SENITMENT ANALYSIS ON WHO SOUTHEAST ASIA REGION ORGANIZATION (WHO SERO) USER COMMENT REVIEW AND OPINION MINING

[1]Thet Thet Aung, [2]Myat Mon Khaing, [3]Khin Shin Thant,[4]Hlaing Htake Khaung Tin
University of Computer Studies, Hinthada, Myanmar[1,2,3,4]
thetthetaung86.htd@gmail.com[1],myatmonkhaing01.htd@gmail.com[2], shinthantkhin10@gmail.com[3,]
hlainghtakekhaungtin@gmail.com4

ABSTRACT

At the present days, COVID-19 (Coronavirus) affections are spreading throughout the world from the beginning of affection of the town of Wuhan in China. Almost the whole world including many countries are facing difficulties. In Southeast Asia Region, schools, universities, industries, workshops, minimarkets and supermarkets are temporarily closed, and instructions are announced such as wash your hands, stay at home and use mask if people go outside on 24.3.2020. So people watch TV Channels, right sites on Social media and WHO information to know this disease's update news without going outside. This paper mines reviews of user comments on WHO Southeast Asia Region Organization (WHO SERO) page on Facebook media from 24.3.2020 to 31.5.2020. Sentiment analysis is a technique to classify user's opinion or emotions about specific domain and identifies phrases and emotions in a text of some sentiment. Sentiment analysis are classified as objective (facts), positive (happiness, satisfaction of the author) or negative (disappointment, dejection) using Information Gain (IG) method.

Index Terms— Opinion Mining, Sentiment Analysis, WHO SERO page, review comments, Facebook media, Information Gain (IG)

INTRODUCTION

Nowadays, the right and update news can be known via right pages, web sites of media. Media is defined as the forth aspect to catch the information all over the world. Media, especially in Facebook media, allows users to convey what they think and feel about each post such as Like, Comment and Share. This is called opinion mining. The aim is to determine the mood of the author or speaker's attitude and it may be positive or negative about these post. People express their emotions (positive or negative) and these positive or negative emotions are known as sentiment. One of the basic tasks in Sentiment Analysis is to predict the polarity of a given sentence, to find out if it expresses a positive or negative feeling about a certain topic.

This paper concentrates on sentiment classification in WHO SERO post domain and considers each post's comments from 24.3.2020 to 31.5.2020 as data set. Features selection in sentiment analysis is an important part. Some features cause the sentiment of analysis slow and less sensitive. Every single unique word and phrase can be assumed as the features. This system uses Information Gain (IG) based on feature selection. It also classifies the comments into positive and negative if most of the features are positive and vice versa.

The ongoing research work related to the Opinion mining and Sentiment analysis are given in this section. Firstly, retrieve the comments from WHO SERO every post as dataset according to each date. Secondly, preprocessing is classified into two parts: word segmentation, stop word removal. Thirdly, Opinion classification whether positive or negative review. Fourthly, present how to mine features in opinion sentences using Information Gain (IG). Finally, Opinion summarization id created depending upon the frequency of occurrences of features.

RELEAED WORK/ LITERATURE REVIEW

In this system, there are many functions. These functions are described.

A. Comments/ Dataset Preprocessing

In the preprocessing process, the following facts are described.

1) Extraction all the features from the given review

During 69 days, all comment reviews which are extracted as dataset before removing. There are about 30000 words in 5068 comment reviews.

2) Removing stop words

The stop words are useless in queries and meaningless with domain review. They can be safely ignored. They are pronouns, articles, conjunction, preposition, verb to be, verb to have and verb to do, etc.

Eg: I, We, They, He, a, an, the, when before, because, how, is , were, have, has, do.

3) Removing Account Name, Links, Phone Numbers, Diseases and Address

In the opinion mining of comment reviews on WHO SERO page of Facebook media, personal name, address, phone no, links, and other words and phrases not concerned with users' view, feelings and opinion can be removed.

4) *Replacing Root Word*

This paper replaces opinion words closet to the target feature because the sentence has the multiple features and distributed emotions.

Eg 1: we don't have anything to fight with COVID-19, We need help of our creator...I pray and hope. Feature = need, help, pray, hope

Eg 2: we are listening the statements from WHO. Request WHO to do something to save lives,, rather than advertisement or giving statement daily on Covid19.

Feature = are listening- listen (root word)

5) Preprocessing Positive, Negative and Neural

a. Example of Positive: good, luck, like, nice, respect, congratulations, bravo, welcome, save, need, help, hero, agree, pray, bless

b. Example of Negative: disappointed, dislike, fear, afraid

B. Feature selection using information gain

Information Gain (IG) is based on the decrease in entropy after a dataset is split on an attribute. Information gain (IG) measures how much "information" a feature gives us about the class. Features that perfectly partition should give maximal information. Unrelated features should give no information. Information Gain is calculated for a split by subtracting the weighted entropies of each branch from the original entropy.

Information Gain (IG) Formula

Entropy of class, H(C)

$$(C) = -\Sigma_{cEC}\,(C) \log p(C)$$

Conditional entropy of class, $(C \mid A)$

$$(C \mid A) = -\Sigma_{cEC}\,(C \mid A) \log p(C \mid A)$$

Information gain, $I\,(C \mid A)$

$$I\,(C \mid A) = H(C) - H(C \mid A)$$

C = one class either positive or negative

A = attributes

(C) = probability of class

$(C \mid A)$ = conditional probability of the class given attribute

1) C. Classification the extracted opinion words as positive, negative and neural

Posts may be one or many in each day, and total comments are extracted and total words modified are classified whether positive or negative or neural using Information Gain (IG).

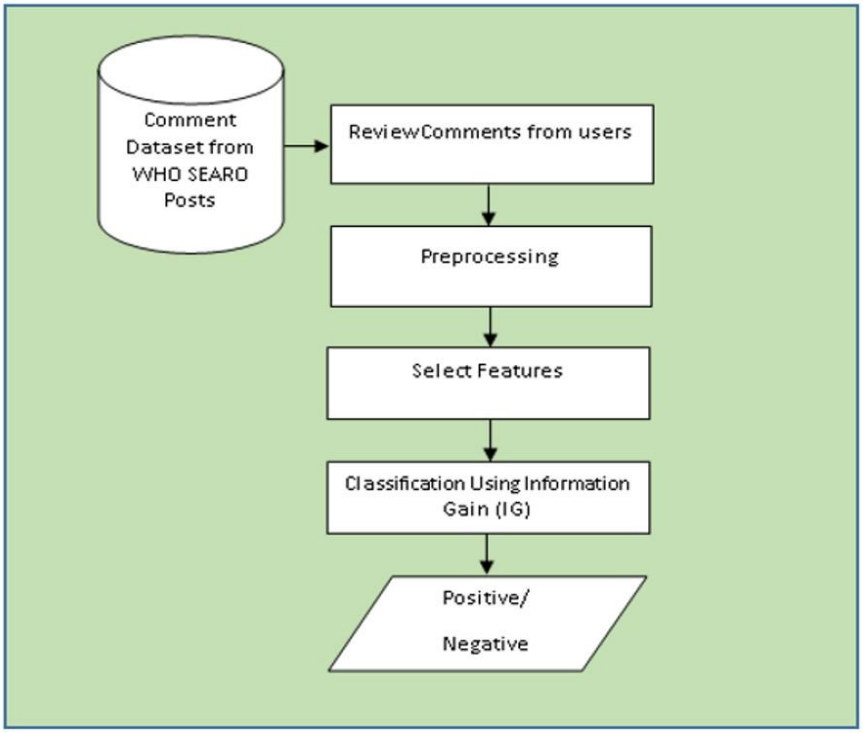

Figure 1: The System Flow Diagram

III. THE OPINION CLASSIFICATION METHODOLOGY

This paper analyses all posts concerned with COVID-19 on WHO SERO page of Facebook between 24.3.2020- 31.5.2020 during COVID-19 epidemics. All comments are used as dataset domain. There are 15 posts and 588 comments for 8days on March, 40 posts and 2857 comments for 30 days on April and 30 posts and 1623 comments for 31 days on May.

A. Retrieving Review Comments

In each month, dates, posts, comments and words are described with tables.

Table 1: Number of Posts and Comments in March

Month	Date	Post	Comment Total	Word Total
March	24.3.2020	6	248	413
	25.3.2020	2	42	539
	26.3.2020	2	98	401
	27.3.2020	1	55	308
	28.3.2020			
	29.3.2020	1	61	369
	30.3.2020	1	38	370
	31.3.2020	2	46	652
Total		**15**	**588**	**3052**

In Table 1, there are 15 posts in 7days, total comment are 588 and total word is 3052 in March. These data record are described.

Table 2: Number of Posts and Comments in April

Month	Date	Post	Comment Total	Word Total
April	1.4.2020	2	45	570
	2.4.2020	2	53	503
	3.4.2020	2	57	514
	4.4.2020	1	30	131
	5.4.2020	1	29	256
	6.4.2020	1	56	498
	7.4.2020	3	120	532

	8.4.2020	1	46	413
	9.4.2020	1	39	279
	10.4.2020	1	51	362
	11.4.2020	1	33	311
	12.4.2020			
	13.4.2020	2	41	202
	14.4.2020	1	65	470
	15.4.2020			
	16.4.2020	1	22	896
	17.4.2020	1	24	230
	18.4.2020			
	19.4.2020	1	23	268
	20.4.2020	1	24	299
	21.4.2020	2	184	330
	22.4.2020	1	34	279
	23.4.2020	3	683	602
	24.4.2020	3	98	233
	25.4.2020	2	617	696
	26.4.2020			
	27.4.2020	1	21	254
	28.4.2020	2	138	372
	29.4.2020	2	236	380
	30.4.2020	1	88	221
	Total	**40**	**2857**	**10101**

In Table 2, there are 40 posts in 30days, total comment are 2857 and total word is 10101 in April. These data record are described.

Table 3: Number of Posts and Comments in May

Month	Date	Post	Comment Total	Word Total

May	1.5.2020	1	94	780
	2.5.2020			
	3.5.2020			
	4.5.2020			
	5.5.2020	2	126	1003
	6.5.2020	1	36	696
	7.5.2020	2	85	487
	8.5.2020	2	107	920
	9.5.2020	2	78	1511
	10.5.2020			
	11.5.2020	1	38	312
	12.5.2020	1	77	747
	13.5.2020	2	140	1306
	14.5.2020	1	70	794
	15.5.2020	1	69	850
	16.5.2020			
	17.5.2020			
	18.5.2020	1	45	499
	19.5.2020			
	20.5.2020			
	21.5.2020	2	88	132
	22.5.2020			
	23.5.2020	1	54	335
	24.5.2020	1	69	390
	25.5.2020			
	26.5.2020	1	44	282
	27.5.2020	3	279	2075
	28.5.2020	1	44	787

	29.5.2020	2	45	1947
	30.5.2020	1	27	698
	31.5.2020	1	8	195
	Total	**30**	**1623**	**16746**

In Table 3, there are 30 posts in 31days, total comment are 1623 and total word is 16746 in May. These data record are described.

Table 4: Total count of word in positive and negative

No	Word	Count In Positive	Count In Negative	Total
1	good	7	3	10
2	luck	1	0	1
3	like	1	12	13
4	nice	2	1	3
5	respect	2	0	2
6	congratulations	21	0	21
7	bravo	11	0	11
8	bless	9	2	11
9	welcome	6	3	9
10	pray	6	2	8
11	agree	1	3	4
12	save	5	1	6
13	need	7	2	9
14	help	3	0	3
15	hero	11	0	11
16	happy	4	6	10
17	glad	1	0	1
18	care	6	0	6
19	dislike	0	5	0

20	fear	0	11	11
21	afraid	0	2	2
22	disappointed	0	6	6

In Table 4 describes the count of word as positive and negative in the comment sentiment.

Calculation of the ***entropy of class***

Assume we have balance comments,

(P) $= -\Sigma_{cEC}\, p(\mathrm{P}) \log p(\mathrm{P})$

$= 3/7 \log 3/7$

$= 0.4(-1)$

$= -0.6$

(N) $= -\Sigma_{cEC}\, p(\mathrm{N}) \log p(\mathrm{N})$

$= 3/7 \log 3/7$

$= 0.4(-1)$

$= -0.6$

(C) $= -\Sigma_{cEC}\, p(C) \log p(C)$

$= -(-0.6-0.6)$

$= 1$

Entropy of class, H(C)

$$(C) = -\Sigma_{cEC}\, (C) \log p(C)$$

Conditional entropy of class, $(C \mid A)$

$$(C \mid A) = -\Sigma_{cEC}\, (C \mid A) \log p(C \mid A)$$

IV. RESULTS AND DISSCUSSION

In this analysis, seven comments are used to evaluate the result of conditional entropy and information gain. These results are described as followed with table 5 and table 6.

Table 5: Result of Conditional Entropy

	Conditional Entropy H(P|A)	Conditional Entropy H(N|A)	Conditional Entropy H(C|A)
Like	-1.603083541825592	0.0014595232540116003	0.001299214899829041

Table 6: Result of Information gain

	Information gain I(P|A)	Information gain I(N|A)	Information gain I(C|A)
Like	1.0001603083541826	0.9985404767459884	0.9987007851001709

Classification

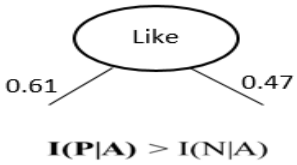

Figure 2: Classification

Positive Count = 3

Negative Count = 0

Result = this comment is positive

CONCLUSION AND FUTURE WORK

In this paper, mining process is performed by retrieving all comments from WHO SERO by preprocessing: word segmentation, stop word removal. Opinion classification using Information Gain (IG) are evaluated. Opinion summarization is created depending upon the frequency of occurrences of features. Finally, results are more positive than negative. So people in Southeast Asia become to know to obey WHO news and instructions according to the recent COVID epidemic and their moods and feelings become positive attitude. As the future work, announcements and instructions posted by WHO will be analyzed and evaluated for better.

REFERENCES

1. Santhosh Kumar K L, Jayanti Desai, Jharna Majumdar, “Opinion Mining and Sentiment Analysis on Customer Review”, 2016.
2. Yi-Ching Zeng, Tsun Ku, Shinh-Hung Wu, ”Modeling the Helpful Opinion Mining of Online Consumer Reviews as a Classification Problem”,2014.
3. Meenambigai B, “An Efficient Surveillanecs of Products Based on Opinion Mining”, 2014.
4. Yoosin Kim, Seung Ryul Jeong, Imran Ghani, ”Text Opinion Mining to Analyze News for Stock Market Prediction”, 2014.
5. Nidhi R. Sharma, Prof. Vidya D. Chitre, “ Opinion Mining, Analysis and its Challenges”. 2014.