

Engineering Project Proposal

2143491 – ICE Pre-project
Information and Communication Engineering
International School of Engineering
Chulalongkorn University

MEMORANDUM

November 19th, 2021

TO:

Dr. Pittipol Kantavat	Advisor
Asst. Prof. Sukree Sinthupinyo	Co-Advisor
Asst. Prof. Nattee Niparnan, Ph.D.	Committee Member
Asst. Prof. Nuttapong Chentanez, Ph.D	Committee Member

FROM:

Pattradana Punvichartkul	Pattradana
Pitawat Sangpaiboon	<i>Pitawat</i>

SUBJECT: Design Project Proposal Submission

Enclosed is our group's design project proposal, Detection and Evaluation of Personal Data on Social Media Platforms. This proposal is submitted in partial fulfillment of the Engineering Pre-Project requirement outlining the plan for the project pursued through the problem formulation with functional requirement, alternative solution generation with engineering approaches, project management and milestones, and task assignments and deliverables. We understand this proposal, in written report as attached, and oral exam scheduled with the committee, will undergo a rigorous assessment, and we are willing to accept recommendations from the committee and modify and resubmit for final approval.

Engineering Project Proposal

Detection and Evaluation of Personal Data on Social Media Platforms

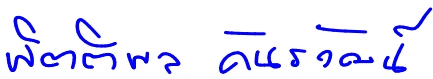
Submitted by

**Pattradanaï Punvichartkul
Pitawat Sangpaiboon**

Approved by:

Senior Project Advisor::

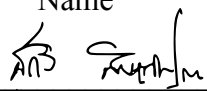
_____ Pittipol Kantavat _____
Name

_____  _____
Signature

_____ 19/Nov/2021 _____
Date

Senior Project Co-Advisor::

_____ Sukree Sinthupinyo _____
Name

_____  _____
Signature

_____ 19/Nov/2021 _____
Date

Detection and Evaluation of Personal Data on Social Media Platforms

1. Introduction

With the rise of the Internet era—a period that took place in the Information Age—in the 21st century, the communication of information and data has immensely increased. The world has become more interconnected than ever before by means of social media platforms, online e-commerce services, different types of entertainment platforms, etc. This is indeed leveraged by flows of information regardless of which type it is. Data collected from various sources allows corporates to conduct data analysis for improving their products and services in a meaningful manner. For instance, information about a customer's age, geography, or gender may be useful for predicting customer behavior within its platform in the future. Social media platforms (e.g., Facebook and Twitter) are capable of collecting critical information from social media posts on their platforms which may include personally identifiable images, a person's whereabouts, or behavioral preferences. Despite the benefits of data analysis obtained by businesses, it can potentially be done without customer consent and can jeopardize customer privacy imperative.

Protecting personal data is necessary to ensure no data will fall into the wrong hands which can further lead to the misuse of personal information. The unprecedented dangers of exploiting users' personal data can be devastating since the data can be used to commit crimes, frauds, and offenses. The social media post is a notable source for gathering personal sensitive data relating to a person because millions of people are using it daily whether to share information or to make communications. Therefore, a user-based tool for analyzing social media posts would be ideal for determining if it contains personal sensitive data.

The following proposal aims to construct the basis and the underlying of an application that manifests the above-mentioned tool. The goal is to demonstrate an attempt to reduce the harmfulness of data on social media platforms. Moreover, the application shall explicitly be designed for the platform users to grant them supervision over their own data. The application shall be capable of extracting data from social media posts of an individual. Specifically, the application shall retrieve and process text, image, metadata (e.g., timestamp and geography), and any information concerning the social media posts. Lastly, the scope of the application is merely within Facebook and Twitter.

To carry out the project meticulously, an effective software development methodology must be implemented; hence, knowledge and skills acquired from the Software Engineering course are essential. Furthermore, problem-solving capability, programming skills, data handling skills, and proficiency with technology stack taught in the Computer Programming, the Advanced Computer programming, the Fundamental Data Structure and Algorithm, and the Database Systems course are vitally important to build the application. In addition to that, the system will implement machine learning which requires an extensive understanding of the materials in the Advanced Mathematics Method course and the Probability and Statistics for Engineers course to formulate an appropriate model.

2. Problem Definition

Social media platforms such as Facebook and Twitter have inevitably become another necessity of life. They are used as a primary source of communication. This result is expected from an increased interconnection of globalization. Even so, many people around the world are not aware of the dangers that come with social media. This is likely because they lack the understanding and the awareness of their own data which brings about risks of disclosing confidentiality and personal data without proper precautions. In most cases, the incident takes place in the form of social media posts, comments, or images. For that reason, social media can be used to track down someone's identity, information, or personal data. Consequently, criminals and attackers can easily identify vulnerability and exploit social media users' personal data.

The types of information that criminals may use against the users can be categorized into three types: personal data (e.g., name, personal phone number, birth date, personal e-mail address, education, gender, occupation, photos, financial information, personal vehicle license plate information, family's names); sensitive data (e.g., ethnicity, race, political views, religious belief, sexual orientation, criminal record, health information, disability, labor union information, genetic information, and biometric information); and dangerous data (e.g., citizen ID, payment card number, bank account number, location, and PIN/password). This classification of data is referred from PDPA (Personal Data Protection Act) which is the law regulating data protection in Thailand. Each type of data comes with different kinds of dangers. For example, attackers may be able to abuse sensitive data such as ethnicity and religious belief for the sake of discrimination, dehumanization, extortion, unethicity, etc. For dangerous data, criminals may use it to commit crimes such as hacking the bank account or stalking a person's location. All of which can lead to property damage, physical assault, or even murder. Although personal data itself may not be prone to any harm, it can be treated as a clue to facilitate wicked acts along with other types of data.

To help prevent unintentional data leak on social media posts for the user, a system that can identify and classify the data within the scope of personal data, sensitive data and dangerous data is needed. Additionally, the system must have a qualitative metric that measures the level of harmfulness to let the user know. Ultimately, the system must follow the requirements of PDPA strictly.

3. Current Status of Research

One of the most common ways to classify texts into various categories is by using predefined rules. In this method, text data are categorized by detecting keywords which are in the word database. For example, if the text data contain the word "basketball," and "basketball" is in the database under the category "Sport," the data would be labelled "Sport." The advantage of this method is that it is easily comprehensible by humans. However, the disadvantage is that it is very time-consuming to write all the keywords for all categories. Since this project has 27 labels, it is infeasible to write all the keywords for each label.

Another way is to use machine learning to learn the pattern of the texts. By using labelled data as training examples, machine learning can relate data with the training data and categorized them with the labels that relate to them the most. The main advantage of this method over a rule-based algorithm is that it can learn implicit associations, so it can label unseen data better. This is more suited to this project, as the result will be more accurate compared to using rule-based text classification.

However, most implementations of text classifications can only be categorized with one label and in English only. Our implementation will support multi-label classifications since a post can have many kinds of personal information. Also, our algorithm can work in Thai language since our target audience are Thai people.

4. Engineering Approach

Python, which we learned from the Advanced Computer Programming course, is the main programming language we will use to extract and analyze the user's data. Selenium, a Python library for web scrapping, is used to extract the data from the social media platform. The user must first login to their social media page so that the application would be able to extract the data from each social media post.

Once the application has access to the user page, each of the user's posts is classified using machine learning. Each post can be classified with multiple labels, which are the types of personal information each post contains (refer to Problem Definition section for all types). Once all posts are analyzed, the user is shown with the posts that contain personal data, the types of personal information that each post contains, and the level of harmfulness of each type. We have not learned machine learning in the previous courses, so we have to make additional research on this topic to come up with solutions.

Just like the data extraction process, machine learning algorithm is done using Python. Because each post can contain many types of personal information, multi-label classification is required. To that end, scikit-multilearn, which is a Python library used for multi-label classification, is used to build the model. In multi-label classification, there are three methods that can be used to train the algorithm: problem transformation, adapted algorithm, and ensemble methods. Ensemble methods tend to give the highest accuracy of the three, so we will use this method as our main solution to the problem. However, in the case that we fail to apply this method, we can use either of the other two methods as the alternative.

4.1 Problem Transformation

In the problem transformation method, the multi-label classification problem is transformed into another type of problem, particularly the binary classification or the multiclass classification. In the binary classification, the data is classified as either "Yes" or "No." For instance, using the binary classification for the label "Name," the post is either labelled "Name" if it contains a name or not labelled if it does not contain any name. We can repeat this process for all labels. By converting into the binary classification problem, we can use well-known algorithms such as Naive Bayes or SVM to train our model.

In the multiclass classification, the data is classified with only one label. For example, if there are two possible labels, "Name" and "Birthdate," the post can be labelled with either "Name" or "Birthdate," but it cannot be labelled as both "Name" and "Birthdate" simultaneously. To be able to classify with multiple labels, we can combine multiple labels as another label entirely. For example, the "Name" label and "Birthdate" label can be combined as a separate label, "Name, Birthdate," so if the post contains only a name, it is labelled "Name," but if the post contains both the name and birthdate, it would be labelled "Name, Birthplace." As a result, we can transform the multi-label classification problem into the multiclass classification problem by finding all combinations of labels. Since we have 27 labels, there are a total of $2^{27} = 134,217,728$ combinations of labels possible. This is way too complex, so this approach is not feasible in this project. Thus, for the problem transformation

method, we will convert the multi-label classification problem into the binary classification problem.

4.2 Adapted Algorithm

Instead of transforming the problem into another kind, we can adapt the existing algorithms so that they can directly be used to train the multi-label classification model. One of the most used adapted algorithms is multi-learn k-nearest neighbors (MLkNN), which is adapted from the k-nearest neighbors algorithm used in a single-label classification. Scikit-multilearn has MLkNN and the other adapted algorithms built-in, so we can directly use them to train the model.

4.3 Ensemble Methods

Ensemble methods divide the label space into subspaces. Four different ensemble schemes can be used. The scheme that provides the highest accuracy will be chosen as the main method.

- **RakelD** – divides the label space into equally sized subspaces of size k . Problem transformation method (from multi-label to multiclass classification) is then applied to each label subspace. The results from each sub-classifier are then combined. Since there are 27 labels in this project, if we let $k = 3$, there would be 9 sub-classifiers in total, each with 3 labels, which would make the transformation into the multiclass classification problem finally feasible.
- **RakelO** – divides the label space into m subspaces of size k , which means a particular label can appear in multiple subspaces. At least more than half (i.e., majority) of the sub-classifiers that contain a particular label must assign that label to the data for the main classifier to assign that label to the data. For instance, if we let $m = 54$ and $k = 3$, a particular label, say “Name,” would appear in 6 subspaces, so there would be 6 sub-classifiers that can assign the label “Name” to a post. At least 4 sub-classifiers must assign the label “Name” to that post in order for that post to be labelled with “Name.”
- **LabelSpacePartitioningClassifier** – divides the label space into subspaces (label overlapping not allowed) just like RakelD, but instead of transforming each sub-problem into multiclass classification, other methods can be used.
- **MajorityVotingClassifier** – divides the label space into subspaces (label overlapping allowed) just like RakelO, but instead of transforming each sub-problem into multiclass classification, other methods can be used. Like RakelO, a particular label is assigned to the data only if the majority of the sub-classifiers containing that label assign the label to the data.

5. Tasks and Deliverables

There are four major tasks in this project:

- Develop a program to extract the user’s posts.
- Gather and label as many sample posts as possible.
- Develop a program to analyze the user’s posts.
- Develop the front-end application.

Out of these four, the second task requires the most time and effort since the accuracy of the prediction depends on the amount of the training data provided to the algorithm. Following are the details of each major task:

5.1 Data Extraction Program

In the first task, we need to develop a program that extracts the contents of the user's social media page so that they can be analyzed by the algorithm. Since our advisor already has this program developed, it will be provided for us. However, we may have to modify part of the program so that it fits more to the project's requirement.

5.2 Collecting and Labeling Training Data

Since the algorithm needs training data to learn the pattern of each type of personal data, we have to collect as many sample texts as possible and label them with appropriate personal data types. The more training data the algorithm receives, the more accurate the result is.

5.3 Text Analysis Program

Development of the analysis program can be divided into three phases: word vectorization, model training, and performance evaluation. First, the text from social media posts must be vectorized so that it can be understood and manipulated by the algorithm. Second, the classification model is trained using the labelled data (refer to Engineering Approach section for candidate algorithms). Finally, the algorithm is evaluated for accuracy using a test set. The algorithm can be tuned by changing hyperparameters and adding more training data until the accuracy reached a satisfactory level.

5.4 Front-end Application

The following items are tasks for the front-end application. First, a dedicated repository for the front-end application must be established. Next, the UX and UI are designed using a design tool such as Figma. After that, the development of the front-end application can be proceeded according to the UX flow and UI wireframe realized from the designing stage. The application must also be integrated with the post extracting and text analysis program, so an appropriate API must be developed. Finally, a user's manual must be documented as a soft copy and must also be accessible via the application.

5.5 Internal Deliverables

The internal deliverables are listed as follows: a Gantt chart to define timeline and milestones, a project scope statement included in the project proposal, a wireframe design of the front-end application, a user -Z flow in UX (User Experience) to visualize user's interaction with the front-end application, and a training data set for the machine learning algorithm.

5.6 External Deliverables

The external deliverables are listed as follows: an application to analyze user's social media posts—can be a web application or a mobile application, and a user's manual to give guidance to users on how to use the application.

6. Project Management

The following is the Gantt chart for this project. Pitawat and Pattradana will work together on all tasks. The project will take four months to finish, from January to April.

Month			January				February				March				April			
Week of month			1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Task	Start date	Due date																
Develop the program to extract user's posts	10-Jan	14-Feb																
Develop the algorithm for analyzing posts	14-Jan	31-Jan																
Find posts and label data	21-Jan	7-Mar																
Evaluate the algorithm performance	8-Mar	14-Mar																
Setup the frond-end repository	9-Mar	14-Mar																
Design UX and UI	11-Mar	22-Mar																
Finalize the UX flow	22-Mar	1-Apr																
Finalize the UI wireframe	23-Mar	7-Apr																
Develop the frontend application	7-Apr	21-Apr																
Connect the front-end application with the text analyzing and posts extracting program	16-Apr	24-Apr																
Prepare user's manual	25-Apr	28-Apr																
Finalize the analyzing application	28-Apr	30-Apr																

Python and its libraries do not have any costs, so this project does not require budgeting. Since we can develop the application using our own computers, no additional resources are required. Because user privacy is of utmost importance, the application will not store any of the user's information.

7. Conclusion

Because there are many people who do not aware of the dangers that come with social media, they are at risk of being the victim of a data breach which can lead to undesirable consequences. Social media posts are one of the targets that criminals and attackers find out information about a person and exploit it. To reduce the risk, the application that analyzes user's posts and evaluate the harmfulness of the data is introduced.

There are three possible methods of developing the algorithm for multi-label classification. The preferred method is ensemble methods since it gives the best accuracy. Ensemble methods divide the label space into smaller subspaces, which reduce the complexity of the algorithm. There are also two alternative methods: problem transformation changes the problem of multi-label classification into binary classification, while adapted algorithm uses the algorithm for single-label classification but modified so that it can work with multiple labels.

Since this project can be developed using Python and its existing libraries, it does not incur any costs. The entire project should take four months to be developed and be finished by the end of April.

8. References

Vajiralongkorn (2020) *Personal Data Protection Act* [e-book] Available from: http://www.ratchakitcha.soc.go.th/DATA/PDF/2562/A/069/T_0052.PDF (Accessed 8th November 2021).

Krungsri Plearn Plearn *มารู้จักกับกฎหมาย “PDPA” และการปรับโทษสำหรับผู้ละเมิด* [online] Available from: <https://www.krungsri.com/th/plearn-plearn/pdpa-law> (Accessed 8th November 2021).

How to deal with sensitive data [online] Available from: <https://www.openaire.eu/sensitive-data-guide> (Accessed 10th November 2021).

Australian Government *What is personal information?* [online] Available from: <https://www.oaic.gov.au/privacy/your-privacy-rights/your-personal-information/what-is-personal-information> (Accessed 10th November 2021).

(2017) *Sensitive Data and the GDPR: What You Need to Know* [online] Available from: <https://gdprinformers.com/gdpr-articles/sensitive-data-gdpr-need-know> (Accessed 12th November 2021).

Privacytrust *Whats the real purpose of the GDPR?* [online] Available from: <https://www.privacytrust.com/gdpr/whats-the-real-purpose-of-the-gdpr.html> (Accessed 12th November 2021).

Charles Kariuki (2021) *Multi-Label Classification with Scikit-MultiLearn* [online] Available from: <https://www.section.io/engineering-education/multi-label-classification-with-scikit-multilearn/> (Accessed 14th November 2021).

Shubham Jain (2017) *Solving Multi-Label Classification problems (Case studies included)* [online] Available from: <https://www.analyticsvidhya.com/blog/2017/08/introduction-to-multi-label-classification/> (Accessed 14th November 2021).

Asmaa M. Aubaid, Alok Mishra (2020) *A Rule-Based Approach to Embedding Techniques for Text Document Classification* [online] Available from: <https://www.mdpi.com/2076-3417/10/11/4009/htm> (Accessed 14th November 2021).

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, Donald Brown (2019) *Text Classification Algorithms: A Survey* [e-book] Available from: <https://arxiv.org/pdf/1904.08067.pdf> (Accessed 14th November 2021).

MonkeyLearn *Text Classification with Machine Learning & NLP* [online] Available from: <https://monkeylearn.com/text-classification/> (Accessed 15th November 2021).

Singapore Government (2021) *PDPA Overview* [online] Available from: <https://www.pdpc.gov.sg/Overview-of-PDPA/The-Legislation/Personal-Data-Protection-Act> (Accessed 16th November 2021).

Abi Tyas Tunggal (2021) *What is Sensitive Data?* [online] Available from: <https://www.upguard.com/blog/sensitive-data> (Accessed 17th November 2021).