IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

Analysis of the impact of social networking sites using web content mining and induction method

Analysis of the impact of social networking sites using web content mining and induction method

Renz Carlyle S. Bernados[1], Joshua O. Ty[2], and Angie M. Ceniza[3]

Department of Computer and Information Sciences, School of Arts and Sciences, University of San Carlos, Cebu City, Philippines

[1] bernadosrc@gmail.com, [2] joshuaty98@gmail.com, [3] amceniza@usc.edu.ph

Abstract. With the prevailing presence of social media, we cannot deny the fact that it influences the life of people. Usually teenagers tend to use this technology either to socialize or entertain themselves. In this paper, we have presented a framework to perform a qualitative analysis of social media networking sites using their social media reviews. The researchers gathered 503 links from Facebook, Twitter, Instagram and other related articles using web content mining. And, induction method was used to restrict and classify the harvested data. Researchers makes use of Word2Vec model that would find the similarity of one word to another. The experimental results show 81.39% precision, 67% accuracy, 70.93% recall and 75.58% F-Measure.

1. Introduction

Social media has gained popularity among young children, teens, and adults. Numerous social networking sites have emerged and has attracted many people into using these sites. The option to create accounts under a pseudonym or obscure username that separates users from their true identity may facilitate a sense of anonymity that allows for the disclosure of sensitive information on these sites [1]. Despite early concerns that adolescents might use the Internet to meet strangers, they use social media to interact with existing friends [2]. These social networking sites are useful as they can increase one’s communication with one another, provide viable information and entertainment. These resources can take the form of useful information, personal relationships, or the capacity to organize groups [3].

The researchers can see and read public posts of people post online on their profiles. The post can be photographs, videos or texts. Some of these posts are considered inappropriate that may cause harm to those who sees these posts and put them at risk to mental illness or may possibly have a good influence to those who had read them. As there are benefits we can gain from using these sites, there are consequences as well. The goal of this research is to determine what are the possible kind of impact these social networking sites have on the people using them. These posts are important to this research since this is the only way of knowing what their activities are or the social interactions of a certain individual to others. We can gather the data we can get from these posts or different sites that have stated comments towards the subject through web content mining and induction method and then we can classify these posts on what are the possible impact of the social networking sites. The only assumptions we can classify them is that they are either positive or negative influences based on the result of the study.

Multiple solutions are presented to solve the problem devised to give classifications of sentiments in text. Most modern machine learning relies on good feature engineering or other high level of domain knowledge to produce good result [4], but in deep learning, there is feature hierarchy which algorithms have to learn automatically, hence the researchers will use Word2Vec since it implements algorithm for that hierarchy and is also computationally efficient.

This research aims to classify the gathered data and determine which social networking sites have a positive or negative impact on the users. By applying the concept of induction mining, there are rules made to extract data easily from the web without having to much "noise" or unnecessary data. This would also help raise awareness to the users to be more careful using social networking sites. To give preferences on what the users would want to use and avoid the less preferred choice of network.

2. Related Works

2.1 Social Networking Sites

In 2016, Stankovska [5] described that majority of children and adolescents in the world have access to the internet at an early age. Most of them use the Social Network Sites for academic purposes and connecting with other people. But the excessive internet usage can lead to negative outcomes such as depression, loneliness, and poor performance in school.

Another study was conducted to relate adults and adolescents having low self-esteem [6][7]. Adolescent who frequently use an SNS have more friends on the site and also more reactions on their profile. Having likes, comments or reactions on one's profile leads to higher self-esteem. Leading to anxiety and depression, thus having a negative impact on one psychological well-being. Furthermore, adolescents tend to socialize more on the internet rather than in social groups leading to depend on social networking friends that there is also a huge percentage where some users are "troll" leading to miscommunication and also depression.

2.2 Web Mining

An existing research focuses on making an algorithm that would extract and find relevant information hidden across a platform is the Web Mining Model and Its Applications for Information Gathering [8]. The unique feature of this research is that instead of using standard web mining techniques, it creates a new algorithm to approximately extract hidden contents in user profiles in certain Web or platform hence an abstract Web mining model. With this feature, the researchers have developed an abstract Web mining model on the semantic Web to apply the abstract Web mining model for information gathering on the Web.

2.3 Induction Method

According to Ceniza and Maderazo [9], since collection of data on a pre-built system considers the domain specific information on a certain individual, the gathered data would then be altered for the user's behalf. Also, when gathering data, some of the data is in different format depending on the user's preference, meaning data might be not organized and it would be more difficult to extract and less time efficiency. Wrapper induction provide the technique to extract effectively the information and it is also easy to learn.

3. Methodology

This section discusses the methodology used in the research as well as the different frameworks used in order to achieve the desired objective of the study.

3.1 Gathering of Data

As shown in Figure 1, the research chose 3 different Social Networking Sites, mainly Facebook, Twitter and Instagram since these social networking sites are currently popular towards the current generations and may be for future generations based on the statistic survey of Statista [10]. Although some of the social network sites are much higher than what the research have stated, the researchers had considered that some of the much higher social networking sites are not available or less used in other countries.

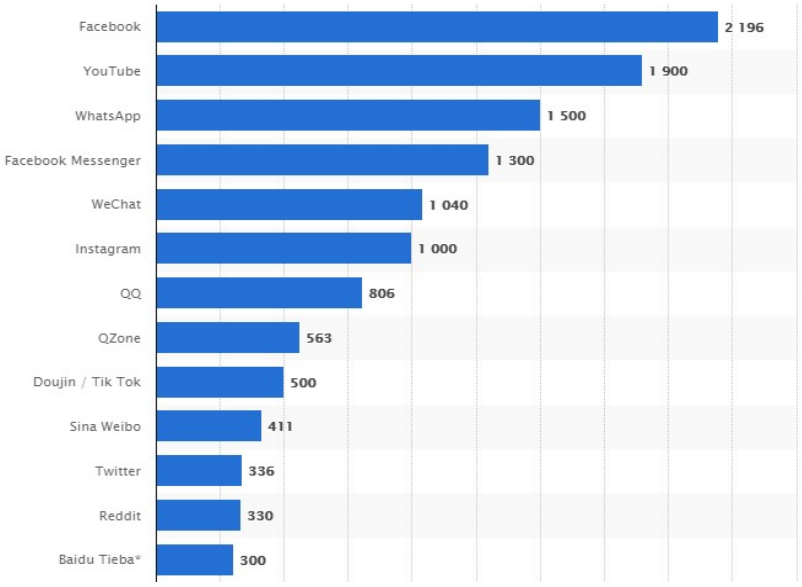

Figure 1. Most famous social network sites worldwide as of July 2018, ranked by number of active users (in millions)

The researchers make use of Google [9] as their platform for collecting data since Google Search Results provide a list of different reviews or topics relating to the subject to be researched. The data are then gathered using a scrapper tool, ScrapeStorm which extracts data or fields, after which those data would be then pre-processed based on the requirements needs of the research.

3.2 Pre-processing of Data

The data is pre-processed by removing unnecessary "stop-words" [11] such as "the", "is", "at", "which", and so on. A total of 709 reviews were used from Google Search Results, with 223(Facebook Reviews), 227(Twitter Reviews), and 259(Instagram Reviews). A word2vec related model was created and trained to identify word embeddings, since most data in the real world are unlabelled which aids in understanding the meaning of those words and also what other words have a relation to that word.

3.3 Continuous Bag-of-Words Model

The research make use of the extracted information based on the result of the pre-processing of the data and with the rules generated when gathering data. It generates "continuous bag-of-word", where in a bag of word model the occurrence of each word will be grouped and be classified as a positive or negative and with continuous bag of word model, there would be clusters of group having word analogies to be reproduced as vector math for prediction and translation. The cluster of words can also be classified as having the same word analogy depending on what the model generated based on the training set.

3.4 Forest Prediction

The research had used scikit-learn, a python library that has the classifier Random Forest in which it will use many tree-based classifiers to make a prediction based collected data. The clustering of the model for each words would be the basis for the many decision tree. Gathered data, pre-processing, sentiment and reactions are needed which aids in giving the prediction to data's whom are unlabelled as shown in Figure 2 below.

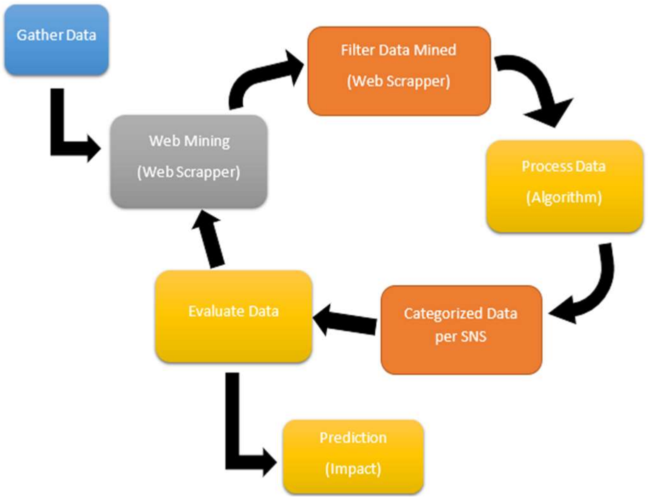

Figure 2. Conceptual framework of data gathering, training and prediction

The figure above shows the process of achieving the result of the study which is giving an analysis on the stated social networking medias. As shown as the title header of each sub-section on Section 3. Where the figure includes the gathering of data, then pre-processing or removing noise, then the classification and the evaluation of the said subjects.

Table 1. Word2Vec Results

Word	Most Similar Word	Similarity
Facebook	Twitter	0.878421
	Access	0.772450
	Media	0.724056
	Users	0.702237
Twitter	Facebook	0.878421
	instagram	0.776290
	Posted	0.689952
	Users	0.696966
Instagram	Twitter	0.776290
	Facebook	0.696815
	followers	0.654215
	Account	0.643138
Related Articles	Social	0.734260
	Public	0.723617
	propaganda	0.680053
	Society	0.705812

4. Findings and Discussions

Based on the data collected, mostly news media or different blogging sites have been collected during the process of data collection. However, not all these sites are giving bias to the said social media based on the different services the platform gives. Though Facebook, Twitter, and Instagram are mainly social networking services, these platforms also give news to different users all over the world. And with that news, the research could give an evaluation of the subject. Furthermore, within the training of the model, it also generated some various word analogy using the Word2Vec approach. Different words were evaluated but we chose those scored ranging from 0.60 and above as shown in the Table 1.

Based on the result on Table 1, we can see that there are words that has the same similarity or word analogy towards a subject. And we can get a general idea of what the analysis of each of the data gathered after training the model of the vocabulary. Table 2. Shows some of the review of each data gathered on each of the social media subjects. Considering the training and the model generated, if there are more negative than positive words from the content of each post, then that posts would then be considered negative and vice versa. Each word from each posts would be clustered to the model and shall be used for the Random Forest classifier.

From the sample result of Table 2, where each post is given sentiment. Table 3 shows the total analysis of the subject after the training had been completed showing the total percentage of each positive and negative words present on each of the social networking subjects corresponding to the final sentiment result of each subject.

Table 2. Harvested data with their corresponding review

Social Media Site	Content	Review
Facebook	Facebook executives have long called the company's top ranks a "family," but after a series of high-profile exits, the family is about to get a lot of...	1
	Section 230 of the Communications Decency Act (CDA), which provides platforms like Facebook with a broad shield from liability when a…	1
	Facebook says it has identified a covert Myanmar military propaganda campaign hosted on its platform, the first evidence that the country's...	0
	Research at Facebook just made it easier to translate between languages without many translation examples. For example, from Urdu to...	0
	Family of dead student criticise Facebook over 'sadist' troll	0
Twitter	A new Twitter update that allows people to see when other users are online has been criticised for leaving people vulnerable to cyber stalking.	1
	WASHINGTON: As Twitter CEO Jack Dorsey gets ready to testify before the US Congress on September 5, reports have surfaced that he is...	1
	This summer was an absolute mess, but at least there was one redeeming beacon of light on the internet: a Twitter thread of kittens growing up.	0
	People across the globe took to Twitter to find out if the issue had affected ... the social media platforms, thousands took to Twitter to complain.	1
	Twitter couldn't help but complain about the amount of adverts they have to endure when they are intently focused on watching the drama...	1
Instagram	Another wrote: "Hang on, Facebook Instagram and WhatsApp are all down." Another said: "#Facebookdown It's down. Instagram appears to be...	1
	My Instagram motto is 'parenting the s--- out of life': what could be more befitting than rear-wiping for comedic effect? It was my most-watched...	1
	A man named Nathan, from Cardiff, Wales, tweeted the picture of Scarlett's post, using it to argue that "Instagram is a ridiculous lie factory made...	0
	Being the guilty Instagram addict that I am, I was intrigued when the New York Public Library (NYPL) announced its latest project, Insta Novels.	0
	AN INSTAGRAM model found dead on a billionaire's luxury superyacht died just hours before a reunion with her family. Sinead McNamara, 20...	0

Table 3. Sentiment Result of Each Social Media Sites

Social Media Site	Positive Words	Negative Words	Total # of Data Gathered	Sentiment
Facebook	0.52470	0.47530	223.0	1.0
Twitter	0.75770	0.24230	227.0	1.0
Instagram	0.73360	0.26640	259.0	1.0

The figure in Figure 3.0 shows the graphical representation of each of the Social media's sentiment result of the gathered posts relating to them. In Facebook, 48% of the posts were evaluated to have a negative review, 52% were positive review and no neutral review was evaluated. Meanwhile in Twitter, 24% evaluated to have a negative review, 76% were positive and 0% were neutral. Lastly Instagram, 27% were evaluated as negative, 73% were positive and 0% were neutral.

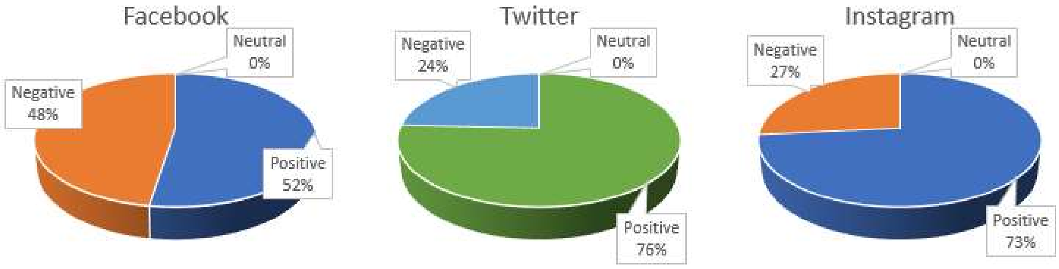

Figure 3. Sentiment result of the 3 social media sites

4.1 Result and Analysis

The measurement of performance evaluation is done by using precision and recall of the sentiment prediction.

$$\text{Recall} \quad = \quad \frac{\text{Number of Relevant Documents Received}}{\text{Total Number of Relevant Retrieved}} \tag{1}$$

$$\text{Precision} \; = \; \frac{\text{Number of Relevant Documents Received}}{\text{Total Number of Documents Retrieved}} \tag{2}$$

$$Accuracy = \frac{\text{Number of Relevant Documents+Numb} \quad \text{of Irrelevant Documents}}{\text{Number of Relevant Documents+To} \quad \text{Number of Documents Retrieved}} \tag{3}$$

$$\text{F Measure} = \frac{2* \text{ Number of Relevant Documents Recieved}}{2*\text{Total Number of Relevant Retrieved+Tot} \quad \text{Number of Documents Retrieved}} \tag{4}$$

The researchers gathered 709 links that have contents towards the subject social media. Each link was then evaluated of the occurrence of positive and negative words based on the model. We have found out that 503(170 – Facebook, 150 – Twitter, 183 - Instagram) links were relevant to the gathered data since some of those remaining links are spam links and are not related towards the subject matter. These data are pre-processed and are being clustered to be evaluated upon the research model. The experimental results show that 81.39% precision,70.93% recall 67% accuracy and 75.58% score

5. Conclusion

In this paper, we have presented a framework to perform qualitative analysis of social media reviews. Using web content mining and induction method to generate rules and harvest data to remove unnecessary links gathered and also reducing the noise for it to be computationally efficient. Word2Vec models shows the similarity or relatedness of one word to another. The clustering of the words to a continuous bag of word model based on the relatedness of each word, is then evaluated to the Random Forest to give result. The result presented were positive results about the public perception of each of the said social networking sites from its current users.

6. Acknowledgments

The researchers would like to extend our deepest gratitude towards the Department of Computer and Information Science and University of San Carlos for supporting this research endeavour.

References

[1] Marwick AE and Boyd D 2011 I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience *New Media & Society* vol **13** no **1** pp 114–33.

[2] Reich SM, Subrahmanyam K and Espinoza G 2012 Friending, IMing, and Hanging Out Face-to-Face *Developmental Psychology* vol **48** no **2** pp 356–68.

[3] Paxton P 1999 Is Social Capital Declining in the United States? A Multiple Indicator Assessment. *American Journal of sociology* vol **105** no **1** pp 88-127.

[4] Deng L and Yu D 2013 Deep Learning for Signal and Information Processing *Microsoft Research Monograph*

[5] Stankovska G, Angelkovska S and Grncarovska SP 2016 Social Networks Use, Loneliness and Academic Performance among University Students *Bulgarian Comparative Education Society*

[6] Ahn J 2011 The effect of social network sites on adolescents' social and academic development: Current theories and controversie *Journal of the Association for Information Science and Technology* vol **62** pp 1435-45

[7] Cavazos-Rehg P A, Krauss M J, Sowles S J, Connolly S, Rosas C Bharadwaj M, Grucza R and Bierut L 2016 An Analysis of Depression, Self-Harm, and Suicidal Ideation Content on Tumblr *The Journal of Crisis Intervention and Suicide Prevention*

[8] Li Y, Zhong N 2004 Web mining model and its applications for information gathering *Knowledge-Based Systems* vol **17** pp 207-17

[9] Ceniza A and Maderazo C 2016 An Analysis of Public Perceptions for K12 Implementation in the Philippines using Web Content Mining Techniques and Wrapper Induction Algorithm *National Conference on IT Education* pp 152- 57

[10] Statista 2018 Most famous social network sites worldwide as of July 2018, ranked by number of active users (in millions) *https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/*

[11] El-Khair IA 2006 Effects of Stop Words Elimination for Arabic Information Retrieval: A Comparative Study *International Journal of Computing & Information Sciences* vol **4** no **3** pp. 119-33