ISSN 2278-3091

Volume 9, No.1.3, 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse2691.32020.pdf

https://doi.org/10.30534/ijatcse/2020/2691.32020

Text Analysis on Instagram Comments to Better Target Users with Product Advertisements

James Arnold E. Nogra
Cebu Institute of Technology – University, jamesnogra@gmail.com

ABSTRACT

Online advertising has been growing steadily over the years as more and more people connect to the internet. Advertising online can reach a large crowd which is both an advantage and disadvantage because a lot of people will see products they don't want. Targeted online advertisements are gaining popularity because it only shows advertisements to people who might be interested in the product based on their preferences and online presence. But these targeted advertisements are still not enough to narrow down the users to sell products. There might be products that might not be needed by certain people at certain times thus reducing the effectiveness of the advertisement. The text analysis tool for Instagram is developed in order to narrow down the most repeated words of media posts based on the type of photo, the gender of who made the comment, and the time the comment was posted. These repeated words are then visualized using a word cloud and word histogram so that businesses can determine which and what products to advertise. A two-week advertising experiment was done wherein during the first week, a random product was advertised on Instagram while in the second week, the text analysis tool was used to determine which product to advertise. The number of sales slightly increased in week two compared to week 1 (18 items vs 16 items sold) while the number of customer inquiries dramatically increased from 92 in the first week to 157 on the second week.

Key words: Instagram Advertising, Text Analysis, Word Cloud, Word Histogram

1. INTRODUCTION

1.1 Background of the Study

Advertising on Instagram has been growing in the past few years [1]. Part of this growth is because of the popularity of the app. There were about 800 million active users on the platform in 2017 and in 2019, there were already 1 billion active users. This is one of the reasons why many marketing agencies are choosing Instagram to advertise or market products or services [2]. For this reason, many independent brands and startup businesses are advertising on this mobile application. Aside from it's cheaper than advertising on TV or print, advertising on Instagram allows business owners to target a specific group of consumers instead of just advertising to everybody.

Target advertising on Instagram has many benefits such as cheaper and it has more impact [3]. But these targeted advertising also has its limitations. One limitation that has plagued on the platform is bot or spam accounts that give likes and generates comments programmatically. An advertisement displayed to a bot or spam account doesn't generate an impact to the business who is advertising. Another disadvantage of advertising on Instagram is that the platform itself doesn't divulge the gender of the user. A marketer would unknowingly advertise beauty products or any products for women to a group of users using the mobile app. A system that would determine if a user is male or female would greatly help in classifying the users of this mobile application. There are many websites and apps that let users report an advertisement which they deem to classify as inappropriate or not for them [4]. This is one of the reasons why targeted advertising must be implemented especially to small business owners who don't have a large budget for marketing.

One way to determine a target audience is to analyze what users or customers are saying to a certain post online. For example, a photo of a person with nice clothes might have comments on where the clothes are bought, brand, or some others relating to the clothing and not to the person. This indicates that some of these people who are talking about this online post are planning to buy the same clothes or something similar. This would be a perfect flag for advertisers, specifically clothing businesses, to advertise on their products to the owner of the media post or even directly to the users who are commenting on it. One way to get representation on what users online are talking about a certain topic or media post is to gather all the comments and determine which words are the most mentioned [5]. From the words that are more commonly mentioned, a histogram and word cloud of the words can be constructed. Word clouds will visualize the words in terms of how repetitive a word is mentioned in a collective text. The more repetition a word has, the larger its font size would be in the word cloud.

1.2 Objectives of the Study

The main goal of this study is to determine which words are mentioned more in Instagram comments depending on the category of the media photo. The comments can be narrowed down to the date and time it was posted and the gender of the user who made the comment. Through this filtering method, advertisers or posting advertisements on Instagram can be properly targeted. Businesses who are just starting out and don't have a lot of budget towards marketing can use this system to advertise their products to specific users. A simple test should be done to make sure if the results can affect the engagement and sales of a small business.

1.3 Significance of the Study

Advertising on Instagram is not that affordable, especially to startup businesses. Having a system that would advertise to an audience based on what they are discussing on Instagram comments will greatly help these startup businesses save money on marketing. The system also records the users who are posting comments and they can be individually targeted with advertisements instead of advertising to a group of people.

1.4 Scope and Limitation

This study is only limited to analyzing words. Analyzing phrases is not part of this study. Another limitation of this study is determining the gender of the Instagram user who made the comment. Instagram doesn't provide the gender of its users thus this study only compares the first name of the user to a list of male and female names. Names that are not on the male or female list is tagged as an "other" gender. The influencers (Instagram users with many followers) are also all Filipinos. The comments of the media posts of these influencers are usually also from Filipino users.

2. REVIEW OF RELATED LITERATURE

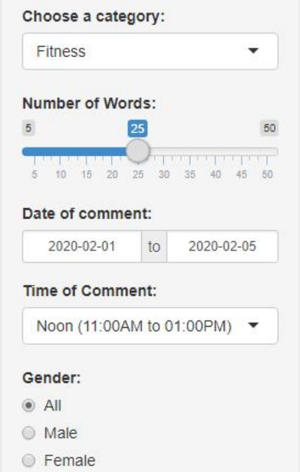
Advertising on print or TV is still an effective way to market a product. Even though advertising online is growing, there are still some businesses that prefer to advertise on the old format, even on the radio. This is because there is still a portion of the Philippine population who still depend on these media formats [6]. The issue of advertising on TV, radio, or print is that the target audience is huge and the variability between ages, gender, etc, are way too high. Even though there are programs on TV or radio that are targeted to certain types of audiences, this would still be not enough most of the time to market a product to a target population.

Since the internet became mainstream, many businesses saw the potential of advertising on websites. Yellow Pages is one of the largest advertising and directory companies to bring their business online [7]. Many advertising companies followed in advertising online because, during the '90s, the internet was growing [8]. Many marketers and businesses saw the potential of marketing online because of its growing popularity. Targeted advertising has been more prominent on the internet because everyone around the globe can access a website or video online. Advertising a local product online might be confusing to other people who can see it across the globe. For instance, beef burgers are very common in the western hemisphere. Advertising a beef burger on a website might offend a Hindu who accesses that particular website. Algorithms and filters have been implemented to avoid issues like these and to better target audiences. Cookies have also been used to store information of a user to determine which advertisement is most suited for that user. User information like these has also been used to develop a machine learning model for targeted advertising [9].

Another way to target users is the use of text analysis for people who are discussing something online. Twitter is one of the places to look for people discussing a topic [10]. Through the discussed topic, text analysis can be done in order to get the main topic users are discussing. The results of the text analysis can be used in order to plan how to market a product or service. Another social media tool that is gaining traction to both ordinary users and business owners is Instagram. This media sharing app or website has over 1 billion active users in 2019. This is the reason why private individuals and businesses are using this media sharing tool. Because of its popularity, businesses are advertising their products or services through their business accounts or paying some influencers (non-celebrity people who have numerous followers or can influence a lot of people) to advertise [11]. But letting influencers advertise a certain product might not target a particular audience properly. Even though these influencers have their own niche, targeting a certain audience might not be suited for them. Filters for Instagram users and comments such as gender, time of comment, and type of media posted might help narrow down a target audience for advertising a service or product. Topic modeling such as Latent Dirchlet Allocation Algorithm might help in order to determine the topic of a large text from an online discussion. This topic modeling is usually used to determine or classify the topic of a very large text document [12]. Word cloud and word histogram of the most repeated words could also be enough to determine what the main topic of a discussion is. The top words mentioned in a collection of texts can also help narrow down a target audience for target marketing.

3. METHODOLOGY

To be able to determine which words are always mentioned in comments, the media tag in which the comment was made, the time that the comment was posted, and the gender of the user who made the comment is stored in the database. These three filters shown in Figure 1 can narrow down which words in comments are mentioned more depending on the category of the photo, the gender of who made the comment, and the time users are commenting on it.



Choose a category:

Fitness

Number of Words:

5 25 50

Date of comment:

2020-02-01 to 2020-02-05

Time of Comment:

Noon (11:00AM to 01:00PM)

Gender:

● All

○ Male

○ Female

Figure 1: Controls or Filters of the text analyzer to determine which comments to process.

Using the Instagram API Librarymgp25 Instagram-API, comments were gathered from the various Philippine-based influencers. A total of 50 influencers' profiles were used to scrape media comments in this study. The influencers involved are Instagram users that have more than 10,000 followers but less than 100,000 followers and these people haven't appeared in any TV show or movie. Aside from the comments being gathered, the media itself is also stored in the database in which it will be manually tagged. The data scraping is done using a PHP script that runs every four hours. When the script runs, it will list all of the defined Instagram influencers. After listing these Instagram influencers, each of their media posts is then listed. Within each of these media posts are comments from users that are then gathered. The media id and comment id are recorded to ensure that no duplicate media and comments are stored in a MySQL database shown in Figure 3.

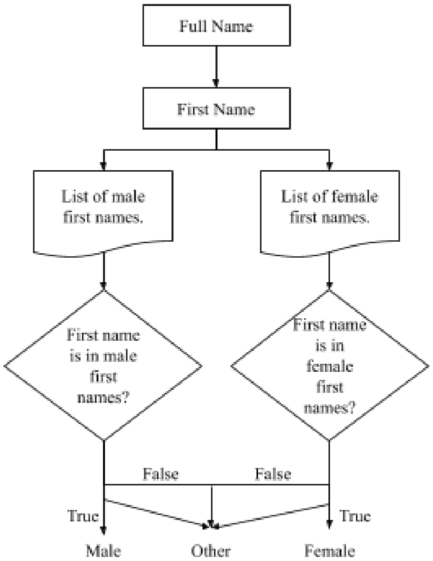
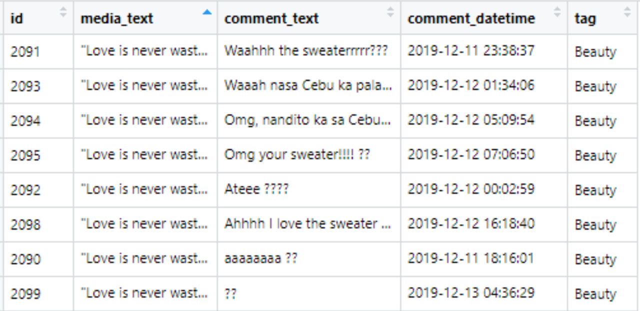

Full Name
First Name
List of male first names.
List of female first names.
First name is in male first names?
First name is in female first names?
False
False
True
True
Male
Other
Female

Figure 2: Gender classifier based on a list of male and female names.

Instagram API doesn't provide the gender of its users. Because of this, a simple name-gender classifier is implemented in this project as shown in Figure 2. The first name is extracted and then compared to a list of male or female names. First names obtained from the Instagram that are not in any of the lists of male or female names are classified as others. Each list of names has more than 1000 names in it. If the first name is found in the list of male names but not on the female list names, then that user will be classified as a male. If a first name is found in the female list names but not in the male list names, then that user will be classified as a female. If a first name is found in both the list of male and female names, then that user will be classified as other. Similarly, if a first name doesn't exist in both lists of names, then that user will be tagged as other.



id	media_text	comment_text	comment_datetime	tag
2091	"Love is never wast...	Waahhh the sweaterrrrr???	2019-12-11 23:38:37	Beauty
2093	"Love is never wast...	Waaah nasa Cebu ka pala...	2019-12-12 01:34:06	Beauty
2094	"Love is never wast...	Omg, nandito ka sa Cebu...	2019-12-12 05:09:54	Beauty
2095	"Love is never wast...	Omg your sweater!!!! ??	2019-12-12 07:06:50	Beauty
2092	"Love is never wast...	Ateee ????	2019-12-12 00:02:59	Beauty
2098	"Love is never wast...	Ahhhh I love the sweater ...	2019-12-12 16:18:40	Beauty
2090	"Love is never wast...	aaaaaaaa ??	2019-12-11 18:16:01	Beauty
2099	"Love is never wast...	??	2019-12-13 04:36:29	Beauty

Figure 3: Sample records of comments retrieved from the database using R.

For the comments, the data that were being stored is the id of the user who made the comment, the comment text, and the date and time in which the comment was posted. Each of the comments is then matched to a specific media post of an influencer. The media information stored includes the user id of the influencer, image or video of the media, date and time the media was posted, the text associated with the media, and the id of the media.

The saved media are then tagged manually by five users. Each media can be tagged up to two categories. These tags are then used to group comments based on the media category. A web application is used to do these media tagging process. The web application was made using PHP, HTML, and MySQL. The styling that was used in this web application is W3CSS which is one of the commonly used CSS frameworks.

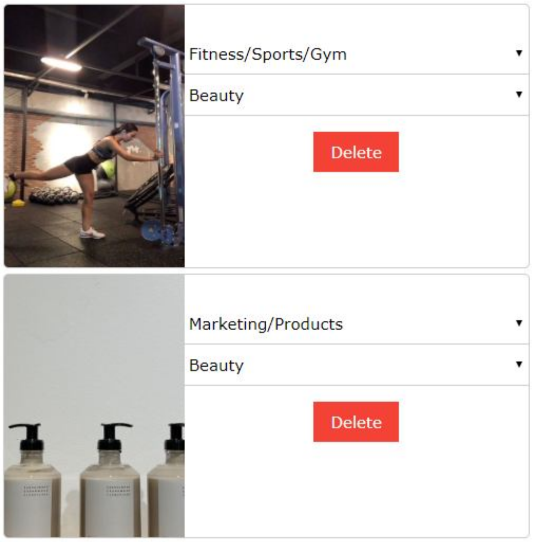


Figure 4: Sample screenshot of the web application that tags media into at most two categories.

Tagging media photos shown in Figure 4 are done on a daily basis. Media photos that are not yet tagged won't appear in the comments that will be processed. For processing the comments, users can filter through the tag of the media in which the comment is attached, the time in which the comment was made, and the number of words to display in the word cloud.

Generating the Word Cloud and Histogram is done by R packages. The wordcloud package is used to generate the word cloud. A word cloud is a visualization tool that displays words and its font size is determined by its frequency, the higher the frequency, the larger its size [13]. The histogram is also generated using the barplot function of R. The barplot will only display the top 10 most commonly used words for a particular query (using media category and time). The barplot is just a supplementary visualization tool to determine the exact number of frequencies each word has because the word cloud doesn't.

In the R code, initially, the comments are pulled from the database depending on what filter is applied (media category and time). From the database table, the results are then stored into a data frame. Then the data frame in which stores the comments are cleaned. These cleaning processes include removing extra spaces, removing special characters, removing emojis, removing stop words, and changing all letters to lowercase. Stop words are words that occur more frequently but are not important such as the words the, is, and, and many more [14]. A total of 40 stop words are used to clean the comments gathered. The languages of the stop words include English, Cebuano, and Tagalog. After these preliminary processes will generate the final corpus using the corpus function from the text mining package of R. A corpus is just a collection of texts from various sources like books or from the internet [15]. In this study, the corpus is composed of comments from Instagram media.

In total, there are 2,544 media posts from Instagram that were analyzed. These media posts have also been tagged manually by at least two categories. For the comments that were analyzed, there were a total of 30,528 records. These comments are a combination of almost all major Philippine languages. The English language, as well as other local languages, have their stop words which were removed in the comments before it can be processed as a corpus. A total of 11,298 unique users can be tied to all of the comments that were recorded. There are about 4,289 users that were classified as female by the system, 3,577 male users, and 3,432 users that can't be classified as either male or female based on their first names. All of these data were gathered from December 2019 to February 2020.

After the system to analyze Instagram comments were developed, a simple test was done. This test will use the results of a particular media tag and then advertise a similar product from a small business. The engagement (user inquiries) and sales in which the advertisement is run is compared to a week where the same product is advertised to a larger audience. The Instagram advertisement filters considered are location, demographics, and interests.

4. RESULTS AND DISCUSSION

One of the main goals of this study is to determine which words are usually mentioned by users for a particular kind of media post. A media post, which has been manually tagged has comments attached to it. Determining which words are mostly used by users for a particular media post will help advertisers better target these users with relevant advertisements [16]. To further narrow down which words are mostly mentioned in specific media posts, additional parameters can be used like date and time, the gender of who made the comment, and the type or category of media posts.

The date filter can narrow down comments made on a specific season to better target users with advertisements that are seasonal such as clothes or travel-related merchandise. The time filter allows to further limit results based on the time the comment is made by a particular user. This would be helpful in advertising products that are time-specific for a certain day [17] such as breakfast buffets or gym memberships.

Few of the comments are removed to ensure that only relevant comments are processed [18]. The research methods of this study have been influenced by [19] and [20].

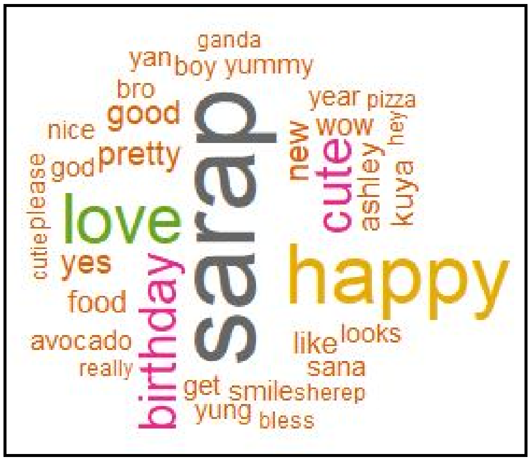


Figure 5: Word cloud for the most commonly used words in comments for all the posts of January 2020 from all media posts tagged as food.

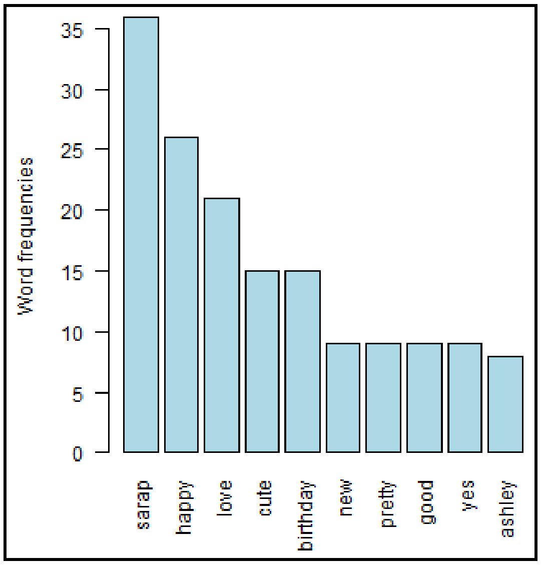


Figure 6: Histogram for the most commonly used words in comments for all the posts of January 2020 from all media posts tagged as food.

Word cloud and word histogram are shown in Figure 5 and Figure 6 can be easily generated based on media tags and date and time of a particular comment. The word cloud is generated by using the wordcloud package in R. The histogram is generated using the histogram function in R. The histogram only displays the top 10 most commonly used words in comments based on the parameters set on a particular query.

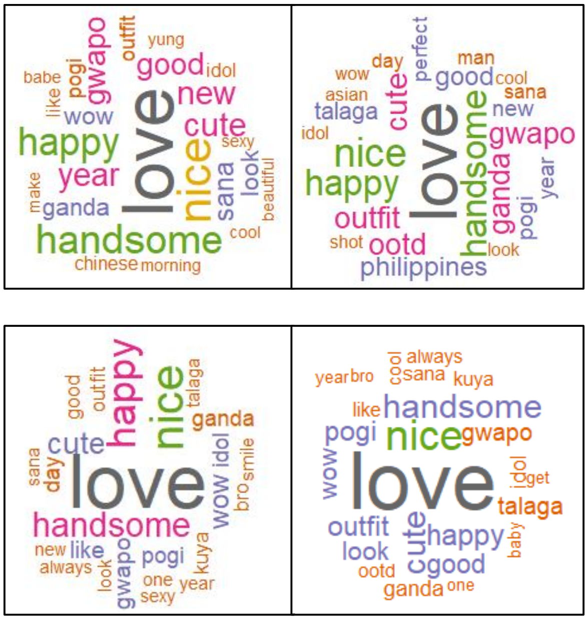


Figure 7: Word cloud for the most commonly used words in comments for media posts tagged as Fashion from December 2019 to early February 2020 at different time frames.

The small business that tested this system is a T-shirt printing business that mostly sells its merchandise online. They haven’t tried advertising online so they are suitable for this simple experiment. The first advertising experiment did was without using the text analysis tool. They created an Instagram account and then posted a random T-shirt with a random print on it. In the Ads Manager on Facebook (Instagram advertisement manager tool), the daily budget is set to ₱1,000.00 and the time frame is from February 2, 2020, to February 8, 2020. For the next advertisement experiment that uses the text analysis tool, the advertisement was run from February 9, 2020, to February 15, 2020. The same daily budget was allocated from the experiment prior to this one. But this time, the product produced and advertised was a t-shirt with cute print hearts printed on the front. The t-shirt print was based on the results gathered from the fashion tag media posts from January 2019 to early February 2020 as shown in Figure 7.

Table 1: Table of the number of customer engagements and sold items between the dates February 2 to 8, 2020 and February 9 to 15, 2020.

	Customer Engagement	**Sold Items**
February 2, 2020 to February 8, 2020	92	16
February 9, 2020 to February 15, 2020	157	18

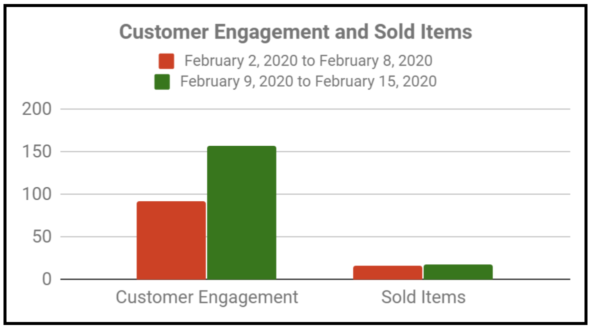


Figure 8: Graph of the number of customer engagements and sold items between the dates February 2 to 8, 2020 and February 9 to 15, 2020.

Based on Table 1 and Figure 8, from the two-week advertising experiment done, it can be inferred that even though the number of sales did not increase dramatically after the text analysis tool was used to influence the products to advertise, the number of customer engagements dramatically increased. The majority of these customer engagements include inquiries about t-shirt color availability, shipment, and bulk orders. The number of items sold on the first week of advertising was 16 items and after the text analysis tool was used the next week, there were already 18 items sold. The dramatic increase comes from the customer engagements wherein there are 92 inquiries were made in the first week and it increased to 157 on the following week. Other aspects of the tool like the gender of the users who made the comment were not used in this two-week experiment.

5. CONCLUSION

In conclusion, using a text analysis tool and visualization like word histogram and word cloud can effectively target users with advertisements. It can also help business owners develop or create new products based on what the users are discussing on Instagram. Discussion can also be narrowed down to the time where the comment is posted, the gender of who made the comment, and the type of the media posted. It was proven during a two-week advertising period that this text analysis tool slightly increased the number of items sold. This tool also dramatically increased the number of customer or user engagements to the business who participated in this experimental advertisement. Facebook and Twitter texts can also be used to further extend this study. These two social networks have millions of users in the Philippines.

REFERENCES

1. C. Muñoz & T. Towner, **The Image is the Message: Instagram Marketing and the 2016 Presidential Primary Season**, *Journal of Political Marketing*, 2017, 16:3-4, 290-318 https://doi.org/10.1080/15377857.2017.1334254
2. N. Evans, J. Phua, J. Lim & H. Jun, **Disclosing Instagram Influencer Advertising: The Effects of Disclosure Language on Advertising Recognition, Attitudes, and Behavioral Intent**, *Journal of Interactive Advertising*, 2017, 17:2, 138-149
3. X. Zhao & L. Xue, **Competitive Target Advertising and Consumer Data Sharing**, *Journal of Management Information Systems*, 2012 29:3, 189-222 https://doi.org/10.2753/MIS0742-1222290306
4. C. Tucket, **The economics of advertising and privacy**, *International Journal of Industrial Organization*, Volume 30, Issue 3, May 2012, Pages 326-329
5. R. Atenstaedt, **Word cloud analysis of the BJGP**, *British Journal of General Practice*, 2012; 62 (596): 148 https://doi.org/10.3399/bjgp12X630142
6. M. Chaves, **Filipino Parents' Attitudes towards TV Advertising and their Controls on Children's TV Viewing**, *IAMURE International Journal of Social Sciences*, 2014, vol. 10 no. 1
7. G. Lohse & D. Wu, **Eye Movement Patterns on Chinese Yellow Pages Advertising**, *Electronic Markets*, 2001, 11:2, 87-96
8. R. Zeff & B. Aronson, **Advertising on the Internet (2nd. ed.)**, *John Wiley & Sons, Inc.*, USA, 1999
9. Z. Jiang, S. Gao, & W. Dai, **Research on CTR Prediction for Contextual Advertising Based on Deep Architecture Model**", *CEAI*, 2016, Vol.18, No. 1, pp. 11-19
10. A. Severyn and A. Moschitti, **Twitter Sentiment Analysis with Deep Convolutional Neural Networks**, *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015, Pages 959–962 https://doi.org/10.1145/2766462.2767830
11. C. Kiss & M. Bichler, **Identification of influencers — Measuring influence in customer networks**, *Decision Support Systems*, Volume 46, Issue 1, December 2008, Pages 233-253
12. R. Krestel, P. Fankhauser, & W. Nejdl, **Latent dirichlet allocation for tag recommendation**, *RecSys '09: Proceedings of the third ACM conference on Recommender systems*, 2009 Pages 61–68 https://doi.org/10.1145/1639714.1639726
13. F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, **Word Cloud Explorer: Text Analytics Based on Word Clouds**, *2014 47th Hawaii International Conference on System Sciences*, Waikoloa, HI, 2014, pp. 1833-1842.
14. W. Wilbur, & K. Sirotkin, **The automatic identification of stop words**, *Journal of Information Science*, 1992, 18(1), 45–55. https://doi.org/10.1177/016555159201800106
15. C. Clifton, R. Cooley and J. Rennie, **TopCat: data mining for topic identification in a text corpus**, *IEEE Transactions on Knowledge and Data Engineering*, 2004, vol. 16, no. 8, pp. 949-964.
16. N. J. Evans, J. Phua, J. Lim & H. Jun, **Disclosing Instagram Influencer Advertising: The Effects of Disclosure Language on Advertising Recognition, Attitudes, and Behavioral Intent**, *Journal of Interactive Advertising*, 2017, 17:2, 138-149.

17. B. Reyck & Z. Degraeve, **Broadcast Scheduling for Mobile Advertising**, *Operations Research*, 2003, 51(4):509-517. https://doi.org/10.1287/opre.51.4.509.16104
18. G. Alcober, T. Revano, & M. Garcia, **E-Safety in the Use of Social Networking Application**, *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 9, No. 1.2 2020, https://doi.org/10.30534/ijatcse/2020/1291.22020
19. R. Dellosa, **An Efficient Position Estimation of Indoor Positioning System Based on Dynamic Time Warping**, *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 9, No. 1.2 2020, https://doi.org/10.30534/ijatcse/2020/0491.22020
20. J. Victoriano & L. Lacatan, **A Geospatial Analysis and Kernel Density Estimation of River Quality Parameter in Bulacan, Philippines**, *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 9, No. 1.2 2020, https://doi.org/10.30534/ijatcse/2020/1191.22020