# Preprocessing

We created a virtual environment for python in R then using my google-api-client, we retrieved comments from a video about eh Samsung Android a14 Model phone, We chose it because cheaper models like this phone are more used by the average user thus the comments would provide more insight into this model because users can comment about their experience with the phone. We acquired these comments through the python code hat was linked in the lab document.

After retrieving, the comments from python, we filtered out any NA rows, and then sampled 1500 rows randomly, these sampled rows were saved in a csv file so that it could be replicated for all the members working on this project together.

A screenshot of a computer code

Description automatically generated

We created the 3 sentiments for us to analyze (positive, neutral, negative) and then created a corpus for each of them including the overall dataset as well.

The first part of the actual preprocessing involved converting emojis into words and getting rid of irrelevant text such as html tags, hyperlinks and unconverted emojis that did not represent an emotion.

A computer screen shot of a computer code

Description automatically generated

Our next step involved using the udpipe library to perform POS tagging for each sentiment, this allowed us to keep the proper nouns capitalized, and also the adjectives that were in all caps in the same state, while lowercasing everything else, We then filtered out all the unnecessary text and only kept nouns, verbs, adjectives and proper nouns

A screenshot of a computer

Description automatically generated

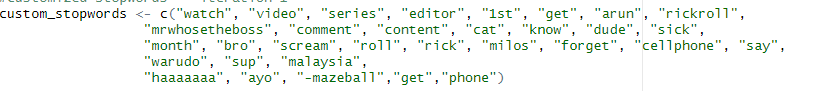
We then performed lemmatization though the pos tagging we performed previously, we chose this over stemming as it helped in narrows down the number of terms in the document term model we will construct later on.

A computer code with black and blue text

Description automatically generated with medium confidence

Our next steps included performing other basic preprocessing methods offered such as getting rid of stop words common in the English language and removing whitespaces between texts.

We then created custom stop words based on out term distribution analysis’ observations:



A group of black and white text

Description automatically generated

A close-up of words

Description automatically generated

A white text with black text

Description automatically generated with medium confidence

A black and blue text

Description automatically generated

Finally, for the document term matrix we used TF-IDF because it helped in removing certain words like phone that were too general and common in every document and then cleaned up each of the dtms to get rid of any empty documents that may have been formed as a result of the preprocessing:

A screenshot of a computer code

Description automatically generated

# Term Distribution Analysis

## Overall Corpus

## 

The top most important words in the overall document represent an overarching positive theme. The users express strong positive emotions but there is also a hint of comparison between different models that is evident via specific model names such as “a13”, which might indicate slight disappointment especially considering the fact that the model is a predecessor to the current model in discussion. Higher frequency terms indicate that by and large people are interested in buying the product hence the product can be considered a success.

## 

## Positive Corpus

# 

The top most important words like ‘love’, ‘best’, ‘good’, indicate a very profound positive sentiment. Users feel very positively about the models a14 and a13 but again suggesting a comparison. It also presents a disposition of the users towards buying this new model, and the model has been successful in sparking interest among the consumers.

## Neutral Corpus

## 

The top words in the neutral corpus suggest a mix of emotions among the users. Words like ‘expensive’, ‘break’, ‘poor’, ‘bad’ suggest an awful user experience. Users are also comparing different models like s23 to the current model. However some terms like ‘buy’, ‘cheap’, suggest that the users might be interested in buying the product.

## Negative Corpus

## 

The top words in the negative corpus express the negative sentiments of the user towards the phone as should be expected. The terms ‘resolution’, ‘pog’, ‘omg’, ‘end’, ‘catch’, indicate the users’ disappoint regarding a variety of features of the phone, moreover, they’ve compared it to previous models like a13, and a12, which also indicates that the users feel that the phone is not worth it.

# LSA

LSA has been used to determine the top terms in the top 5 features for the entire corpus, the positive corpus, the negative corpus, and the neutral corpus to analyze and determine the overall theme under each of these corpuses, which could prove very useful in improving product specifications, in this case the mobile phone under discussion, thus leading to better customer satisfaction.

We have used tf-idf to construct the term document matrix, because raw tf simply tells us which words appear the most in the corpus which may not be the most important terms while tf-idf gives us term importance by offsetting the number of times the word occurs in one document as compared to the entire corpus.

## Overall Corpus

|  |  |
| --- | --- |
| nice | nice (1.0000000) , pretty (0.7925549), border (0.4201106), giveaways (0.2938917), slick (0.2938917), grin (0.2655983) |
| nice | nice (1.0000000) , pretty (0.7925549), border (0.4201106), giveaways (0.2938917), slick (0.2938917), grin (0.2655983) |
| cool | cool (1.0000000), compromise (0.3981813), dive (0.3981813), affordable (0.2121287), computers (0.2121287), extend (0.2121287) |
| cool | cool (1.0000000), compromise (0.3981813), dive (0.3981813), affordable (0.2121287), computers (0.2121287), extend (0.2121287) |
| first | first (1.0000000), baby (0.6434994), corny (0.6434994), sleep (0.5711029), niece (0.4923963), color (0.3632975) |

The above top terms and features highlight the positive aspects of the product, which means that, by and large, the phone appeals to the users. For the top term of the first feature, which is “nice”, the neighbors such as “pretty”, ”border”, “giveaways”, “slick”, and “grin” suggest that users love the aesthetic qualities of the phone, they also love the design of the borders, which implies that users love the minimalist design. Furthermore, users value the promotional items given with the phone which may possibly include earphones, phone cover, a power bank etc. Additionally, the overall design is described as ‘slick’, highlighting the users’ positive reaction to the intricate yet minimalist design, and interface. Finally, the user's overall emotional response may be indicated by the term ‘grin’ depicting consumer satisfaction.

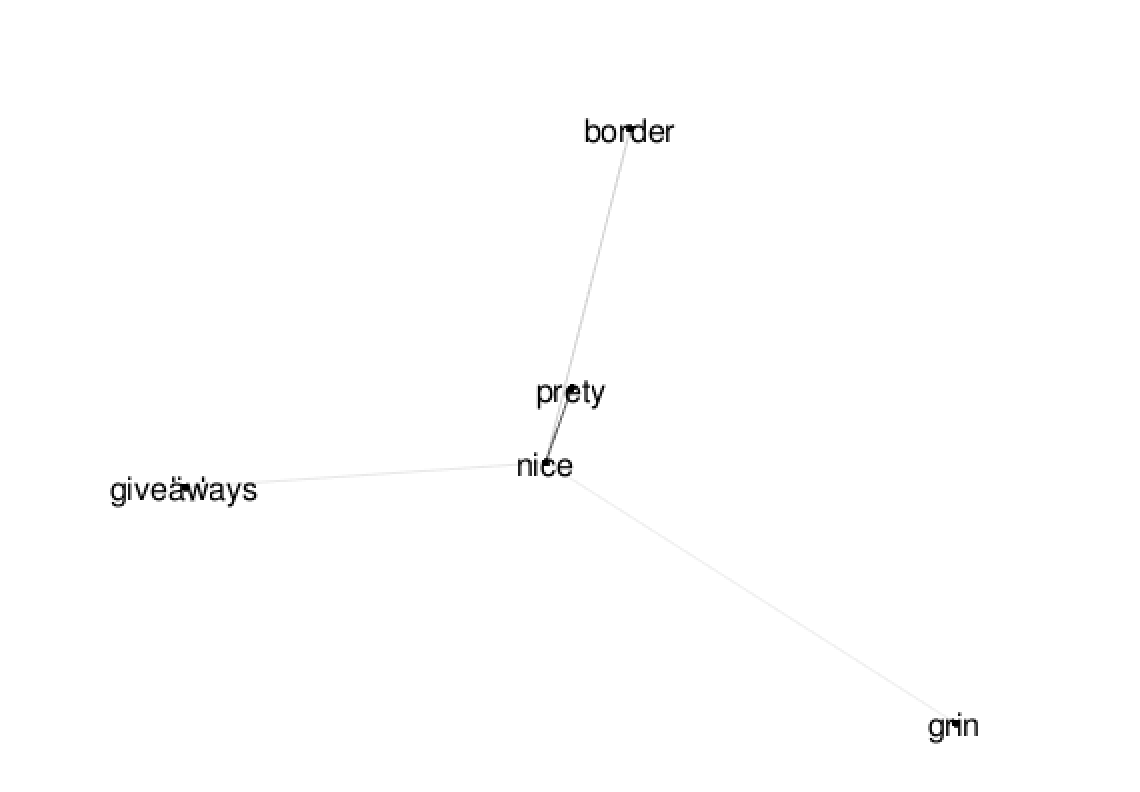
The top term of the second feature is the same as the top term of the first feature, that is, “nice”. Thus it can be inferred that the term ‘nice’ highlights an overall prominently positive theme that repeats itself across the entire corpus.

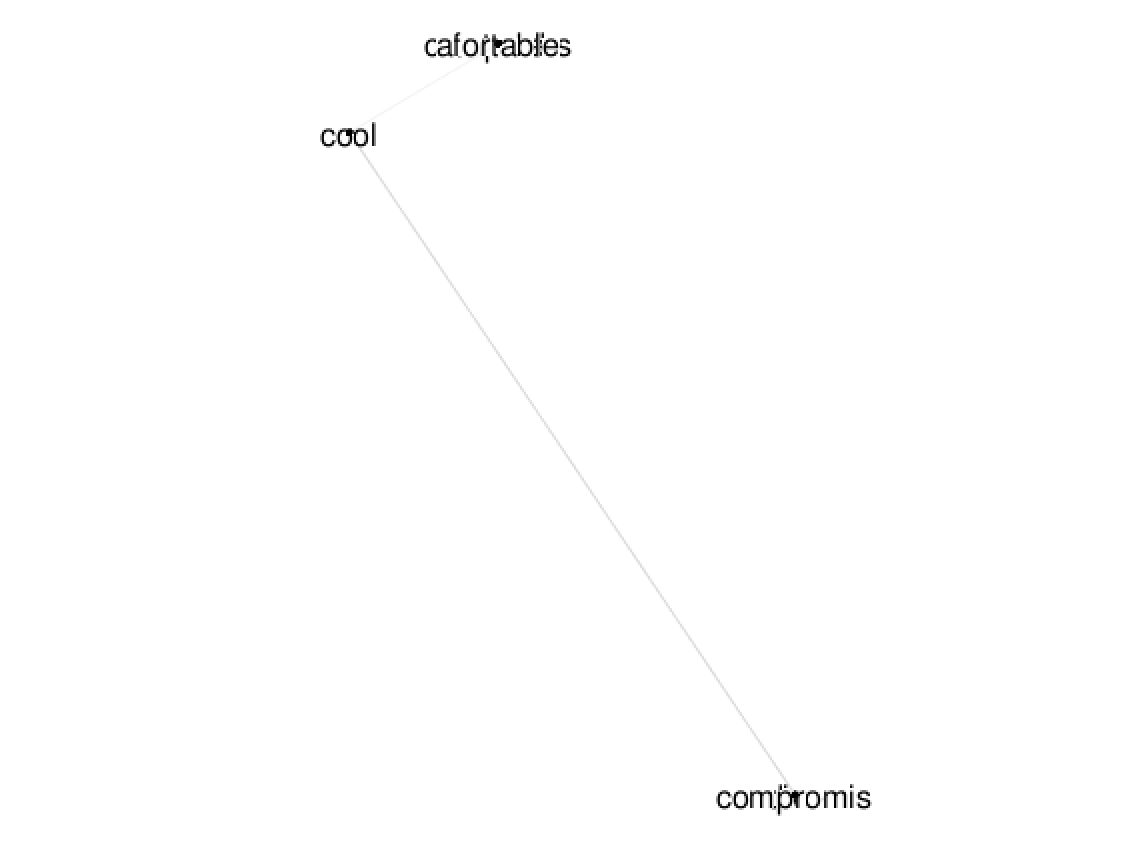
The top term of the third feature, ‘cool’, the neighbors, “compromise”, “dive”, “affordable”, “computers”, and “extend”, suggests that the phone comes along with a well balanced set of features that the users find particularly interesting such as long battery life, high quality display, extensible storage etc. Moreover, they also suggest that the phone has an engaging interface, thus streamlining user experience, and that the phone can be easily afforded and is well worth the price. Additionally, the computing prowess of the phone is linked to the word ‘cool’ highlighting its exceptional processing power and specifications such as RAM or secondary memory etc. Finally, the phone’s extensible characteristics such as batter life, and storage capabilities appeal to the users.

The top term of the fourth feature is the same as the top term of the third feature, that is, “cool”. Thus it can be inferred that the term “cool” highlights an overall prominently positive theme that repeats itself across the entire corpus.

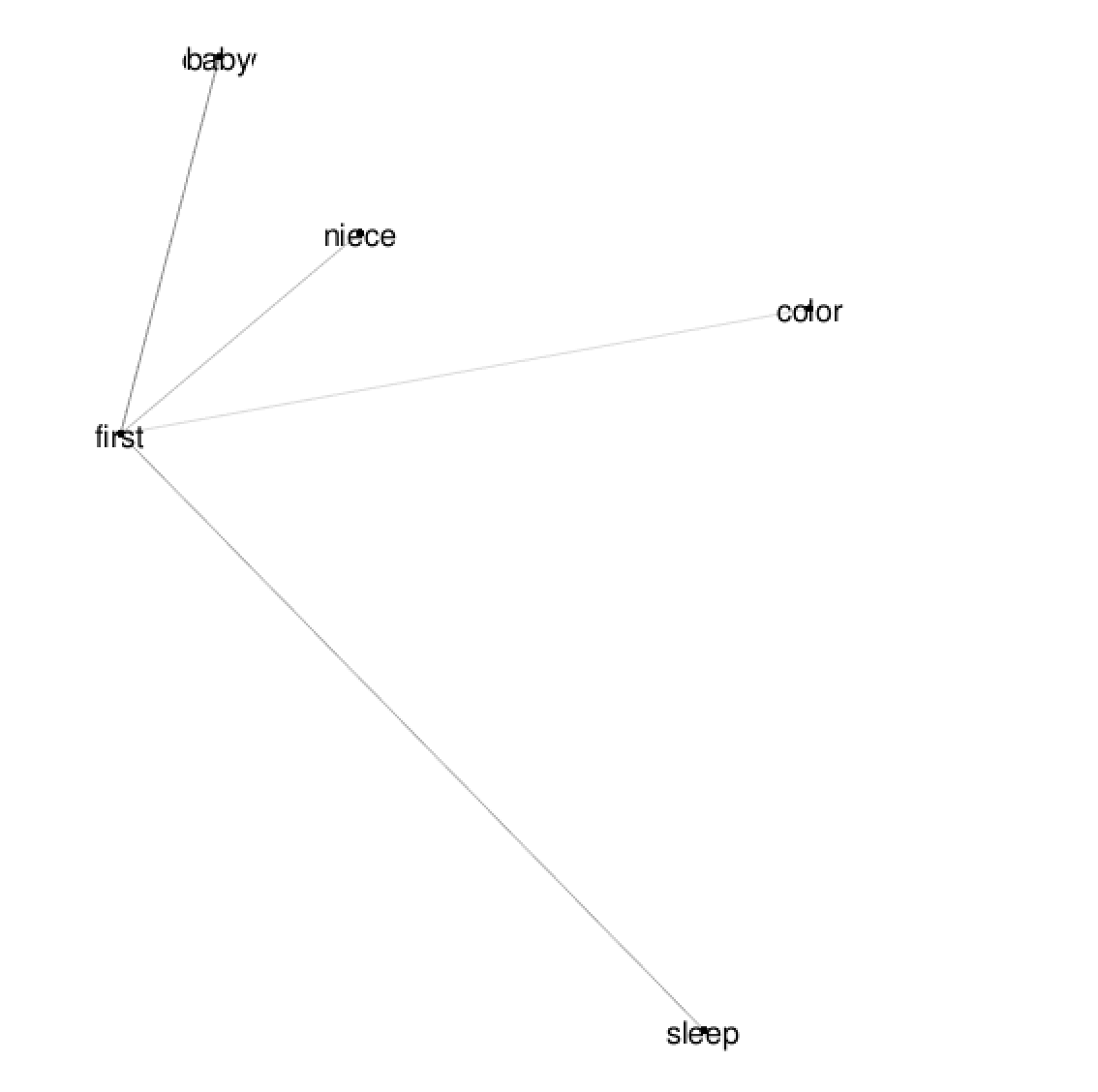
Ultimately, the top term of the fifth feature is ‘first’, which may imply that this is the users’ or the gift to a loved one as would be discussed now. Its neighbors, “baby”, “corny”, “sleep”, “niece”, “color”, suggests that the phone appeals to the younger generation which might also be why users usually gift it to younger people. Furthermore, the sleep mode and battery saver characteristics particularly appeal to the user and are of great importance to users. Furthermore the colors that the phone is available in might be very impressive, and the display color of the phone are also appreciated by the users.

Below, we provide some graphs that were used for visualization of top terms with their neighbors





The words compromise and dive, and affordable, computers, and extend, overlap and hence cannot be seen clearly.



The words baby and corny overlap and hence cannot be seen distinctly

## Positive Corpus

|  |  |
| --- | --- |
| nice | nice (1.0000000) , pretty (0.7925549), border (0.4201106), giveaways (0.2938917), slick (0.2938917), grin (0.2655983) |
| nice | nice (1.0000000) , pretty (0.7925549), border (0.4201106), giveaways (0.2938917), slick (0.2938917), grin (0.2655983) |
| cool | cool (1.0000000), compromise (0.3981813), dive (0.3981813), affordable (0.2121287), computers (0.2121287), extend (0.2121287) |
| cool | cool (1.0000000), compromise (0.3981813), dive (0.3981813), affordable (0.2121287), computers (0.2121287), extend (0.2121287) |
| first | first (1.0000000), baby (0.6434994), corny (0.6434994), sleep (0.5711029), niece (0.4923963), color (0.3632975) |

The top terms of the top 5 features of the positive segment and the overall corpus are the same, which implies that the overall theme of the entire corpus is predominantly positive, and the product has been received well by the customers.

## Neutral Corpus

|  |  |
| --- | --- |
| unable | unable (1.000000e+00), vote (1.000000e+00), full (4.057261e-14), app (4.040139e-14), system (3.297634e-14), picture (2.741087e-14) |
| break | break (1.0000000), apart (0.7515825), bang (0.7515825), buck (0.7515825), offer (0.7515825), scam (0.7515825) |
| break | break (1.0000000), apart (0.7515825), bang (0.7515825), buck (0.7515825), offer (0.7515825), scam (0.7515825) |
| poor | poor (1.0000000), entry (0.4960295), game (0.4960295), level (0.4960295), play (0.4960295), soc (0.4960295) |
| unable | unable (1.000000e+00), vote (1.000000e+00), full (4.057261e-14), app (4.040139e-14), system (3.297634e-14), picture (2.741087e-14) |

The top term of the first feature, ‘unable’ and its neighbors, ‘vote’, ’full’, ‘app’, ‘system’, ‘picture’ underscore an overall negative aspect of the product, they imply that the user is unable to perform certain functions with the product, and are probably unable to vote certain features which is an indispensable aspect of user feedback. Furthermore, the users are probably encountering storage issues, with the storage being limited preventing users from downloading important applications, and also that some applications do not work properly and crash frequently. Additionally, the operating system seems to have bugs which cause all of the above issues and the picture capturing or storage capability are also bothering the users.

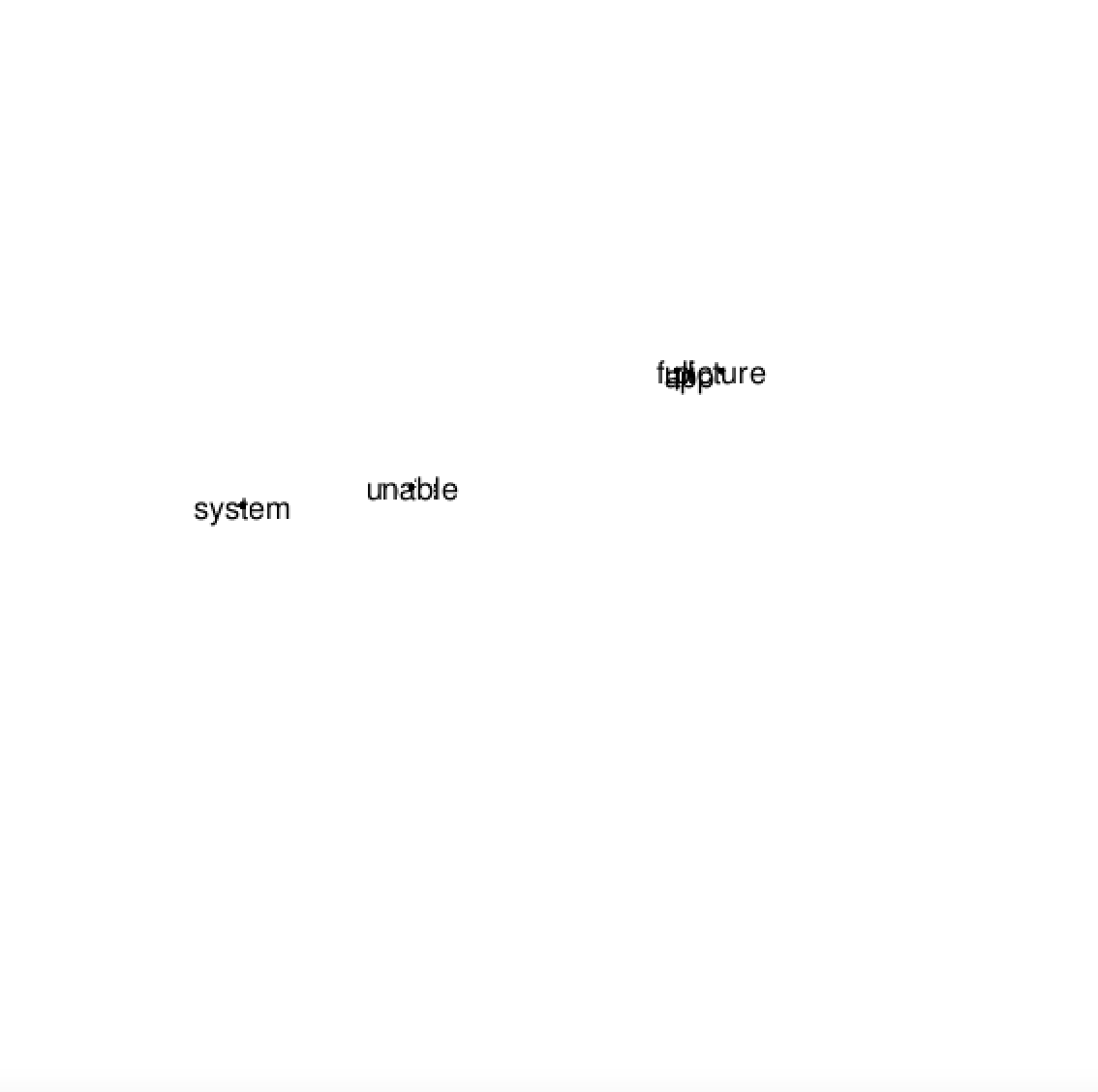
The top term of the second feature, ‘break’ , and its neighbors, ‘apart’, ‘bang’, ‘buck’, ‘offer’, ‘scam’, convey negative sentiments towards the product. Some users have expressed their concerns over the phone's durability, misleading promotions and feeling deceived, and that the phone is not worth the value and that they are not satisfied with the promotional offers.

The top term of the third feature is the same as the top term of the second feature. Thus it can be inferred that the term “break” highlights the negative themes that repeat itself across the neutral segments of the corpus.

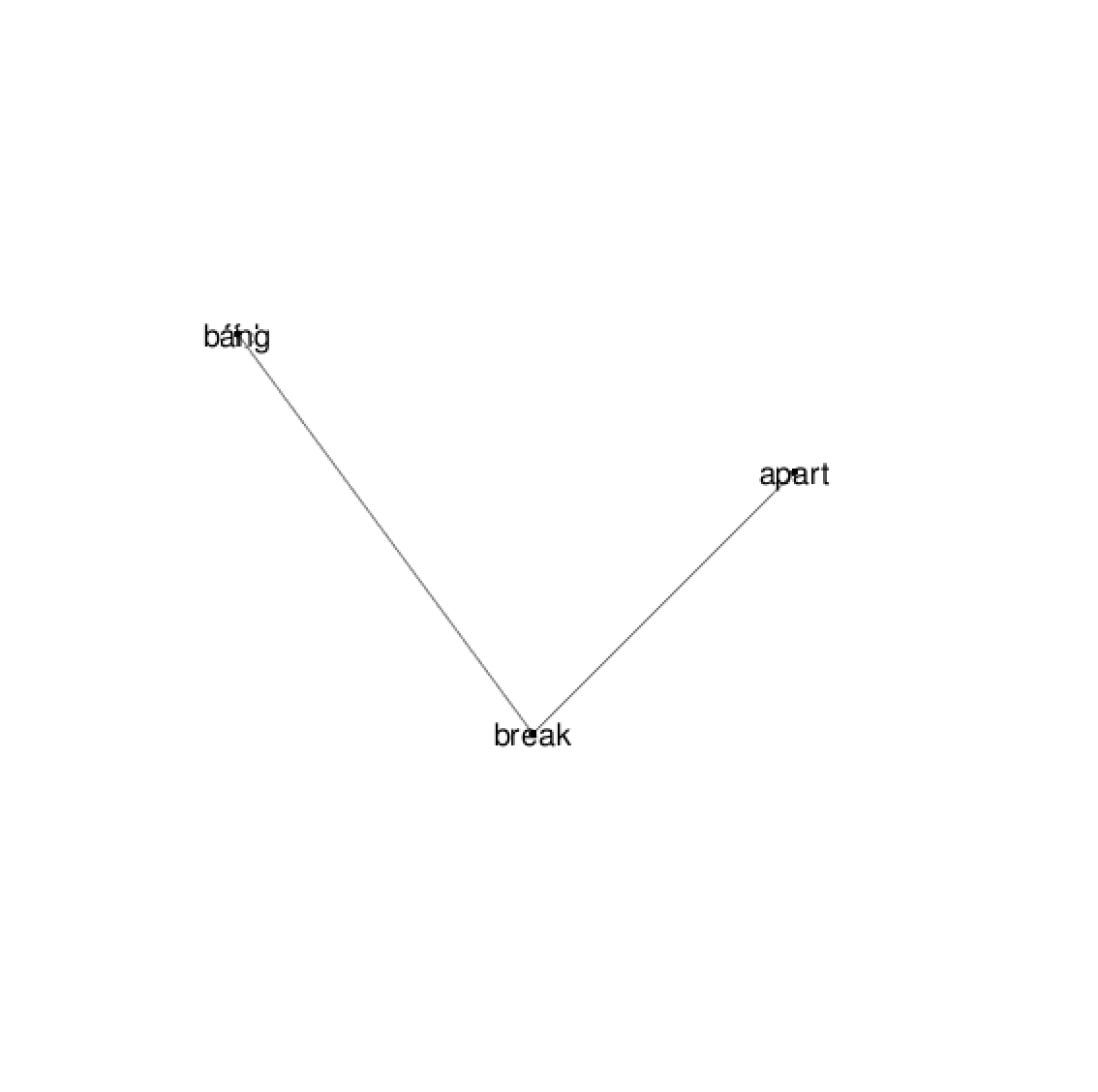
The top term of the fourth feature, ‘poor’, and its neighbors ‘entry’, ‘game’, ‘level’, ‘play’, ‘soc’, represent an overall negative sentiment towards the gaming experience provided by the phone. The phone might possess limited processing capabilities to handle games and other resource intensive applications. The users are also concerned about the System on a Chip, a component of the phone which impacts the phone’s ability to support games and other applications.

The top term of the fifth feature is the same as the top term of the second feature. Thus it can be inferred that the term “unable” highlights the negative themes that repeat itself across the neutral segments of the corpus.

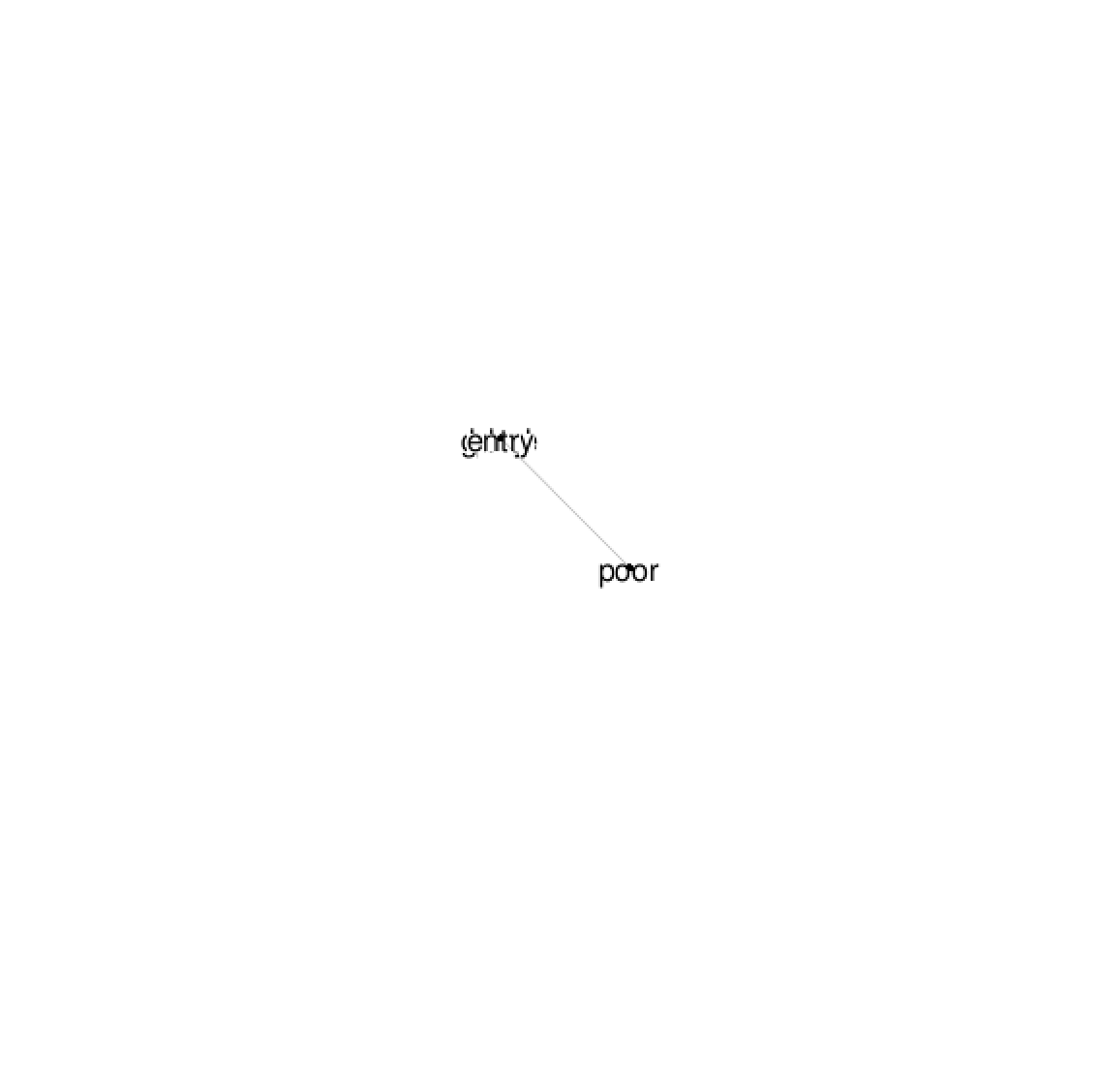
Below, we provide some graphs that were used for visualization of top terms with their neighbors



The lines in the above picture are not visible because the neighbors are very far away



All the words are at the same distance, hence some may not be visible



All the words are at the same distance, hence some may not be visible

## 

## Negative Corpus

|  |  |
| --- | --- |
| minority | minority (1.00000000), mercury (.98208238), redmagic (0.98208238), pull (0.33911975), feel (0.05352196), say (0.03986845) |
| minority | minority (1.00000000), mercury (.98208238), redmagic (0.98208238), pull (0.33911975), feel (0.05352196), say (0.03986845) |
| a14 | a14 (1.0000000), whole (0.9954487), thing (0.9714360), new (0.8268757) similar (0.5794922), downside (0.5449414) |
| resolution | resolution (1.00000000), feature (0.99666539), new (0.54422834), store (0.08104581), walk (0.08104581), call (0.05203809) |
| a13 | a13 (1.0000000), handset (0.9711351), quote (0.9711351), bezel (0.9674572) smaller (0.9674572), surprise (0.9674572) |

The top term of the first feature “minority”, along with its neighbors “mercury”, “redmagic”, “pull”,

“feel”, “say”, conveys an overall negative sentiment towards the project. It suggests that the phone is not widely popular and probably appeals to a limited bunch of people, and that the phone is being compared to another brand which is “Redmagic Mercury”. Additionally it suggests that the phone does not perform as well, and that the overall feel and the speech related features are not up to the mark.

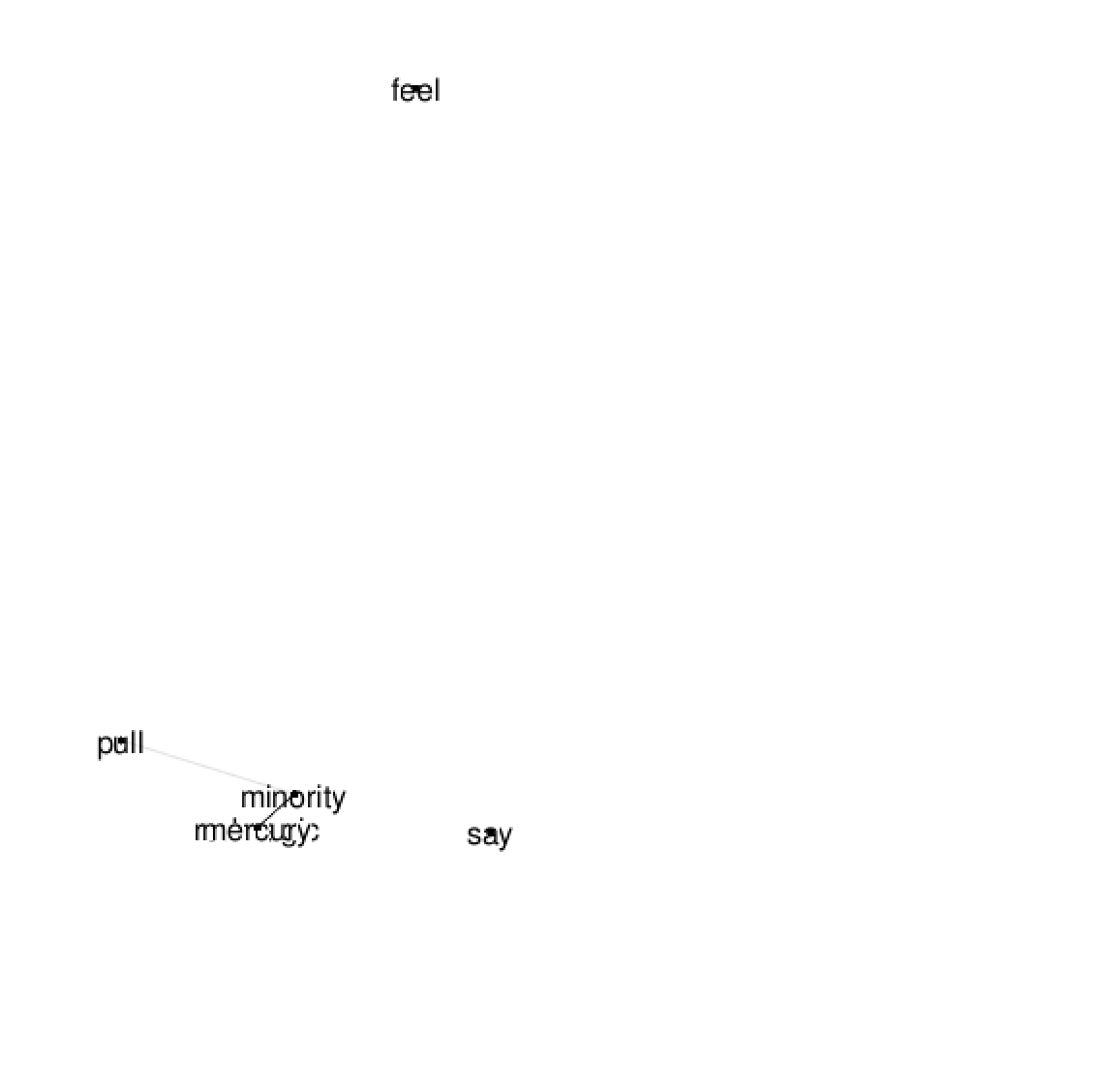
The top term of the second feature is the same as the top term of the first feature. Thus it can be inferred that the term “minority” highlights the negative themes that repeat itself across the negative segments of the corpus.

The top term of the third feature, “a14” along with its neighbors, 'whole', 'thing', 'new', 'similar', and 'downside', further reinforce the negative response of the users towards the phone. It feels like the phone is usually not received in its entirety and maybe damaged or does not include the entirety of features required or expected by the users. Furthermore despite the fact that it is new it is very similar to older models and hence is not worth it and has many downsides or disadvantages and many negative aspects.

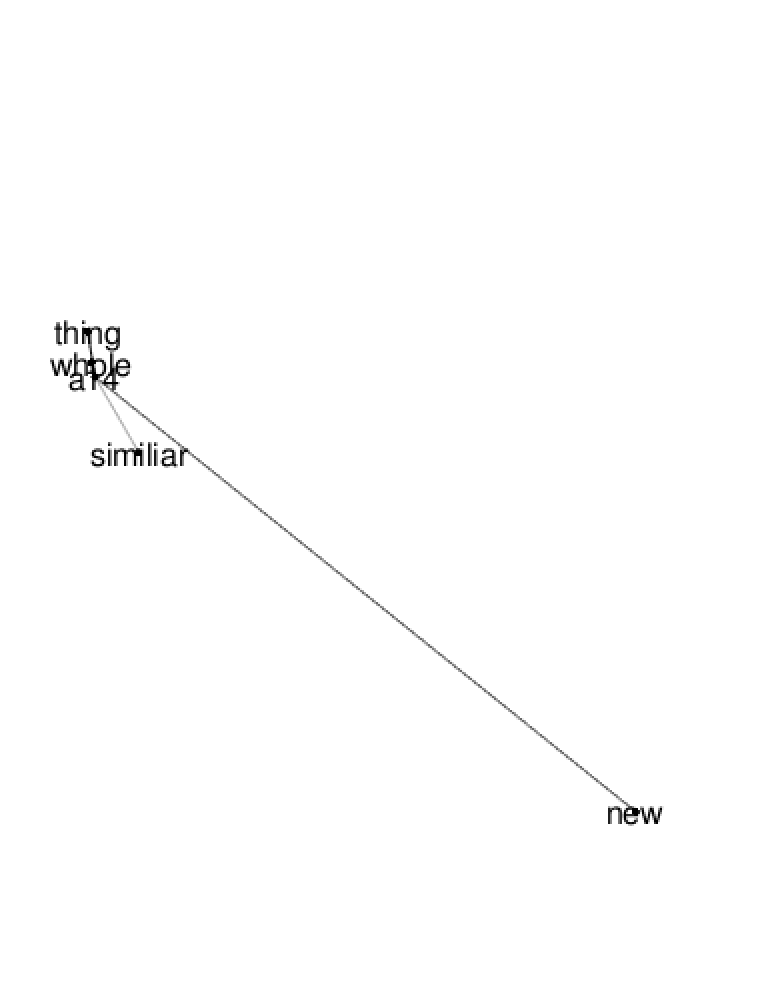
The top term of the fourth feature, “resolution” along with its neighbors 'feature', 'new', 'store', 'walk', 'call', also provide insights into negative aspects of the product as experienced by the user. It indicates a poor resolution of the phone, impeding user experience, and the fact that the user does not like the new features and the fact that the user is not available online and the user has to walk up to the store to get the phone. Also the call quality and call related features are not received well by the user.

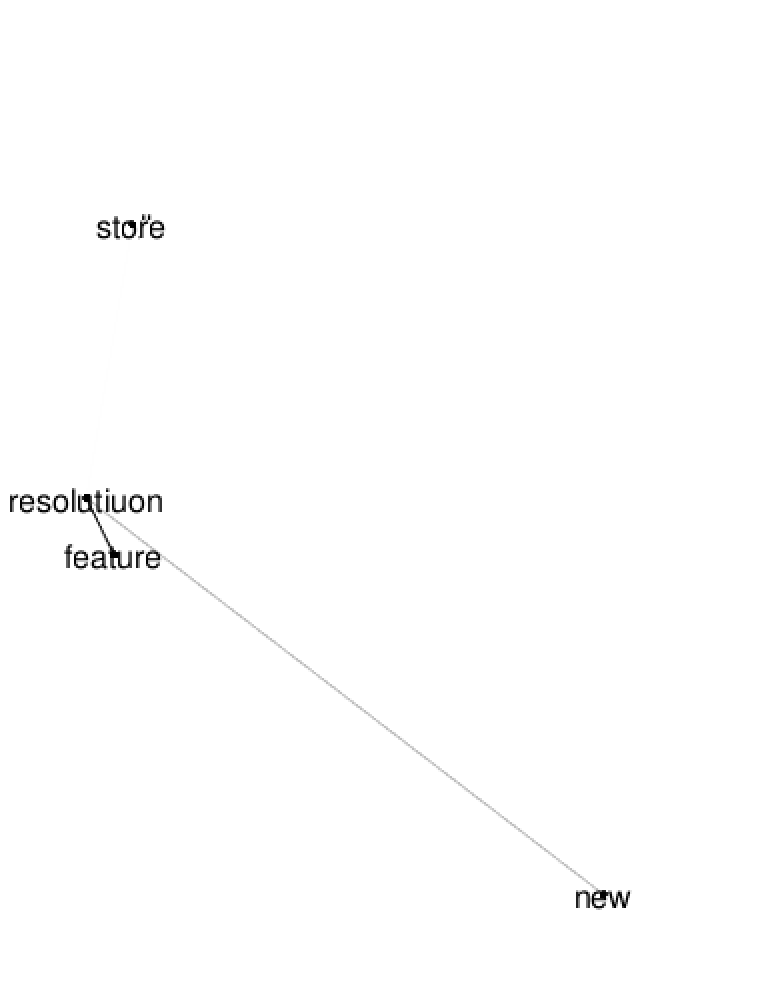
The top term of the fifth feature, ‘a13’, along with its neighbors "handset", "quote", "bezel", "smaller", "surprise" involve a comparison between the current product and the model a13 in various features indicating that the user liked the previous version better. The holding experience of the phone, along with the fact that a14’s frame is too big is seen as a potentially negative aspect of the phone. The users might have probably expected a sleeker design, but have been disappointed due to the phone's overall physical features.

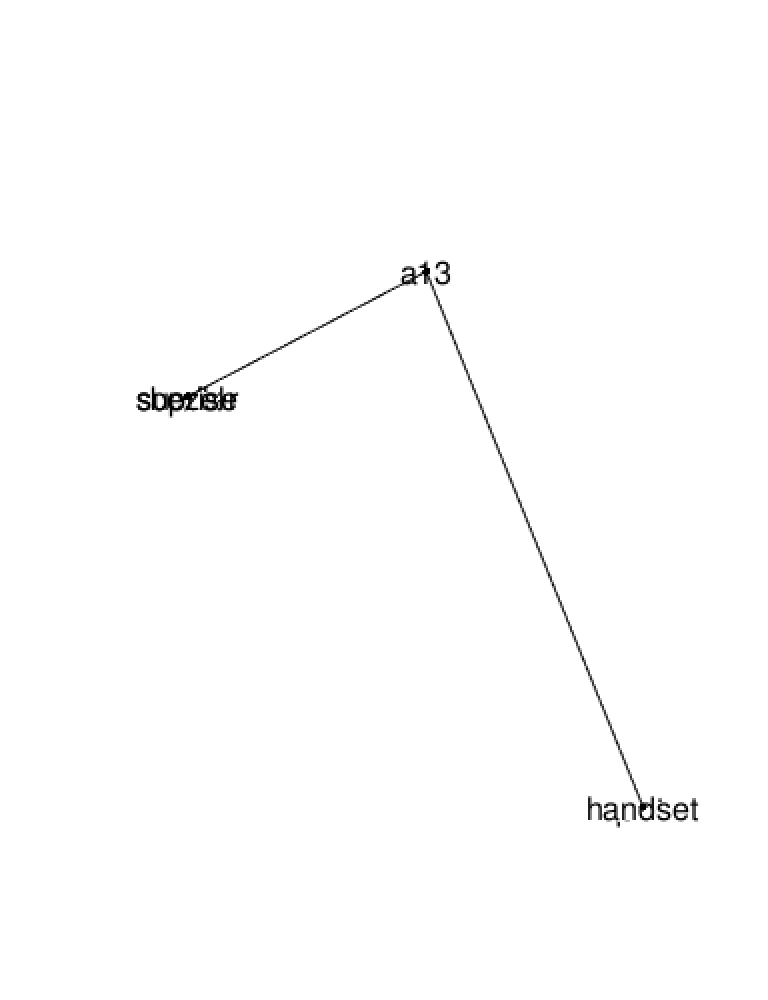
Below we present some graphs that were used for visualization



Some of the words are probably too far away and hence the connecting lines are not visible







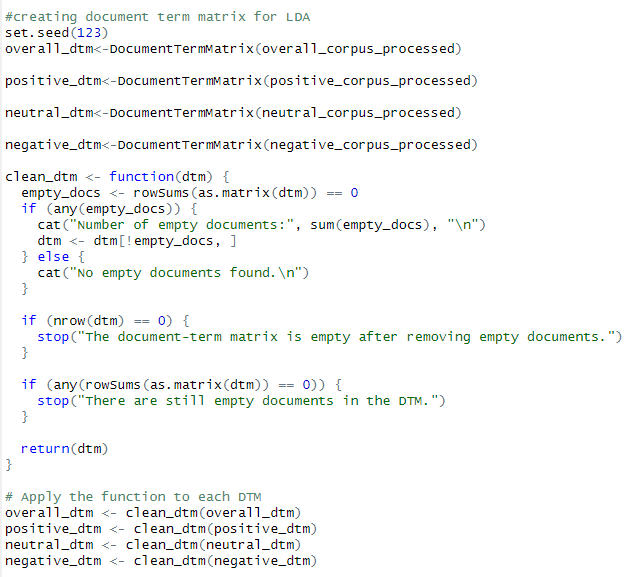
Some of the words are at the same distance and hence are probably not visible.

***\*It must be noted that sentiment tagging has not performed as well as expected because it has tagged most of the stronger negative sentiments under neutral \****

# Latent Dirichlet Allocation (LDA)

LDA helps in discovering the structure within a large corpus of text. It automatically identifies a set of topics present in the documents and associates each document with a mixture of these topics. By modeling topics, LDA is helpful as it helps in understanding a large corpus of unstructured texts through providing insights into the major themes and trends within our dataset.

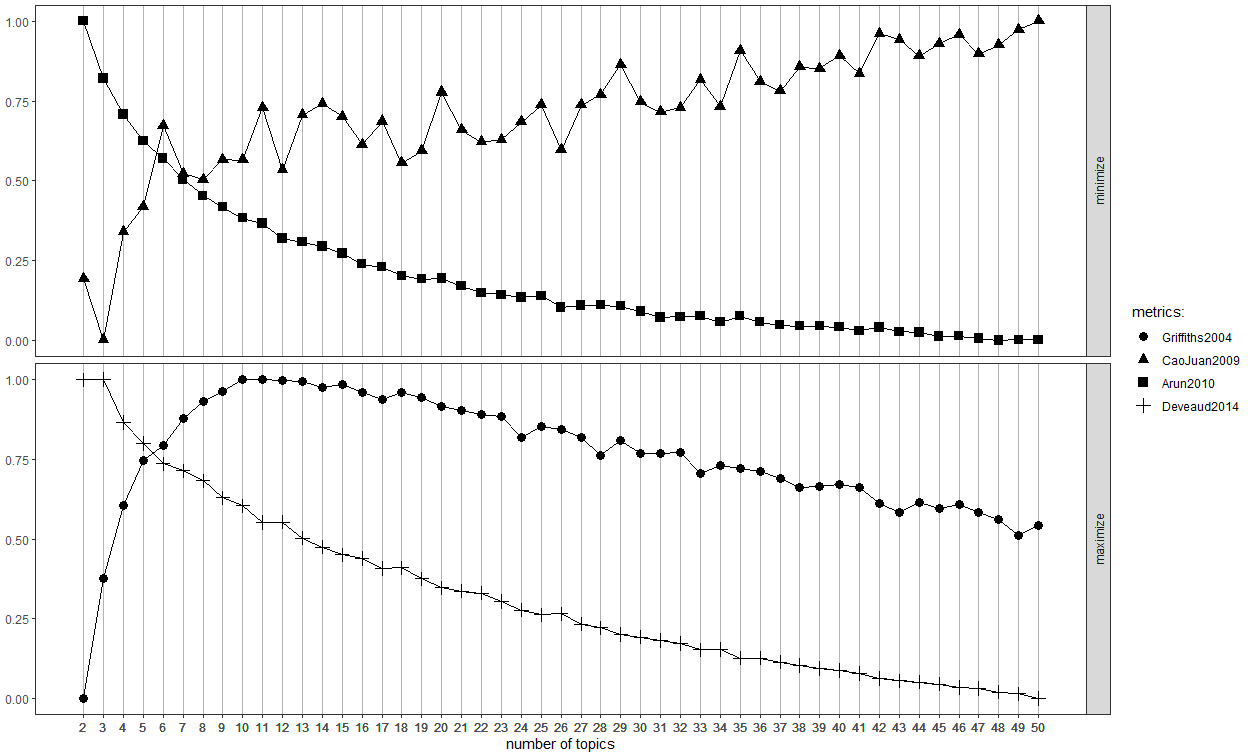
When creating the document term matric for the LDA model, it does not work with any other type of method than the Term Frequency weighting so the TF-IDF model used for LSA cannot be applied here and we need to create new document-terms matrices:



So we begin with installing the relevant packages then figuring out the ideal number of topics for each document-term matrix, we use 4 metric:

* Griffiths2004: Measures the likelihood of the model.
* CaoJuan2009: Measures the density of topics.
* Arun2010: Measures the symmetry of topics.
* Deveaud2014: Measures the coherence of topics.

Over here we have an example from the positive dtm: in the minimize graph we choose the lowest points and in the maximize we chose the highest, we choose points by trying to balance these metrics, so for positive\_dtm 10 or 12 topics appear to be the best choice



These were my findings for number of topics for each dtm:

* #13 or 7 for overall
* #12 or 10 for positive
* #11 or 15 for neutral
* #19 or 16 for negative

We use a deterministic method for the lda model because it is faster and more reliable for re-testing of the model. Based on topic distribution observations was then performed for each sentiment and the answers were

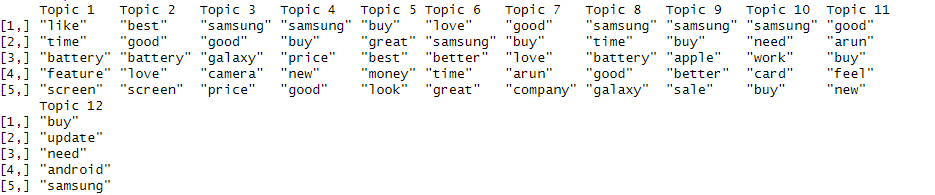
* #13 for overall
* #12 for positive
* #15 for neutral
* #16 for negative

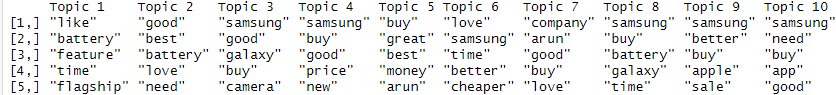
We then did the comparisoon of different number of topics for each dtms’ top 5 most common words in each topic to get a better understanding of which number of terms is better, We performed this for each of the dtms.

Here we performed comparison for 12 vs 10 topics for positive dtm, 12 had a better representation of words, so we choose it:

A black and white text

Description automatically generated

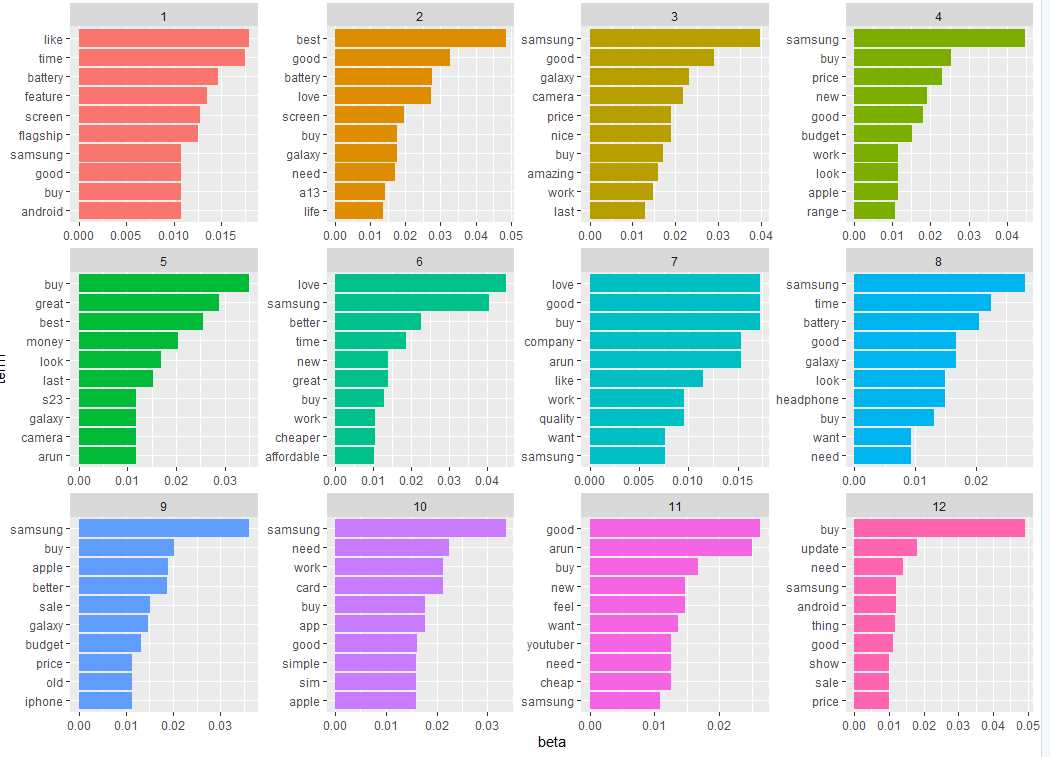




Let’s now look at the topic term analysis for each sentiment.

## Topic Term analysis:

This is the topic term analysis for the positive LDA, it has many unique terms that you wont find in other sentiments such as words related to pricing like “cheap, “good, “budget , “affordable”, “want. We can also identify aspects of the phone model here, such as the screen, battery, range, and software occurring in multiple topics, since this is a positive LDA model, we can safely assume the phone excels in these departments:



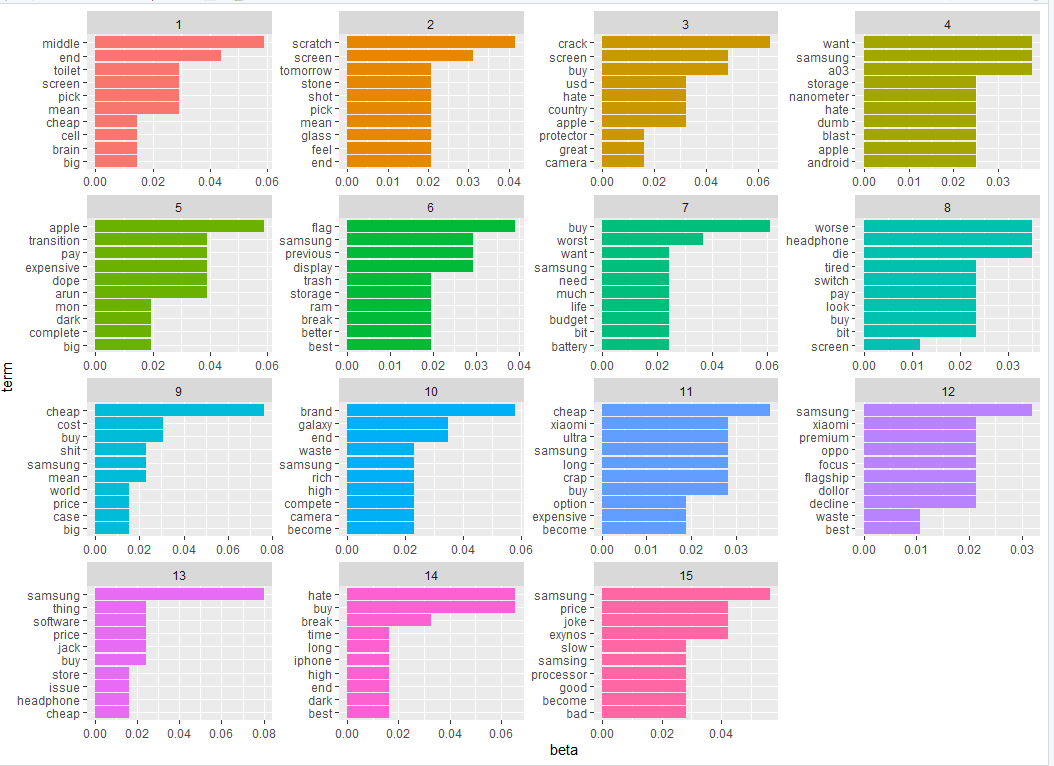
Lets look at a topic wordcloud to get a better idea of the topic term analysis.

This is topic 3 and we can see it mentions some of the positive qualities of the phone model being reviewed and the average users thoughts on the phone.

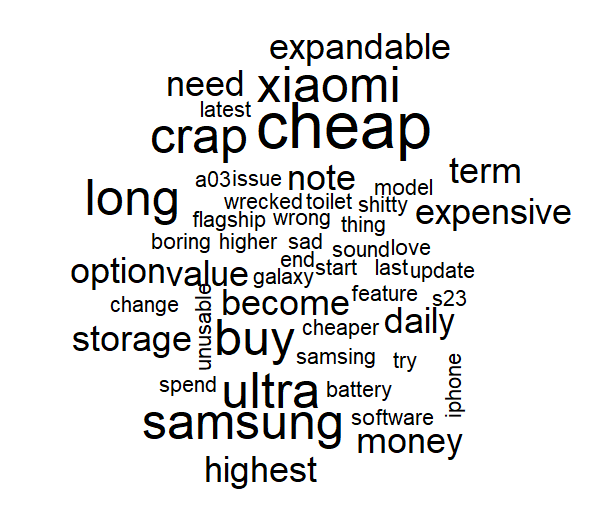
A close up of words

Description automatically generated

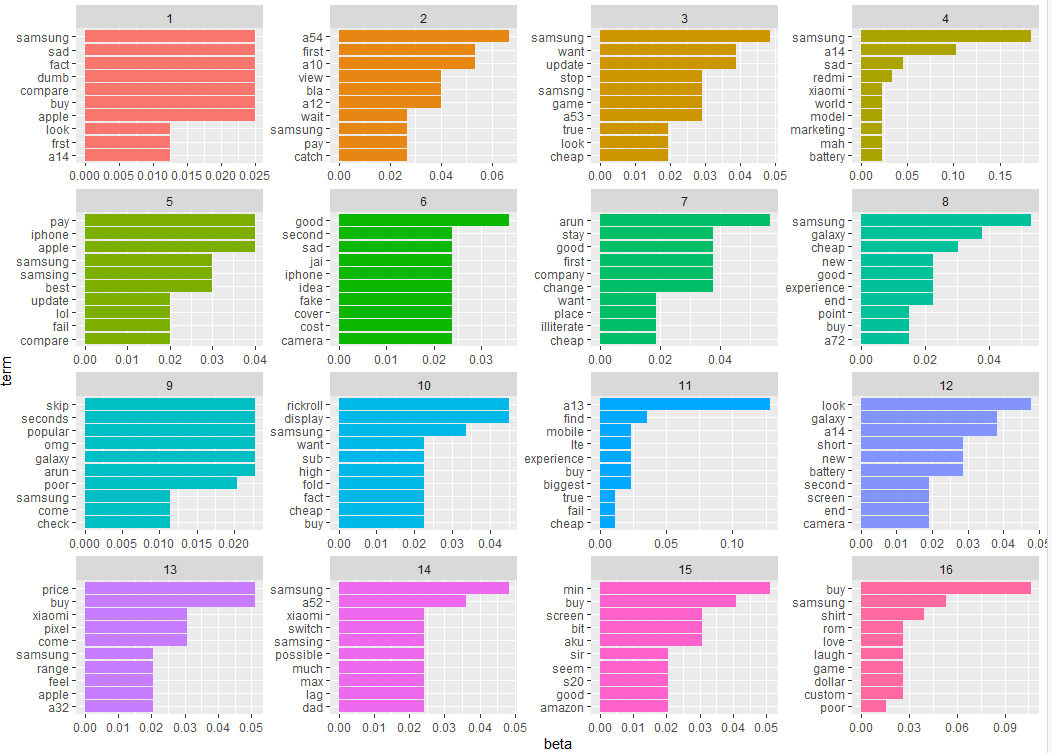
For the neutral LDA model, the number of terms for it is much smaller and their beta value, meaning the probability of term in each topic is similar for most terms, which makes sense for neutrality, this topic/term analysis seems to heavily lean towards negativity, especially towards swear words. Perhaps this model considers those words as neutral because their meaning can be changed based on the context, or perhaps because it could not identify whether the words were talking about the phone or youtuber or other related subjects.



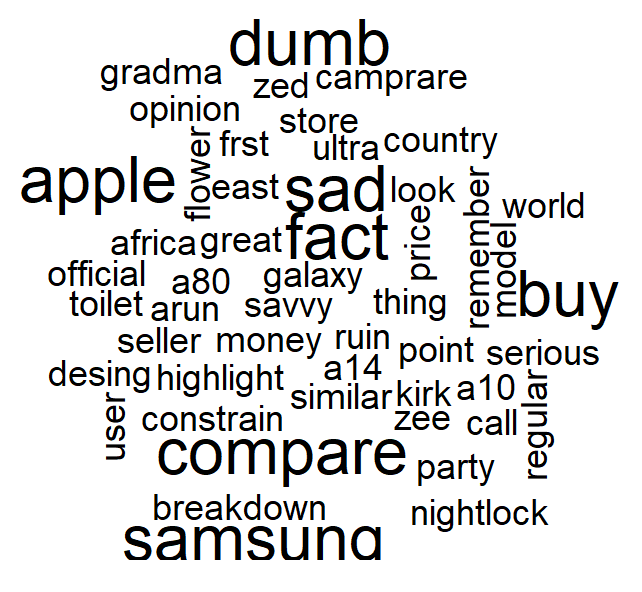
A wordcloud of topic 11 shows us that neutral topics lean heavily towards negative comments but they also seem to contain some positive/neutral words such as expandable, storage, samsung.



In the negative Lda model, the terms in the topics seems to lean towards the comparison of the a14 model, which is the main phone reviewed in the video, to other phone models such as the m72, iphone, xiamoi, a52, you can see in topic 8, the chart show that that comparisons to decide whether this phone is worth buying or for example waiting for the next phone.



In the 1st topic for negative lda model we talked about, we can take a closer look at the words that make up the topic and identify the key trends in negative results, such as comparisons with other phones or problems with the phone itself like battery and the size.

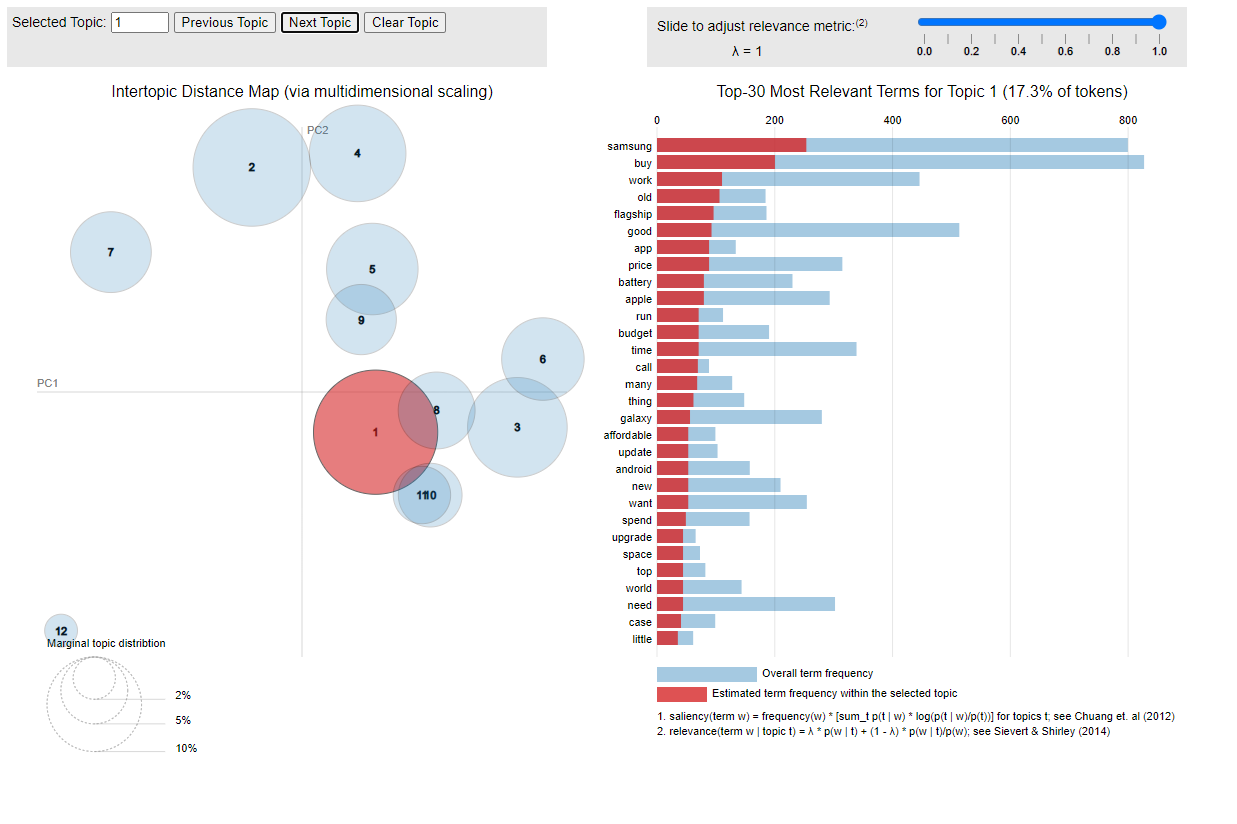


To finish of this segment, the overall and positive model works well in distributing the terms correctly to the topics and the top comments per topic per segment seem to support this well. However, for the neutral and negative lda models, for some reason the negative model leans towards more neutral terms and the neutral model leans more towards negative terms. Me and my teammate could not figure out the reason for this problem, we speculated the sentiment tagging method perhaps does not work well with YouTube comments.

## LDA Visualization

This visualization, includes a cluster chart that’s shows related topics, a bar chart that shows the relevant terms per topic, and a relevance metric, where the higher the number it shows terms that are more unique and specific to the topic and lowering it would indicate terms that are more common throughout the topic. I will use 1 image per lda visualization but these observations are based on all the topics, It would be overloading of information to include screenshots of each topic for each sentiment as that would be over 30 screenshots.

Positive LDA Model:



* Topic 1 is the most prominent topic, as indicated by its large red circle it accounts for 17.3% of tokens.
* Topics 2(15.4% of tokens), 3(11.1% of tokens), and 4(10.4% of tokens) are also significant but less dominant compared to Topic 1.
* The smaller circles represent less significant topics, suggesting a more focused discussion around certain themes.
* Many of the topics seem to group up showing similar terms that usually discuss the price, screen, battery, software of the phone model to varying degrees.
* Topic 7 appears isolated as many of th terms there lean more towards being neutral and topics unrelated to the phone model such as religion or the youtuber himself.

Overall, the LDA visualization highlights that the primary positive sentiment themes in the YouTube comments revolve around the android a14's features, value for money, and comparisons with other major brands like Xiaomi and Apple.

Neutral LDA Model:

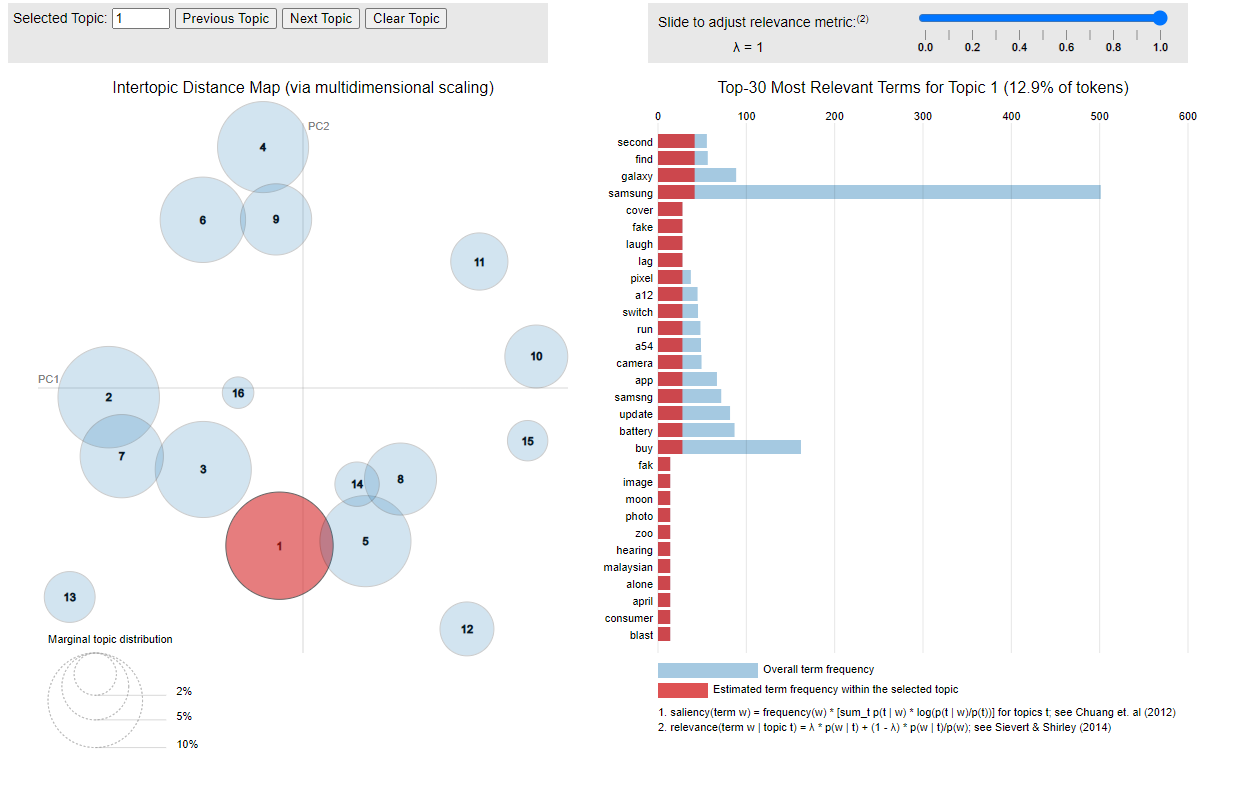
A screenshot of a graph

Description automatically generated

* Topic 1 is the most prominent topic, but it contains less of a percentage of tokens than positive model, as it accounts for 13.9% of tokens.
* Topics 3,10,12 have a lot of overlap suggesting these topics discuss similar terms, they talk about negative terms that indicate problems with the model, such as problems with the display, speed of the device and its battery/
* Topics 2,8 also have overlap but these instead discuss some positive terms that have been included the neutral sentiment such as its feel and price for value.

Overall, the LDA visualization highlights that the neutral sentiment themes in the YouTube comments are centered around hardware features, considerations for purchasing, comparisons with other brands, and broader discussions about the product's performance and marketing. However, it also seems to lean far more on the negative side containing many swear words and words usually associated with negative connotations implying that this sentiment tagging may not be accurate.

Negative LDA Model:



* Topic 1, being the most prominent, likely captures some of the negative sentiments and issues with the phone model, such as poor battery, bad speakers through words like “hearing” and some broader comparisons with other phone models buy Samsung indicating how the a14 is a weaker model
* Topics 5, 8, and 14, being closely grouped near Topic 1, also go more into comparisons with other models such as the s21 and the a52, talking about our a14 model’s shortcomings. It also mentions problems with reliability and availability talked about through mentioning different locations like “Africa”, “disappoint”, and “reliability”.
* Topics 2, 3, and 7 mainly discuss technical problems being discussed in the comments such as overheating and e-waste production. It also mentions other related problems with the phone in terms of its presentation and shipment problems.
* Topics 4, 6, 9, focus more on the comparisons than 1,5,8,14, however the majority of the terms are nonsensical irrelevant terms suggesting that these topics are not relevant to the discussion.
* 11,10 and 15 seem to focus less on the actual a14 phone model and more on negative terms associated with the comparison models for iPhones, and for problems with the video itself

In summary, the LDA visualization for negative sentiment reveals problems with the a14 phone model though a comparison with other brands such as iPhone, and other models produced by Samsung. It is also far less accurate in terms of its sentiment tagging as opposed to the positive model which succeeds in capturing the right terms for the explanation of the LDA model.