5. Sampling:

Explain possible sampling bias through Weibo, such as gender bias, etc.

1. First of all, With 40 million active users on Weibo, 55% of them are people young people that born after 1980. Then weibo marketing may have a problem reaching out to those older customers.

2. Weibo has a lot of powerful Agent accounts may repost some weibo automatically, which could mislead your conclusion.

6. What are some possible algorithms to identify users who showed interest in Michael Kors over Kate Spade? What are some data points that can be used to illustrate the algorithm's utility? Give a couple examples and discuss pros and cons. (You do not have to code here)

Sentiment based on BAG:

1. Get and cleaned data from Weibo API or web crawling.
2. Performed Sentiment Analysis on both Michael Kors and Kate Spade

   One Naive Score method could be scoring each weibo related to their brand by :

$$\frac{num\ of\ positive\ word - number\ of\ negative\ word}{total\ words}$$

   Score>0: positive weibo

   Score<0: negative weibo

3. Then calculate the mean and STD for each brand based on the scores of Weibo. Also perform a t-test to see if their means are significantly different.

Pros: easy to implement

Cons: Bag of method to score sometimes could not be accurate, for example, “I do not like Mac book” may have a positive score because it has a “like”.

Sentiment based on Naive Bays method:

1. Manually label at least 2,000 brand-related weibo or comments by 1 if positive, 0 if negative by human being.(Training data)
2. Build a Naïve Bays classifier using Naïve Bays statistics method to predict the sentiment of future weibo.
3. Apply this classifier to all the weibo to score the data.
4. Calculate the ratio of positive weibo for each brand.

Pros: More accurate than bag of words

Cons: Manual work needed.