

MS4252 Big Data Analytics

Project Presentation



Group Members

Chow Sat Yat Morris

Agenda

- ❖ Introduction
- ❖ Data Description
- ❖ Objective 1 - Evaluate Helpful Reviews
- ❖ Objective 2 - Determine the sentiment of Fire Tablet comm



Company Background



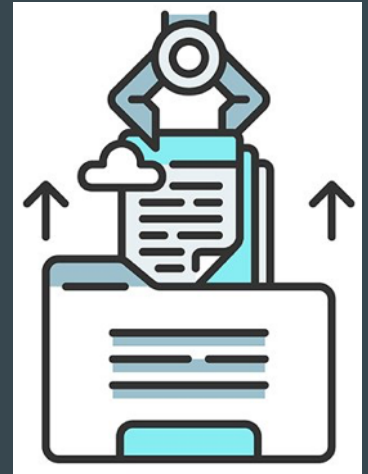
- Largest online e-commerce company in the US
- One of the earliest companies to start e-commerce
- 3 subsidiaries
 - Alexa Internet
 - a9.com
 - Internet Movie Database (IMDB)



Dataset Description

- > 34,000 **consumer reviews** for Amazon products
 - Kindle, Fire Tablet, etc
- Basic product & customer information, rating, review text, etc for each product

→ Extract **different variables** for each objective



Objective 1 - Evaluate Helpful Reviews



Method: Text Cluster, Topic Extraction

Aim: Find out things that people value when choosing a product & characteristics of a product

Dataset Description

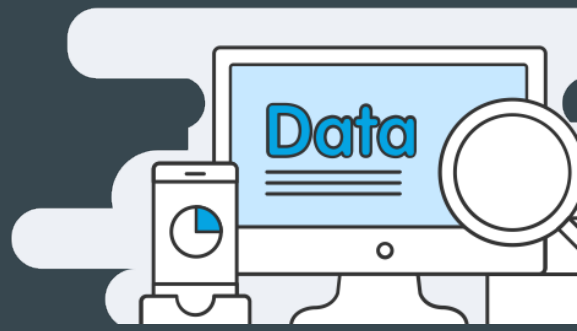
Analyze reviews `where reviews_numHelpful > 0;`

(Total obs: 3,311)

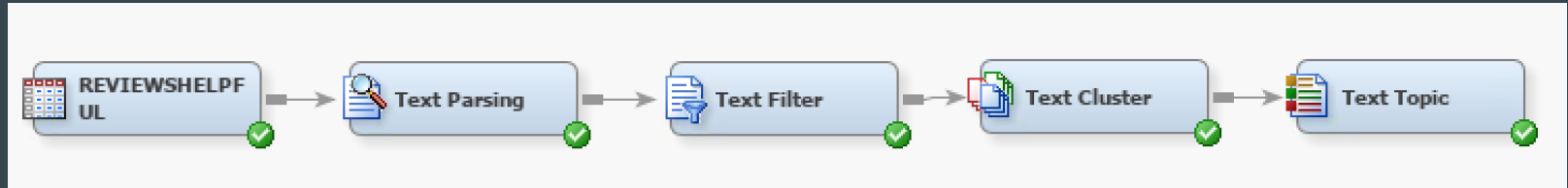
→ At least 1 person find the review helpful

Variables Used: reviews_numHelpful, reviews_Rating, reviews_Text

	reviewsNumHelpful	reviewsRating	reviewsText
1	814	5	IMPORTANT UPDATE (3/8/17): As you read my original review, please be aware that Amazon announced in early February (2017) a new firmware update that allows Amazon Tap owners to enable a hands-free option! By going into Settings in the Alexa app, you can now choose the hands-free option and have your Tap respond to voice commands without touching it, just like the Echo and Echo Dot! This is a fabulous development and something Tap owners have been hoping for. Thank you, Amazon! Since the lack of hands-free operation was the one major shortcoming of the Tap, this wonderful little device has now gotten a whole lot better. One thing to keep in mind, though, especially if you use your Tap on battery power: The hands-free mode will drain the battery more quickly, because it requires the microphone to be on at all times. But since you can easily toggle between hands-free or regular tap mode using the Alexa app, you can pick and choose how you want to use your Tap for maximum battery life and convenience. ————— Since the Fire TV stick was announced, I've heard a lot of people wonder if it was something worth picking up, essentially asking what does this thing do that my TV doesn't do already. The answer is going to be different for different people, but here's an easy way to think about it: Amazon Fire TV Stick is an excellent purchase if: -Your TV is not a smart TV. Meaning your TV or dvd player did not come pre-loaded with apps such as Netflix, Youtube, Pandora etc... -Your TV is a smart TV, but apps do not perform well. Some smart TVs have painful performance when streaming video, apps can be difficult to update, wifi connection can be spotty etc... -You want to access Youtube, Netflix, Amazon Prime, Pandora, and other such apps on your TV - You are an Amazon Prime Member and want to access your Prime Video, Prime Music, and Prime Photo accounts - You want to participate in casual gaming on your television - You want to listen to music on your TV - You want to have portability. In other words, Television access to the above
2	780	5	Hey Alexa. Hey Alexa. - Night and day it's Hey Alexa. Make the fire, fix the breakfast, Wash the dishes, do the mopping! Ok, so I haven't figured out how to make Alexa clean my house like Cinderella... yet. But I'm working on it. In the meantime, I have figured out that she is much more than a glorified alarm clock or DJ. Although, you should know that alarms are not currently supported on this device, like they are on the Echo. This is a more basic version of Alexa. It doesn't have her full functionality. ABOUT ME! I'm not sure what the target demographic for this product is, but I am a professional female in my early thirties, who is in a committed relationship. That is a nice way of saying that my partying days are behind me and my boyfriend and I spend a lot of time binge watching stuff on the sofa. Yes. We are suburban couch potatoes. No shame here. Also, I am an Amazon Prime member, but do not have a 4k TV. I used to have a Roku that I used for Netflix and Hulu. However, I was more interested in free shipping.
3	744	4	UPDATE! 11/11/15: More than a year now of using the stick literally every single evening for netflix, and it has been overall very reliable. Bottom line: This is the best streaming device for the dollar (in my humble opinion). It's cheap, it handles netflix/prime/rboGo/plex/games/a lot more doesn't seem much slower than the full firetv box. I started with a Roku2 in the living room which got pushed to the bedroom when I got the firetv box, then the firetv stick replaced the roku2. I tried the roku stick but returned it because it was just too damn slow. I'm completely firetv now and 100% happy with it. What it does: it adds internet/streaming functionality to any tv/monitor with an HDMI port. You can use any of the thousands of paid and/or free apps such as netflix, amazon prime, hulu, plex, news, digital fireplace, etc... There's a ton you can do. With some extra effort you can sideload things like Kodi and 3rd party apps. It's super simple to use, and the image quality can be great if your internet connection is
4	740	5	I owned a Kindle Keyboard for about a year, and I purchased the leather lighted case for that unit. While the corner light left a little to be desired as far as even lighting, overall I was very pleased with the case. Based on that experience, I went ahead and purchased the lighted cover for my new Kindle. I can't say this experience has as been as positive. Let's start off with what's good about the case. The Kindle battery powered light is a big improvement. The light is now centered on the screen, so illumination is more even now. I don't use the light often, but in the cases where I do it will be nice to have more balanced coverage. One positive on the case itself is that I appreciate how much lighter this one feels than the previous generation. The hinge area is smaller, which means less of the spine-folding effect the other one took on after heavy use. I never took my Kindle Keyboard out of the case after it was put in, so the snug fit I have seen complaints about is not an issue for me. It IS light, tho
5	730	3	



Model Description



1. Text Parsing



- **Breaking down** the review text into tokens (words)



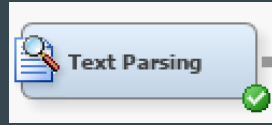
Treat all terms that have the same root form as **equivalent**



Synonyms	
Stem Terms	Yes
Synonyms	SASHELP.ENGSYNMS ...
Filter	
Start List	...
Stop List	SASHELP.ENGSTOP ...
Select Languages	...

<input checked="" type="checkbox"/>	love	793	689	<input checked="" type="checkbox"/>	3,263	Verb	Alpha
<input type="checkbox"/>	loving	9	9	<input type="checkbox"/>		Verb	Alpha
<input type="checkbox"/>	love	532	459	<input type="checkbox"/>		Verb	Alpha
<input type="checkbox"/>	loved	77	75	<input type="checkbox"/>		Verb	Alpha
<input type="checkbox"/>	loves	175	168	<input type="checkbox"/>		Verb	Alpha

1. Text Parsing



Synonyms	
Stem Terms	Yes
Synonyms	SASHELP.ENG SYNMS ...
Filter	
Start List	...
Stop List	SASHELP.ENG STOP ...
Select Languages	...

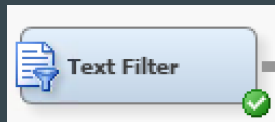


Stop List-EMWS2.TextParsing_stopList

Term	Role
'd	
'll	
'm	
're	
's	
've	
a	
aboard	
about	
above	
according	
accordingly	
across	
actually	
after	
afterwards	
again	
against	
ago	

Default stop list is used
→ remove meaningless/
low information words

2. Text Filter



- ↓ total number of parsed terms analyzed

Inverse Document Frequency (IDF)

→ **importance** of a word to a document

Train	
Variables	...
Spelling	
Check Spelling	Yes
Dictionary	...
Weightings	
Frequency Weighting	Default
Term Weight	Inverse Document Frequency
Term ...	
Minimum Number of Documents	5
Maximum Number of Terms	.
Import Synonyms	...

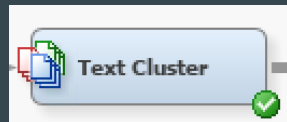


Manual Filtering

→ keep desired/**interested** topics only

TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
⊕ tablet	1600	1080	<input checked="" type="checkbox"/>	2.614	Noun	Alpha
⊕ great	1212	1028	<input checked="" type="checkbox"/>	2.685	Adj	Alpha
⊕ good	935	767	<input checked="" type="checkbox"/>	3.108	Adj	Alpha
⊕ buy	835	713	<input checked="" type="checkbox"/>	3.213	Verb	Alpha
⊕ kindle	1019	694	<input checked="" type="checkbox"/>	3.252	Verb	Alpha
⊕ love	793	689	<input checked="" type="checkbox"/>	3.263	Verb	Alpha
⊕ price	639	585	<input checked="" type="checkbox"/>	3.499	Noun	Alpha
⊕ easy	582	511	<input checked="" type="checkbox"/>	3.694	Adj	Alpha
⊕ fire	706	498	<input checked="" type="checkbox"/>	3.731	Noun	Alpha
⊕ read	629	471	<input checked="" type="checkbox"/>	3.811	Verb	Alpha
⊕ book	518	379	<input checked="" type="checkbox"/>	4.125	Noun	Alpha
⊕ screen	478	378	<input checked="" type="checkbox"/>	4.129	Noun	Alpha
⊕ work	400	360	<input checked="" type="checkbox"/>	4.199	Verb	Alpha
⊕ purchase	361	327	<input checked="" type="checkbox"/>	4.338	Verb	Alpha
⊕ device	432	312	<input checked="" type="checkbox"/>	4.405	Noun	Alpha
⊕ apps	381	306	<input checked="" type="checkbox"/>	4.433	Noun	Alpha
⊕ product	348	304	<input checked="" type="checkbox"/>	4.443	Noun	Alpha
⊕ want	353	284	<input checked="" type="checkbox"/>	4.541	Verb	Alpha
⊕ like	298	274	<input checked="" type="checkbox"/>	4.593	Verb	Alpha
⊕ recommend	258	253	<input checked="" type="checkbox"/>	4.708	Verb	Alpha
⊕ time	275	239	<input checked="" type="checkbox"/>	4.79	Noun	Alpha
⊕ game	254	236	<input checked="" type="checkbox"/>	4.808	Noun	Alpha
⊕ play	248	226	<input checked="" type="checkbox"/>	4.871	Verb	Alpha

3. Text Cluster



- Cluster documents into different **groups**
- Find out **descriptive terms** for clusters



Train	
Variables	
Transform	
SVD Resolution	Low
Max SVD Dimensions	100
Cluster	
Exact or Maximum Number	Maximum
Number of Clusters	40
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	15

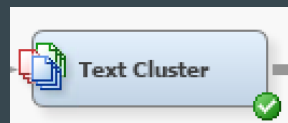
Not a big dataset → Max **100** dimensions

✓ Capture most of the information in document collection

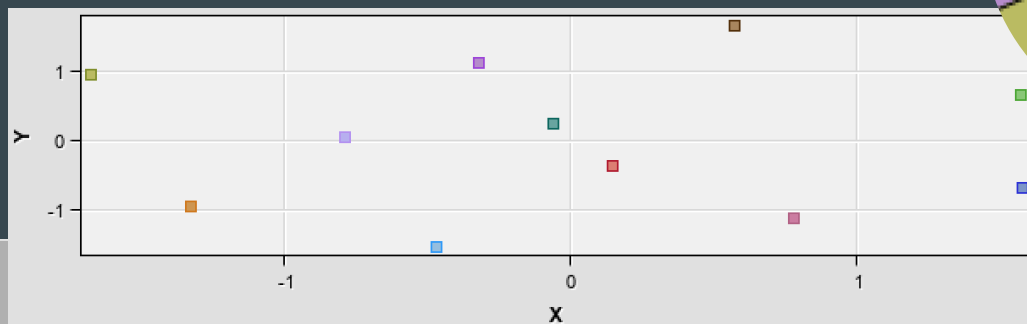
(Default)

- ☐ Soft-Clustering (**X** boundary)
- ☐ Iterative method

3. Text Cluster

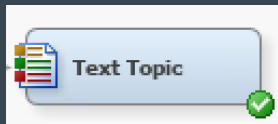


❖ 11 clusters



Cluster ID	Descriptive Terms	Frequency	Percentage
5	+tablet +great +price +game +kid +'great tablet' 'great price' +play +son +beat +gift +buy +child +work +little	478	14%
2	+battery +life 'battery life' +charge +long +screen +device +nice +size ipad +time +hold +day +return +fast	438	13%
3	+good alexa +echo music quality +speaker +voice +question weather home +control +play +add setup +want	422	13%
8	+kindle +light reading paperwhite +page voyage +reader +turn +button +upgrade +hold night +backlight +eye +screen	390	12%
11	apps +store google +card +app. memory sd +game +download +video +play +tablet free +movie +add	322	10%
1	+book +read +library +'reading book' +kindle reading able +download +carry +size +want +eye +reader perfect +bright	275	8%
4	+love +buy +gift christmas +daughter mom +birthday +granddaughter +niece +kid +son +friend wife +child +happy	263	8%
7	+fire +tv +'fire tv' +stick +remote 4k roku +cable +service +stream +content prime +box +fast +upgrade	242	7%
10	+easy 'setup' setup +operate +carry +read +navigate +great +bright +light +simple +love +tablet easy +item	197	6%
6	+product 'great product' buy +great best +'good product' +recommend 'great price' +price +return excellent +big +operate +gift +fast	171	5%
9	+tv +cable +stream +box kodi netflix +show roku +connection +movie 4k +remote +content internet prime	113	3%

4. Text Topic



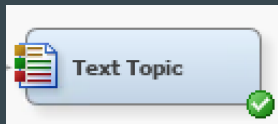
- Automatically associate terms & documents to topics

Topic: Collections of terms that **describe** & **characterize** a main theme/idea

❖ **10** Text Topics → majority are **positive** (“good”, “great”)

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.130	0.029	+great, +price, +tablet, +great tablet, great price	39	439
Multiple	2	0.118	0.032	+kindle, +screen, +page, paperwhite, reading	112	485
Multiple	3	0.113	0.031	+tv, +fire, +fire tv, +stream, +cable	75	298
Multiple	4	0.137	0.029	+love, +buy, +gift, +kid, +daughter	57	507
Multiple	5	0.111	0.032	+tablet, +good, apps, +price, +screen	91	498
Multiple	6	0.093	0.032	+echo, alexa, music, +device, +speaker	114	225
Multiple	7	0.118	0.031	+book, +read, +game, +play, +download	70	468
Multiple	8	0.112	0.030	+easy, +tablet, +read, set up, +good	52	457
Multiple	9	0.105	0.030	+battery, +life, battery life, +long, +great	45	209
Multiple	10	0.106	0.030	+product, great product, +good, +recommend, +buy	45	339

4. Text Topic



Amazon Echo

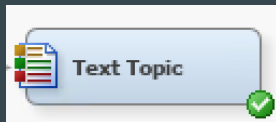
- Hands-free speaker (control with your voice)
 - Play **music**
 - Ask questions
 - Make calls
 - Control lights at home, etc

6th

Topic
+great,+price,+tablet,+great tablet,great price
+kindle,+screen,+page,paperwhite,reading
+tv,+fire,+fire tv,+stream,+cable
+love,+buy,+gift,+kid,+daughter
+tablet.+good.apps.+price.+screen
+echo,alexa,music,+device,+speaker
+book,+read,+game,+play,+download
+easy,+tablet,+read,set up,+good
+battery,+life,battery life,+long,+great
+product,great product,+good,+recommend,+buy



4. Text Topic



Fire Tablet

- Good **gift** for kids
- Care about **apps**, **price** & **screen**
- Read books, download & play games



4th

5th

7th

Topic

+great,+price,+tablet,+great tablet,great price

+kindle,+screen,+page,paperwhite,reading

+tv +fire +fire tv +stream +cable

+love,+buy,+gift,+kid,+daughter

+tablet,+good,apps,+price,+screen

+echo.alex.a.music.+device.+speaker

+book,+read,+game,+play,+download

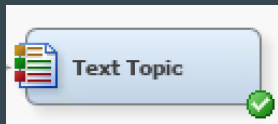
+easy,+tablet,+read,set up,+good

+battery,+life,battery life,+long,+great

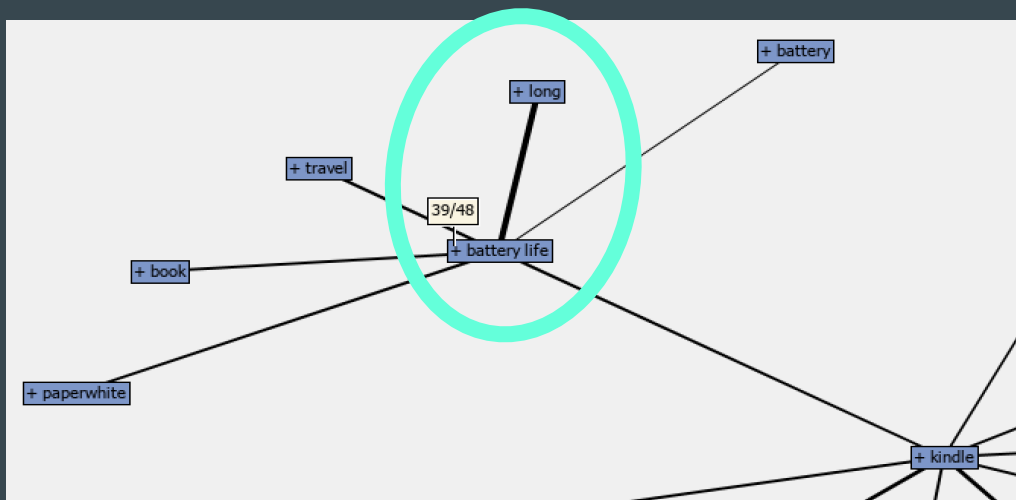
+product,great product,+good,+recommend,+buy



4. Text Topic



- “Battery Life” & its variants → 48 documents
 - 39 (81%) also contain “Kindle”
- Easy to set up



Topic

+great,+price,+tablet,+great tablet,great price

+kindle,+screen,+page,paperwhite,reading

+tv,+fire,+fire tv,+stream,+cable

+love,+buy,+gift,+kid,+daughter

+tablet,+good,apps,+price,+screen

+echo,alexa,music,+device,+speaker

+book,+read,+game,+play,+download

+easy,+tablet,+read,set up,+good

+battery,+life,battery life,+long,+great

+product,great product,+good,+recommend,+buy

8th

9th



Conclusion & Recommendation

- Answer **questions** that they have in mind
- Represents their **thoughts** on the product



Amazon Echo

- Make sure music are **up-to-date**
- **Cooperate** with different music platforms

Fire Tablet

- **Bundle selling** (other kids item)
- Come with access to **Google Play Store**

Kindle

- Keep up with **long battery life**
- Balance between easy set up and diversified functions

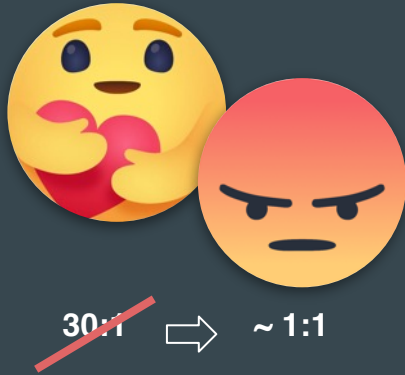
Objective 2 - Determine the Sentiment of Fire Tablet comments



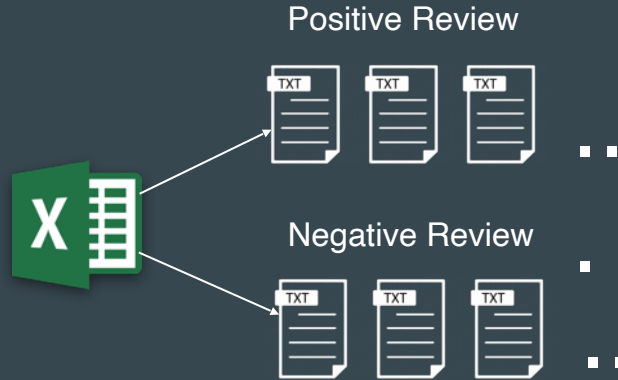
Method: Sentiment Analysis
Aim: To discover what features should
Amazon improve the Fire Tablet

Dataset Preparation

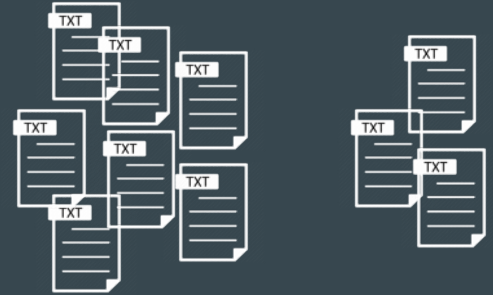
Best Selling Product (Fire Tablet) Text Review



Step1: **Random sample** for balanced review dataset



Step2: Convert each text review cell in Excel to individual text documents by sentiment in **Corpora Tab**



Step3: Data Partition **70/30**

Statistical Model Description

Simple Statistical Model

Text Result | Graphical Result |

With text normalization algorithm [Smoothed Relative Frequency] and feature ranking algorithm [No Feature Ranking]:
Overall precision: 70.06%
Positive precision: 72.38%
Negative precision: 65.79%

With text normalization algorithm [Smoothed Relative Frequency] and feature ranking algorithm [Risk Ratio]:
Overall precision: 70.37%
Positive precision: 75.24%
Negative precision: 61.40%

With text normalization algorithm [Smoothed Relative Frequency] and feature ranking algorithm [Chi Square]:
Overall precision: 66.05%
Positive precision: 54.76%
Negative precision: 86.84%

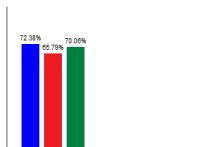
With text normalization algorithm [Smoothed Relative Frequency] and feature ranking algorithm [Information Gain]:
Overall precision: 66.98%
Positive precision: 56.19%
Negative precision: 86.84%

BEST MODEL is Smoothed Relative Frequency and Risk Ratio

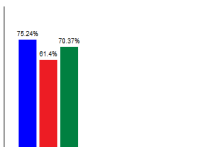
70.37
%

Overall Best Model:
(Smoothed Relative Frequency and Risk

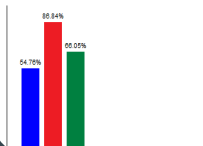
Smoothed Relative Frequency No Feature Ranking



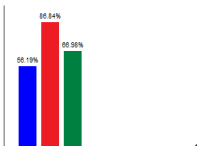
Smoothed Relative Frequency Risk Ratio



Smoothed Relative Frequency Chi Square



Smoothed Relative Frequency Information Gain



Advanced Statistical Model

Text Result | Graphical Result |

With text normalization algorithm [Relative Frequency] and feature ranking algorithm [No Feature Ranking]:
Overall precision: 68.83%
Positive precision: 73.33%
Negative precision: 60.53%

With text normalization algorithm [Relative Frequency] and feature ranking algorithm [Risk Ratio]:
Overall precision: 66.67%
Positive precision: 80.95%
Negative precision: 40.35%

With text normalization algorithm [Relative Frequency] and feature ranking algorithm [Chi Square]:
Overall precision: 66.98%
Positive precision: 55.71%
Negative precision: 87.72%

With text normalization algorithm [Relative Frequency] and feature ranking algorithm [Information Gain]:
Overall precision: 66.05%
Positive precision: 54.29%
Negative precision: 87.72%

BEST MODEL is Relative Frequency and No Feature Ranking

68.83
%

Relative Frequency
%

Text Result | Graphical Result |

With text normalization algorithm [Smoothed Relative Frequency] and feature ranking algorithm [No Feature Ranking]:
Overall precision: 70.37%
Positive precision: 72.38%
Negative precision: 65.79%

With text normalization algorithm [Smoothed Relative Frequency] and feature ranking algorithm [Risk Ratio]:
Overall precision: 70.37%
Positive precision: 75.24%
Negative precision: 61.40%

With text normalization algorithm [Smoothed Relative Frequency] and feature ranking algorithm [Chi Square]:
Overall precision: 66.05%
Positive precision: 54.76%
Negative precision: 86.84%

With text normalization algorithm [Smoothed Relative Frequency] and feature ranking algorithm [Information Gain]:
Overall precision: 66.98%
Positive precision: 56.19%
Negative precision: 86.84%

BEST MODEL is Smoothed Relative Frequency and Risk Ratio

70.37
%

Smoothed Relative Frequency
%

Text Result | Graphical Result |

With text normalization algorithm [Okapi BM25] and feature ranking algorithm [No Feature Ranking]:
Overall precision: 69.14%
Positive precision: 72.86%
Negative precision: 62.28%

With text normalization algorithm [Okapi BM25] and feature ranking algorithm [Risk Ratio]:
Overall precision: 66.36%
Positive precision: 80.00%
Negative precision: 41.23%

With text normalization algorithm [Okapi BM25] and feature ranking algorithm [Chi Square]:
Overall precision: 66.67%
Positive precision: 54.29%
Negative precision: 89.47%

With text normalization algorithm [Okapi BM25] and feature ranking algorithm [Information Gain]:
Overall precision: 66.05%
Positive precision: 53.81%
Negative precision: 88.60%

BEST MODEL is Okapi BM25 and No Feature Ranking

69.14
%

Okapi BM25
%

Text Result | Graphical Result |

With text normalization algorithm [Pivoted Length Normalization] and feature ranking algorithm [No Feature Ranking]:
Overall precision: 64.20%
Positive precision: 99.05%
Negative precision: 0.00%

With text normalization algorithm [Pivoted Length Normalization] and feature ranking algorithm [Risk Ratio]:
Overall precision: 64.20%
Positive precision: 99.05%
Negative precision: 0.00%

With text normalization algorithm [Pivoted Length Normalization] and feature ranking algorithm [Chi Square]:
Overall precision: 66.98%
Positive precision: 71.90%
Negative precision: 57.89%

With text normalization algorithm [Pivoted Length Normalization] and feature ranking algorithm [Information Gain]:
Overall precision: 67.28%
Positive precision: 72.38%
Negative precision: 57.89%

BEST MODEL is Pivoted Length Normalization and Information Gain

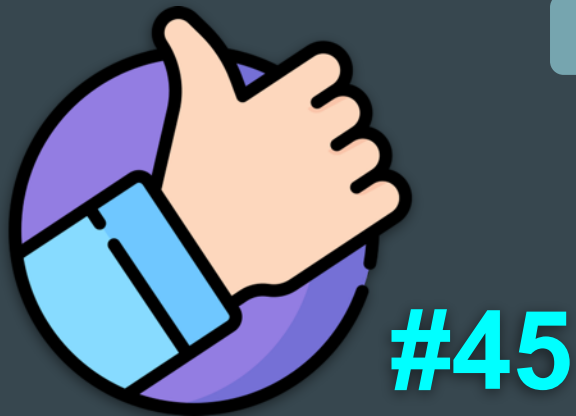
67.28
%

Pivoted Length Normalization
%

Bayes Method | 0.6 Prob Threshold

Rule-based Model Description

Tonal Keyword



Wonderful

Awesome

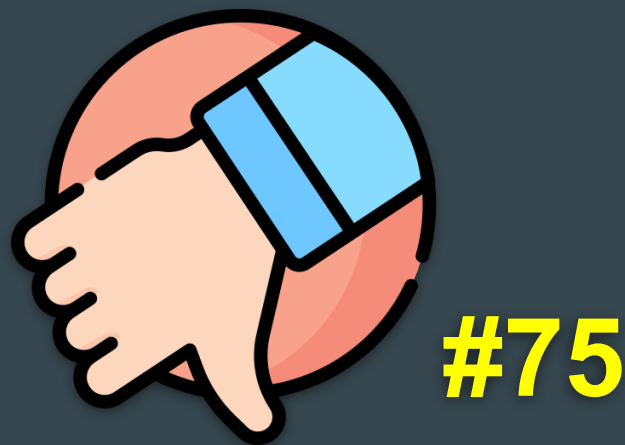
Able to

Pleased

Impressed

Perfect

Easy



Weak

Terrible

Problem

Disappointed

Lack of

Worse

Difficult

Rule-based Model Description

Features & Definitions

1. Function

Navigate | functions | operate

2. Internet

Mb/min | internet | connection

3. Apps

Youtube | Amazon store | games

4. Price

Cost | bucks | budget



5. Screen

Screen | lcd | size | resolution

6. Camera

Cameral lens | len

7. Audio

Audio | sound | speakers

8. Memory

Gb | internal memory | space

9. Processor

Cpu | speed | processor

Rule-based Model Description

Rule-based Body

General Formula



Concept_Rule | (SENT, “_c{TonalKeyword**Positive/Negative**}”, “_def{FireTablet**Features**}”)

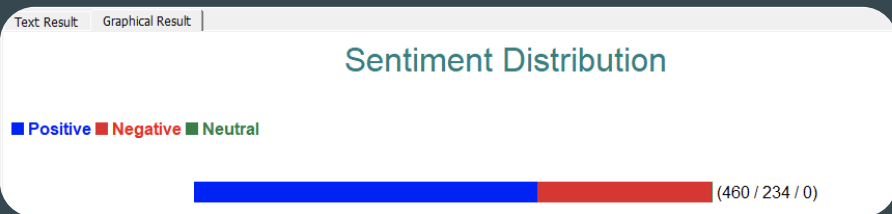
Specified



```
-ve [Internet] | Concept_Rule | (DIST_3, "_c{not locate}", "_def{FireTabletInternet}")
-ve [Apps] | Concept_Rule | (SENT, "_c{not supported}", "_def{FireTabletApps}")
-ve [Memory] | Concept_Rule | (DIST_6, "_c{need}", "more space")
-ve [Memory] | Concept_Rule | (AND, "_c{out of}", "_def{FireTabletMemory}")
-ve [Price] | Classifier | expensive
+ve [Price] | Concept_Rule | (AND, "_c{short}", "_def{FireTabletPrice}")
```

Model Testing Result

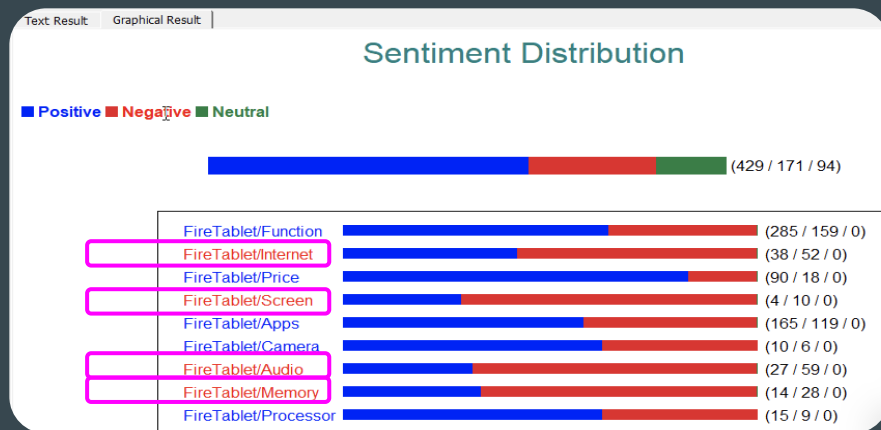
Simple Statistical Model



Text Result Graphical Result

Results for selected folder:
This directory is Positive
Number of articles:694
Number of positive articles:460
Number of negative articles:234
Number of neutral articles:0
Positive percent:66.28%.

Rule-based Model



Text Result Graphical Result

Results for selected folder:
This directory is Positive
Number of articles:694
Number of positive articles:429
Number of negative articles:171
Number of neutral articles:37
Positive percent:61.82%.

Recommendation

Internet (42.2% +ve)

It is an **ok tablet**. As I am using it, I **lose** my **page** quite often and have to go **back to** find it again. Also, I **lose internet connection** or am **unable** to **connect** at all. For the **money**, I find it **acceptable** but **not ideal**.

everything started out **great**, but then about 9 days into owning the **fire**, could **not locate wifi**, **doesn't work without wifi**

Stronger receiver

Memory (33.33% +ve)

I purchased this as a after thought. The **tablet** has **8gb** but from the moment you start to **use** it almost five is **already used by** Amazon I was totally **disappointed**. I **love** how the **screen savers** changed though it was so **clear** and **unique** but the **memory** took the cake for me!

Nothing special and **not much memory** but for \$40 it is **worth** the **price**

Extended Memory

Screen (28.57% +ve)

The **resolution** on the **games** is **not good**.

Bought it accidentally. **Meant to** get an 8 **inch screen**.

The touch **screen** is very **poor** quality. It's **hard** to move things around on it when he is playing a **game**.

Large inch LED/OLED screen

Audio (31.40% +ve)

Cheap tablet but **solid** built. **Weak audio** and **poor camera**.

Downloaded **movies** for my son to watch and the **audio** is **terrible**, you could **barely hear** it so I **returned** it

Speaker with larger amplitude waves

The End. Thank You~

Appendix - Rule-based Model

Pos Combination

[+] Concept_Rule I (SENT, “_def{FireTabletFeatures}”, “_c{TonalKeywordPositive}”)

[++] Predicate_Rule I (SENT, “_def{FireTabletFeatures}”, “_a{TonalKeywordPositive}”, “_b{TonalKeywordPositive}”)

[--] Predicate_Rule I (SENT, “_def{FireTabletFeatures}”, “_a{TonalKeywordNegative}”, “_b{TonalKeywordNegative}”)

Neg Combination

[-] Concept_Rule I (SENT, “_def{FireTabletFeatures}”, “_c{TonalKeywordNegative}”)

[-+] Predicate_Rule I (SENT, “_def{FireTabletFeatures}”, “_a{TonalKeywordNegative}”, “_b{TonalKeywordPositive}”)