Analysis of Celebrity Reputation Based on Metadata from Internet Tabloids

Daewon Kim
School of Computing
20180061

Youngil Yoon
School of Computing
20180416

Hyunho Lee
School of Computing
20190531

Abstract

This document contains the overall information, including research objectives, research process, and anlysis on celebrity reputation from metadata from internet tabloids. This research team used metadata from internet tabloids to analyze celebrity reputation. Specifically, we analyzed attention from the public and impression celebrity gives using *R*. Statistical techniques such as regression, moving average, STL, etc. are used to analyze and predict. Results suggests that analysis using internet tabloids metadata is meaningful.

1 Introduction

In the 20th and the early 21st century, newspapers were the main media to give people information. In modern days however, internet tabloids have taken their place. Internet tabloids does more than publish news, but they also offer some functions for people to react to the news, and those reactions that people gave are shown to other people as the form of metadata.

This research team has started this research based on a simple question while observing the metadata of internet tabloids; Can we Analize the reputation of a celebrity from the metadata of internet tabloids?

Figure 1: Article from Internet Tabloid *NAVER*

We drew up some research plans, including crawlling the internet tabloids for the metadata, pre-processing those metadata with natural language processing techniques, and statistical research methods involving *R*, which we learned from class.

2 Objectives

Our first goal is to plot the impression certain celebrity gives, and the interest from public verses time and observe the transition.

Our second goal is to present the magnitude of issues which gives dramatic changes to the impression that certan celebrity give.

Our Third goal is to present the keywords related to certain celebrity in the form of wordcloud.

Our fourth goal is to predict the transition of impression certain celebrity gives, and the interest from public in the form of time.

3 Background Knowledge

This research team has mainly used the internet tabloids website *NAVER Newspapers* to extract metadata(crawlling). We used natural language processing techniques to extract meaningful data(preprocessing), and made statistical approaches via *R*.

The following illustrates how this research team approached each components.

3.1 NAVER Newspapers

NAVER Newspapers is the biggest internet tabloids website in Korea. Thousands of posts and articles are uploaded everyday in NAVER Newspapers, including the three biggest newspapers: Chosun, Donga, and Joongang. Articles are categorized in 'politics', 'cultural', 'society', and 'entertainment' etc. Most significantly, *NAVER Newspapers* offers readers some functionalities to express and react to the article.

Comment Section Firstly, there is comment section, where readers write down their expressions freely, and people can either 'like' or 'dislike' the comment. Due to regretful incidents happend in entertainment industry, involving suicide of celebri-

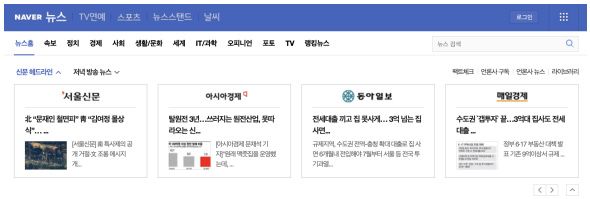

Figure 2: Internet Tabloids Platform *NAVER*

ties, comment sections has been removed from those articles of categories entertainment.

Emotional Expressions Secondly, and most importantly, there is a emotional expressions where people can pick from 'like', 'warm', 'sad', and 'angry' etc. emotional expressions are most easily accessible as they are exposed at the top of the article, right below the title.

This research team has used *NAVER Newspapers* to scrap data as it contains variaty of metadata including comment section, emotional expressions, and the date of time the article is released etc.

3.2 Bigkinds

Bigkinds is a website which offers easy access to internet tabloid platforms such as *NAVER Newspapers*. The website makes it easy to search internet tabloid articles with keywords such as person, the journalist, event etc.

This research team used *Bigkinds* to access *NAVER Newspapers* with celebrity name as a keyword.

3.3 koNLPy

koNLPy is a python package to process korean literal informations. This package offers high-precision text pre-processing, such as word tokenizing, part-of-speech tagging etc.

This research team used this package to preprocess the title of the article, hence signifying meaning of the title.

4 Assertions

Before starting our research, this research team had to make some assertions in order to make our analysis valid. As concepts such as 'reputation', 'attention', 'opnion', 'preference' are abstract and possibly vague, we had to define or assume those concepts in the form of statistical variables.

Impressions We assume that people have impression for a certain celebrity as a number between 1 and 0, where the prefrence gets more intense as the number gets larger.

Emotinal Expressions We assume that people with negative impression of a certain celebrity will give negative emotional expressions, such as *Angry*, *Sad*, *surprized*. We also assume that people with positive impression of a certain celebrity will give positive emotional expressions, such as *Like*, *Congratulations*, and *Looking forward to*. These are emotional expressions *NAVER* provides: Like, Warm, Sad, Angry, Want further inquaries, Cheer, Congratulations, Looking forward to, Surprized

Interest We assert that any individual either has interest in certain celebrity, or not. Those with interest in certain celebrity will give a emotional expression to the celebrity related article, those with no interest will not.

Press We assume that press does not give any influence to individual reader's any choice.

Independancy We assume that every individual's opinion is independant.

etc We assume that most of the inflow to any article is through the most-searched keyword section from the website.

5 Dataset

As we mentioned earlier on the material part, we used *NAVER News* and *Bigkinds* to gather the data we need. Also, we had to some preprocessing on the dataset to handle exceptions. In this section, we will introduce with some of our methods and preprocessing.

5.1 Dataset Gathering

As we need specific articles related to specific celebrities, keyword searching for articles were necessary. For keyword searching, using *Bigkinds* was one option, as it provides keyword searching for various articles. We coded a crawler to gather all the articles related to celebrities, containing the date the article was posted, article title, and emotional expressions.

5.2 Dataset

The dataset contains approximately 200 csv files containing the date the article was posted, article title, and emotional expressions for past year

searched on celebrity name. We additionaly acquired 3years amount of dataset for those celebrities with significantly resulting data.

5.3 Pre-processing

There were some anomalities where negative article about a celebrity was given positive emotional expressions, mainly for mocking purposes.

Figure 3: Internet Tabloids Platform *NAVER*

A pre-processing based on NLP(natural language processing) technique was used to handle this exception. We used a python NLP package *koNLPy*, iMDB package, kkma class, and tensorflow to determine the positive/negative sentiment in the article title, together with a manualy made list of negative words.

6 Methods

These are some of the methods we used to achieve our objectives. Firstly, we categorized the emotional expressions such as Like, Warm, Want further inquaries, Cheer, Congratularions, Looking forward to as positive or neutral expressions, and Sad, Angry, Surprized as negative expressions.

6.1 Analyzing Attention from Public

- Observe the transition from attention from public by using moving average and STL.
- Find correlation in idol groups and its members by using scatterplot
- Find the member in a group that influences the group's public profile the most.
- Level the public profile of idol groups by using clustering.

6.2 Inpression Celebrity Gives

- Observe the transition of attention from public by using plot()
- Present the magnitude of the issues that makes significant change in the transition from observation.
- Categorize the negative articles and use the data to analyze idol fandom.

6.3 Wordcloud

- Present the most deeply connected words recognized by public in a visible scale.
- Change the threshold for attention for which we collect the articles and observe the change.

7 Analysis

7.1 Attention from Public

Techniques We computed the attention from public as the net number of emotional expressions.

Transition We used moving average for 15 weeks and STL(Seasonal and Trend decomposition using Loess) to effectively observe the transition of attention from public. The observation was meaningful based on the issues that appeared on significant weeks and days. (Folder : Trend Graph)

Correlation in Team We used scatterplot to observe the correlation between an idol group and its members. The observation turned out to be meaningful, as the transitions were highly alike. From this observation, we could make a justification about our methodology; detaild analysis on group members can be deducted from analyzing in units of groups as we did in this research. (Folder : Scatter)

Member with high influence We tried to find the member from group whom influences the group's public profile using multiple regressions.

In trying regression on a idol group by the data of each member from the group, we figured out that

as the data has a trend we have to remove the trend and then do a regression. In which case, we can try two different options: first-difference method and doing a regression after subtracting trend acquired by STL.

We could deduce the member who influences the BTS's public profile. The significance probability was observed in both first-difference method and STL. The common significant members are observed as Jungkook, J-Hope, Jin, RM. The STL using method suggests that V is significant too. We analyzed based on articles that Sugar is a member who does individual work a lot, and Jimin has a large individual fandom.

The result presents that ACF is very high, which suggests that the regression was not appropriate. There will be a need to add informations other than metadata from internet tabloids. (Folder : multi)

Clustering We used clustering for the attention versus time in the unit of groups. For idol groups, it turned out a meaningful result of groups clustered by their public profile. (Folder : Cluster)

7.2 Impression Celebrity Give

Techniques We computed the Impression certain celbrity gives from the ratio of Positive/Negative article titles for time and the ratio of emotional expressions per net number of emotional expressions for time. We also computed severity with negative ratio with net number of emotional expressions.

As we mentioned in dataset pre-processing, for some controversial subjects there were a lot of positive expressions even for the negative titles, mostly for mocking purposes. We handled the case by processing natural language on the titles.

Issues We lined up those issues that showed great downward change in impressions celebrities give. (Folder : list)

Transition We observed the transition of impression versus time plot. Interestingly, the change in the transition of related celebrities was more than significant in the week when the critical issues such as 'Me Too'movement and 'Debt Too'movement. Also, in the week the 'Burning Sun Gate' was exposed, the related celebrities' impression literally plummeted. (Folder : Plots)

Categorizing As we assumed three of the expressions *Angry*, *Sad*, and *Surprized* as negative expressions, we could categorize the negative articles in to three categories.

Fandom For idol groups and their individual members, the existence of large-scale fandom gives a lot of influence to the impression. For non-idol celebrities, there are a very few number of negative expressions unless there is a highly severe incident. However for idols, the change in the transition for emotional expressions are very significant even for a minor incident; which makes it possible for us to verify the characteristics of idol fandoms. Thus, fandom anlysis will enable idol agencies to understand and predict how the fandom will react. (Folder : Fandoms)

7.3 Wordcloud

See folder 'Wordcloud' for images.

Techniques We wordclouded those articles related to a certain celebrity and over a certain threshold of attention. We seperated the article title to 9 lexical categories, and stored them as tibble. For nouns, we excluded a stopword and stored them as frequency distribution. We used same technique for those articles with bad issues about celebrity, resulting in the wordcloud of negative keywords.

keywords Through wordclouding, we could get gather the keywords which the public recognize them of in a visible scale, and those keywords which the public got most interested when a bad issue happend to the celebrity.

Idol Groups For idol groups, we could distinguish the member who got the most interest from the public in a visible scale.

Threshold We initially set the threshold for the attention at 30, but as we changed the threshold to 50 and 100, the generated wordcloud barely changed, which must mean we have sampled the dataset pretty well.

8 Conclusion

Through the series of analysis, we have discovered it is not only possible to analyze the metadata from internet tabloids, but we can also find the pattern in the processed data and predict the transition.

9 Github

The github repository this research team has worked apon is as follows.

`git@github.com:askme143/CS492E_R.git`

10 Appendix

A Trend Graph

All of Figures and plots in a seperated zip file