Aim: In South Korea, group conflict between males and females has been severe since military service incentive emerged as an issue in 1999. Hate speech reflecting the gender conflict has pervasively used as SNS becomes popular. This research gathered tweets from September to December of each year between 2010 to 2018, and has attempts to track the change of frequency of misogyny and misandry vocabulary, *kimchi-nye* and ‘*han-nam*’ respectively. This research will help to understand how gender conflict is developed.

Background: South Korean male citizens between the ages of 18 and 35 are required to perform 21 months of compulsory military service, as the country still remains at war with North Korea after Korean war ended in 1953. As part of effects to compensate those who completed the mandatory service, a bill was proposed in 1999 to give incentives to test-takers with military background for a government job. In concert with the tight job market and high unemployment rate in Korea, the military service incentives led to conflict of interest between males and females (Kim 2016). Around this time, a new misogyny word ‘*toinchang-nye’* was created, but got substituted by ‘*kimchi-nye*’ around 2008. These words were pervasively used on SNS,

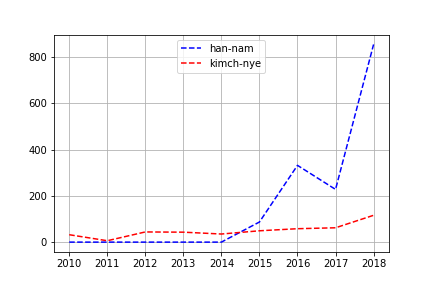
The literal translation of the *kimchi-nye* is ‘a kimchi woman’. Kimchi as representative food in Korea, *kimchi-nye* refers to Korean women and has the similar meaning as ‘gold-digger’. There had been attempts to create misandry vocabulary by females as counter-attacks, but none of them lasted as long as the misogyny word. However, since a tragedy in 2016, which is unofficially considered as a hate crime towards women, a misandry word *han-nam*,which is an acronym of *hankuk namca* ‘Korean man’, has been created and its use has spiked on SNS.

Data-gathering: I gathered historical tweets containing two key words‘*kimchi-nye* and ‘*han-nam*’ from September to December of each year between 2010 to 2018. I could not get historical tweets using Twython library because twitter API has limitations in collecting historical tweets. I found an automated crawling by keywords using Selenium, which automates web applications, and Beautifulsoup, a python package for parsing HTML and XML. The number of tweets gathered by each keyword showed big difference. Tweets gathered with a keyword *kimchi-nye* was 660 while tweets with *han-nam* was 2201 tweets. I am not sure whether this difference could be due to the user gender ratio or some other reasons.

Data-cleansing: The data collected by Selenium and Beautifulsoup still included HTML code, so I used regular expression to extract tweets only. After that, I used Mecab class of Konlp package for tokenizing and POS tagging, which is a Python package for natural language processing of Korean. Since the ultimate goal was to look at the frequency change according to timeline, I tagged each tweet with written year, and categorized them into two groups by the search keywords.

Sentiment-Analysis: the target vocabulary had more than one sense involved. For instance, *kimchi-nye* was used in a non-offensive way when the word started to emerge. Sometimes, it was used to refer to a woman who eats kimchi, or it meant bourgeois, Therefore, it was necessary to distinguish whether the same string was offensive or not. I thought that one way to achieve it is to look at whether each tweet has positive reading or not. If it is positive, I thought there would be less chance of the target word to be used offensively. To achieve this, I built a crude version of sentiment analyzer by using a Korean sentiment dictionary (<http://word.snu.ac.kr/kosac/lexicon.php>), which includes 16363 words with its polarity score marked. My plan was to check the polarity of each tweet, and only count the tweets with negative polarity, in order to ensure that the vocabulary is used offensively. To check the accuracy rate, I manually marked the polarity score of 280 tweets of each group by the keywords. The accuracy was not high enough. Both of them were 62 percent accurate. I could not improve the accuracy at this point due to some limitations provided under “Challenge and Limitation” at the end of this document.

Results: I drew plot of frequency of each word by year using Matplot library. According to the data gathered in 2010 and 2011, *kimchi-nye* did not carry offensive meanings. It either meant women who eat kimchi or bourgeois. However, since 2012, it has been only used to mean ‘gold-diggers’, and its frequency has gone up until 2018. A misandry word against *kimchi-nye* was created in 2015 after about 4 years after *kimchi-nye.* Since then, its frequency has spiked up. Compared to the increase of *kimchi-nye,* the use of *han-nam* has dramatically increased despite the fact that the misandry word created much later. The result is represented in the plot below.



This dramatic increase seems to reflect *Gangnam* murder[[1]](#footnote-1) happened in 2016, which was considered a hate crime towards women by the public. The frequency of *han-nam* decreased in 2017, but I am not sure whether it is because of the low accuracy of the sentiment analyzer or whether the frequency was really dropped in that year.

Challenge and Limitation: The sentiment dictionary was not inclusive enough to evaluate every vocabulary in the gathered tweets, leading to undermining accuracy. Moreover, due to the nature of the data, many words include newly created words, slangs, and swearing words, which are also not included in the sentiment dictionary. I found over 900 Korean swearing words in Korean contemporary dictionary website, but I could not crawl these with my current knowledge. To get this data, I have to know how automatically click a advanced search button, type “curse word” in keyword section, which gives 99 pages of words. I have to know how to automatically click those pages and scrap the data. Otherwise, I have to get the data manually, which will take long time. Though adding this data in the sentiment dictionary will not still be inclusive, it will improve accuracy.

2. Normalizing: Since Korean is an agglutinative language, conjugation is very complex, which makes normalizing hard.

Reference

Kim, S.(2016) yesenghyemo, feminismui say sitaylul kacyeota. *Kyoyookpipyeng,* (38), 163-188.

1. This is the link, explaining the murder case. <https://en.wikipedia.org/wiki/Seocho-dong_public_toilet_murder_case> [↑](#footnote-ref-1)