**5. Sampling: Explain possible sampling bias through Weibo, such as gender bias, etc.**

The most straightforward bias from weibo is the nationality, because only Asian people use weibo whereas the users in other countries seldom use this platform. So if we only sample the data from weibo, this will lead to nationality bias to the brand.

The other one is age bias. Even this kind of social network is easy for young people. However, it is a little bit complicated for our parents or even grandmother or grandfather. So we are not able to extract the information from them through weibo. And usually the people in this generation have more capacity to buy the product as compared with young people. This will lead to a serious bias for the brand.

**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.**

For this problem, if we only calculate the total brand-mentioned counts per day and use this indicator to conclude the one with more counts is more popular, this will raise a problem. Because some comments are positive for this brand and some comments might be negative for this brand. If we only count the total mentioned numbers to compare, it seems like we transfer those negative comments into positive comments. We will get wrong result from this model.

As far as I am concerned, we can divide dataset into two groups first: One for dataset mentioning Michael Kors, One for dataset mentioning Kate Spade. And then we can follow the Phrase Finding algorithm developed by Takashi and Matthew *(A Language Model Approach to Keyphrase Extraction , 2003)* to find the meaningful phrase from these two groups. Briefly to say, this algorithm can assign the score for each phrase and the higher score means the phrase is more meaningful for this group. We can look at this algorithm on the other side: these scores can be served as weight for each phrase.

For example, after the executing phrase finding algorithm, we find the top 5 phrases for the two groups.

| **Michael Kors** | **Kate Spade** |
|---|---|
| 1. great product | 1. easily break |
| 2. huge discount | 2. bad quality |
| 3. high price | 3. huge discount |
| 4. soldout | 4. awesome |
| 5. buy together | 5. love it |

Like the example above, we might know customers prefer to choose Michael Kors instead of Kate Spade. We can maybe list the top 40 phrases and compare two groups. This method can more accurately judge which brand is more popular as compared with pure word count method. However, based on this scenario, we still have to define which word belongs to positive and which belongs to negative. This might be a little subjective or we can research on other method to quantify it.

Overall, the proposed method can firstly filter out non-meaningful phrases for each group and then we only need to focus on these “important” phrases to compare with each group.