Voice of Customer - a text mining application to understand the opinions of cellphone users in the US market

Abhinav Bajpai
Department of Data Science
Indiana University – Bloomington
USA, 47405
abbajpai@iu.edu

Abstract

Voice of Customer can be captured to gain insights into what customers value, to improve product design, to understand competitive landscapes and to penetrate markets deeper. In this paper, we analyze a large sample of reviews on the Amazon site expressed by buyers of cellphones (US market). We explore 14 predetermined topics related to common features of a cellphone and determine customer sentiment (positive, neutral, negative) across these features. We then compare the top 5 brands in the US cellphone market and provide a comparative analysis of their features. Instead of discovering the discussion topic in the text corpus, our approach assigns the discussion topic based on context relevant words, making it a more objective approach for comparing products with each other.

1 Introduction

A business entity's most important asset is its customer. The customer's perception of the product affects its performance in the marketplace - sales volume, market share and profitability. It is without a doubt that businesses are always eager to hear what customers have to say because they not only drive profitability, but also impact future product innovations that keep the business as a perpetual operation. Up until a decade ago, traditional methods such as customer surveys, interviews, and Internet surveys were among the most common ways to gather customer feedback. Lately, easy, and cheaper access of internet and the arrival of social media platforms have brought the customers and businesses closer. The voice of customer is now available real time - a tweet complaining a bad service is immediately noticed and ratified, a negative comment on the product is quickly compensated. This change has presented its own challenges as the voice of the customer is not buried inside a market research document that can be debated; it is freely available on every possible social media platform, vulnerable to be exploited by competitors. Therefore, it needs to be quickly gathered, analyzed, and responded to in order to maintain the brand's reputation.

Continuing previous research on this topic, this study aims to develop a methodology through which business owners can measure the voice of the customer via the intensity of their liking and disliking of various product features. For this study, we have focused on cellphone as the product which is ubiquitous and is used by customers of all demographics. We analyze the online text reviews of top 5 cellphone brands by number of customer reviews available on Amazon.com between 2015 and 2018. With Amazon's star rating system, customers can give a product a rating from 1 to 5, providing a sense of the product's likeability. In our data sample between 2015 to 2018, percentage of 1-star rating has grown from 15% in 2015 to 31% in 2018. On the other hand, 5-star ratings have dropped from 57% in 2015 to 43% in 2018. This trend is evident across all 5 brands, indicating that customers have become more demanding in terms of product features and customer service. The ratings are a good proxy indicator of likeability, but don't provide a more robust measure of intensity with which product is liked or disliked. A 1-star rating might be less critical of a product in their review comments when compared to a 2-star rating and similarly a 4-star rating might be more appreciative of the product than a 5-star rating. Therefore, we use a normalized sentiment intensity score between -1 (most extreme negative) and +1 (most extreme positive) to measure a product's likeability based on its reviews rather than relying on ratings.

User Comment	**User Rating**	**Intensity Score**
Similar issue but different rating and intensity of expression.		
"The phone does not stay charged"	1-star	0.1511
"The phone has to be plugged in 24/7. I'm very disappointed because I really thought it wouldn't be like that."	2-Star	-0.6708
4-star rating is more positively expressed than 5-star rating		
"Good phone just used but it works and looks fine"	5-Star	0.4854
"Great phone... In great condition no issues... Phone's awesome"	4-Star	0.9022

Table 1. Variation In User Ratings

The product's sentiment intensity score captures variations across ratings and also distinguishes within each rating, making it a more robust measure of the product's likeability among users. To determine whether the intensity score is directionally correct, it is also validated against the rating data. We find that 1-star ratings generate an average negative score of -.13, and 5-star ratings generate an average positive score of 0.74

Year	**1-Star**	**2-Star**	**3-Star**	**4-Star**	**5-Star**
2015	-0.13	0.06	0.28	0.61	0.76
2016	-0.13	0.07	0.26	0.58	0.75
2017	-0.13	0.04	0.24	0.54	0.71
2018	-0.13	0.01	0.21	0.52	0.68
Avg. Score	**-0.13**	**0.05**	**0.26**	**0.58**	**0.74**

Table 2. Intensity Score Across User Ratings

Further, through our topic assignment approach, we examine the product feature with highest positive and negative intensity across all 5 cellphone brands. The main themes of our study are battery life, network issues, phone freezing issues, screen quality, volume clarity, cellphone memory, cellphone camera and return policies.

2 Related Work

Feedback is an important aspect of product improvement and market planning for a product manufacturer. Almost all online retailers let customers express their opinion of a product through reviews, which benefit other users and can help manufacturers and sellers improve their products. Past studies have shown the usefulness of product reviews and their impact on driving sales for the business. Rather than relying on star ratings, (Ghose & Ipeirotis, 2011) considered readability, subjectivity, and linguistic correctness in product reviews to gauge the impact on sales. Similarly, in their study concerning three competing New York hotels, (Amadio & Procaccino, 2016) discovered that product reviews can also be used to gather competitive intelligence. They used TripAdvisor comments to extract feature/opinion pairs that expressed customer sentiments for strength, weakness, opportunity, and threat (SWOT) analysis of the three hotels.

Several researchers have taken advantage of the advancement in text mining approaches to analyze large corpora of online reviews to determine customer opinions. (Leem & Eum, 2021) applied a supervised learning methodology (Naive Bayes) to segregate customer reviews of a mobile banking app into positive and negative opinions. They further utilized these labeled opinion outputs to measure apps' performance by computing service quality scores across various quality dimensions like security, privacy, design, etc. Lastly, they summarized customer complaints by grouping keywords derived from negative reviews into 4 categories - technology, interaction, customer convenience, and process. A similar approach was developed by (Mandhula et al., 2020) who used Latent Dirichlet Allocation (LDA) and fuzzy C-Means to extract keywords from Amazon reviews and then built a convolutional neural network on top of the keywords to assign a positive, negative, or neutral label to the reviews. Their study is not industry-specific and considered reviews of eight products (amazon instant video, books, electronics, home and kitchen, movie review, media, kindle, and camera) for the experimental investigation.

Another study in this area has been done by (Rust et al., 2021) who promote the idea of measuring brand reputation as the voice of all stakeholders and monitoring it in real-time using social media. On 100 global brands' tweets, they implement a lexicon-based approach through a collection of positive and negative words grouped by brand drivers and subdrivers. The ratio of positive to negative words is used as a brand reputation score, with drilldowns at the driver and sub-driver levels to enable comparisons among brands. They further examine the relationship of brand drivers with stock returns. The study does have limitations since the lexicon-based approach places a limit on the number of words for assessing polarity, which can result in information loss.

3 Data

The review dataset utilized for analysis is a subset of Amazon Review Data compiled by (Ni et al., 2019) during their research study. As this study aims to understand Voice of Customer for cellphone users, only reviews about cellphone were extracted from the larger dataset. There were two types of datasets that were downloaded and merged to make a complete dataset for analysis.

Cell Phones and Accessories – Ten million reviews about various brands of cellphones and their related accessories from Oct 1999 to Oct 2018. The dataset also provides star ratings, review time, and reviewer IDs along with the review text.

Cell Phones and Accessories Metadata – This dataset contains information about products in the review data, including product descriptions, categories (cellphone, accessories, etc.), similar items, rank, prices, and brands.

Both datasets identify products using the Amazon Standard Identification Number ("ASIN") field, which is used to combine the dataset. The combined dataset is then further subset on year range, product category and brand information. We also dropped all duplicate reviews and only considered reviews with at least 10 words. Below are the steps followed in data processing and sampling:

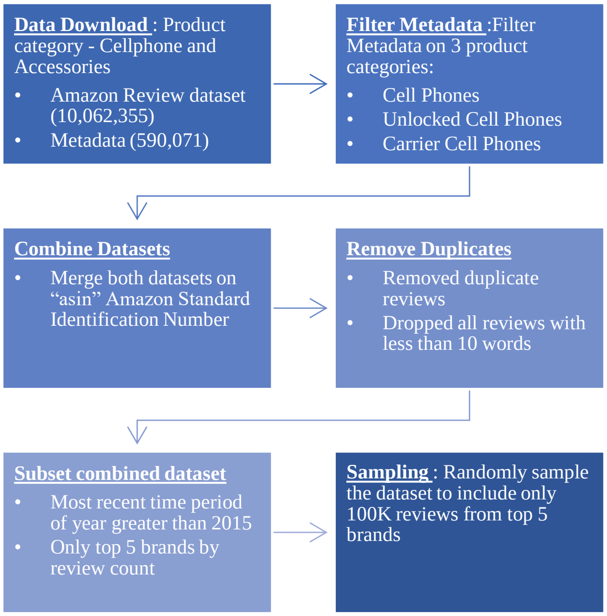
Data Download : Product category - Cellphone and Accessories
• Amazon Review dataset (10,062,355)
• Metadata (590,071)
Filter Metadata :Filter Metadata on 3 product categories:
• Cell Phones
• Unlocked Cell Phones
• Carrier Cell Phones
Combine Datasets
• Merge both datasets on "asin" Amazon Standard Identification Number
Remove Duplicates
• Removed duplicate reviews
• Dropped all reviews with less than 10 words
Subset combined dataset
• Most recent time period of year greater than 2015
• Only top 5 brands by review count
Sampling : Randomly sample the dataset to include only 100K reviews from top 5 brands

Figure 1. Data Processing Steps

In the data sample of 100K reviews covering the most recent available period of 2015 to 2018, Samsung, Apple, Blu, LG, and Motorola ranked as the top 5 brands by count of reviews. Figure 2 below shows the trend of reviews available for analysis across the top 5 brands. As data for the year 2018 is only partially available, we see a sharp decline in the number of reviews.

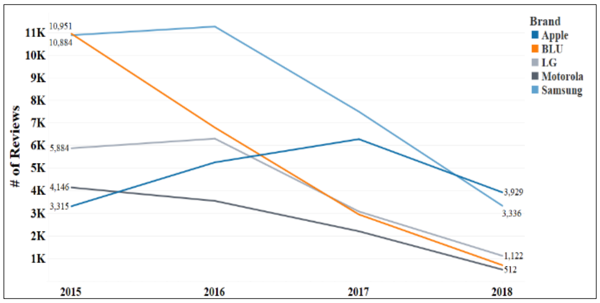

Figure 2. Trend of Available Reviews by Top 5 Brands

On average, approximately 39% of the ratings are 5- star across all 5 brands, and 1-star rating ranges from a minimum of 24% (Motorola) to a maximum of 34% (Apple). The average 39% 5-star rating is as expected because all 5 brands under study own a sizeable market share and can be regarded as market leaders. The 2-star rating ranges between 8% and 10% and is the least common rating given by customers. Similarly, 3-star rating also ranges between 8% and 12%. These observations suggest that compared to 2, 3, and 4-star ratings, customers are more likely to assign a 1-star or 5-star rating.

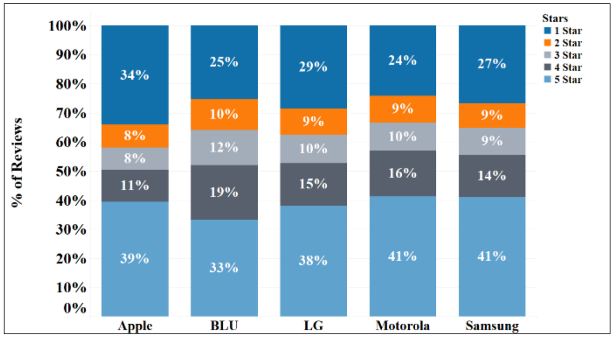

Figure 3. Star Rating Distribution of Top 5 Brands

In the sample dataset, the average length of reviews is 62 words. Star ratings 2-4 are fairly above average, 1-star and 5-star ratings, however, are below average with 58 and 57 words per review respectively. (Judith & Dina, 2006) also observed a similar pattern where length of 2-, 3- and 4-star reviews were longer than 1- and 5-star reviews. In their view, a "longer" review might indicate that the reviewer has expended more effort or may be necessary to support mixed feedback.

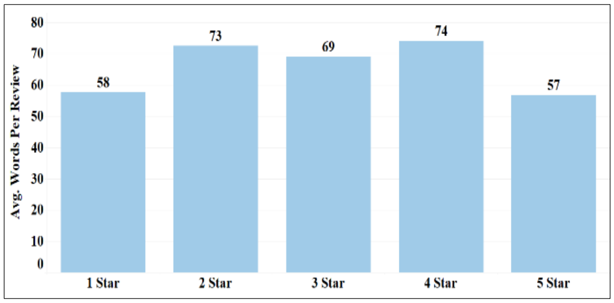

Figure 4. Average Length of Text Reviews

4 Methods

Various features of a cellphone are common knowledge, so instead of discovering topics, we follow a semi-supervised approach for understanding the topic of a review. We compiled a list of 14 predetermined topics that relate to features of a cellphone and then searched for context-relevant words (preceding and following words) in the corpus to assign predetermined topics. Figure 5 provides an example of context-relevant words for the topic Battery.

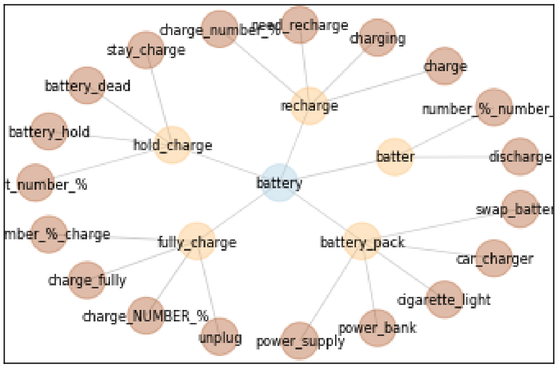

Figure 5. Example of Context related words – Topic Battery

To find context-relevant words we utilized the Genism[1] package to train a Word2Vec (W2V) model after scrubbing the review text from impurities like URLs, numbers, whitespaces, emojis, etc., Once the model was trained, we used the "most_similar" method to locate similar words related to our predecided topic. As per the genism documentation[2] - "*The most_similar method computes the cosine similarity between a simple mean of the projection weight vectors of the given words and the vectors for each word in the model*". Table 3 below lists downs the pre-decided topic and top 5 similar words suggested by the trained Word2Vec model.

Topic	**Top 5 Similar Context Words/** **(Cosine Similarity Score)**
1. Battery	batter (0.84) , recharge (0.72), hold_charge (0.68), battery_pack (0.66), battery_replaceable (0.66)
2. Screen	display (0.83), lcd_screen (0.67), lcd (0.67), lcd_display (0.64), dot (0.63)
3. Volume	speaker_volume (0.88), ringer_volume (0.86), volume_low (0.83), max_volume (0.83), volume_loud (0.82)
4. Memory	internal_memory (0.9), storage (0.88), storage_space (0.86), internal_storage (0.83), space (0.81)
5. Return-Policy	return_period (0.83), return_window (0.79), refund_replacement (0.77), warranty_period (0.74), period (0.69)
6. Customer Service	customer_support (0.83), customer_care (0.74), tech_support (0.71), service_rep (0.7), costumer_service (0.69

[1] https://radimrehurek.com/gensim/

[2] https://tedboy.github.io/nlps/generated/generated/gensim.models.Word2Vec.most_similar.html

Topic	Top 5 Similar Context Words/ (Cosine Similarity Score)
7. Value	value_money (0.79), bargain (0.77), number_-$_number (0.76), price_point (0.76), price_tag (0.73)
8. Durable	sturdy (0.82), beating (0.76), construction (0.72), rugged (0.71), pretty_durable (0.7)
9. Network	mobile_network (0.86), cellular_network (0.83), at&t_network (0.82), at&t_tower (0.79), tower (0.78)
10. Phone Freeze	constantly_freeze (0.86), freeze_time (0.84), randomly_start (0.82), freeze_lot (0.82), keep_freeze (0.82)
11. Phone Camera	photo_quality (0.85), rear_camera (0.84), picture_quality (0.83), cam (0.83), low_light (0.82),
12. Design	premium_feel (0.77), unique (0.72), ergonomic (0.72), overall_design (0.72), hand_feel (0.7)
13. Interface	ui (0.84), samsung_touchwiz (0.83), apple_io (0.82), interface (0.82), customizable (0.8)
14. Security	encryption (0.77), encrypt (0.7), incorporate (0.65), guest_mode (0.64), enable (0.63),

Table 3. Similar Word Output from Word2Vec Model

Using cosine similarity scores greater than .5, we filter the set of the most similar words for each topic and search for those words in the review text to assign a topic label, e.g., if the word encryption is found, we assign the topic security. In our 100K sample, 77% of the reviews were covered by this approach. More pre-defined topics can cover the remaining reviews, but for this study, we limit ourselves to the 14 pre-defined topics listed in Table 3 and filter out reviews that do not have a topic assigned to them. A customer can like or dislike more than one feature of a phone, so a review can have one or more topics. Below Table 4 provides the distribution of reviews as per the number of topics assigned to them.

# of Topics	# of Reviews	% of Reviews	Avg. Words in Review	Intensity Score
1	26,023	34%	30	0.30
2	18,961	24%	42	0.28
3	11,884	15%	61	0.31
4	7,292	9%	85	0.38
5	4,409	6%	113	0.46
> 5	8,912	12%	256	0.66
Total	**77,481**	**100%**	**74**	**0.35**

Table 4. Distribution of Reviews by Number of Topics

In the next step, we use the VADER[3] (Valence Aware Dictionary and Sentiment Reasoner) package built by (Hutto & Gilbert, 2014) which is a lexicon and rule-based sentiment analysis tool to compute a compound sentiment intensity score for each review in the corpus. The score is derived by adding up the valence scores for each word in the lexicon. It is then adjusted according to the rules and normalized to be between -1(most extreme negative) and +1(most extreme positive). The intensity score is further grouped as Positive (compound score greater than or equal to .5), Neutral (compound score between -.5 and .5) and Negative (compound score less than or equal to -.5) for easier comprehension and conclusion.

5 Evaluation

Using the customer ratings as a benchmark, we found that the sentiment intensity scores were relatively consistent. On average, ratings with 1 and 2 stars showed a negative intensity score whereas those with 3 to 5 stars showed a positive intensity score. The findings are summarized in Table 2.

Our manual review of a few negative reviews in Table 5 serves to validate the assignment of topics and intensity scores as well as nature of customer grievances cited in reviews.

Topics	Review Text	Intensity Score
Battery	*not as expected it should be in good condition but* ***the iphone battery is dead*** *:(:(:(:(:(*	-0.98
Phone Freeze	*worst phone i've ever had/owned/used. it's so bad i came to write this review just to say that.* ***it freezes/crashes/will get stuck on random keyboard buttons*** *or just delete everything you say. i don't know why the hell samsung made this or how it's legal to sell it. i wouldn't give this phone to my worst enemy.*	-0.97

[3] https://vadersentiment.readthedocs.io/en/latest/

Topics	Review Text	Intensity Score
Network	*i got this phone on april 20 2015 its okay phone but* ***ur calls drop and not a good bluetooth connection*** *with my lg bluetooth. product phone not that good. and for the network omg its not all that good bad bad is all i can say. but if u just looking for cheap r texting its ok for that*	-0.95
Value	***for the price, it's hard to complain. it does the job****. i hate texting on this phone, though. it's a bit on the small side, and even with my teeny tiny fingers, i'm always hitting the wrong letter. the predictive text is oddly useless, too. i had a blu tank before, and i preferred that phone.*	-0.91
Return-Policy	*dead on arrival.* ***but was refunded quickly*** *so no foul, just disappointed.*	*-0.89*
Screen	*this smartphone has good look, but has a failure.* ***the touchscreen have a dead zone*** *that block two letters. to jump this failure i have to turn the phone and change the position of the letters.*	-0.97
Phone Freeze, Battery	*"refurbished", more like "take it, it's broken but cheaper!".* ***battery is in its worst condition****, and there are scratches everywhere. the* ***touch id is non-responsive****. bad product.*	-0.92
Phone Camera	*bad, very bad month i burn the plate and* ***the camera is very bad*** *not see anything if you take pictures of letters are blurred*	-0.90
Customer Service, Return-Policy	*the worst possible phone! spent so much money when it came out, and after the phone died with known problem,* ***i cannot claim the warranty!*** *shame on you lg!*	-0.94
Volume	*revived a defective possibly refurbished phone.* ***echo really bad*** *and gps test failed.*	-0.87

Table 5. Topic Assignment Examples

We also compared the performance of Word2Vec (W2V) model with an alternative word embedding model, FastText (Joulin et al., 2016), and found only minor variations in the output of similar words. Figures 8 below show the top 5 words by topic for the FastText model.

6 Discussion and conclusion

Table 6 provides the list of top 10 topics by number of reviews counts. Three of the most discussed topics in the review corpus relate to phone batteries, freezing problems, and networks.

All Topics	2015	2016	2017	2018	Grand Total
Battery	1111	1249	1070	613	**4043**
Phone Freeze	1318	1273	970	461	**4022**
Network	1321	1150	708	337	**3516**
Value	1234	1045	611	221	**3111**
ReturnPolicy	677	775	630	291	**2373**
Screen	506	549	467	200	**1722**
Phone Freeze, Battery	415	541	484	249	**1689**
Phone Camera	634	579	308	100	**1621**
Customer Service, Return-Policy	469	456	398	185	**1508**
Volume	464	422	279	124	**1289**

Table 6. Top 10 Topics by Review Count

For each topic, we compare the 5 brands based on the percentage of positive, neutral, and negative reviews. By comparing brands, we can determine which is most popular for a specific feature. Figure 7 below provides the comparison by each topic. Further, we also look at the trend charts of intensity scores to understand whether customers' perception of the performance of a feature is improving or declining. Cellphone battery dying quickly is one of the major concerns for customers of all top 5 brands. As observed in the trend chart below, there is a consistent decline in the intensity score for all the brands except Apple which is recovering from its known battery issues during 2016-2017[4].

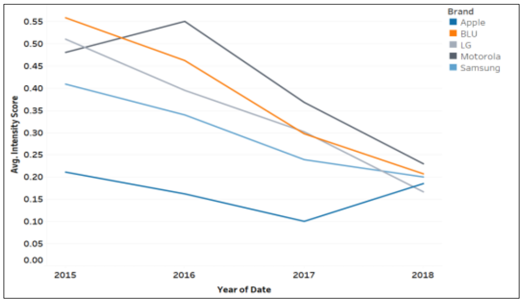

Figure 6. Intensity Score Trend Chart - Topic Battery

[4] https://www.wsj.com/articles/apple-facestwo-federal-probesover-iphone-battery-issue-1517348203

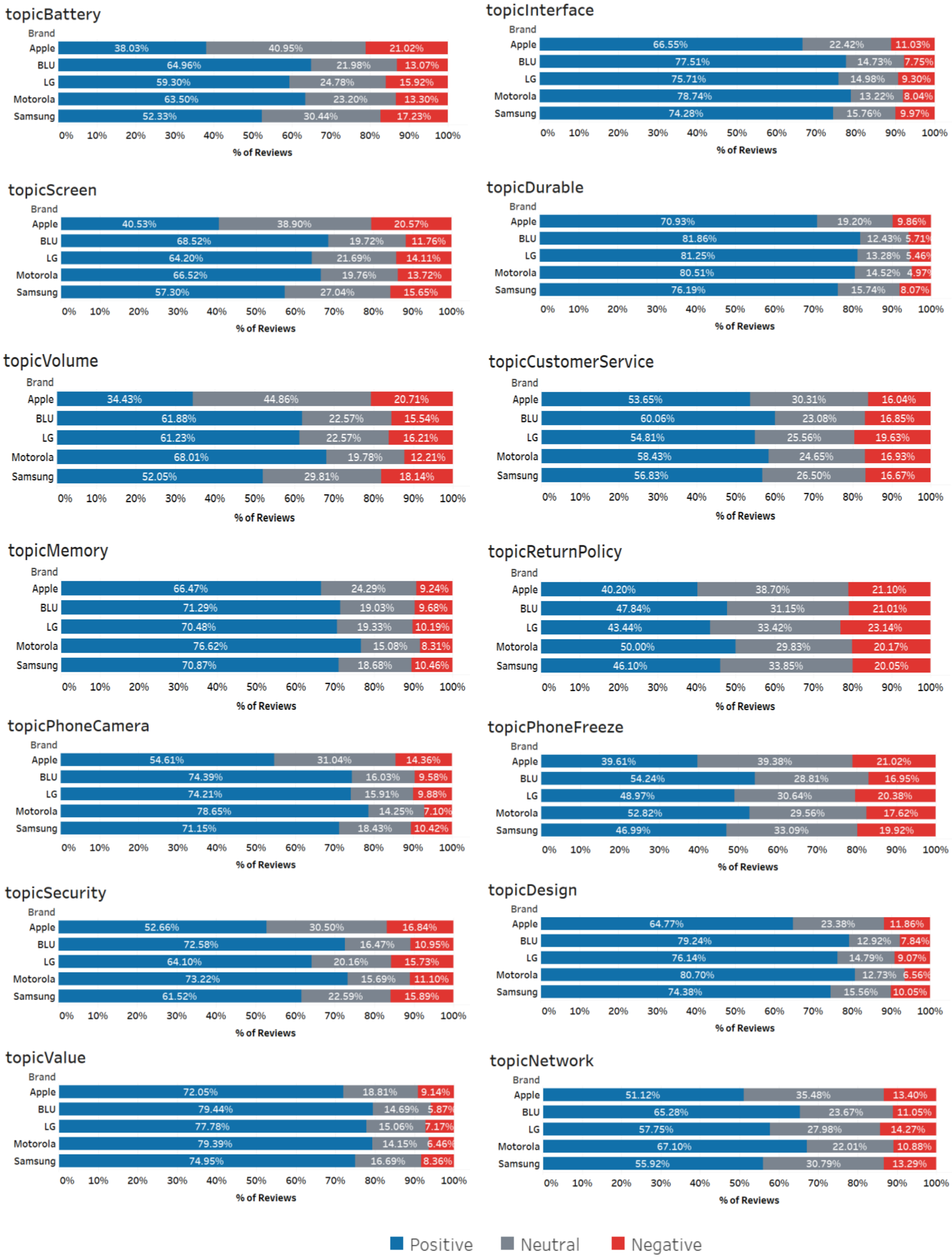

Figure 7. Sentiment Intensity Distribution of Reviews Across Top 5 Brands

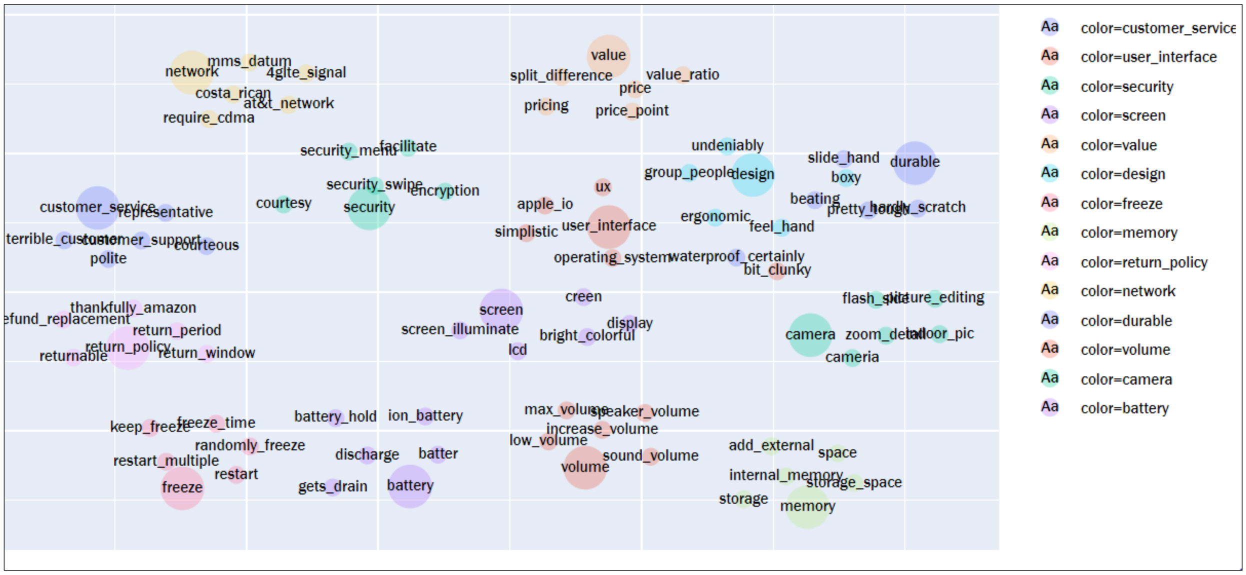

Figure 8. Similar Word Output from FastText Model

In our approach, we used various hyperparameters like cosine similarity to filter the context words or set the range of the sentiment intensity score for positive, neutral, and negative sentiment. These can vary the results and their interpretations significantly. Therefore, a degree of manual curation is required while reviewing the results and deciding the most appropriate fit of hyperparameters to tune the model. Additionally, we have ignored possible fake reviews that might exist in our sample data and can create bias in the final result. A fake review filter will improve the robustness and accuracy of our results.

References

Amadio, W. J., & Procaccino, J. D. (2016). COMPETITIVE ANALYSIS OF ONLINE REVIEWS USING EXPLORATORY TEXT MINING. *Tourism and Hospitality Management-Croatia*, *22*(2), 193-210. https://doi.org/10.20867/thm.22.2.3

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *Ieee Transactions on Knowledge and Data Engineering*, *23*(10), 1498-1512. https://doi.org/10.1109/tkde.2010.188

Hutto, C. J., & Gilbert, E. (2014, 2014 / 01 / 01 /). VADER: A parsimonious rule-based model for sentiment analysis of social media text.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. In.

Judith, A. C., & Dina, M. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews [research-article]. *Journal of Marketing Research*, *43*(3), 345-354. https://doi.org/10.1509/jmkr.43.3.345

Leem, B. H., & Eum, S. W. (2021). Using text mining to measure mobile banking service quality. *Industrial Management & Data Systems*, *121*(5), 993-1007. https://doi.org/10.1108/imds-09-2020-0545

Mandhula, T., Pabboju, S., & Gugulotu, N. (2020). Predicting the customer's opinion on amazon products using selective memory architecture-based convolutional neural network. *Journal of Supercomputing*, *76*(8), 5923-5947. https://doi.org/10.1007/s11227-019-03081-4

Ni, J., Li, J., & McAuley, J. (2019). *Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects*. https://doi.org/10.18653/v1/D19-1018

Rust, R. T., Rand, W., Huang, M.-H., Stephen, A. T., Brooks, G., & Chabuk, T. (2021). Real-Time Brand Reputation Tracking Using Social Media. *Journal of Marketing*, *85*(4), 21-43. https://doi.org/10.1177/0022242921995173