Joseph Hartman

IST-736

03/21/2019

Final Project: Analyzing Facebook Posts to Optimize Reaction

**Introduction**

Whether it be for a personal brand or a company, being able to understand one’s audience for social media is very important in order to market oneself and draw huge viewership numbers. In the social media marketing sphere, a big emphasis is on how to draw more “likes”, “comment”, and/or “retweets” to one’s posts. High viewership posts are targeted by the internal algorithms of social media sites to be featured on a network’s news feed, so being able to create posts that can have this draw on a consistent basis is important.

For this project, we’re going to analyze 100 posts from the closed Facebook group, Subtle Asian Dating. Subtle Asian Dating is a sort of dating group, where users can “auction” off their friends to prospective partners. Usually posts are humorous, somewhat long, with lots of emoticons (somewhat reflective of the age group that are the primary users 18-26). Each “auction” usually consists of the following:

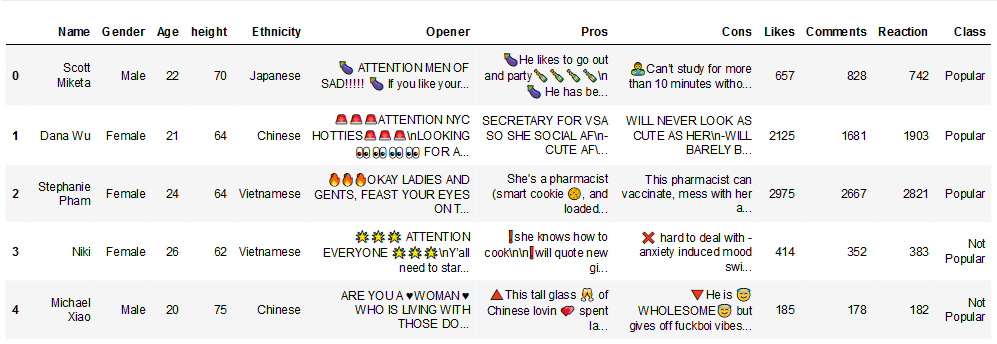
* An opener to grab attention, an example would be “🚨🚨🚨ATTENTION NYC HOTTIES🚨🚨🚨”, followed by some more explanatory texts.
* The auctioned person’s demographic info (name, age, location, ethnicity, etc.)
* A list of positives attributes or “pros” that prospective partners might find attractive
* A list of negative attributes or “cons” that prospective partners might not find attractive. However, these cons are often listed sarcastically or are just positives listed in a sarcastic way. An example would be on an auctioned male’s post, “Probably has a longer skin care routine than you.”
* (Optional): A link to the auctioned person’s Instagram profile or to the person that wrote the post, often called a “shameless plug”.

From this data we look to answer two questions:

* What are common topics that we see across all of these "auction" posts? What interests or qualities do people deem to be desirable?
* Can we use informative features from an SVM classifier to determine words or language choices to determine what distinguishes popular posts from not popular posts?

**Method**

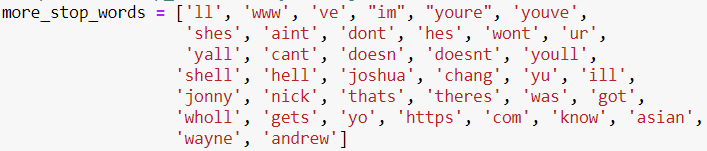
The posts were manually scraped by a friend and I and then broken up into a table that could be exported into a Pandas data frame. The main text body of the posts were broken up into their own attributes (Opener, Pros, and Cons). The number of likes and comments that each post received was also recorded. Since the numbers of likes and comments were relatively consistent across posts, an average of the two were taken to derive a reaction score from which a class label could be created. In this case, posts with a reaction score above the median were classified as “Popular”, while those below the median were classified as “Not Popular”.



**Figure 1: The first 5 lines of posts that have been organized into the dataframe, note that the body of the text has been broken up.**

An overview of our text mining process now that the data has been loaded:

1. Extract and remove the emojis from the text so that analysis can be run on the emojis
2. Vectorize our data using TF-IDF vectorization, the posts are relatively long so TF-IDF seems like a good choice to make sure those words that can distinguish between “Popular” and “Not Popular” have more weight. Some parameters during this process to note:
   1. Lowercasing and setting a minimum document frequency to 2 was done for vocabulary reduction purposes and to remove obscure words.
   2. A custom stopword list was created to remove web terms, names, and certain slang that’ll appear in our informative features. The process of adding to the stopword list was an iterative process where the output of our models was shown and then words that had no intuitive value were added to the stopword list. This custom stopword lsit was added to the built in list from sci-kit learn.



**Figure 2: The custom stopword list added to sci-kit learn’s built in stopword list.**

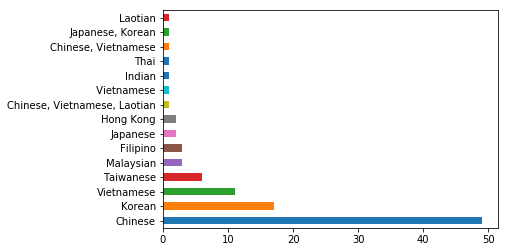
* 1. Unigrams, bigrams, and trigrams were used. This was done to hopefully capture more of that sarcasm that can’t be captured with unigrams.
  2. Emojis were removed. We tried LDA models that used the name representations of emojis included into the dataset but it threw off the models and nothing really insightful could be gained. Instead, these emojis were extracted and then counted to look for frequency of emojis overall.

1. A total of 6 LDA models were trained and fitted over the whole dataset for anyone that was labeled “Popular” (3 for male Opener, Pros, and Cons, and 3 for female Opener, Pros, and Cons).
2. For SVM classification, two different vectorizations were used: one with unigrams and one with unigrams, bigrams, and trigrams. The rest of the vocabulary reduction from LDA modelling were kept the same.
3. A 60/40 train test split was done for hold-out testing
4. A total of 6 models were trained (a unigram SVM for Opener, Pros, and Cons, and an N-gram SVM for Opener, Pros, and Cons)
5. Models were tested, a confusion matrix and Precision, Recall, and F1-score report was outputted for each of the 6 models
6. Informative features were outputted for each of the models to look for words or N-grams that distinguish “Popular” from “Not Popular” posts.

Note all relevant code will be included separately in a IPython notebook

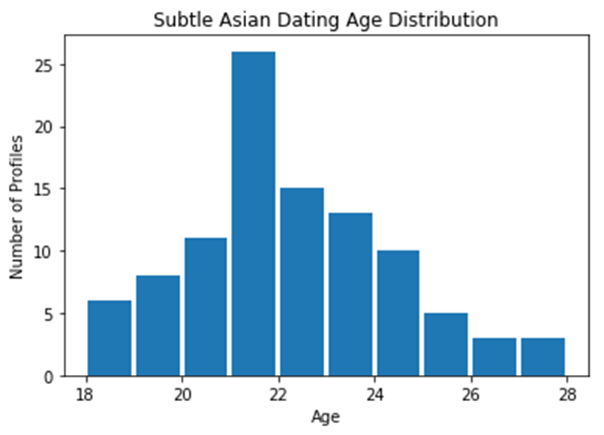
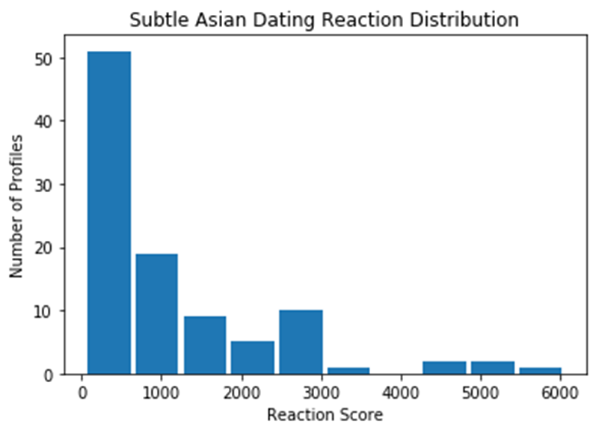
**Results**

**Exploratory**



**Figure 3: Breakdowns of Gender and Ethnicity**

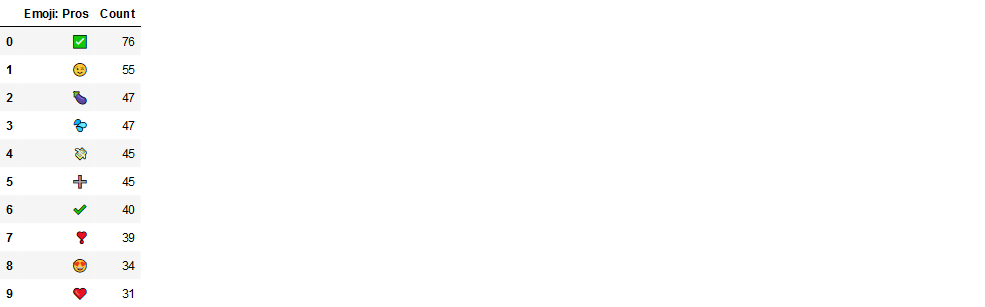
We see that among our sample, there are slightly more male than female users, though it’s not very unbalanced. If we breakdown, our users by ethnicity, we see that almost 50% of our users are Chinese or Chinese Americans, followed by Korean or Korean Americans, and Vietnamese or Vietnamese Americans. This order somewhat matches with population estimates from the Pew Research Center, though in their case, Taiwanese residents were grouped with Chinese Residents. Knowing who the major participants are might give us more insight when we run our topic models and extract our informative features.



**Figure 4: Breakdowns of Reactions and Age**

Again, we see that the primary age group is in the range of 18-26, with the distribution centered around 21-22. In terms of reactions we see it highly skewed to the left, hence why we chose to use the median as our splitting point instead of the mean.

The counts of emojis show trends we expect:



**Figure 5: Counts of Emojis for Openers, Pros, and Cons**

Openers tend to include items that grab attention, such as the “fire”, “double exclamation”, “or alarm emoji”. Pros include indications of positive emojis like the “checkmark”, “plus”, and variations of “heart” emojis, while cons include list of “cross”, “minus”, and the “poo” emoji. There is some carryover between lists such as the “caution sign” emoji or the “eggplant” emoji. In the case of the Pros and Cons emojis, these are often used as the bullet points for the posts, hence why they have higher count frequencies.

**LDA – Topic Modeling**

Topic models were broken down into analyzing auctioned male posts and auctioned female posts. The intuition behind the decision is that males and females value different things in their partner and the attributes they highlight would be different. Topic numbers were chosen mostly based on intuition and if they made the most sense in terms of the breakdown.

For Male Openers (number of topics = 3):

1. *Wholesome Boy:* looking attention ladies man oppa need handsome wholesome new attention ladies life boi tired young korean like home make years 20
2. *University Bachelor:* like georgetown look university georgetown university looking single girls wasian georgetown university right buy drinks bachelor peter aka buy drinks year namjas right like like korean namjas namjas right old studying government journalism wasian georgetown
3. *Postgrad Professional:* boy hello heart come big ny york wanna girls attention studying friend young year 23 comfort white new york single sad share

For Female Openers (number of topics = 3):

1. *Wholesome Girl:* nyc wholesome girl cause cuff really season make attention need ass boys don calling want af 22 time bae cute
2. *Other Wholesome Girl:* time wholesome look sad looking friend attention girl little present make good time beautiful ya guy bois introducing really stop covered
3. *Free Spirited Woman:* tag hot looking okay post bitch wholesome bay area area bad bitch don winter just originally friends love tag single year free lookin

Based off the opener we can sort of see what kind of archetypes sort of appear on the posts. For auctioned woman, it seems that they either fall into wholesome or free-spirited archetypes, where men fall into ­a more amorphous category but seem to be distinguished between whether or not they’re still a student at a university. Also, interesting thing to note is that for males, we see a lot of Korean words, such as “oppa” and “namjas”. Koreans being the 2nd most populous users, are also, from personal observation, seen as being very desirable. This may be due to the popularity of Korean pop music and dramas and the aesthetic and style in those forms of media being accepted as a standard of beauty. Next, we’ll dive into a few of the topics that we see in the Pros and Cons list for each of the genders. We won’t list all of the topics as some of them aren’t easily recognizable as well as to keep this report concise.

Some examples of topics for Male Pros (number of topics = 8):

* *Talented:* play knows computer make better piano plays use business finance meals btw likes cs trilingual english ig kpop songs humble anime songs
* *Family Man:* man life family time loves wholesome school friends knows best things family man university smart cultured like good come super dental

Some examples of topics for Male Cons (number of topics = 15):

* *Subpar Appearance:* look just pretty makes big literally smaller photos stolen fall like body ugly head little mans salad enjoys bad romcoms leaving younger
* *Bad Food:* mcdonalds younger smaller flow free drink instagram eats 60 eat money everyday drunk boba actually like probably soon photos nonstop week

Some examples of topics for Female Pros (number of topics = 8):

* *Good looks:* eat good want need loves gym great wholesome order amazing stay easy talk humor super face best sense model looks dance
* *Personality:* gonna good pretty smart sense actually probs sweet guess type loves personality humor sense humor kpop game love meals lookin eater

Some examples of topics for Female Cons (number of topics = 15):

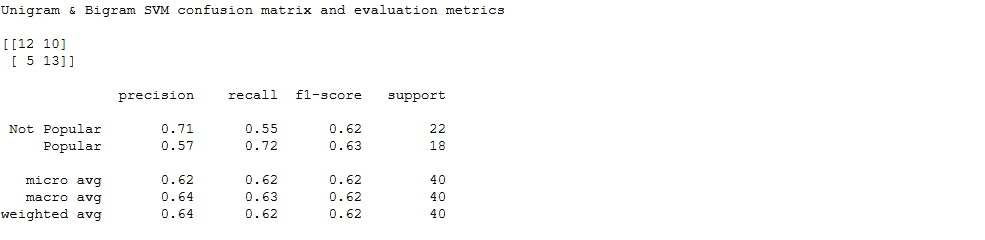
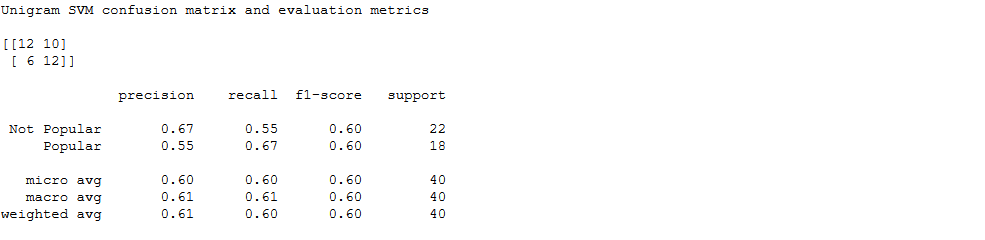
* *Shallow:* ig make clout handle 10 high like bc week big game hoodies saying gram pics gram loves lol girl lot probably
* *Partier:* ass gotta free drink literally does dms week eating bc handle game 10 rave people guys pics better lift best

Distinguishing topics was rather difficult in these sections as there were many words that appeared in multiple topics despite changing the number of topics parameters. We can see, in general, what users are interested in, such as when we see the word “raves”, “club”, and “dance” we can assume that the topic is about dancing/clubbing. However, words like “food” will also be intermixed into the topic. A future implementation I would like to use would be to find something that can use the ll/token method for finding the best number of topics for sci-kit learn.

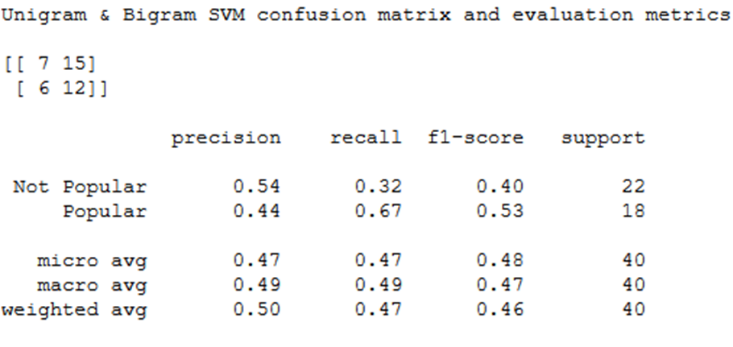
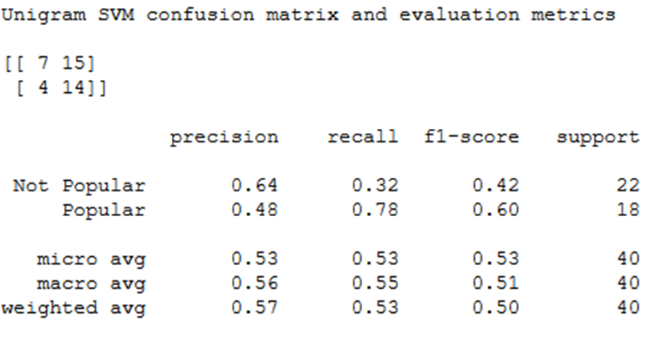
**Linear SVM Classsification**

(Note in the reports it’s “Unigram & Bigram SVM Confusion matrix…” where it’s supposed to say “N-gram Confusion Matrix…”)

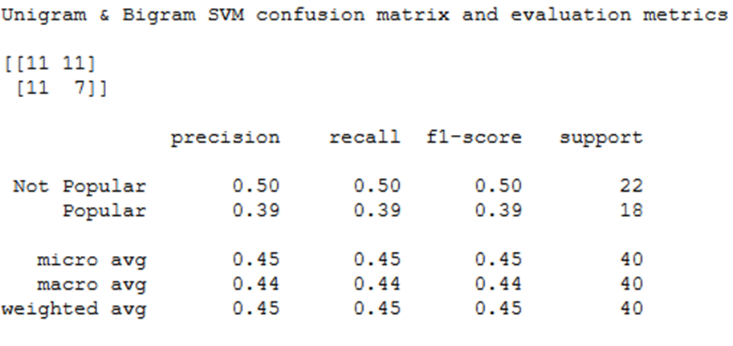
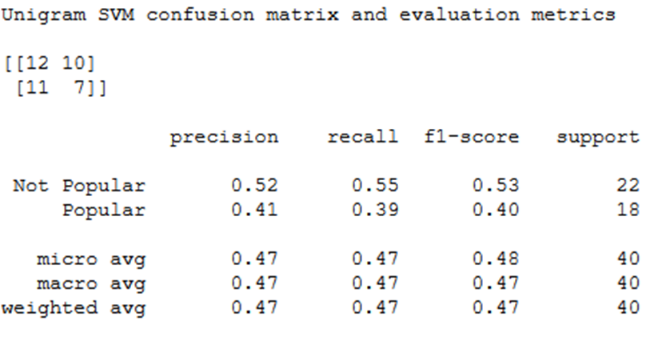
Results from our Linear SVM opener where C= 0.8 for our unigram model and C = 1 for on N-gram model:



Results from our Linear SVM Pros where C= 0.8 for our unigram model and C = 1.5 for on N-gram model:

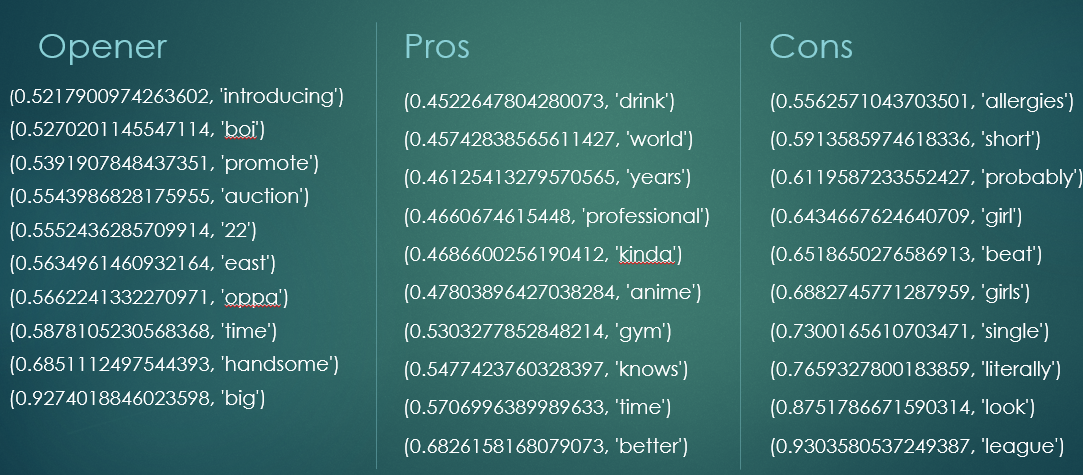


Results from our Linear SVM Pros where C= 1.5 for our unigram model and C = 2 for on N-gram model:

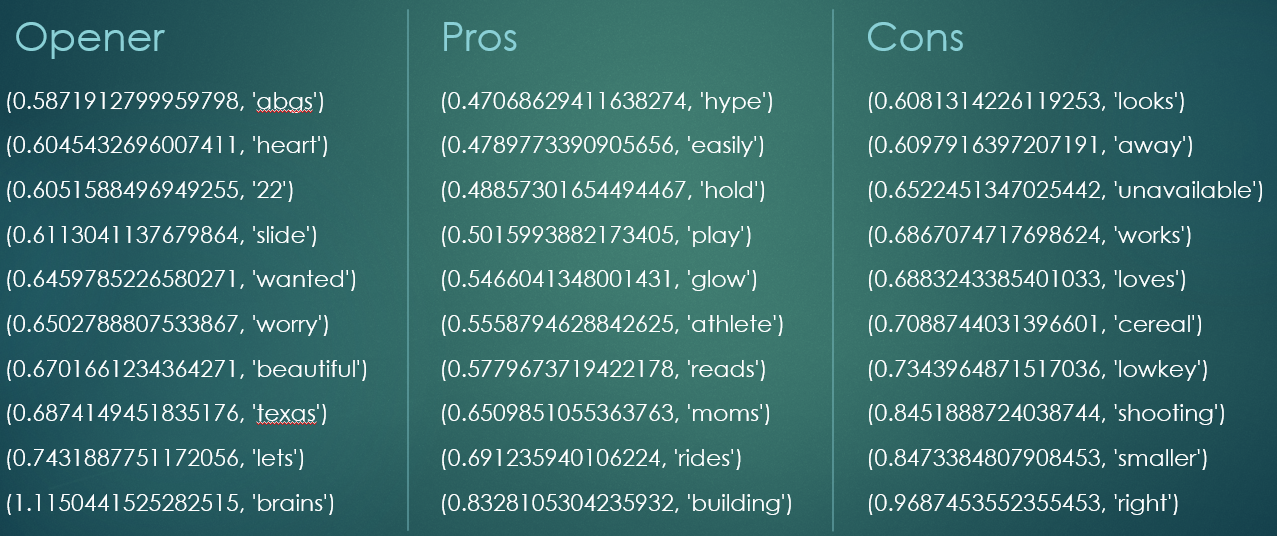


Note that the cost parameters increase as we go from “Opener” to “Pros” to “Cons”, while this does make the models less robust, the f1-scores of the models were atrocious. A possible explanation for the reasoning behind why a high cost value is needed for “Cons” is due to the introduction of sarcasm into the “Cons” list. Another interesting thing to note is that on average, the “Pros” and “Cons” list are longer than the “Opener”, yet we can build more robust models on the opener. Finally, the last thing to note is that we only have 100 examples, more examples would probably make a better model.

Finally, let’s look at the informative features for our Unigram models:



And the informative features for our N-gram models:



These are the words that are used to distinguish “Popular” from not popular. We see again some Korean words that appear like “oppa”. Also there is some slang such as “abgs” meaning “Asian baby girl” used to describe a specific style since amongst the Asian American millennial community. Overall, it seems that users are advertising the whole package on these posts. Someone who not only is a working professional and good looking but is also athletic and has a good personality.

**Conclusion**

Some Takeaways:

* While it was difficult to distinguish topics from the LDA modeling, there are common topics that appear overall. It seems that each topic is a mix of a bunch of topics. This aligns with an article from Jonathan Chang and colleagues when discussing Topic Intrusion.
* Emoji usage is very popular in these posts and it’s possible to extract them to run their own analysis on frequency. However, implementing them into our LDA and SVM modeling is rather difficult due to the nature of how emojis interact with text in a way that is atypical of standard English conventions.
* Our SVM models seem to show that users describe partners that are the “whole-package”. Though it seems that Korean auctioned people seemed more favored due to the use of Korean words. Also, it seems that the ABG aesthetic is also seen as desirable.

# References

GUSTAVO LÓPEZ, N. G. (2017, September 8). *Key facts about Asian Americans, a diverse and growing population.* Retrieved from Pew Research Center: https://www.pewresearch.org/fact-tank/2017/09/08/key-facts-about-asian-americans/

Jonathan, C., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems*.