Research Project Outline

TaeYoung Kang (IT Management, KAIST)

It is well known that the behavior and decision making of an agent is affected by emotional state. (George and Dane, 2016) Among various types of emotions, *anger* is considered to be a catalyst for the attitude of evaluating others (Wiltermuth and Tiedens, 2011), and hence, the comments section on online news platform has been an ideal place to observe such relation. Since it provides affluent opportunities to handle the key concepts of social science including emotion, user interaction, and opinion formation, *online trolling and swearing effects* has been a main concern in computational social science and its adjacent fields. (Kwon and Gruzd, 2017; Ksiazek, 2018; Weber, 2013)

This research also aims on verifying the variables that could influence the *incivility level* of users based on online news platform comments data. The users' internal process of writing comments and reflecting their emotions through it is complicated. The negative sentiment level of news article's title and content would affect the users' comment at the first hand. At the same time, however, we should not neglect the effect of attached image, called *thumbnail*.

The *human salience level* in thumbnail image is a key concern of this research. According to previous research, the *numerical salience level*, or *the number of an object* tends to present a unique relationship with the emotion provoked from that object. (Hsee and Rottenstreich, 2004) Their relations are illustrated in the plot below.

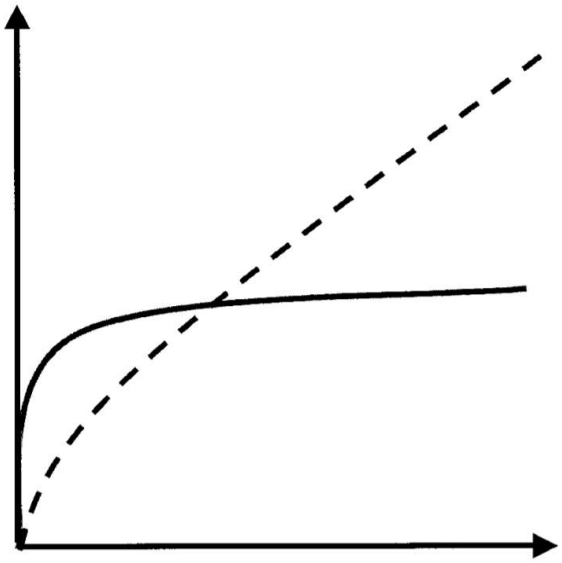


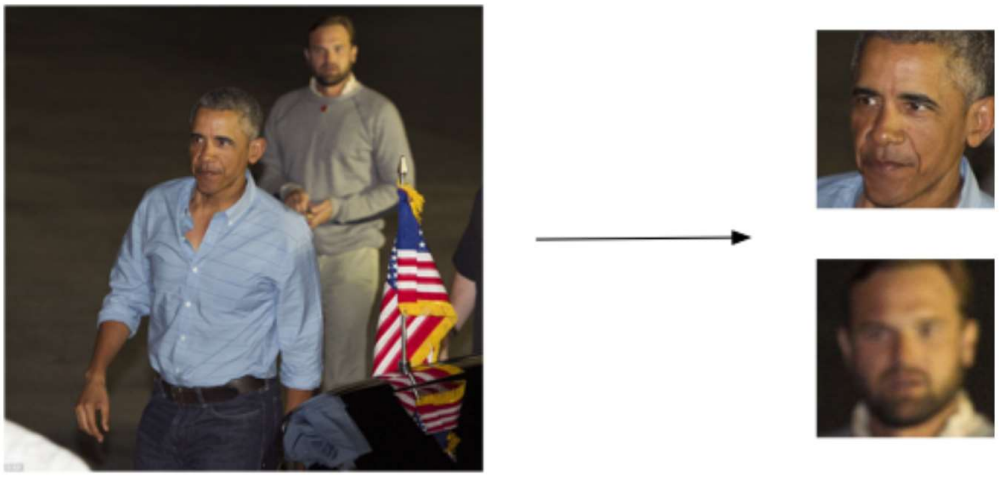
Figure 1. Value functions based on calculation (dotted line) and based on feeling (solid line). The x-axis of the function is the scope of a stimulus, and the y-axis is subjective value.

Such pattern could give us a hint on interpretating the possible outcome of human salience. Reporters can adopt distinct thumbnail images even when dealing with the identical issue as they have different level of human salience. In case of news article on economic downturn for instance, the reporter can use either the picture of ministry of economy building, GDP time trend visualization plot, or the specific photo on the minister which is a concrete human figure. In sum, the percepted human salience level of image would show the relations below.

$a < b \leq c$
where
a = image with no human figure
b = image with a single human figure
c = image with multiple human figures

The *Attribution Theory* takes an important role as a theoretical base on the relations between online trolling and human feature salience. Social psychologists have argued that angry agents are prone to express their emotion more intensely when the source of the anger is provided, since it enables the attribution to that source. Such source salient circumstance is highly similar with online news platform environment. When positing the news commenters with an explicit political predisposition, they would be more likely to show harsher attitude to the article when the human feature is provided as a specific source in the image.

Aside of theoretical issues, this research can also enrich the empirical analysis by adopting latest computer vision and natural language processing techniques. The angle, color, and aesthetic quality of an image is known to affect the customer's information acquirement and emotion formation (Hagtvedt, H., V.M. Patrick. 2008; Miller, E.G., B.E. Kahn. 2005; Larsen, V., D. Luna, L. A. Peracchio, 2004.). During past several years, deep neural network showed a drastical improvement and aided researchers by enabling the measurement of more complicated internal aspects of an image. (Zhang, Shunyuan, et al, 2016) Facial recognition is not an exception. Simply by adopting existing computer vision packages, verifying the location and size of human faces is an easy task these days.



Input Output

Measurement of swearing comments, however, is a way more complex issue. As users are aware of dictionary based automatic swear words censoring algorithm, they try not to swear directly. In English, for example, users would type *"$hit"* or *"5hit"* instead of *"Shit"*. It becomes even more complicated when it comes to *Hangul*, the Korean alphabet.

(a) chosung (b) joongsung (c) jongsung

As can be seen above, each character is composed of three subcharacters(or *jamo*) called *chosung(초성)*, *joongsung(중성)*, and *jongsung(종성)*. To convert these unique character system similar to latin alphabet, we can simply split character into consecutively listed subcharacters. (ex.해달→ㅎㅐㄷㅏㄹ) (Park, Sungjoon et al, 2018) Although it hinders the human interpretability of words and models derived from them, its merits on word-vector embedding generation could exceed such trivial dimerit, especially when designing swear words classification model.

By adopting FastText, an algorithm which generates word-vector embeddings based on subcharacter n-grams (Bojanowski et al, 2016; Joulin et al, 2016), on *Hangul* subcharacter preprocessing trick, we can maximize the classification performance. Let's look at more specific swear example. *병신*, a common Korean swear word, can be variously modified into other forms such as *빙신*, *비융신*, *븅신*, *병슨*, *병1신*, and *븅1쉰*. If we use swear words dictionary, it would be hardly possible to detect the correlation between these variations. Such predicament would still not be alleviated when using POS(Part of Speech) tagged words, since there are almost infinite possibility to generate new variation simply by adding or substituting a single subcharacter. However, if we split subcharacters of words and then apply FasatText, the shared subcharacters and subsequent n-grams shared among these words such as *{ㅂ, ㅇ, ㅅ, ㅣ, ㄴ}*, *{ㅣ, ㅇ, ㅇㅅ, ㅣㄴ}*, and *{ㅣ, ㅇㅅ}* are now detected. When tested this methodology on pilot data, it showed intuitively enhanced performance.

In [19]:

```
ftmodel.wv.most_similar(jamo_split('개새끼'))
```

Out[19]:

```
[('_ㄱㅏㅣㅅㅏㅣdㄱㅣ', 0.9877994060516357),
 ('ㄱㅏㅣㅅㅏㅣdㄱㅣㄱ', 0.9872113466262817),
 ('ㄱㅏㅣㅅㅏㅣdㄱㅣㄴ', 0.9870151281356812),
 ('ㄱㅏㅣㅅㅏㅣdㄱㅣ_', 0.9867669343948364),
 ('ㅣㄱㅏㅣㅅㅏㅣdㄱㅣ', 0.9831262230873108),
 ('ㄱㅏㅣㅅㅏㅣdㄱㅣㅇ', 0.9806551933288574),
 ('ㅇㄱㅏㅣㅅㅏㅣdㄱㅣ', 0.9803736209869385),
 ('ㅏㅣㅅㅏㅣdㄱㅣㄱㅏ', 0.9734358787536621),
 ('ㄱㅏㅣdㅅㅏㅣdㄱㅣ', 0.973200798034668),
 ('ㅗㅏㅣㅈㅣㅅㅏㅣdㄱ', 0.9714223146438599)]
```

In [20]:

```
ftmodel.wv.most_similar(jamo_split('갯1쉭키'))
```

Out[20]:

```
[('ㅏ_ㄱㅏㅣㅅㅏㅣㄱg', 0.984258234500885),
 ('ㅣ_ㄱㅏㅣㅅㅏㅣㄱg', 0.9773533940315247),
 ('ㅜ_ㄱㅏㅣㅅㅏㅣㄱg', 0.975597620010376),
 ('ㅅㅜ_ㄱㅏㅣㅅㅏㅣㄱ', 0.9745869040489197),
 ('ㄱㅏㅣㅅㅣㅂㅅㅏㅣㄱ', 0.970960259437561),
 ('ㄱㅏㅣㅅㅏㅣㄱㅎㅣㄷ', 0.9684487581253052),
 ('ㅗㅏㅣㅈㅣㅅㅏㅣㄱg', 0.9662880301475525),
 ('ㄷㅗㅏㅣㅈㅣㅅㅏㅣㄱ', 0.9652479887008667),
 ('_ㄱㅏㅣㅅㅏㅣㄱgㄱ', 0.9650763273239136),
 ('ㄴ_ㄱㅏㅣㅅㅏㅣㄱg', 0.9591826796531677)]
```

To summarize, this research aims on tessting the hypothesis with the model below.

Hypothesis :
The more salient the human features in the thumbnail image of an article, the higher the average swearing level of all the comments it has.

$$\sum_{j=1}^{m} s_{i_j} = \beta_1 x_i + \beta_2 h_i + f(\sum_{j=1}^{m} s_{i_{j,t-}}) + \beta_3 \sum_{j=1}^{m} d_{i_j} + \beta_4 \sum_{j=1}^{m} g_{i_j} + \epsilon_i$$

s_{i_j} = incivility level of comment written by j^{th} commenter of i^{th} article
x_i = control variables generated from i^{th} article (ex. number of comments, topic sentiment, article sentiment, article length)
h_i = human feature salience level of i^{th} article's thumbnail image

$s_{i_{j,t-}}$ = incivility levels of ***previous comments*** written by j^{th} commenter of i^{th} article
d_{i_j} = j^{th} commenter of i^{th} article's diversity of news consumption
g_{i_j} = j^{th} commenter of i^{th} article's average time gap on writing new comment (= proxy variable for the acivity level of user)

Methodological Issues and Questions

(1) How should we handle the term $f(s_{i_{j,t-}})$ in the second model?
(2) How should we control the selection bias issue of commenters?