Chapter 1

Topic Modelling

In this chapter, we use a topic model to find themes/topics that exist in our dataset. Our input dataset is a set of relevant tweets as determined by the classifier in the previous chapter. We use Latent Dirichlet Allocation as our topic model as described in Section **??** on page **??**.

Tables 1.1 and 1.11 on pages 5 and 19 respectively, show a list of 30 and 40 topics. It also contains their respective topic-tokens distribution. For the purpose of this study, a token is either a unigram or bigram. Each row comprises of a list of tokens that try to explain a topic and they are ordered by their level of influence. While it is helpful to have our tokens ordered by level of influence, the respective influence values are excluded from the table because we will not pay much attention to them during our analysis and evaluation.

1.1 Pre-Processing

Before running LDA on the dataset, we have to cleanup our dataset and what we do is very similar to what we did in Section **??**. In summary, we perused our tweets and removed the occurrence of special characters like new lines, all links and stop words from each tweet. Finally, we converted the tweets into features(unigrams and bigrams). For topic modelling, we perform the same operations and more. Specifically, we remove anything that does not add value to the topic model.

Firstly, we remove usernames from our data. Our topic model would try to find relationships

in tweets by using the words that occur in them. Usernames are also technically words but are semantically irrelevant for our use case. Fortunately, usernames on twitter follow a specific pattern so they can be easily removed with a regular expression. The regular expression used can be found in Appendix **??**. Secondly, our dataset is Apple centric and as a result, words like "iphone", "ipod" and "ipad" are bound to occur in almost every tweet. For this reason, we add these words to our stop words list so they are remove from every tweet. Appendix **??** has a complete list of these words.

1.2 Evaluating Topic Models

In this section, we analyse two separate models one of which comprises of 30 topics and the other, of 40 topics. They both use a mixture of unigrams and bigrams in their token distribution. This was inspired by our experiments in Section **??** where the classifier showed better performance when using a mixture of unigrams and bigrams. We analyse and evaluate a few topics from the 30 topics model and have a look at some of the tweets that fall under those topics. We the compare topics generated from the 30 topics model to that of the 40 topics model.

1.2.1 Evaluating a 30 Topics Model

Table 1.1 shows a list of 30 topics and their respective token distribution. This table can be used to get an abstract view of the topics but to get an even better view of distribution across all topics, we refer to Figure 1.1 on page 3 which is a word cloud of all tokens on our table. The frequency of a word determines its size in the cloud.

We can see words like "android" and "ios" which are mobile operating systems built by Google and Apple, respectively. We can also see words like "app", "issues", "5s","mini", "google" and "samsung" which can be in some way related to Apple. For instance, "app" might refer to applications on any of Apple's platform while "5s" could refer to the new mobile phone released by Apple around the time our data was gathered. Google and Samsung are competitors with Apple, so these words could have been gotten from tweets that compare either their products or companies as a whole.

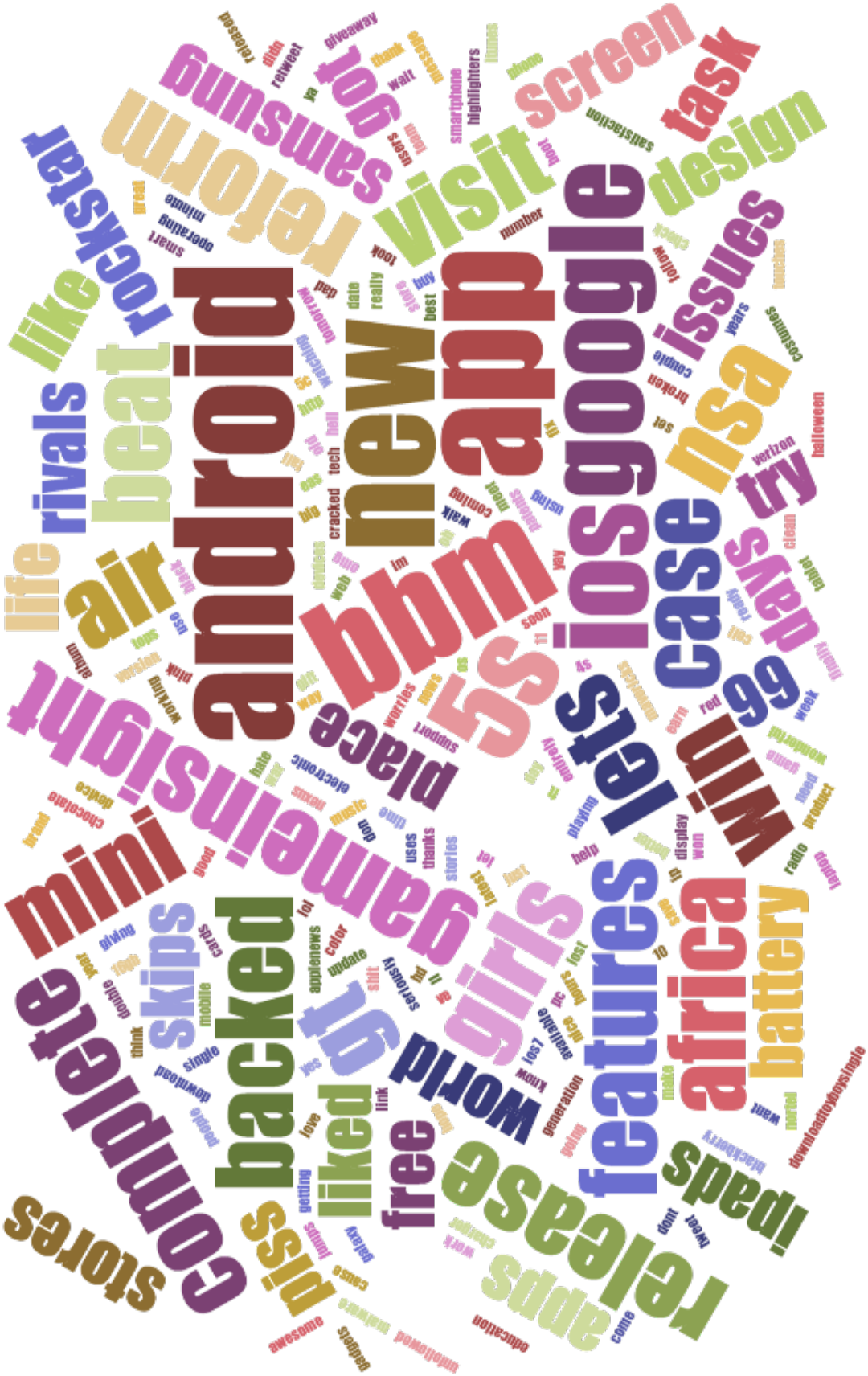

Figure 1.1: A word cloud of all tokens from all topics in our 30 topics model

To get a more detailed insight into what these words represent and what the topics represent, we analyse a few topics in detail by making a few assumptions about the topics and using the tweets to verify our assumptions.

Topic	Topic-Tokens Distribution
0	app, latest, generation, version, galaxy, won, set, oh, save, minute
1	walk, watching, unveils, today stories, beat rivals, ipads beat, missed unveils, revamped, revamped ipads, rivals
2	perfect, case, 16gb, black, gt, giving, clean, smartphone, ya, pink
3	screen, place, http, better, place visit, visit gameinsight, entirely, electronic
4	video, love, app android, yay, tomorrow, let, liked, liked video, operating, single
5	complete, follow, managed, week, having, girls, task, complete task, managed complete
6	ios, lets, bbm ios, ios lets, lets apps, features, game, tech, missed
7	google, backed, google samsung, mobile, nortel, patents, microsoft backed, rockstar, backed rockstar, uses
8	using, world, best, hate, beat, skips, skips africa, africa, africa world, release 5s
9	music, 10, mavericks, number, yes, coming, soon, earn, cards, support
10	new, app, store, available, design, playing, stargazing, stargazing app, new design
11	really, os, gift, stores, hours, news, fail, stores piss, piss, broken
12	day, battery, good, shit, display, thanks, battery life, omg, seriously, ios7
13	android, use, blackberry, update, web, 11, issues, devices, hd, fix issues
14	official, 5s, 5c, life, models, cracked, color, worries, nexus, highlighters
15	know, pc, 4s, old, does, releases, air saywhatnow, releases air, saywhatnow, did
16	time, ll, date, wonderful, charger, working, nice, message, cell, took
17	got, today, ve, going, smart, help, protector, hell, screen protector, got new
18	samsung, don, download, updated, people, radio, meet, ft, updated ios, malware jumps

19	nsa, microsoft, facebook google, google substantial, nsa surveillance, reform, reform nsa, substantial, substantial reform, surveillance
20	lol, education, chocolate, team, double, wait, boot, white girls, girls like, couple days
21	air, free, visit, power, satisfaction, 99 free, app 99, power tablet, tops, war
22	release, like, new, facebook, big, make, new 5s, gadgets, product, brand
23	just, white, app, gold, users, gold 5s, awesome, way, unfollowed, link
24	verizon, laptop, finally, say, dont, years, getting, ready, costumes, lost
25	apps, phone, check, ip, cause, thank, hope, dad, im, gt gt
26	gameinsight, halloween, try, retweet, giveaway, win, work, try gameinsight, days, tweet
27	bbm, bbm android, official release, android official, android bbm, perfect features, features bbm, want, buy, need
28	mini, win, chance, come, chance win, case mini, kickstand, kickstand case, mini models
29	itunes, released, think, year, great, touches, album, downloadtoyboysingle, didn, eas

Table 1.1: 30 topic-tokens distribution with unigrams and bigrams

1.2.1.1 Topic 6

ios, lets, bbm ios, ios lets, lets apps, features, game, tech, missed

The most common theme in the above distribution is "ios" and "lets". "bbm" also seems to have a considerate amount of relevancy as it is the third most relevant token. The other tokens seem random but we should be able to explain them better after taking a look at some of the tweets with a fair proportion of this topic.

No	Proportion	Tweet
1	81%	Fantastical 2 for iPhone gets bold iOS 7 redesign, many new features #tech
2	75%	Limbo for iOS is now even cheaper at $0.99
3	80%	Don't miss out on loads of great iOS game sales from...
4	70%	What new features does Apple's iOS 7 boast?
5	63%	get the BBM on iPhone iOS, lets get the apps here now
6	89%	time guessing is hard! test yourself on ur iPhone it's #cool tweet your results from the app #game #ios #app

Table 1.2: Tweets classified under topic 6

The tweets in Table 1.2 all seem to have at least 63% of topic 6. While they might look like carefully selected tweets, they were actually chosen at random. Our dataset contains a large fraction of the third, fourth and fifth tweet. Approximately 90% of them are retweets which explains why our topic model extracted "ios lets", "lets apps", "game" and "tech" as relevant tokens. The sixth tweet arguably does not really have anything to do with "iOS" but because it has been tagged with three words that our topic model finds salient, it gets tagged as having 89% proportion of topic 6.

1.2.1.2 Topic 7

google, backed, google samsung, mobile, nortel, patents, microsoft backed, rockstar, backed rockstar, uses

From a simple scan through the tokens in this topic, we can postulate that the tweets in this topic will mostly be about Google, Samsung, Nortel and the Rockstar Consortium. Nortel was a communications and networking equipment manufacturer that went bankrupt and the Rockstar Consortium was formed to negotiate licensing for patents they owned. However, we do not expect our whole dataset to be about patent war between these companies. Samsung and Google are large technology companies and we might encounter tweets that simple compare their products with that of Apple. We can make this assumption because we know our dataset is Apple-centric.

No	Proportion	Tweet
1	86%	Apple files patent for slim solar-powered technology via GigaOM
2	78%	Google, Samsung, and others sued over search patents by Apple-backed Nortel group
3	78%	Apple, Microsoft-Backed Rockstar Consortium Sues Google, Samsung Over 7
4	83%	Apple, Microsoft-backed Rockstar uses Nortel patents to sue Google, Samsung and others
5	78%	Google Fiber comes to iPhone, iPod touch with DVR functions
6	80%	Google Replacing Android ID With Advertising ID Similar To Apple's IDFA
7	68%	Nexus 4 will get the updated "in the coming weeks". If Apple can offer to update the (almost) entire install base on day 1, why not Google?
8	88%	Google smartwatch: Will it be an "iPhone" moment for wearables?
9	84%	I wonder who's richer Google or Apple?!
10	90%	Apple earned more than Samsung, LG, Nokia, Huawei, Lenovo & Motorola's mobile shipments combined

Table 1.3: Tweets classified under topic 7

Table 1.3 is a list of 10 randomly selected tweets with a fair proportion of our topic. The first four tweets are about patents and lawsuits and our dataset contains a number of variations of those tweets, each of which have been retweeted many times. The fifth, sixth and seventh tweets all talk about products by Google while the last three tweets compare Apple's products to that of other companies.

1.2.1.3 Topic 9

music, 10, mavericks, number, yes, coming, soon, earn, cards, support

This topic, compared to previous topics, is a little tougher to analyse. “music” is th most salient token in the distribution but we also have unrelated tokens like “mavericks”, “cards”, “earn” and “support”. While they might actually be closely related, it is not very obvious by merely looking at the topic-token distribution. To get a better understanding of this topic, we take a look at some tweets with a fair proportion of the topic.

No	Proportion	Tweet
1	63%	Just updated my mac - in loveee #Mavericks
2	83%	nahhhh, save up some cash to buy that freaking iPhone 6 thats coming out soon :)
3	77%	iPhone Battery Always Running Low? 10 Tips To Prolong The Battery Life.
4	70%	I wish I could type my mood into my iPhone and it would make a playlist for me.
5	67%	I would like to tag items on my iPhone and then be able to search tags, like in Mavericks.
6	75%	Apple needs to make a iPhone thats bigger than 64gb! My music has just about filled mine
7	70%	iPhone has to let me record videos while music playing on my phone. LET ME BE GREAT APPLE!!!
8	76%	if you have an iPhone you can block the number
9	58%	My phone just erased everybody messages number with an iPhone.
10	78%	Earn gift cards, flier miles and more with Perk - download for iphone now!

Table 1.4: Tweets classified under topic 9

Apple released a new operating system called Mavericks around the time our dataset was gathered which is what the first and fifth tweets are about in Table 1.4 and it also explains what the token “mavericks” means. The second tweet talks about the arrival of an iPhone 6 but our dataset only contains less than 10 tweets with the word “coming”. Tweets 4, 6 and

7 all talk about music and our dataset contains a large amount of those type of tweets while tweets 8 and 9 talk about mobile numbers. Without having to dig deeper, it is obvious that our topic is not well formed as it is a combination of multiple topics that do not complement each other.

1.2.1.4 Topic 12

day, battery, good, shit, display, thanks, battery life, omg, seriously, ios7

The most salient token we have is "day" which does not mean much to us. Our token-distribution also seems to have a number of adverbs, adjectives and slang like "good", "seriously" and "omg". These all seem to represent sentiments towards a certain topic. Ignoring them, we are left with "thanks", "battery" and "battery life". At this point, we could postulate that most of the tweets in this category will have a "battery" theme. To confirm this, we take a look at some of the tweets with a fair proportion of the topic.

No	Proportion	Tweet
1	81%	iPhone battery just went from 23% to 3% in the space of five minutes. Thanks again, iOS7.
2	76%	iPhone battery is so crap
3	80%	iPhone battery dropping in 5% increments. Can't be a good sign.
4	64%	Considering the amount of times I have to plug my IPhone in to charge a day it might as well be a fucking landline
5	50%	FORGOT MY IPHONE CHARGER oh shit man. My poor battery :(
6	66%	Apple was considering making an iPod for kids but apparently, the name iTouch Kids didn't sit too well.
7	50%	so your apple store doesn't get stock on release day?
8	50%	can I have an iphone with bbm and the battery life of a nokia please

Table 1.5: Tweets classified under topic 12

All tweets(except the 7th) in Table 1.5 refer to the iPhone battery life. A very large number of tweets that have a good proportion of this topic take in some way, the shape of tweets 1–4 in our table and fortunately, our topic model is able to find the relationship between these tweets.

The seventh tweet looks out of place as it has nothing to do with battery life but going back to our topic-token distribution, the most salient token as described by the model is "day". Luckily, our dataset also contains a large number of that tweet and this is because that particular tweet was retweeted a lot of times.

1.2.1.5 Topic 13

android, use, blackberry, update, web, 11, issues, devices, hd, fix issues

At first glance, the main themes that stand out in the above distribution are "android", "blackberry" and "issues". Our dataset is Apple-centric so we could assume that the tweets with a large proportion of this topic will have some sort of comparison between Apple, Android and Blackberry with respect to issues that occur with their products. If this is not the case, it is possible that our topic model has merged two different topics into one topic. To verify our assumption, we analyse a few tweets that have a reasonable proportion of this topic.

No	Proportion	Tweet
1	51%	Wow hello typos. . . cracked iphone screen problems
2	50%	Check out WhatsApp Messenger for BlackBerry, Android, iPhone, Nokia and Windows Phone. Download it today from. . .
3	50%	Pandora finally comes to Chromecast via Android and iPhone apps
4	80%	I just connected with friends on #BBM. Invite your BlackBerry, Android and iPhone friends at. . .
5	80%	Develop iPhone, Ipad and Android Apps creatively with Mawaqaa.
6	75%	Was curious if that was the culprit. I have had tons of issues since upgrading my Apple devices to latest
7	57%	Apple testing Mail update for OS X Mavericks to fix several issues
8	78%	Manufacturing issue causing battery problems in some iPhone 5s devices

Table 1.6: Tweets classified under topic 13

Tweets 2–5 in Table 1.6 all have either an android or blackberry theme in them which is expected. They all talk about applications on all three platforms. On the other hand, tweets 1, 6, 7 and 8 all have an "issues" theme in them. Unfortunately, these two topics do not really complement each other and should arguably be splitted into two separate topics. Some of the latter mentioned tweets do also talk about applications on Apple's platforms which may be the reason why our topic model observed a relationship between these topics.

1.2.1.6 Topic 14

official, 5s, 5c, life, models, cracked, color, worries, nexus, highlighters

An initial scan of our token distribution above does not tell us that this topic might be mostly about the 5s and 5c. In September 2013, a month before our data set was gathered, Apple released two models of its mobile phone and they were called iPhone 5S and iPhone 5C. With this knowledge, we could hypothesise that tweets with a large proportion of this topic will

mostly be about these new phones. We can also rely on the fact that our distribution contains tokens like “5s” and “5c”. Tokens like “models”, “cracked” and “worries” could also be used to describe a state of the phones.

No	Proportion	Tweet
1	88%	Cracked iPhone No worries. Color it in with highlighters!
2	68%	Apple discovers manufacturing defect causing iPhone 5S battery woes for some customers
3	88%	#Apple Admits Defect with Some #iPhone 5s Batteries
4	33%	Everyone who bought the iPhone 5S or iPhone 5C is dumb. The iPhone 6 has already been announced lol.
5	76%	iPhone 5S, 5C debut in India today - Customers get the new models in the price range of Rs 41,900 to Rs 71,500
6	88%	$AAPL Apple’s yellow iPhone 5C is a lemon
7	51%	Well the battery life on the iPhone 5c is great. . . I need a charger in every room
8	22%	What Are The Most Popular iPhone 5s and 5c Colors? Space Gray And Blue.

Table 1.7: Tweets classified under topic 14

The first tweet in Table 1.7 does not really refer to the iphone 5S or 5C but is genrally about the iPhone which is acceptable. Tweets 2–8 however are all about either the 5S or 5C. They each also use terms like “color” and “models” to describe the phone in some way. Unfortunately, our topic model has tagged the last tweet with only 22% of our topic which is unexpected because the tweet does actually talk about both the 5S and 5C models.

1.2.1.7 Topic 22

release, like, new, facebook, big, make, new 5s, gadgets, product, brand

After a first scan, it is not very clear what this topic is really about. Our most salient token is “release” and in combination with other tokens like “new”, “new 5s” and “product”. We could assume that this topic could be mainly about the new release of devices. However, we also have tokens like “facebook”, “like”, “big” and “make” which we cannot really explain. We could have a mixture of topics or just a poorly formed one.

No	Proportion	Tweet
1	75%	Retina Display iPad mini 2 Release Date Tipped By Target’s Online Product Page
2	74%	Ubisoft Releases “Rabbids Big Bang” for iOS!
3	75%	Do you fancy a brand new #iPhone 5s? Like the #busuu Facebook page for your chance to #win!
4	67%	I’m not a big fan of the screen ratio of the iPad mini. Also the Nexus 5 choice is easy now, all gone.
5	84%	my iphone fell while I was tweeting and I stepped on it then i heard a big crack, I paused for 5 mins and prayed it was ok
6	80%	I have a major craving to make a Loki/Sigyn video. Blah, stupid dead back light on my Mac.
7	79%	like my #OpenTouch #OTC #iPhone #App from @ALUEnterprise to see the phone presence before I make a call
8	70%	iBoobies case for Apple iPhone 4 - make your phone even sexier

Table 1.8: Tweets classified under topic 22

From Table 1.8, the first two tweets do refer to the release of a product and mobile aplication, respectively. Unfortunately, other tweets have nothing to do with products/application release. The third tweet is a promotional tweet, the fourth is about devices and its features, the fifth and sixth refer to features of the iPhone and Mac(Apple’s laptop). Without going any further, it is fairly clear that these tweets are not related. It is possible that our topic model has incorrectly merged multiple topics.

1.2.1.8 Topic 27

bbm, bbm android, official release, android official, android bbm, perfect features, features bbm, want, buy, need

At first glance, we could say this topic has a lot to do with *android*, *bbm*, and *features*. The last three tokens also seem out of place. At the time of data gathering, there was a lot of chatter on social media about the BlackBerry Messenger(BBM) application coming to the iOS and android platform. To be certain of this, let's take a look at some of the tweets that have a fair proportion of this topic.

No	Proportion	Tweet
1	56%	More perfect features, BBM android, BBM iPhone
2	50%	BBM on android and iPhone, official release - get it here
3	50%	BBM Now on Android and iPhone.
4	72%	Just got bbm chat for the iPhone, feel free to add me if you want :)
5	60%	Apple should create the option of removing yourself from a group chat
6	68%	Anyone want to buy a black 64GB ipad 2 from me in excellent condition?
7	25%	someone buy me an iphone ugh
8	76%	I need a iphone 5 asap

Table 1.9: Tweets classified under topic 27

Table 1.9 is a list of tweets that fall under topic 27. We can see that the first four tweets have a lot to do with the new BBM for iOS and Android. The fifth tweet is a little tricky as it says nothing about BBM. However, it does in fact talk about a chat application which is what BBM is. While the user might not have been referring to BBM in particular, our model was able to pickup on the relationship between both subjects.

The last three tweets in our table explain the last three tokens in out topic-word distribution for topic 27. This means that topic 27 is actually a combination of two different topics.

1.2.1.9 Topic 28

mini, win, chance, come, chance win, case mini, kickstand, kickstand case, mini models

There are a lot of promotions/giveaways on Twitter that involve Apple products. Tokens like "win", "chance" and "chance win" in our distribution tell us that this topic is about these promotions. Our dataset is Apple-centric so we could hypothesise that tweets with a large proportion of this topic might be offering users a chance to win Apple products like the iPad/Mac Mini, hence the "mini" in our distribution. We also have "case" occurring in the distribution which might refer to iPad cases up for promotion. To confirm that our hypothesis is valid, or not, we analyse some tweets with a fair proportion of this topic.

No	Proportion	Tweet
1	75%	Win an $800 Mac Mini for FREE from MacTrast the perfect addition to any home or office!
2	82%	RT to WIN! - #Win an iPad Mini to celebrate the start of #50at50
3	56%	15,000 Facebook Fan Giveaway happening now - Win a Lens, iPad Mini, $500 Amazon gift card and LOTS more! #colorvale15k
4	94%	WIN an iPad mini plus a chance to win a Williamson Tea Elephant Tea Caddie
5	67%	Win an #ipad follow and RT for a chance to win
6	66%	Targus Kickstand Case for Apple iPad Mini all models - Red
7	86%	Cooper Dynamo Apple iPad Mini Kids Play Case review
8	78%	Fab Purse Moschino IPhone Cases. Come in Lots of Colours
9	50%	Retina iPad Mini may be launched Nov. 21
10	20%	iMore – iMore show 373: iPad Air and Retina iPad mini buyers guide

Table 1.10: Tweets classified under topic 28

From Table 1.10, we can tell that Tweets 1–5 are all promotional tweets offering users a chance to win Apple products like the iPad mini. 60% of the tweets with a fair proportion of this topic

are all variations of those tweets. Tweets 6–8 all refer to iPad and iPhone cases and contrary to what we previously assumed, these cases are not part of the promotion. This means that our topic model has merged two topics that do not complement each other. The last two tweets also have nothing to do with the promotions as well as the cases. These tweets are general tweets about the iPad mini. Fortunately, our topic model has tagged them with low proportions(50% and 20% respectively) which is acceptable.

1.2.2 Evaluating a 40 Topics Model

In Section 1.2.1, we analysed a topic model that generated 30 topics from our dataset. In this section, we analyse a model that generates 40 topics from our dataset. We use the same dataset for both models so we expect a few overlapping topics. We briefly look at these topics and then we look at a few new topics.

Table 1.11 on page 19 shows a list of 40 topics and their respective token distribution. This table can be used to analyse all topics abstractly but to have a general overview of all tokens in our table, we use a word cloud as shown in Figure 1.2 on page 20. In our cloud, we see prominent words like “app”, “bbm”, “google”, “device” amongst others which also appears in our 30 topic word cloud on page 3.

Most of the tokens used in the 40 topics model also appear in the 30 topics model and as a result, it is difficult to draw a sane comparison between these models from either the word clouds or tables. For this reason, we take a more detailed look at each topic and it’s token distribution. We attempt to find similar topics in the 30 topics model and analyse some new topics.

Topic	Topic-Words associations
0	big, year, chocolate, mobile, gadgets, double, boot, girls like, white girls, touches
1	app, store, know, available, does, design, stargazing, stargazing app, ft, new design
2	gift, version, 25, liked, liked video, months, earn, watch, gift cards, knew
3	follow, ip, place, walk, laptop, electronic, web, dont, competition, costumes
4	using, hate, tweet, using app, reason, bout, thing, unfollowed using, case like, girl

5	white, education, giveaway, tablet, comes, brand, brand new, way, halloween giveaway, win plink
6	visit, place visit, visit gameinsight, ipads, place, unveils, power, rivals, stories case
7	video, download, 16gb, black, watching, nowplaying, clean, smartphone, eyes, sprint
8	great, giving, yay, miss, post, feeling, turns, maps, keyboard, canine evil
9	number, tfbjp teamfairyrose, retweet tfbjp, sougofollow, teamfairyrose, autofollow, 06, cards, 90sbabyfollowtrain, interesting
10	ios, official, 5s, like, world, game, facebook, new 5s, cool, africa
11	new, gameinsight, try, updated, try gameinsight, achievement, new achievement, jobs, achievement 10, areas
12	android, need, use, blackberry, devices, hd, problems, card, art, issue
13	bbm, bbm android, official release, android official, got, want, gold, macbook, gold 5s, smart
14	apps, lets, bbm ios, ios lets, lets apps, phone, don, best, people, meet
15	missed, having, app android, case missed, windows, 12, daily, cell, apps android, decided
16	perfect, android bbm, perfect features, features bbm, love, collection, pink, did, let, website
17	samsung, google samsung, backed, entirely, nortel, patents, share, beat, uses, really entirely
18	mini, win, chance, chance win, models red, kickstand case, mini models, targus, targus kickstand, case mini
19	air, http, better, young, news, mirror, market, mirror world, souvenirs mirror, envy complete
20	retina, device, red, electronic device, device using, ur, inch, 13, fix, wallet
21	pc, life, old, releases, saywhatnow, releases air, air saywhatnow, cases, away, battery life
22	check, music, 10, mavericks, generation, os mavericks, ready, fix issues, mail, mail update
23	case, buy, cover, 99, end, 2013, case 4s, case cover, smart cover, cover case

24	haha, isn, line, allowed, oh, solar, price, android phone, guy, basically
25	app, just, radio, updated ios, users, young hitta, yhr app, yhr, radio yhr, hitta radio
26	os, stores, playing, latest, hours, fail, piss, stores piss, issues, protector
27	release, halloween, 5c, work, models, days, online, product, color, pics
28	battery, lol, good, shit, thanks, tomorrow, ios7, minutes, minute, november
29	itunes, released, think, thank, album, downloadtoyboysingle, itunes link, downloadtoyboysingle itunes, saw, sister
31	ll, charger, didn, personal, later, nice, brother, movies, butt, butt away
32	features, tech, make, cracked, worries, color highlighters, cracked worries, highlighters, worries color, say
33	complete, 4s, , managed, week, display, girls, task, managed complete, complete task
34	free, enter, win, 99 free, app 99, iphone5s, navigation, comp, fucking, network
35	really, update, broken, damn, 11, family, seriously, point, message, took
36	time, retweet, date, wonderful, hell, edtech, learn, screwed, winners chosen, chosen tonight
36	today, ve, going, verizon, couple, couple days, team couple, dressed, came, online store
37	screen, finally, look, years, getting, weekend, remote, siri, finally got, ad redesigned
38	google, nsa, microsoft, reform, reform nsa, substantial, substantial reform, surveillance, facebook google, nsa surveillance
39	day, easy, link, older, candy, browser, os browser, otterbox, defender, otterbox defender

Table 1.11: 40 word-topic distributions with unigrams and bigrams

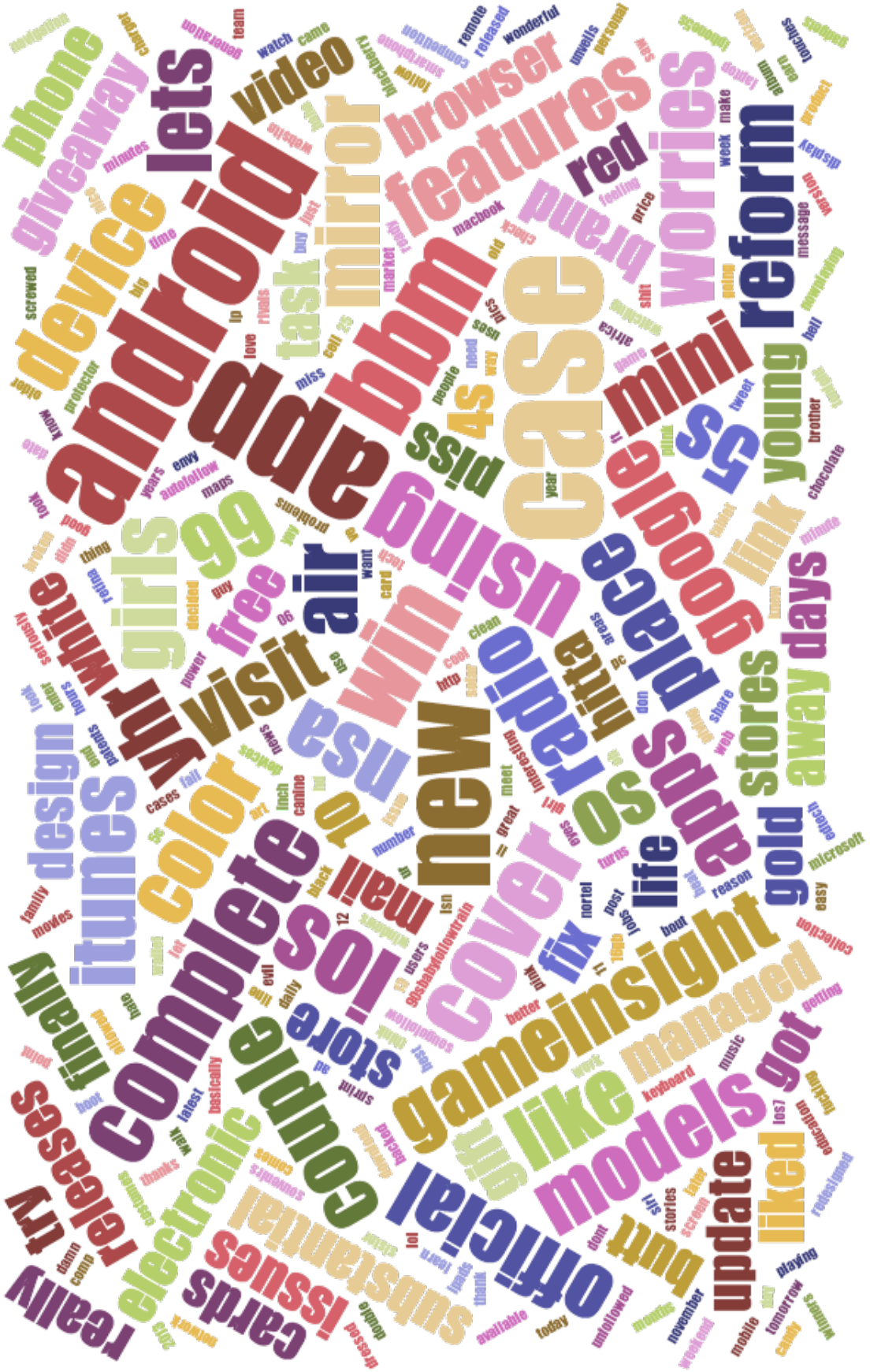

Figure 1.2: A word cloud of all tokens from all topics in our 40 topics model

No	30 topics id	40 topics id
1	6	14
2	7	17
3	9	22
4	12	28
5	19	38
6	27	13
7	28	18

Table 1.12: List of similar topics from our 30 topics and 40 topics model

1.2.2.1 Analysing Similar Topics

As previously mentioned, we expect our 40 topics model to have some similar topics with the 30 topics model. Fortunately, there are a few similar topics and Table 1.12 is a list of such topics. Each row contains a topic id from our 30 topics model and its corresponding similar topic from our 40 topics model. The ids correspond to the "Topic" attribute on Tables 1.1 and 1.11. Both tables can also be found on Pages 5 and 19 respectively.

In other to confirm that these topics are actually similar, we analyse the token distribution for some of the topics and also compare tweets with a fair proportion of the respective topics. Specifically, we analyse rows 2 and 6 from Table 1.12

Row 2

1 → *google, backed, google samsung, mobile, nortel, patents, microsoft backed, rockstar, backed rockstar, uses*
2 → *samsung, google samsung, backed, entirely, nortel, patents, share, beat, uses, really entirely*

We can infer that both distributions above are trying to explain the same topic because they both use tokens like "google", "samsung", "patents", and "nortel" to explain their topics. With this knowledge, we can also expect our 40 topics model should have tagged the same tweets with this topic.

No	Proportion	Tweet
1	76%	Apple Continues To Lose Tablet Market Share But Should Rebound With The Air
2	72%	iPad market share dips to less than 30% while tablet market grew 7% overall
3	57%	would you suggest the IPad or galaxy note 10? What do you like best?
4	77%	Apple, Microsoft-backed Rockstar uses Nortel patents to sue Google, Samsung and others
5	77%	Google, Samsung, and others sued over search patents by Apple-backed Nortel group
6	80%	Apple, Microsoft-Backed Rockstar Consortium Sues Google, Samsung Over 7

Table 1.13: Tweets classified under topic 17(40 topics model)

Comparing Table 1.13 above and Table 1.3 on page 8, we can see that both models have very similar tweets. The former seems to have a combination of two topics and so does the latter. Section 1.2.1.2 gives a more detailed explanation of what the common tweets represent.

Row 6

1 → *bbm, bbm android, official release, android official, android bbm, perfect features, features bbm, want, buy, need*
2 → *bbm, bbm android, official release, android official, got, want, gold, macbook, gold 5s, smart*

After an initial scan, we could infer that the above distributions are both trying to explain the same topic. They both have their first four salient topics in common, in the same order. This increases the chances of our assumption being valid. To confirm the validity of our assumption, we analyse the tweets categorised under this topic by our 40 topics model.

No	Proportion	Tweet
1	21%	More perfect features, BBM android, BBM iPhone
2	50%	BBM on android and iPhone, official release - get it here
3	65%	I just connected with friends on #BBM. Invite your BlackBerry, Android and iPhone friends
4	47%	BBM. Now on Android and iPhone.
5	82%	Just got bbm chat for the iPhone, feel free to add me if you want
6	70%	i want an iPhone 5
7	32%	white girls be like: i want a iPhone 5c with THAT case!!
8	60%	I want the gold iPhone 5s so bad

Table 1.14: Tweets classified under topic 13(40 topics model)

Comparing Table 1.14 above and Table 1.9 on page 15, we can see that both models have very similar tweets. It also tells us that the 40 topics model merged two topics as our 30 topics model did. Unfortunately, the second topics are not similar but this is acceptable because both models are slightly different which makes a 100% similarity highly unlikely. Section 1.2.1.8 gives a more detailed explanation of what the common tweets represent.