Using LDA, SVM, Naïve Bayes Models In Filtering and Tracking Information on Social Networks

Toan Nguyen Chi[1], Trung Nguyen Hoang[2 *] and Nhi Yen Tran Thi[3]

[1]HUTECH – Ho Chi Minh City University of Technology, Thu Duc District, Ho Chi Minh City, Vietnam

[2]PVCFC, Ward 1, Ca Mau City, Ca Mau Province, Vietnam

[3]VLU – Van Lang University, District 1, Ho Chi Minh City

Abstract

Collecting and analyzing big amounts of data to extract useful data is a big challenge that we face in modern society. This presents great opportunities and challenges in the field of computer science research. If successfully analyzed and made sense of, data can help determine market trends, growth trends of an organization or stop the spread of information on social media. In this paper, the authors will conduct research on the theories of the Latent Dirichlet Allocation model (LDA), algorithmic the Gibbs sampling, Support Vector Machine (SVM), Naive Bayes theorem and the Waikato Environment for Knowledge Analysis (Weka). The authors also analyze and design the research system. This research will construct an empirical system to aid in the qualification and control of information on social media, detect implicit themes and potentially negative messages, trace the spreader of this news, and determine the speed of this news spreading. We aim to finalize a support system to aid with decision-making in the research that focuses on hot topics and development trends in the future and stop the spread of negative information and fix it.

Keywords: LDA, SVM, Naïve Bayes, Filtering Information, Tracking Information.

1. Introduction

Research on the spread of information and news filtering on social media has been conducted since the 2000s. In these researches, researchers have developed a viral marketing strategy and analyzed the effectiveness of this strategy through data. Developing the model to maximize the influence on social media (Influence Maximization) was considered also to be an optimization problem too. The first research focused on the independence cascades model that follows a linear threshold and came up with a general model for these two models. The research on Influence Maximization was conducted in the context of detecting an outbreak. The researchers especially looked into the node files on the internet to detect the outbreak as fast as possible. Information overflowing on mass media without careful deliberation is harmful, affects the mental health of people, causes social disharmony and public arguments. Developing an empirical system is important, and it allows the researchers to detect implicit themes, identify negative messages, contagion level through time.

The scientific journal focused on the LDA, SVM and Naïve Bayes model as well as Weka and organizational strategy and operating mechanism of social media. From this research, the researchers hope to develop an empirical system to aid in the categorization and control of the information flow on social media. Research plays a critical role in viewing social media, sharing

information in a more new way, especially in the context of the growing importance and influence of social media.

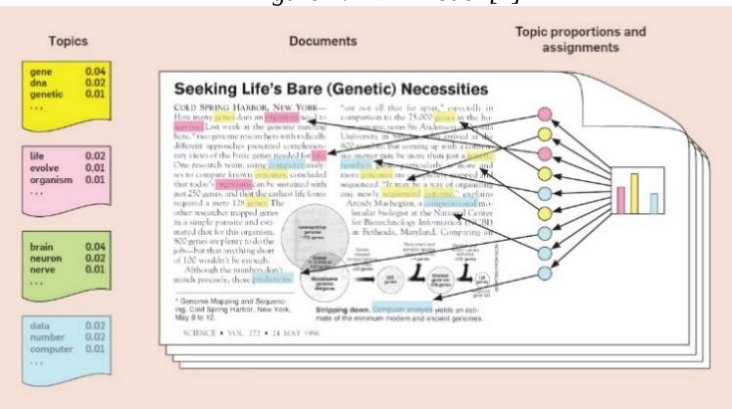
2. Theories

2.1 Theories

2.1.1 LDA model

The model [1] suggests finding an implicit theme for a dataset. This model is based on the idea that every dataset is a combination of different themes, and every theme is a combination of different words, every dataset is related to different themes with different probabilities and LDA essentially is a Bayesian model with 3 levels (dataset, text, and words). In that model, every part of the level performs as a finite mixed model on the grounds of the set of themes probability.

Figure 1: LDA Model [2]



Below are the steps to create a document from a list of themes. Every theme is a collection of specific words.

1) Determine the amount of N words in the document

2) Choose the amount of topics for the document based on polynomial distribution.

3) Generate the words for this document by:

- Pick the theme based on the polynomial distribution as determined above.

- Use the theme to generate words according to the probability of each theme as determined above.

Algorithm to find Gibbs sampling for LDA model

To find the theme from the text, we use posterior reasoning. These attributes are calculated through the 2.1 expression:

$$p(\theta, \emptyset, z|w, \alpha, \beta) = \frac{p(\theta, \emptyset, z, w|\alpha, \beta)}{p(w|\alpha, \beta)} \quad (2.1)$$

But in reality, we can not compute with exact certainty p(w|α,β) therefore we will use algorithm to gather the Gibbs sample [3]

Using an algorithm to find Gibbs sampling is one of the many algorithms of the Markov Chain Monte Carlo. This algorithm creates the Markov chain with steady posterior distribution. This means that with repetition, the sample from that distribution should be alike with the sample from the desired posterior statistical probability.

Getting the Gibbs sampling will depend on the sampling process of conditional distribution of the variables during this posterity statistical probability process.

The Algorithm is executed as below:

Determine the parameters for the algorithm:

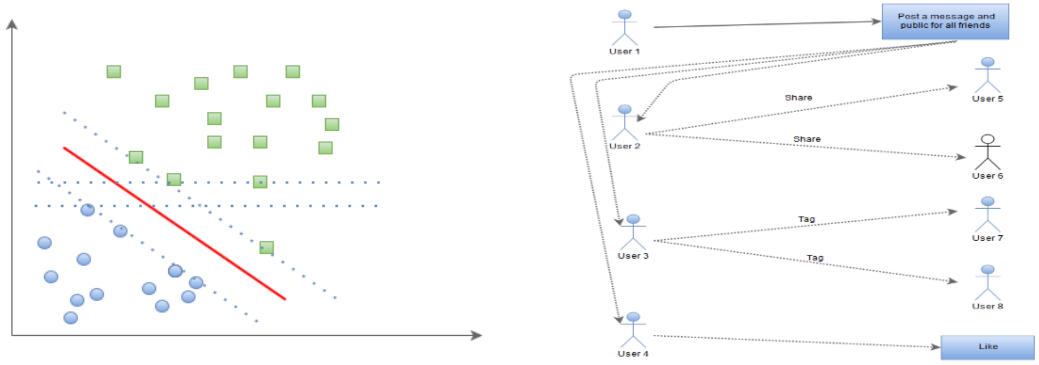
- D is the text folder
- d is a text document in that folder
- k is the number of topics
- w is a word
- $n_{d,k}$ is the number of words that gets attached to k amount of topics in the d document
- $n_{k,w}$ is the number of times the w word is attached to the k topic
- n_k is the total of times every single word is attached to k topic

Algorithm is created with variables that are counted randomly and are run in a loop with desired repetitions (usually from 1000-2000). In every loop, topics will be sampled one by one for words in each text document. At the end of the loop, an implicit distribution will be computed according to the counting variables.

2.1.2 SVM Algorithm

Support Vector Machines (SVM) is a categorization technique that originated from statistical study and the computer science field, and it is for a set of methodologies that are observed and related to each other in terms of categorization and regression analysis. The idea for this SVM algorithm is based on previous learnings in the vector space, every document will be a point. The idea of this technique is to map the dataset that needs categorization into a vector space in which the optimized hyperplane f will be found to categorize the data from 2 different layers.

Figure 2: SVM Model (left) and Post and share (right)



In Figure 2 (left), the bold red line is called the best hyperplane line. The points that are surrounded by rectangles are the points that are nearest to the hyperplane; they are called supporting vectors. The dashed lines are called borders. The purpose of this algorithm is to find the vector space F and the hyperplane f so that the margin of error is at the lowest.

Given the sample space **D={(x_1,y_1), (x_2,y_2), ……(x_i,y_i)}** with $x_i \in R^n$

With $y_i \in \{-1,1\}$ is the class label of x_i. (-1 is the label of "-", 1 is for "+").

We have a hyperplane equation that has the x vector in the space vector:

$$f(\overrightarrow{x_l}) = Sign(\overrightarrow{x_l}.\vec{w} + b = 0) = \begin{cases} +1, \overrightarrow{x_l}.\vec{w} + b > 0 \\ -1, \overrightarrow{x_l}.\vec{w} + b < 0 \end{cases}, with\ \overrightarrow{x_l}.\vec{w} + b = 0$$

$f(\overrightarrow{x_l})$ is expressed the categorization of $\overrightarrow{x_l}$ either as label + or –

We conclude: $\mathbf{y_i = +1}$ if $\overrightarrow{x_l}$ belongs to the class + and $\mathbf{y_i = -1}$ if $\overrightarrow{x_l}$ belongs to the class –.

2.1.3 Naïve Bayes Algorithm

The Naïve Bayes [4] algorithm is based on the Bayes theorem:

$$P(Y \mid X) = \frac{P(XY)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)} \quad (2.2)$$

Y is the hypothesis. Y is reached when we have evidence X.

P(X): probability of X happening.

P(Y): probability of Y happening.

P(X|Y): probability of X happening if Y happened first.

P(Y|X): posteriorly probability of Y happening if X is identified.

In the categorization algorithm:

D: dataset that is vectorized as $\vec{x} = (x_1, x_2, \dots, x_n)$

C_i: class label with i = {1,2,…,m}.

Attributes that are conditionally independent that parallels with each other.

According to the Bayes theorem:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (2.3)$$

According to the conditional independence:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad (2.4)$$

Assuming as above, the classification of new documents X^{dnew} = {x_1,x_2,x_3 ….. ,x_n} is identified as below:

$$\max_{C_i \in C}(P(C_i) \prod_{k=1}^{n} P(x_k|C_i) \quad (2.5)$$

- $P(C_i|X)$ is the probability of belonging to class i given sample X.
- $P(C_i)$ is the probability of belonging to class i.
- $P(x_k|C_i)$is the probability that an attribute with k order will have a value of x_k given that X belongs to class i.

Algorithm:

Step 1: Data training and compute $P(C_i)$ and $P(x_k|C_i)$

Step 2: Classify $X^{new}=(x_1,x_2,\ldots,x_n)$, compute the probability of each classification given X^{new}. X^{new} is assigned to the class with the highest probability based on the formula.

$$\max_{C_i \in C} \left(P(C_i) \prod_{k=1}^{n} P(x_k|C_i) \right) \quad (2.6)$$

2.1.4 Weka

Weka [5] is a data mining software that is researched and developed by Waikato University, New Zealand. It includes a collection of algorithms that helps with data mining. Algorithms might be applied directly into the data or provide an API that can be called forth with Java.

Weka has these main functions:

- A diverse tool to change, analyze, algorithmic and review data.
- User-friendly interface.
- An environment to compare algorithms

Environment:

- Simple SLI: simple interface to run Command Line commands.
- Explorer: graphic interface to do data mining.
- Experimenter: an environment to experiment and run statistics of machine models.
- Knowledge Flow: a drag and drop environment to execute the steps of the experiment.

2.1.5 Social Media Spread

Social media spread is the transference of information between one user to the next. The information could be status posts, news, or advertisements that companies aim at consumers.

2.1.6 Facebook Social Media

Right now, Facebook is using the Edge rank [6] algorithm to spread data.

$$\sum_{edges\ e} u_e\, w_e d_e \quad (2.7)$$

In the 2.7 formula:

- u_e: is measured through the relationship that you have with users that created the "Edge".
- w_e: these are the factors that you can easily exploit in the EdgeRank algorithm. There are 2 types of Weight: first is post (Photo, video, link, text, text + link + photo, …), two is interaction (share, comment, like).
- d_e: is the factor that determines the likelihood of your content appearing on the newsfeed.

2.1.7 Enron Email

Enron Email network has a specific structure that is used to exchange emails between users in the system. This exchange is executed through a number of actions: reply, reply all, send, cc, bcc, forward.

Once an email is sent then there are multiple actions you can take with that email: forward, cc, bcc, reply, reply all. In this situation, the spread in real-time is determined by the time at which these actions are taken.

Here are the steps to determine the time it takes to spread an email.

Step 1:

Initialize *listEmail*

Initialize listNode

Step 2:

listEmail ← Split input data to individual emails

Step 3:

For i = *listEmail.length*-1 → 0 **do**

Initialize count as number, t as datetime parameter

Initialize node list: *listTo, listCC, listBcc*

listTo ← Separate recipients (to)

listCC ← Separate recipients (cc)

listBcc ← Separate recipients (bcc)

t ← sending datetime

listChildren.add(*listTo*)

listChildren.add(*listCC*)

listChildren.add(*listBcc*)

For j = 0 → *listChildren.length -1* **do**

count ++

Initialize node = (count, t)

listNode.*add(*node*)*

Step 4:

Plot the propagation graph over time with listNode as input parameter

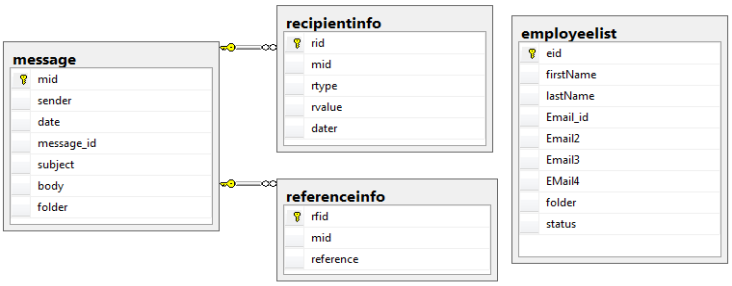
2.2 Analyze and design the system

2.2.1 Database of the system

Software uses the Enron Email network. A diagram of the Enron Email network is demonstrated in figure 3 Enron Email Database

- Elaboration on the connection between **Message – Recipientinfo:** Each message can be sent to many people by using different types of: cc, bcc, to. The **Recipientinfo** table has the ID of the message table.
- Elaboration on the connection between **Message – Referenceinfo:** Each message when getting sent to many people can be forward or reply, reply all. Besides the content of the message that the user wants to forward, reply, reply all, the email also includes the content of the previous message. The table **Referenceinfo** includes the ID of the table message.

Figure 3: Enron Email Database

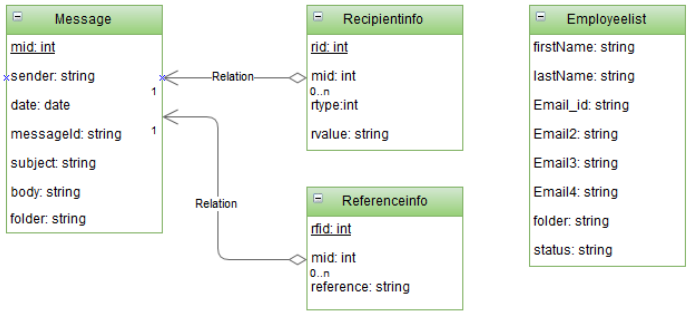


2.2.2 Model of objects

The system is analyzed to be object-oriented (ORM - Object Relation Mapping). To assist with the task of data tracing, the mapping mechanism is expressed as follows:

- With every data table, the system will create a corresponding Class.
- With every attribute of the table system, the system will create a corresponding attribute with the appropriate storage data type.
- With every data line, the system will create an object from a Class.
- With the data table that has a foreign key, the system will create a Collection to store.

Figure 4: Enron Email Database model

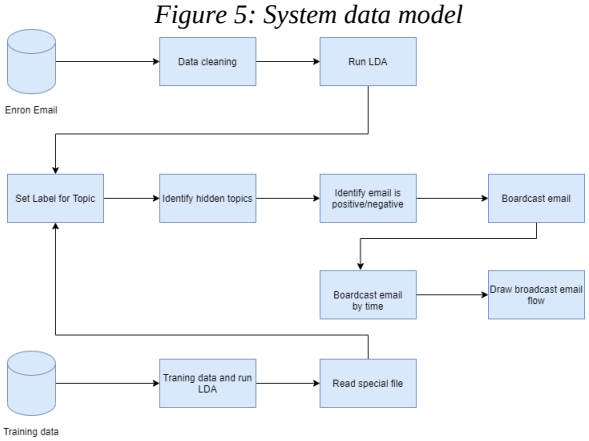


2.2.3 System model

3. The steps to execute the system in Figure 5: Run LDA on the dataset Enron Email. Run LDA on the dataset that is used for training. Label them and find the implicit topics of each email. Determine negative intent and the spread (of both recipient and sender). Determine the spread through time.

4. We end the process of running the LDA on the input dataset and based on the model-final.theta document, we can determine: words that belong in topics and their corresponding probability. One email belongs to topics and corresponding probability.

Figure 5: System data model



2.3 Label the topic and find the implicit theme of each email

The purpose of the labelling is to find the topic (topic of analysis) that fits with the topic that is in training. We will combine the index of Jaccard, Cosine, Overlap, Mutual, Dice and Tanimoto to label. *Output result of this step: determine the topic in section B corresponding to which topic in section A.*

Table 1: Topic and keyword

TopicId	TopicName	Keyword
1	Topic 1	couple|0.004447955671955152#pop|0.003906841843250146#lot|0.0037
2	Topic 2	witter|0.003349455192350456#mr|0.003349455192350456#series|0.00
3	Topic 3	#culture|0.004509571335078534#funding|0.004305055628272251#indu
4	Topic 4	3069#work|0.005330067219956023#thing|0.005277837261896679#thing
5	Topic 5	1#work|0.006086485219582221#photograph|0.0060132423168195#janu
6	Topic 6	13145598565#show|0.003510656905124809 8#published|0.00351065690
7	Topic 7	070351503#eyes|0.0034689332070351503#find|0.0032129233762576486
8	Topic 8	58305#funny|0.004378158471855830 5#bit|0.004298700060388393#joke
9	Topic 9	4550752#twitter|0.004995991474550752#royal|0.004995991474550752#
10	Topic 10	775#women|0.003285720322866923#american|0.003285720322866923#

After finishing this algorithm, we can find the corresponding topic.

Table 2: List of topics (left) and Identify topic (right)

Topic Id	TopicName
1	culture
2	economics
3	lifestyle
4	politics
5	society
6	technology

TopicName	TopicName
Topic 1	technology
Topic 2	technology
Topic 3	politics
Topic 4	technology
Topic 5	technology
Topic 6	society
Topic 7	society
Topic 8	technology
Topic 9	technology
Topic 10	society

2.4 Find implicit topic

The purpose of this step is to determine which email (on the trial dataset) corresponding to which topics (the list of topics that is in training). Input data is the LDA result and labelling

result. Output is successfully determining the main topic of each email. Pinpoint the probability of each email matches with a corresponding topic (topic of analysis). The one with the highest probability will be the main topic. After that, we can determine the topic (topic of analysis) matches with what topic (of training). We can determine an email that has which topic (training) based on transitive attributes.

Example: When an email has a probability, it will correspond with 10 topics that are in training when we run LDA to analyze our data.

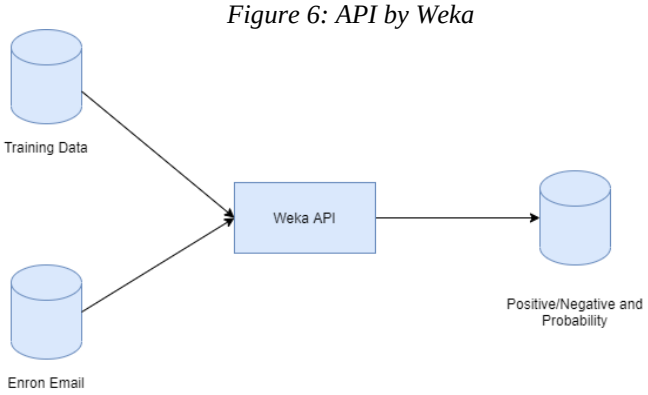
0.075	0.025	0.025	0.025	0.025	0.075	**0.425**	0.025	0.025	0.275

At this step, we will pick an email that is of **topic 7**. From that result, we can conclude that this email is about the topic **society.**

2.5 Determine negative message

To determine if a message is negative, we use the API that is given by WEKA. Input data will be the training dataset and the email dataset. Output result will be an email will either be flagged good or bad.

Figure 6: API by Weka



In the above picture, the training dataset and the Enron email data set are inputted into Weka. Weka will export the result that each email in the Enron email data set will either be positive or negative.

The training dataset that is put into training is collected from:

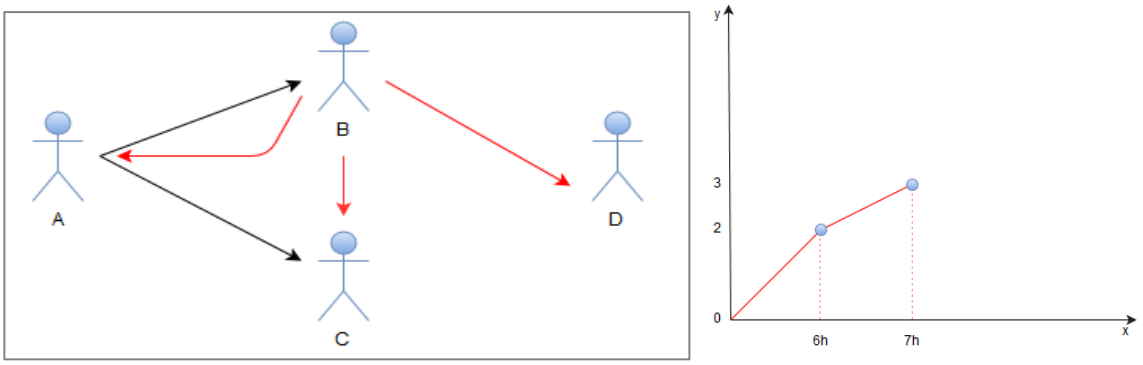
- http://www.newsnow.co.uk
- http://positivenews.org.uk

Finishing this step, we can determine whether an email is good or bad and its corresponding probability.

2.6 Determine the spread

Determining the spread is finding the first person who spread it, how many people has it reached, and who is continuing spreading it. In this part, we will trace the first spreader of the email. The input data will be the email list, and it will be trained. The output result will be the contagion model (based on spreaders) of each email. With the structure of the social media Enron Email, we can easily determine how the flow of information moves based on the history of an email.

Figure 7: Contagion model (left) and Spread throughout time (right)



2.7 Determine spread through time

Determining the spread through time is determining, in a timeframe, how many people receive one email.

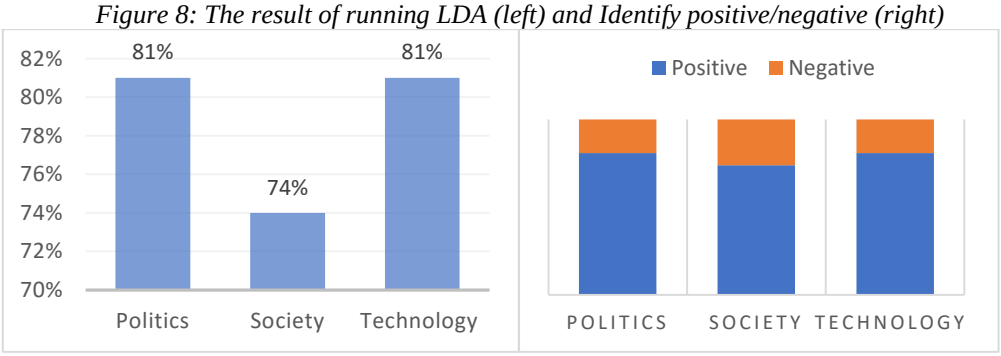
Example:

- At 6pm, A sends B and C an email
- At 7pm, B forwards to A, C, D

The above information will be expressed through the figure 10 (right).

2.8 Result

Figure 8: The result of running LDA (left) and Identify positive/negative (right)



The Enron Email dataset is the email dataset that is used for the CALO project. This dataset consists of 150 users and 517,431 emails. It is published by the Federal Energy Regulatory Commission. After that, Leslie Kaelbling has bought this dataset. After that, Jitech Shetty and Jafar Abidi cleaned up the data and converted them into MySQL. Running LDA on the Enron Email dataset. After finishing this step, we will have collected the collection of words that matched with each topic. We also determine the probability of an email belonging to a specific topic. After that, we run the LDA on the training dataset. At this step, we will do machine learning on 6 topics: Culture, Economics, Lifestyle, Politics, Society, Technology.

Read the unique document corresponding to each topic. The document consists of the unique meaning for a topic. Figure 12 (right) Determine the negative topic

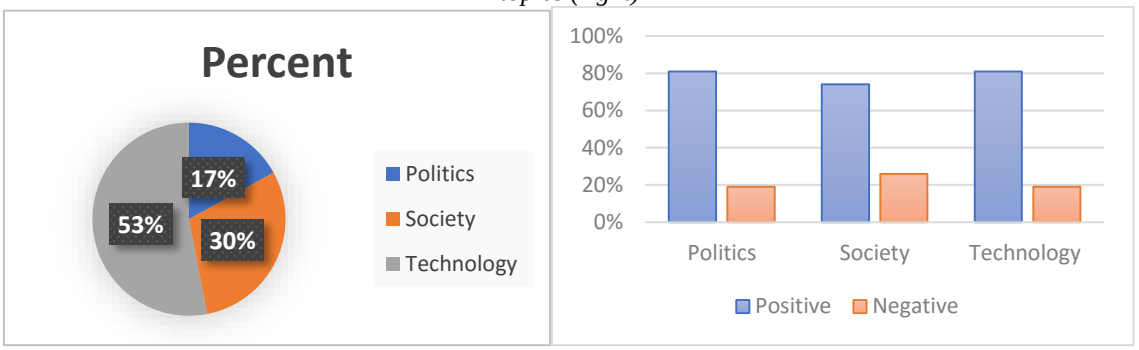
Label each topic. At this step, we will use the Jaccard, Cosine, Overlap, Mutual, Dice and Tanimoto index to determine the topic (out of a list of topics of analysis) that corresponds with the topics that are picked for training.

Table 3: Result after labelling

Mid	Content	Topic	Probability
79	Bob, I m able to open the document. Here it is again. Do you need ...	Technology	0.3
842	Robert,Per the email below, access granted to ERMS and TAGG. Pl...	Politics	0.40625
845	Monaco,Whats up with you? Great SuperBowl. I hate the Rams, so ...	Politics	0.5740741
851	Kevin,Here are the rates that we came up with. Please keep in min...	Society	0.5
852	Kevin,Thanks for the clarification. Can you tell me if the following as...	Society	0.6333333
855	Kevin:Per you request I left a phone message on 5/7/01 covering th...	Technology	0.28846154
869	Kevin,Dean McCallister said he would be the point for this project. Y...	Society	0.5416667
870	All, please see message below. James already has this in a datab...	Technology	0.43939394
875	Co. 0366 E.002780 - Alaska Pipeline Project has been set up to cap...	Society	0.36666667
881	Eric,FYI--One of the Iowa congresswomen told us BP was out in her...	Society	0.375
886	Bill,It sounds like we should proceed with separate meetings and th...	Technology	0.3482143
888	Scott,I received your attached letter reflecting your "out-of-scope" exp...	Society	0.41489363
889	Gentlemen, this is the cost estimate I was trying to send you previou...	Society	0.40625
891	No. Because of the problems that occurred with the swing loads, w...	Society	0.3125
892	Eric,There are a multitude of cases run on this project over time and...	Society	0.40384614
893	Ron,Thanks for the note. Several comments/questions-It is not nec...	Society	0.33673468
902	Not sure why they just sent it to me. At any rate, here is the estimate...	Politics	0.25
908	Eric,What we have done was use meteorological data from the inter...	Society	0.31976745
910	Station 1 on TW is about 30 miles east of Kingman, Arizona and is ...	Society	0.29545453
915	Kevin Howard is the best contact on item 1. wrt item 2, will APS prov...	Technology	0.45833334
916	The points listed below are not currently active on any IT of FT contra...	Society	0.2777778
920	FYI--Meant to include you in the original distribution. Thought you m...	Technology	0.30952382
929	KevinAs we discussed today, since we don t know anymore today th...	Society	0.25
934	Kevin:Sorry it has taken so long to get back to you on this. I have be...	Technology	0.64285713
939	Kevin,That is the scenario where we installed all 36" pipe and adde...	Society	0.45652175

2.9 Diagram of data statistics

Figure 9: Diagram of topics demonstrated as statistics (left) and Statistical diagram of negative and positive topics (right)



Data that consists of over more than 1000 emails is run and trained.

- 17% is about politics. With 81% positive and 19% negative.
- 30% is about society. With 74% positive and 26% negative.
- 53% is about technology. With 81% positive and 19% negative

The speed of the program: The input dataset consists of 1000 emails used for training and a training dataset that consists of 6 topics with 300 news for each topic. The program takes around 15 - 20 minutes. The bulk of the time is spent in data training. After we have the result, we can store it and the next time we run the program, it will only take 1 minute for 1000 emails.

5. Conclusion

The purpose of this scientific journal is to research the attributes of the models: LDA, Support Vector Machine (SVM), Naïve Bayes and Weka in determining if news is negative or positive. We also research about the structure and mechanism of social media.

From this research, the group wants to create an experimental system that aids in the categorization and control of information flow on social media and network environment of

universities in Vietnam. The experiment also helps: identifying implicit topics, determining if a message is negative, finding the spreader and determining the rate of spreading through time.

With that being said, the research still has some limitations in that we do not have sufficed dataset to implement the training with different topics. The insufficiency of the dataset prevents us from running the experiment, comparing the results in each dataset.

After achieving the initial results, the research team will continue researching and developing the system: allow doing training on more dataset like Twitter and Facebook, support Vietnamese data, build as model, independent API to support other systems, divide the steps of LDA running into smaller steps, labelling, and running at multi-process mode to speed up the process, change up the database system and utilize the cache mechanism to optimize algorithm.

After that, the research group, based on the knowledge and experiences accumulated, will develop a network that monitors information to manage forums, social media created by the school for students and lecturers. These platforms are created so learners and lecturers can discuss the topics they care about and exchange learning resources. From the data collected from social media, I will create a support system that facilitates research on topics that are most discussed, hosts Q&A events for students, prevents and fixes the spread of bad information.

References

D. M. Blei, "Latent Dirichlet Allocation," *Journal of Machine Learning Research 3 (2003)*, vol. 3, pp. 993-1022, 2003.

D. M. Blei, "Introduction to Probabilistic Topic Models," *Computer Science*, 2010.

W. M. Darling, (2011). "*A Theoretical and Practical Implementation, Tutorial on Topic Modeling and Gibbs Sampling*," School of Computer Science, University of Guelph.

D. Berrar, "Bayers' Theorem ang Naive Bayes Classifier," in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, Tokyo, Tokyo institute of technology, 2019, pp. 403-412.

Wikipedia, "https://en.wikipedia.org/wiki/Weka_(machine_learning)," [Online]. Available: https://en.wikipedia.org/wiki/Weka_(machine_learning). [Accessed 10 6 2018].

K. H. H. S. N. Dokyun Lee, "Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook," *Management Science*, vol. 64, pp. 4967-5460, 2018.

E. P. L. J. J. a. L. N. T. Byung-Won On, "Engagingness and Responsiveness Behavior Models on the Enron Email Network and Its Application to Email Reply Order Prediction," *The Influence of Technology on Social Network Analysis and Mining*, vol. 6, pp. 227-253, 2013.

A. Y. N. a. M. I. J. David M. Blei, "Research Collection School of Information Systems," *Latent Dirichlet Allocation. J.Mach. Learn. Res.*, vol. 3, pp. 993-1022, 2003.

R. C. Tolman, "Consideration of the Gibbs Theory of Surface Tension," *The Journal of Chemical Physics 16*, vol. 1, p. 55–353, 2004.

M. A. DeVito, "From Editors to Algorithms," *Digital Journalism*, vol. 5, no. 6, pp. 753-773, 2017.

T. Bucher, "Want to be on the top? Algorithmic power and the threat of invisibility on Facebook," *SAGE Journals*, vol. 14, no. 7, pp. 1164-1180, 2012.