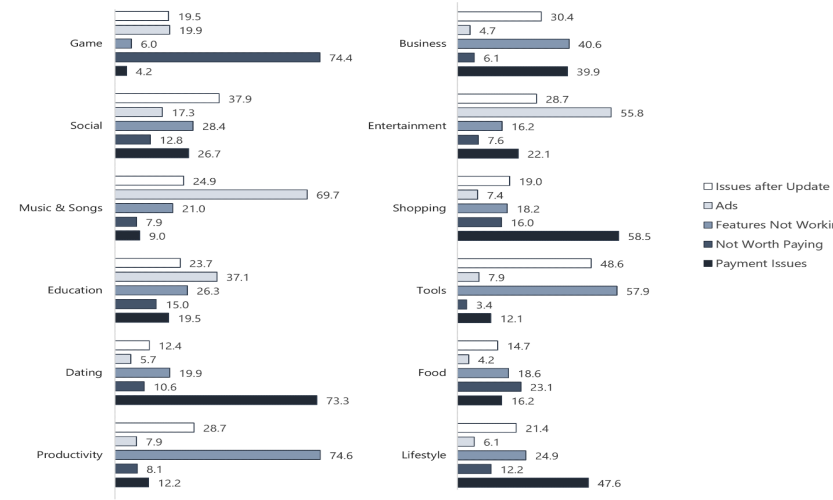

Figure 4. Percent distribution of topics by category in **Negative** App reviews.

Advertisements appear to constitute a large share of negative reviews among Music & Song Apps (69.7%), Entertainment Apps (55.8), and Education Apps (37.1), all of which are media-related Apps, and it is possible that the ads interrupt the sound or video, causing distractions to the media experience. It can also be observed that a bad interface design is commonly mentioned in negative reviews among Apps under Productivity (74.6%), Tools (57.9%), and Business (40.6%). In Game and Food Apps, undesirable purchases or payments appear in 74.4% and 23.1% of negative reviews, respectively. Lastly, subscription or account

issues that may also be related to payments are found in the majority (73.3%) of negative reviews in Dating Apps, Shopping Apps (58.5%), and Lifestyle Apps (46.7%).

VI. CONCLUSION AND RECOMMENDATION

An empirical study of topic modeling using Google Play user reviews was carried out in this paper. The final number of topics in the topic model was assessed using a perplexity-based approach and human interpretation. The topic model was validated by the classification model with topics as outcome variables and reviews as predictors. To sum up, the 5-topic LDA model was able to capture a coherent and meaningful set of themes from the extracted positive and negative reviews. It is important to note that this interpretation and set of topics may only reflect the data that was extracted and does not generalize the millions of reviews currently available in Google Play. After all, the field of topic modeling is still prone to serious issues in optimization, sensitivity, and instability which can result in reproducibility issues.

For future work, there is reason to believe that a higher number of topics may improve the granularity of the themes extracted. An eight to fifteen-topic model might be incorporated in a separate paper. Other topic models might also be used such as CTM using variational-expectation-maximization (VEM) algorithms and comparing the results with the LDA model.

VII. LITERATURE CITED

BLEI, D. M., NG, A. Y., and JORDAN, M. I., 2003, Latent Dirichlet Allocation, *J Mach Learn Res*, Vol. 3: 993–1022.

CHANG, et al., 2009, Reading Tea Leaves: How Humans Interpret Topic Models. Available at https://proceedings.neurips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf

GRUN, B. and HORNIK, K., 2011, topicmodels: an R Package for Fitting Topic Models, *Journal of Statistical Software*, 40(13): 1-30. Available at: https://ro.uow.edu.au/cgi/viewcontent.cgi?article=2408&context=commpapers

HONG, L. and DAVISION, B. D., 2010, Empirical study of topic modeling in Twitter, *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, Washington D.C., District of Columbia, 80–88.

PRITCHARD JK, STEPHENS M, and DONNELLY P., 2000, Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945-59.

SILGE, J. and ROBINSON, D., 2017. Text Mining with R. 1st Ed. O'Reilly. Available at https://www.tidytextmining.com/index.html

STEYVERS, M. and GRIFFITHS, T., 2011, Probabilistic Topic Models, Handbook of Latent Semantic Analysis, 1st ed., T. K. Landauer, Ed. New York, NY, USA: Routledge, 2011, pp. 427–440.

VAYANSKY, I. and KUMAR, S. A. P., 2020, A Review of Topic Modeling Methods, *Information Systems (2020)*, Available at: https://doi.org/10.1016/j.is.2020.101582

WANG, J., 2015, Predicting Yelp Star Ratings Based on Text Analysis of User Reviews, Available at: https://pdfs.semanticscholar.org/1445/e46d8bf48f5739246c290340fbb15113902e.pdf

ZHAO et al., 2015, A Heuristic Approach to Determine an Appropriate Number of Topics in Topic Modeling, *BMC Bioinformatics*, 16(13), p. S8. Available at https://doi.org/10.1186/1471-2105-16-S13-S8