Understanding Anonymous Social Media Posts using Topic Modeling

John Daniel M. Valencia, Al Joseph T. Laure, Niño Mark R. Centino, Bernie S. Fabito, Joseph Marvin R. Imperial, Ramon L. Rodriguez, Angelica H. De la Cruz, Manolito V. Octaviano Jr., Marilou N. Jamis
College of Computing and Information Technologies
National University - Manila
Manila, Philippines
{valenciajdm, laureajt, centinonmr}@students.national-u.edu.ph, {bsfabito, jrimperial, rlrodriguez, ahdelacruz, mnjamis}@national-u.edu.ph

***Abstract*— Social Media holds a substantial amount of text data that can help organizations better understand their clients. For students of National University (NU) – Manila, Facebook serves as a medium to express their opinions and create topics for discussion that may generally speak about the University. Through Topic Modeling using Latent Dirichlet Allocation (LDA), various experiments were conducted to identify the topics discussed by the students based on the highest coherence score value obtained. From these experiments, a total of twenty (20) topics with Alpha and Beta values set to one (1) revealed the highest coherence. The topics were labeled and revealed interesting insights. Personal relationships and school-related concerns were the common topics posted on the two Facebook pages. To further improve the study, a chronological approach for topic modeling is recommended.**

Keywords*—*Topic Modeling, LDA*, *Social Media*, *Freedom Wall

I. Introduction

Social Media holds a substantial amount of text data that can help organizations better understand their clients. For students of National University (NU) – Manila, Facebook serves as a medium to express their opinions and create topics for discussion that may generally speak about National University. As of writing, two Facebook group pages namely, the "NU Freedom Wall" and the "Nationalian Files" exist permitting the creation of various topics and sentiments to NU. Extracting information from these posts can likewise help the University understand its students thereby establishing better policies and services to its clients.

Topic Modeling is a powerful statistical machine learning approach that helps identify topics given a large set of corpus data [1]. Although graphical models have been previously used for topic discovery, the literature dictates inefficiency due to the noise and diversity brought about by social media posts [2]. In Machine Learning and Natural Language Processing, Topic modeling has been used to explore and unravel topics in a collection of text through the hidden parts emanating in the text corpus. Topic Modeling is helpful mainly in discovering hidden topical patterns present across the collection of data, annotate documents according to the topics, and use annotations to manage the texts [3].

For this study, an unsupervised topic modeling using the LDA algorithm was used to discover the topics created and discussed in the two Facebook pages. The objective is to explore the various topics that are shared among the students of National University from the text corpus obtained from Facebook posts. The Latent Dirichlet Allocation (LDA) algorithm is a "generative probabilistic model for collections of discrete data" [4]. The LDA is the simplest and most often used type of topic modeling algorithm [5].

In the literature, LDA has been used extensively for topic modeling [5] [6] [7]. In a recent study in the US, the LDA was used to automatically reveal consumer issues from the complaints submitted to the Consumer Financial Protection Bureau (CFP) while the extracted topics from the model revealed interesting chronological insights to the financial community [8]. Another study that has used LDA determined prominent topics in an online black market and how it evolved over time. The result has led to the conclusion that users of the dark web are increasingly becoming security-minded.

Though LDA is not efficient for Twitter due to its limited text [9], Facebook may surpass this limitation as users are able to use a bigger number of characters.

II. Methods

The methodology used for the study followed the following phases:

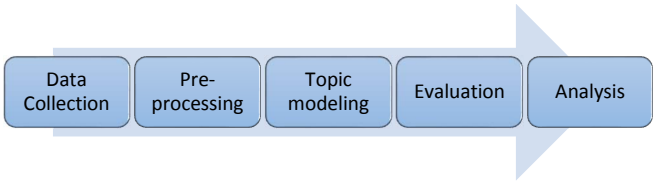

Fig. 1. Research Methodology

A. Data Collection

The data collection was made for the two Facebook pages namely, the Nationalian Files and NU Freedom Wall. Texts were collected manually from September of 2017 to January of 2019 and October 2018 to January of 2019 respectively. A total of 1,002 posts were collected for both pages. The posts collected contained English and Filipino texts.

B. Pre-Processing

Pre-processing includes data cleaning by removing unnecessary words that may affect the accuracy of the model. By using a regular expression, unnecessary words like English [10] and Filipino [11] stop words were removed. Subsequently, symbols, (e.g. number sign, dollar sign, percentage, parenthesis, asterisk) and extra spaces were removed. From a total of 127,225 words collected, it was reduced to 42,731 words.

C. Topic Modeling

Topic models can generate topics automatically based on important words from unlabeled texts in an unsupervised way. This method allows unstructured texts to become structured. However, the topics generated may not be guaranteed as interpretable. Hence, the need to use coherence measures to distinguish good and bad topics.

978-1-7281-3044-6/19/$31.00 ©2019 IEEE

Since the coherence score was the basis for the accuracy of the model, several experiments on the number of topics to be retrieved were conducted. The experiment started with five topics and ended up to 20 topics with an increment step of 5 topics (5, 10, 15, and 20).

After the cleaning stage, the experiment started using HDP-LDA. LDA goes through each of the words in each of the posts, randomly assigning the word to one of the topics. Though this random representation is not optimal, LDA betters this representation by analyzing the percentage of words within a post assigned to a topic. Subsequently, LDA will analyze all the posts by identifying the number of times (percentage) that word has been assigned to a topic. Hence, LDA will therefore calculate:

1.) p(topic t | document d) = percentage of words in document d (post) that are currently assigned to topic t

2.) p(word w | topic t) = percentage of times the word w was assigned to topic t over all documents (posts).

Aside from the K topics, the hyperparameters of LDA (Alpha and Beta) were also subjected to the experiments. Values were changed starting from a lower value moving up to a higher value to obtain the highest coherence score. For an asymmetric distribution, a low alpha value would mean that each document or post may likely to contain a mixture of most of the topics and not to a single topic alone. A higher value (greater than one) may mean that a post may contain a few or only one topic. [12].

Below describes the summary of the experiments conducted.

TABLE I. EXPERIMENTS

Topics	Iterations	Alpha	ETA
5…20 inc of 5	10,000…30,000 inc of 5,000	0.001…10 inc of (0.009 x 10)	0.2…1 inc of 0.2

The experiments were done using python with installed Gensim library. Gensim [13] is an open-source library for unsupervised topic modeling and natural language processing. Gensim contains known functionalities for topic modeling including LDA.

D. Evaluation

A total of 120 experiments were conducted for the study. The topic models generated were evaluated using coherence scores. The topic coherence measure is a good way to compare different topic models based on their human-interpretability. The K topics with the highest coherence score will serve as the basis for the result of the study.

Topics	Iteration	Alpha	eta	C_V
20	15,000	0.001	0.2	0.5105

E. Analysis

The analysis includes interpreting the accuracy of the models based on the coherence scores of the experiments. This is followed by interpreting and labeling the topics from the topic words.

III. RESULTS AND DISCUSSION

Tables 2 – 5 represents the results of the experiments conducted for each parameter of the LDA with the highest coherence score.

TABLE II. 1ST SET OF EXPERIMENTS (TOPICS AND ITERATIONS)

Topics	Iteration	Alpha	eta	C_V
20	20,000	1	none	0.4896

For the first set of experiments, the number of topics was selected starting with 5 topics to 20 topics with an increment of 5 for each experiment. Similarly, the iterations started with 10,000 up to 30,000 with an increment of 5,000. The alpha and Beta (eta) values contained default values. From the experiments, the 20 topics with 20,000 iterations had the highest coherence score (C_V) of 0.4896.

For the second set of experiments, the Beta (eta) was set from 0.2 to 1 with an increment of 0.2 and alpha retaining its default value. From the experiments, 20 topics with 20,000 iterations, and an alpha value of 1 gained the highest coherence score (See table 3). The third set of experiments included changing the eta values from 0.001 to 10 with the alpha set to its default value. The result with the highest coherence score can be observed in table 4. The fourth and last set of experiments combined altering the values of both the alpha and eta values. Table 5 shows the highest coherence value for the last set of experiments.

TABLE III. 2ND SET OF EXPERIMENTS (ETA)

Topics	Iteration	Alpha	eta	C_V
20	20,000	1	none	0.4896

TABLE IV. 3RD SET OF EXPERIMENTS (ALPHA)

Topics	Iteration	Alpha	eta	C_V
20	20,000	1	1	0.5198

TABLE V. 4TH SET OF EXPERIMENTS (ALPHA)

Topics	Iteration	Alpha	eta	C_V
20	30,000	0.01	none	0.4919

TABLE VI. 5TH SET OF EXPERIMENTS (ALPHA)

From all the experiments conducted, the coherence score with the highest value can be found from the second set of experiments (c_v = 0.5198). Looking at all the results, it seemed that the highest number of topics may yield contributory to a better coherence score.

From the experiment that generated the highest coherence score, the following are the topic words obtained for each topic:

TABLE VII. BAG OF WORDS FOR EACH TOPIC

K	Topic Words
Topic 1	0.001*"higher" + 0.001*"good" + 0.001*"kasalanan" + 0.001*"talk" + 0.000*"inip" + 0.000*"uaap" + 0.000*"nagbayad" + 0.000*"totoong" + 0.000*"future" + 0.000*"pipila"
Topic 2	0.002*"girl" + 0.001*"learned" + 0.001*"sarap" + 0.001*"problema" + 0.001*"sabay" + 0.001*"classmate" + 0.001*"mahal" + 0.001*"cr" + 0.001*"matalino" + 0.001*"nakakapagod"
Topic 3	0.001*"love" + 0.001*"tickets" + 0.001*"galit" + 0.001*"hate" + 0.001*"puke" + 0.000*"chu" + 0.000*"inuman" + 0.000*"sakit" + 0.000*"jacky" + 0.000*"kesa"
Topic 4	0.002*"mama" + 0.001*"grade" + 0.001*"mahal" + 0.001*"gf" + 0.001*"hold" + 0.001*"tagal" + 0.001*"final" + 0.000*"break" + 0.000*"juice" + 0.000*"pera"
Topic 5	0.003*"year" + 0.003*"student" + 0.002*"mama" + 0.002*"sabay" + 0.001*"subject" + 0.001*"population" + 0.001*"college" + 0.001*"grade" + 0.001*"anak" + 0.001*"prof"
Topic 6	0.011*"nu" + 0.002*"masaya" + 0.002*"college" + 0.002*"school" + 0.001*"cr" + 0.001*"students" + 0.001*"side" + 0.001*"engineering" + 0.001*"group" + 0.001*"university"
Topic 7	0.001*"jacky" + 0.001*"masarap" + 0.001*"lalong" + 0.000*"mukha" + 0.000*"staffs" + 0.000*"labi" + 0.000*"trisem" + 0.000*"freshies" + 0.000*"remaining" + 0.000*"enrollment"
Topic 8	0.001*"mahal" + 0.001*"gaano" + 0.001*"jacky" + 0.001*"deserve" + 0.000*"gitna" + 0.000*"ulo" + 0.000*"drta" + 0.000*"pta" + 0.000*"party" + 0.000*"pre"
Topic 9	0.005*"masaya" + 0.004*"araw" + 0.003*"mahal" + 0.003*"taong" + 0.002*"time" + 0.002*"buhay" + 0.002*"guy" + 0.002*"chat" + 0.002*"isip" + 0.002*"shirt"
Topic 10	0.002*"athlete" + 0.002*"nu" + 0.002*"ate" + 0.002*"lalaki" + 0.001*"aral" + 0.001*"university" + 0.001*"aaral" + 0.001*"magaling" + 0.001*"kamay" + 0.001*"room"
Topic 11	0.001*"nusg" + 0.001*"uweek" + 0.001*"center" + 0.001*"shit" + 0.001*"review" + 0.001*"inis" + 0.000*"fuck" + 0.000*"night" + 0.000*"athletes" + 0.000*"tamad"
Topic 12	0.001*"puso" + 0.001*"pinili" + 0.001*"barkada" + 0.001*"pagbagsak" + 0.001*"alon" + 0.000*"nagmamahal" + 0.000*"buhay" + 0.000*"pamilya" + 0.000*"guys" + 0.000*"happiness"
Topic 13	0.002*"sakit" + 0.001*"oras" + 0.001*"boyfriend" + 0.001*"confession" + 0.001*"bsa" + 0.001*"week" + 0.001*"kailan" + 0.001*"natutong" + 0.001*"girlfriend" + 0.001*"kwento"
Topic 14	0.002*"accounting" + 0.002*"pila" + 0.001*"units" + 0.001*"elevator" + 0.001*"magulang" + 0.001*"management" + 0.001*"enrollment" + 0.001*"man" + 0.001*"pumila" + 0.001*"bukas"
Topic 15	0.006*"time" + 0.004*"man" + 0.004*"school" + 0.003*"loob" + 0.003*"day" + 0.002*"sir" + 0.002*"love" + 0.002*"friends" + 0.002*"babae" + 0.002*"saya"
Topic 16	0.002*"kuya" + 0.002*"hope" + 0.002*"girlfriend" + 0.001*"kuyang" + 0.001*"professors" + 0.001*"bagsak" + 0.001*"year" + 0.001*"prof" + 0.001*"hirap" + 0.001*"sayang"
Topic 17	0.001*"jacky" + 0.001*"ateng" + 0.001*"freshmen" + 0.000*"boy" + 0.000*"anonymous" + 0.000*"unknown" + 0.000*"kim" + 0.000*"nu" + 0.000*"magreklamo" + 0.000*"magulang"
Topic 18	0.002*"concert" + 0.001*"hanap" + 0.001*"gwapo" + 0.001*"mundo" + 0.001*"admin" + 0.001*"prof" + 0.001*"talo" + 0.001*"good" + 0.001*"ben" + 0.000*"times"
Topic 19	0.004*"crush" + 0.001*"subject" + 0.001*"bakla" + 0.001*"team" + 0.001*"salamat" + 0.001*"nu" + 0.001*"nawala" + 0.001*"shs" + 0.001*"advance" + 0.001*"lalake"
Topic 20	0.002*"isip" + 0.001*"bahay" + 0.001*"online" + 0.001*"quizzes" + 0.001*"response" + 0.001*"school" + 0.001*"sarap" + 0.001*"tila" + 0.001*"grades" + 0.001*"ticket"

From the topic words, the researchers labeled them based on the story attributed to the words which were also verified by professors who have been with NU for the longest time. This is shown in table 8.

TABLE VIII. TOPIC LABELS

Topic Number	General Description	Label
1	UAAP Tickets	Sports Event
2	Unrecognizable pattern	Unrecognizable
3	Unrecognizable pattern	Unrecognizable
4	Relationship concerns with mother and girlfriend	Relationship
5	Grade concerns	Grade concerns
6	College life of engineering students in NU	College life
7	Lust for women	Lust
8	Unrecognizable pattern	Unrecognizable
9	An emotional state to a guy	Relationship
10	Student-Athlete	Student-Athlete
11	Disgust to athletes or events	Student-athlete concerns
12	Happiness and fall from a relationship with friends	Relationship concerns
13	Relationship confessions	Relationship concerns
14	Enrollment matters	Enrollment concerns
15	Unrecognizable pattern	Unrecognizable
16	The regret of getting failed grades	Grade concerns
17	Unrecognizable pattern	Unrecognizable
18	Unrecognizable pattern	Unrecognizable
19	Falling to a gay guy	Relationship Concerns
20	graded online exam/activity	Grade Concerns

From the given topic words, a total of six (6) from the twenty (20) topics do not have a recognizable pattern. Hence, only fourteen (14) topics have been labeled. Furthermore, it can be concluded that most of the facebook posts relate to personal relationships and services provided by the School. Topics also include frustrations of the students during enrollment and buying tickets for a regional sports event where NU is also a member. The sentiments and topics seem to be true for other Universities in Manila that have existing Facebook pages [14].

Algorithmically, the lower alpha parameter value would lead to some topic words to spread across other topics. For this, the researchers and the consultant faculty members had difficulty labeling the topics. Based on the experiments

conducted, it was realized that the lesser the alpha value, the more comprehensible the words are for each topic.

IV. RECOMMENDATIONS

The experiments conducted using LDA provided a glimpse of the topics discussed by the students of National University-Manila. The topics that were labeled provides an overview of the issues raised to the University. One notable topic may be on enrollment and grade concerns. This topic modeling can be a good tool to help NU provide better services to its clients.

Though the timeline to which the Facebook posts obtained were only for around two years, it would be interesting to know the pattern of discussion of students chronologically. Studying this could help address whether the topics are the same or do it evolve over time.

V. ACKNOWLEDGMENT

The authors would like to acknowledge the support provided by the NU - College of Computing and Information Technologies through the Human and Computer Interaction (HCI) Research Laboratory.

REFERENCES

[1] V. A. Rohani, S. Shayaa and G. Babanejaddehaki, "Topic modeling for social media content: A practical approach," in *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*, Kuala Lumpur, 2016.

[2] N. F. Rajani, K. McArdle and J. Baldridge, "Extracting topics based on authors, recipients and contents in microblogs," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, Queensland, Australia, 2014.

[3] G. Nair, "Text Mining 101: Topic Modeling," KDNuggets , July 2016. [Online]. Available: https://www.kdnuggets.com/2016/07/text-mining-101-topic-modeling.html. [Accessed 17 June 2019].

[4] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research ,* pp. 993-1022, 2003.

[5] L. Hagen, "Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?," *Information Processing & Management,* vol. 54, no. 6, pp. 1292-1307, 2018.

[6] K. Porter, "Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling," *Digital Invistigation,* vol. 26, pp. 587-597, 2018.

[7] H. Nabli, R. Djemaa and I. A. B. Amor, "Efficient cloud service discovery approach based on LDA topic modeling," *Journal of Systems and Software,* vol. 146, pp. 233-248, 2018.

[8] K. Bastani, H. Namavari and J. Shaffer, "Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints," *Expert Systems with Applications,* vol. 127, pp. 256-271, 2019.

[9] M. Hajjem and C. Latiri, "Combining IR and LDA Topic Modeling for Filtering Microblogs," *Procedia Computer Science,* vol. 112, pp. 761-770, 2017.

[10] "Alir3z4/stop-words," GitHub, [Online]. Available: https://github.com/Alir3z4/stop-words. [Accessed January 2019].

[11] "stopwords-iso/stopwords-tl," GitHub, [Online]. Available: https://github.com/stopwords-iso/stopwords-tl. [Accessed January 2019].

[12] "Natural interpretation for LDA hyperparameters," StackExchange, [Online]. Available: https://stats.stackexchange.com/questions/37405/natural-interpretation-for-lda-hyperparameters/37444#37444. [Accessed February 2019].

[13] "gensim: topic modeling for humans," gensim, [Online]. Available: https://radimrehurek.com/gensim/. [Accessed January 2019].

[14] "DLSU Freedome Wall: Speak your mind but be kind," TheLaSallian, 18 December 2018. [Online]. Available: http://thelasallian.com/2018/12/17/dlsu-freedom-wall-speak-your-mind-but-be-kind/. [Accessed 2019 January 2019].