A MUSIC VIDEO RECOMMENDER SYSTEM BASED ON EMOTION CLASSIFICATION ON USER COMMENTS

BY

PHAKHAWAT SARAKIT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF MASTER OF
ENGINEERING (INFORMATION AND COMMUNICATION
TECHNOLOGY FOR EMBEDDED SYSTEMS)
SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2015

A MUSIC VIDEO RECOMMENDER SYSTEM BASED ON EMOTION CLASSIFICATION ON USER COMMENTS

BY

PHAKHAWA SARAKIT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF MASTER OF
ENGINEERING (INFORMATION AND COMMUNICATION
TECHNOLOGY FOR EMBEDDED SYSTEMS)
SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2015

A MUSIC VIDEO RECOMMENDER SYSTEM BASED ON EMOTION CLASSIFICATION ON USER COMMENTS

A Thesis Presented

By

PHAKHAWAT SARAKIT

Submitted to

Sirindhorn International Institute of Technology

Thammasat University

In partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING (INFORMATION AND COMMUNICATION TECHNOLOGY FOR EMBEDDED SYSTEMS)

Approved as to style and content by

Advisor and Chairperson of Thesis Committee ____________________
(Prof. Thanaruk Theeramunkong, Ph.D.)

Committee Member and
Chairperson of Examination Committee ____________________
(Choochart Haruechaiyasak, Ph.D.)

Committee Member ____________________
(Assoc. Prof. Ekawit Nantajeewarawat, Ph.D.)

Committee Member ____________________
(Prof. Manabu Okumura, Ph.D.)

APRIL 2016

Abstract

A MUSIC VIDEO RECOMMENDER SYSTEM BASED ON EMOTION CLASSIFICATION ON USER COMMENTS

by

PHAKHAWAT SARAKIT

Master of Engineering, Sirindhorn International Institute of Technology, 2015

Along with the concept of collaborative intelligence, user comments are a useful information source for adding more value on online resources, such as music, video, books and other multimedia resources. While several works have been conducted to utilize user comments for sentimental analysis, i.e., LIKE or UNLIKE, there are still few exploitations of such comments for detecting emotion (user mood) on online resources. With emotion recognition, it is possible for us to understand the content of online resources and use the recognition result for further value added service, such as product recommendation. This thesis proposes a two-step method to perform emotion classification using user comments and utilize the result for music video recommendation.

In the first step, the emotion filtering tags user comments with three label types of emotional comments, non-emotional comments, and unrelated junk comments. As the second step, the emotion classification aims to classify the emotional comments into six emotion types, including anger, disgust, fear, happiness, sadness, and surprise. With the YouTube API, the total of 85 video clips with 12,000 comments are collected and used for emotion filtering and classification. The emotion filtering detects that 5,345 comments are emotional comments and the emotion classification categorizes them into six emotional classes using 7,722 features (word types) extracted from the dataset. For the classification method, three alternative machine learning algorithms are considered; (1) multinomial naive Bayes (MNB), (2) decision tree induction (DT), and (3) support vector machine (SVM), where the best SVM method obtained at most 76.41% accuracy for the filtering task and 75.68% for the classification task. However, with error analysis, it was found that such low accuracies may be caused by class imbalance where happiness and sadness classes suppress other remaining classes.

To improve performance of the emotion classification using unbalanced data, a sampling based algorithm called SMOTE (Synthetic Minority Over-Sampling Technique) is applied. With this preprocess, the SVM classifier yielded the best classification accuracy of 93.30% on the emotion filtering task and 89.43% on emotion classification task. The result improves by 16.9% and 13.76% for filtering and classification, respectively. Finally, a music video recommendation system is developed by suggesting a set of video clips which have a similar emotional profile with the currently focused clip.

Keywords: Emotion Classification, Machine Learning, Text Analysis, Text Categorization

Acknowledgements

First of all, I would like to express my deepest appreciation to my advisor, Prof. Thanaruk Theeramunkong from Sirindhorn International Institute of Technology, Thammasat University (SIIT) and my co-advisor, Dr. Choochart Haruechaiyasak from National Electronics and Computer Technology Center (NECTEC) who provided me the possibility to complete this thesis. Their guidance helped me a lot in all time of study and writing of the thesis.

Besides my advisor and co-advisor, I would like to thank my thesis committees: Assoc. Prof. Ekawit Nantajeewarawat from SIIT and Prof. Manabu Okumura from Tokyo Institute of Technology for their useful comments and help me to improve my research work.

Last but not least, I would like to thank my family for their encouragement and supporting me, my friends especially ICTES batch 6 who help me to pass any problems and gave me a lot of moral supports.

This research has been financially supported by Thailand Advanced Institute of Science and Technology (TAIST), National Science and Technology Development Agency (NSTDA), Tokyo Institute of Technology, National Research University Project, Thailand Office of the Higher Education Commission, and Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU) under the TAIST Tokyo Tech program.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

Nowadays, internet and social network usage are rising. We cannot deny that social network is a part of our life, such as communication, E-commerce, work or activity in daily life. Moreover, the internet usage factor is rising because the price of electronic devices, especially smart phone is cheaper. Everyone, even children or adults, can own the devices. Highly popular activities of internet users is social networking then searching and reading respectively. According to a survey conducted by ETDA [45], top five most used social media in Thailand are Facebook, Line, Google+, Instagram and Twitter, respectively.

While so far music has extracted people's interest by the channels of radio, music player, television, and web music, recently the YouTube, the most popular media sharing, has become an alternative channel for music listening. YouTube is a kind of media sharing websites, which provide online video services. Users can upload, watch, share, and comment the videos, the content of which can be a music, television show, advertisement or tutorial videos.

Along with the concept of collaborative intelligence, user comments are useful information source for adding more value on online resources, such as music, video, books and other multimedia resources because user comments can tell us many things such as, complacence, user emotional or opinion. While several works have been conducted to utilize user comments for sentimental analysis, i.e., LIKE or UNLIKE, there are still few exploitations of such comments for detecting emotion (user mood) on online resources. With emotion recognition, it is possible for us to understand the content of online resources and use the recognition result for further value added service, such as product recommendation.

1.1 Social media content

Social media is a form of communication on the internet. It allows users post and share their activities; an experience to the public or to a group of people in their network. Nowadays, social media play an important role in our daily life and business, especially to promote product and online marketing activities.

Social media can be divided into several categories such as

1. Blog : Blogger

   The Blogger is a popular type of websites, which support diary writing. Two main features of blogs are reader-supported and writer-supported systems, where readers and writers are allowed to provide comments and share their experience via the shared blackboard concept. Several existing works studied a method to classify emotions in text by using web blog as a dataset. Inrak and Sinthupinyo [15] proposed an approach to utilize of bi-words occurrence to classify emotion that hiddens in a short sentence. Nivet [16] proposed emotion classification of Thai texts using term weighting and machine learning algorithms.

2. Microblog : Twitter

   The Twitter have a similar structure to the blog, but smaller. It allowed users to type a short message with 140 characters at must, called "Tweet". Recently a number of researchers have purposed techniques to classify emotion on Twitter [3, 4, 5, 6]. Due to a short message for one Tweet, the Tweet message is hard to be classified its emotion with wrong meaning or no meaning as output.

3. Social Networking : Facebook

   The Facebook is a popular social network in Thailand. These style support users to connect with other users. It facilitates and promotes user communication during activities with other users; post comment, picture, write article or live chatting with friends.

4. Media Sharing : YouTube

   The YouTube is a popular video sharing website in the world. Users can upload, watch and share the video to the people at no cost. Moreover, users can rate the video i.e., LIKE or UNLIKE and give the comment express their feelings to the video. We have some works conducted to utilize user comments for sentimental analysis. Orimaye et al. [2] detected sentiment on Yoruba movies using YouTube comments.

5. Social News and Online Forums.

One interesting and useful information source on the web is online news. Its content can be varied from simple news to complicate news. The news can be national and international. Most news is a text and video. Some news website allowed users give their comment to the news item. There are several works detecting emotion on headline news [11-12], but the experimental results separate into two ways depend on reader i.e., global oil prices drop in 2014, this news might be a good news for buyer, but bad news for seller.

1.1.1 YouTube

YouTube is a social media of a type of media sharing. It was founded in February 2005, to allow billions of people to discover, view, comment, give a feedback and share videos. YouTube has forum for people to connect to each other, share and inspire others around the world, as well as serve as a platform for video publishing by original content creators and advertisers. According to statistics of website, YouTube is a video service web that have viewer more than 100 million times per day or equivalent to about 29% of all videos viewed in US. Each month there are more than 65,000 videos has been uploaded. [1] YouTube are suitable for any kind of marketing whether for products, services or business by creating their network for their clients by using VDO as a main key. Keeping a lot of data publicly, this platform provides a new opportunity to add value on the online resources, services, product such as automatic emotion classification of a media clip and product recommendation.

Figure 1.1: An example of music video in YouTube

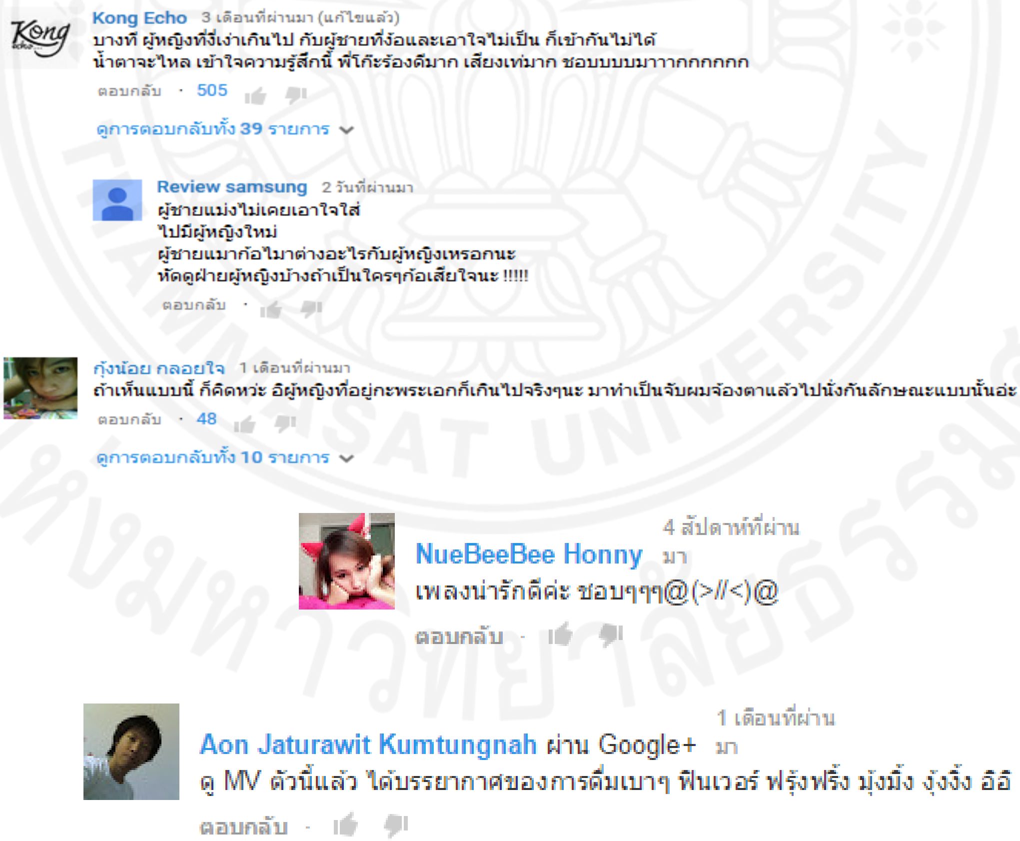


Figure 1.2: An example of user comments

Figure 1.2 shows and example of user comments on YouTube that express their own emotion to the music video clips they had listened. Some of those feel happy when they are listening the music. Some of those feel sad and so on.

1.2 Motivation

With raising usage of social media, automatic sentimental analysis triggers a chance of a company to catch feelings of its customers or prospective customers about its products and services by analyzing their reviews. YouTube have many ways to recommend music video clips, for example by top user views, tags, newest video, similar content or same artist, but a recommended music may have different mood. YouTube has not recommend similar emotion music video to user yet (e.g. happy music, sad music and so on.) Recommending music video by tags should be the most accurate than other ways, but it needs users to give the tag. Disadvantage of this recommendation is if a user give wrong tag, recommendation will wrong too. By this background, this research aims to study a method to classify emotion on text rather than three basic sentiment classification types: positive, negative and natural. Generally music recommender system may recommend music based on band, genre, record label and lyrics. According to our hypothesis, if we use incorporating user comments to recommend music video. It will more effective than use lyrics to give actual emotion perception.

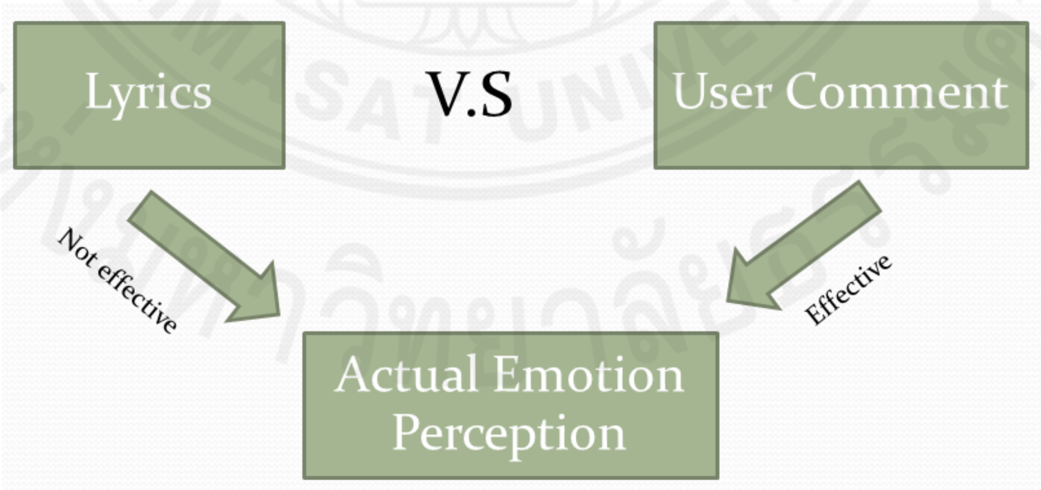


Figure 1.3: Hypothesis

Figure.1.4 shows an example song named "ผีหลอก" (Ghost haunt) performed by Magenta, the lyric describes the person in lyric faced with ghost haunting every day and he is scared ghost. At first glance, this song occupies fear emotion. However, the comments showed some contrast with its lyric. User feel happy and fun not fear in this song. Another example, one music sung by multiple singers might have different mood from the original version. Therefore, using incorporating user comments are the one choice for gives better music recommendation than old recommendation technique.

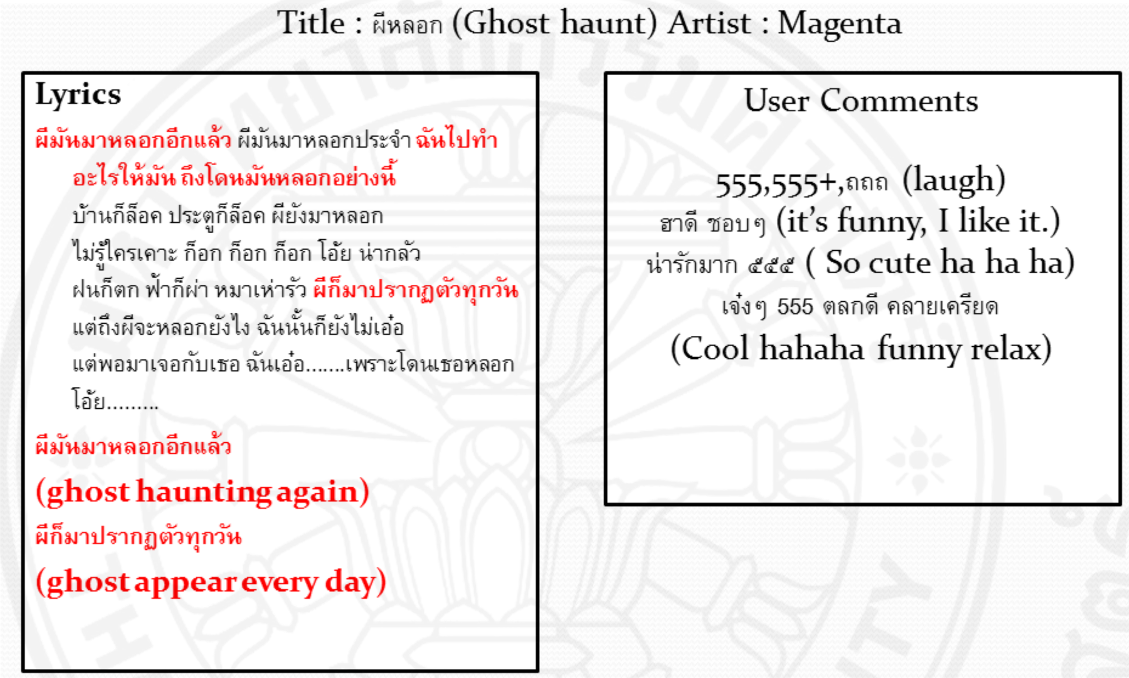


Figure 1.4: An example of lyrics

1.3 Objective

There are three objectives of the thesis:

1. To compare algorithms on classifying emotion from YouTube comments.
2. To compare performance of different features on constructing an emotion classification model.
3. To implement a music video recommender system by exploiting comments from users in classification.

1.4 Structure of the Thesis

In this thesis, there are totally five chapters including

Chapter one presents introduction, it gives you about Thailand internet users besides, describes social media content. Then, describing more detail about YouTube, motivation and objective.

Chapter two discusses the research background of music recommender system based on emotion classification. Then, the reviews of related work, recommender system, emotion classification, and imbalanced data problem.

Chapter three describes the proposed methodology. This chapter has four part: Section 3.1 and 3.2 describes data collection and annotation respectively. The building of the emotion classifier is shown in section 3.3. Finally, implementation of prototype music recommender system is shown in section 3.4.

Chapter four describes the experimental results. Section 4.1 describes the experimental setup. While, section 4.2 describes a dataset used in this research. The experimental results shown in section 4.3 can be divided into two parts. Moreover, prototype recommender system of this thesis is discussed. Finally, the discussion is indicated.

Chapter five gives conclusion and future work.

Chapter 2
Literature Review

This chapter describes the research background of the music recommender system based on emotion classification. These are emotion classification and recommender systems. The related works in this thesis consists of three topics; emotion classification, imbalanced data problem and recommender system.

2.1 Emotion Classification in Text

The current number of users on social network sites such as web blogs, web board, Facebook, Twitter as well as YouTube has expanded rapidly in Thailand. The amount of content and opinion dramatically increase every day. Currently, the web technology has entered to the web 2.0 era also known as web social networks. It see users as important and allow users to interact, share and exchange information collaboration of user data. For example, information like comments or message posts, comment on posts and social network pages. Emotion are both prevalent in and essential to all aspects of our lives. It influences our decision-making, affects our social relationships. Dr. Paul Ekman is an American psychologist who is a pioneer in study of emotion and facial expression. He is the one of top 100 psychologists of the twentieth century. His research findings led him to classify six basic emotions: anger, disgust, fear, happiness, sadness and surprise. Currently, he is a manager of the Paul Ekman Group, LLC (PEG); a company that provides training on emotional skills and doing research related to the national security organization. There is a great opportunity to identifying and analyzing people's emotions expressed in texts.

2.2 Recommender system

Nowadays, a recommender system has become popular and used widely in its many applications such as E-commerce. In the past, an old system provided static recommendation in which users search for interested products. A recommender system automatically analyzes recommendations for individual users based on user preferences, past activities, rating and other users' behavior that may have the same interests. Now, a recommender system has been widely applied in many applications such as movies, music, books and many more. Most recommender systems take either of two basic approaches: collaborative filtering or content-based filtering.

2.2.1 Collaborative filtering

Collaborative filtering works by comparing and collecting large amount of information on user's activities, preferences or behaviors instead of items and predicting what users will like based on their similarity to other users. The model can be constructed from single user's information or more from the other users who have similar traits. When, create the model can be achieved two types; explicit (e.g. asking a user to search.) and implicit (e.g. observing users behavior in an online store what item user like or what item user view.) forms of data collection.

A key advantage of the collaborative filtering approach is that it does not require item attributes or understanding of the item itself. The system focuses on predictions about the interests of other users with similar interests who have rated highly because it doesn't depend on the content of items to build recommendations.

Many algorithms have been used in measuring item similarity or user similarity in recommender systems (e.g. cosine similarity, the k-nearest neighbor (k-NN) and Pearson Correlation.). Collaborative filtering has two types which are model based collaborative filtering and memory-based.

Music	Toon	Tong	khim	Win
Hello/ How are you	-	-	?	+
Watch me	+	+	+	-
Yappari Thailand	?	-	-	+
Sugar	+	+	+	+
Skyfall	+	-	-	?

Table 2.1: An example of collaborative filtering

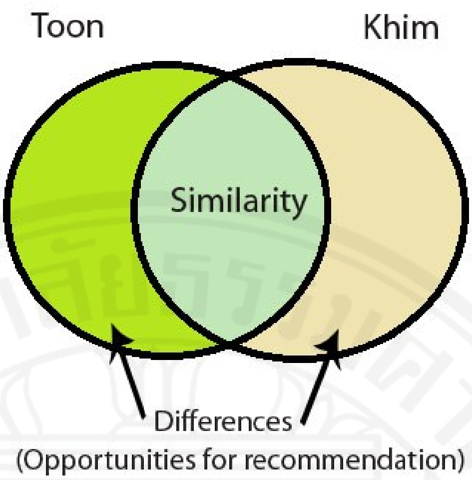


Figure 2.1: Similarities and differences used in collaborative filtering

2.2.2 Content-based filtering

Content-based filtering creates a recommendation based on user's behavior or user’s preference. For example, this approach might use past activity information, such as which website the user often visits and the characteristics of those websites. If a user usually visits music related websites about music, leave comments on website about technology or give a rating for specific item, then this technique can use this information to analyze and recommend similar item (e.g. music video clips or article about technology). The system will compare those items to other items in the system. This technique tries to recommend items that are similar to user’s preferences. In particular, items with a high similarity score are presented as recommendations. In order to estimate the probability of user’s preferences on the items, various algorithms have been used to do the measurement, such as cluster analysis, decision trees, Bayesian Classifiers, and artificial neural networks. Content-based filtering technique does not suffer from rating problem, but suffers from the problems of overspecialization and cold start. Moreover, Content-based filtering technique ignores the preferences of other users.

In addition, overspecialization occurs when these systems judge single item based on a small amount of features. This affects the system in recommending items that are either too similar to what the user has seen in the past or items which are the same content to other recommendations presented at the same time. Sometimes, this problem can be solved by introducing a small amount of random recommendations.

Finally, cold start problem occurs when the system has yet to receive enough user profile information to give reliable recommendation.

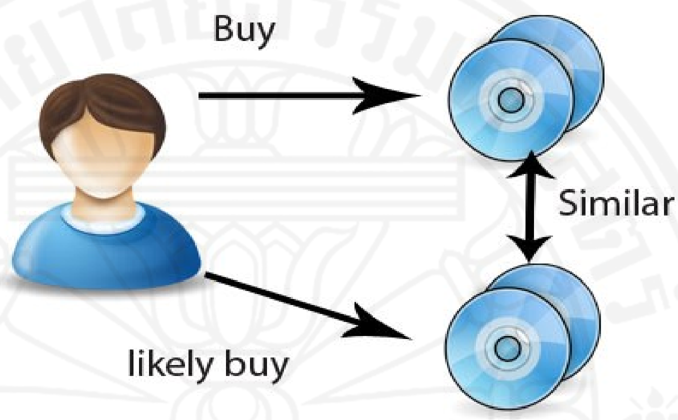


Figure 2.2: An example of content-based filtering

2.3 Emotion Classification

In the past, a number of sentiment analysis works can be divided into two groups. The first group, sentiment analysis focuses on classification of a text into neutral, positive and negative sentiments. Orimaye et al. [2] presented sentiment analysis algorithm on Yoruba movies by using YouTube comments using SentiWordNet and Yoruba sentiment lexicon, and a corpus containing 15,000 Yoruba YouTube comments. They manually annotated comments with three alternative labels; positive, negative and neutral.

Bravo-Marquez et al. [3] presented an approach for sentiment classification on twitter messages based on combination of several existing lexical resources and sentiment analysis methods. That is, OpinionFinder Lexicon (OPF), AFINN Lexicon SentiWordNet 3.0 (SWN3), SentiStrength Method, Sentiment140 Method and NRC Lexicon, and learning algorithms they apply Logistic regression, RBF SVMs, J48, Naive Bayes, and CART.

On the other hand, the second group, the emotion analysis is done by classifying a text into several emotions such as happiness, sadness, anger, disgust, fear and surprise. Strapparava et al. [4] presented automatic annotation of emotions in texts. Through the comparison of evaluations of knowledge based and corpus based methods with a dataset of 1,000 headlines, they claimed that their method works well for the emotion annotation.

Nagarsekar et al. [5] classified emotion on twitter data into six emotions; joy, sadness, anger, disgust, fear and surprise. Two machine learning algorithms; naïve Bayes and support vector machine are used and evaluated with three datasets.

Roberts et al. [6] analyzed how emotions are distributed in the data annotated and compared it to the distributions in another emotion-annotated corpus. Also, they used the annotated corpus to train a classifier that can automatically discover the emotions in tweets.

Esmin et al. [7] proposed an approach of the sentiment and emotion classification on Twitter by using hierarchical classification algorithm and compared it with flat classification. The data sets are collected from Twitter in topic of the Brazilian Soccer League 2011.

Burget et al. [8] classified the emotion of headlines news in Czech newspapers. In this work, a total of nine different learning algorithms were used and evaluated. The results showed different accuracies of different emotion labels; surprise 71.09%, fear 81.32%, sadness 75.4%, disgust 95.01%, joy 71.6%, and anger 87.3%.

Gao et al. [9] proposed joint learning on emotion and sentiment classification. Their approach is to utilize emotion and sentiment classification to classify the data for both classifications. They obtain same input vectors that annotated with both emotions and sentiment labels. Then, they calculate the transformation probability between these two kinds of labels. The experimental result shows that, their proposed classification can achieve much better results compared to individual classification.

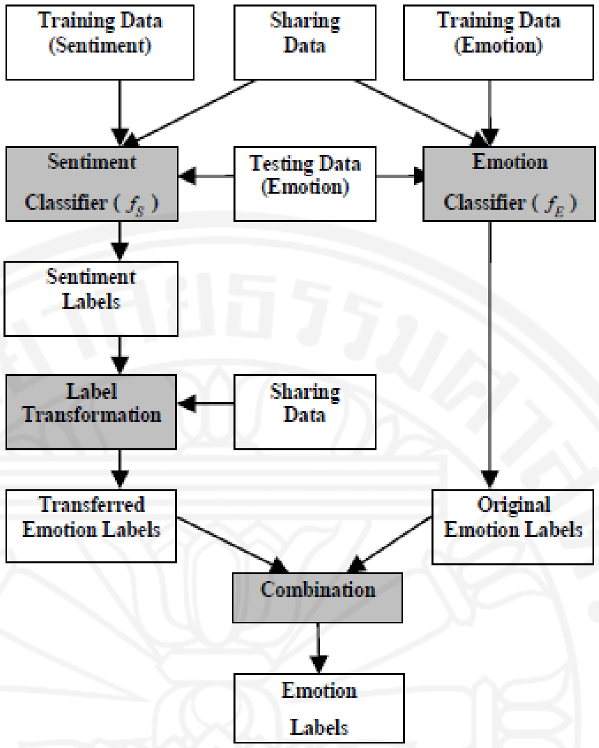


Figure 2.3: The framework of joint learning [9]

Wang et al. [10] created large dataset approximately 2.5 million emotion tweets, covering seven emotion categories for automatic emotion identification by gathering related hashtags available in the tweets.

Lin et al. [11] focused on classifying emotion on online news articles from reader's perspective into reader emotion categories. To do the research, they collected Chinese news articles from Yahoo! Kimo news and they use Yahoo! eight emotions: sad, boring, angry, surprise, happy, heartwarming, useful and awesome. Then, applied five feature selection; affix similarities, Chinese words, Chinese character bigrams, news metadata, and word emotions. Their experimental results show that five feature selection with SVM gives highest accuracy of 76.88%

Jia, Y. et al. [12] proposed emotion classification on headline news that related reader. They tried to check what emotions in text provokes their readers. Headline news including a few words are usually written with the desire to provoke emotions, and consequently to attract the readers' attention. They collected news article and emotion data from www.chinanews.com.cn. Then, they defined eight emotion classes for users to choose for further reading. SVM was applied to predict emotion. Four classes of features were considered, including part-of-speech, Chinese characters, words and news domains. The experimental results indicated that certain feature combinations achieve good performance.

For emotion classification in Thai language, Lertsuksakda et al. [13] proposed a methodology to construct a Thai sentiment resource based on sentic computing.

Haruechaiyasak et al. [14] presented a comparison study of a variety of approaches for Thai word segmentation. The word segmentation approaches could be classified into two specific types, machine learning based (MLB) and dictionary based (DCB) approaches. The performance of MLB approach relied on a training model from a corpus by using machine learning techniques. While the performance of dictionary based approach depends on the quality of the word set and quantity in dictionary. They used the LEXiTRON Thai-English electronic dictionary which contains about 30,000 words as the main dictionary. They compared between four algorithms from the machine learning based approach: decision tree, Naive Bayes (NB), Support Vector Machine (SVM) and Conditional Random Field (CRF), and two algorithms from the dictionary based approach: longest-matching (LM) and maximal matching (MM). They used ORCHID corpus as a data set, which contains about 100,000 manually tagged words. From the experimental results, the best performance was obtained from the CRF algorithm with the precision and recall of 95.79% and 94.98%, respectively. However, the dictionary based approach yielded better performance than the NB, decision tree and SVM algorithms from the machine learning based approach.

Inrak and Sinthupinyo [15] proposed an approach to utilize bi-words occurrence to classify emotion that is hidden in a short sentence. In this research, they classified Thai text into six basic emotions based on a latent semantic approach.

Nivet [16] classified emotion in Thai texts using different weighting schemes. He study a comparison of a variety of weighting schemes. The experimental results showed that the Boolean weighting with a SVM classifier was the most effective in his experiments.

Haruechaiyasak et al. [17] proposed a new feature to improve the keyword-based search, which is called category browsing for their Thai-language news article search engine, called Sansarn News Search.

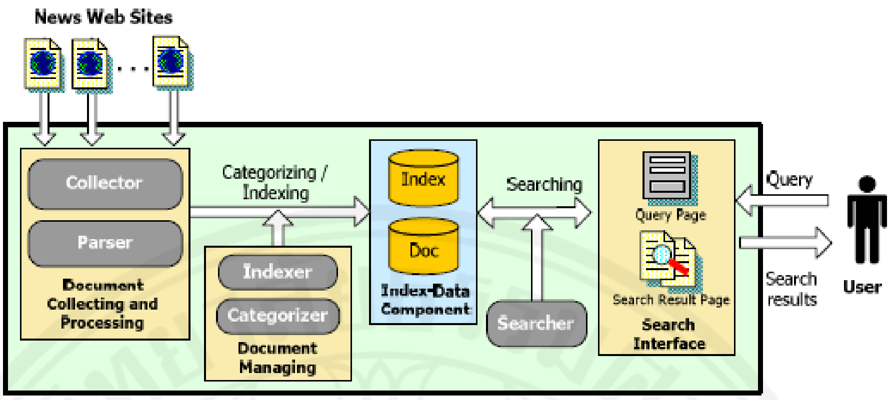


Figure 2.4: Sansarn News Search Engine: System Architecture [17]

Figure 2.4 shows the system architecture of the Sansarn search engine. Their news search engine is implemented by using "Sansarn Look!" which is tool to create an information retrieval system on Thai texts.

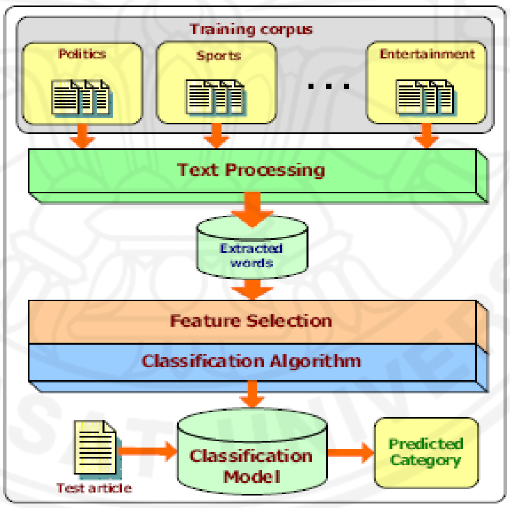


Figure 2.5: News article categorizer [17]

Figure 2.5 illustrates the news article categorizer. Their corpus is a set news articles that pre-sort into a set of news categories such as politics, sports and entertainment. Then, they execute segmentation to separate texts into word tokens by using LexTo, a word segmentation program. They use TF-IDF as a weighting scheme.

To improve the performance of text categorization, they apply feature selection techniques and compare with several methods including Information Gain (IG), $\chi 2$ (CHI) and Document Frequency thresholding (DF). For their experimental result, the SVM algorithm with the Information Gain gives the highest performance of 95.42%.

2.4 Imbalanced data Problem

For real-world datasets, there are many situations that the number of instances in one class is much lower than the number of instances in the other classes. This problem is known as the imbalanced dataset, the classification performance usually degrades in several data mining applications including medical pattern recognition, telecommunication management, bioinformatics, and text categorization. Most of the conventional methods assign the majority class to data and ignore the minority class due to data skewness. One of the most popular solutions to this problem is by apply a kind of sampling methods [18-20].

There are three types of the sampling-based approaches towards the imbalanced datasets. The first type is oversampling, by adding more of the minority class. However, it could lead the classifier to overfitting to a few examples. The second type is undersampling, by removing some of the majority class, but could remove some useful information of the majority class. The third type is hybrid, a mix of oversampling and undersampling. The hybrid still inherits both advantages and drawbacks of both approaches.

Lokanayaki and Malathi [21] have proposed a new setting of the missing data imputation and compare performance by using SVM classifier with oversampling, undersampling and SMOTE technique. The information is obtained from UCI Machine Learning Repository. The result showed that SMOTE obtained the highest accuracy.

Rivera et al. [22] performed an experiment to study the relationship between SMOTE and Propensity score matching (PSM). Introducing their approach called over-sampling using propensity score matching (OUPS). The OUPS approach is an algorithms to solve imbalanced data problem, which is a combination between SMOTE and PSM. Using propensity score to match criteria. While, using SMOTE for creating synthetic sample without randomly picking nearest neighbors. Four data sets obtained from the Machine Learning Repository. Next, they applied four machine learning algorithms; logistic regression (LR), support vector machine (SVM), neural network (NN), and linear discriminant analysis (LDA) by using R statistical software. The result showed that OUPS outperformed SMOTE and PSM in terms of accuracy and sensitivity. SMOTE outperformed all other sampling techniques in specificity and F-measure.

Chen [23] applied three re-sampling techniques that is SMOTE algorithms, over-sampling and under-sampling. Then, evaluated performance by using three classification algorithms: decision tree, Naïve Bayes and neural network. The corpus consists of 40,000 examples, each of which involves 20 real-valued features. Among these, 36,364 of them are annotated with negative labels, 3,636 of them are annotated with positive labels. The experimental result shows that, for naïve Bayes classifier and neural network classifier with SMOTE technique gives the best accuracy compared to other classifiers and sampling techniques in their experiment.

Li et al. [24] presented an experiment with five datasets from the UCI datasets, their method increases the number of the minority and enormously improves the identity of the minority with no bad impact to the majority. The experimental results with the predictive model shows that the method using the random-SMOTE method can improve the classification performance essentially.

Cheawsakunwattana S. et al. [25] constructed a decision tree for an imbalanced data set using C4.5 basis with a new entropy adjustment called AMOTE in order to enhance the minority classification. The result showed that AMOTE method obtained higher overall performance.

Li et al. [26] focused on identifying several manners of data imbalance including, class overlap, class size and text distribution.

Wang et al. [27] improved the SMOTE algorithm by incorporating the locally linear embedding algorithm (LLE). The LLE technique is a technique for mapping the original data into a new linearly separable feature space. While the three datasets are collected from chest x-ray image databases in automatic computerized detection of pulmonary. The experimental results are evaluated, based on leave-one-out validation tests and applying three standard classifiers; K-NN classifier, naïve Bayesian and support vector machine. Finally, the experimental result indicated that their algorithm obtained better performance than traditional SMOTE.

Naseriparsa et al. [28] presented a method to improve prediction rate in a lung cancer dataset by using incorporate principal components analysis (PCA) with SMOTE technique. PCA is a method that provides a sequence of best linear approximations to a given high dimensional observation. Next, the data set is collected from the UCI Repository of Machine Learning databases. While naïve Bayes classifier is applied on five different datasets which are obtained from application of five methods. The lung-cancer dataset contains 56 features and 32 samples, which is classified into three groups. Their proposed method consists of two steps; first step is compacts the dataset feature space by applied PCA, second step oversampling minority classes by using SMOTE. Finally, the experimental results show that performance improved. While feature space reduced more than a half and this leads to lower the cost and complexity of classification process.

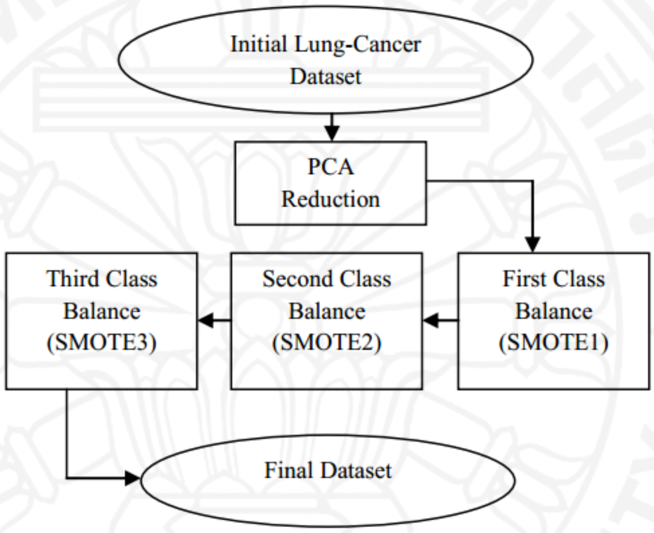


Figure 2.6: Naseriparsa et al. proposed method

2.5 Recommender System

The recommender System is a popular system that used in electronic commercial (E-Commerce) for suggest product and service for client and increase product sales in E-Commerce system. Ponchob and Nitsuwat [29] created a recommender system for notebook purchasing using content based filtering. Tanmuangjai and Utakrit [30] proposed a recommender model for credit product using collaborative filtering.

Tewari A.S. et al. [31] proposed book recommender system based on hybrid filtering and association rule mining to give stronger recommendations to the users. Their book recommender system tries to recommend book to the user based on user preference. Their system works in offline mode and stores recommendation into the user's web profile. This system uses collaborative filtering and associative model to give stronger recommendations.

Iwahama et al. [33] proposed music recommender system based on content-based filtering. The target music data of their system are written in MIDI format. First, they analyze feature parameters in music data. Then, they select feature parameters which appropriate for their filtering system. Next, they created a filtering method based on the feature parameters that they had selected. Finally, they implement the prototype system in Java language.

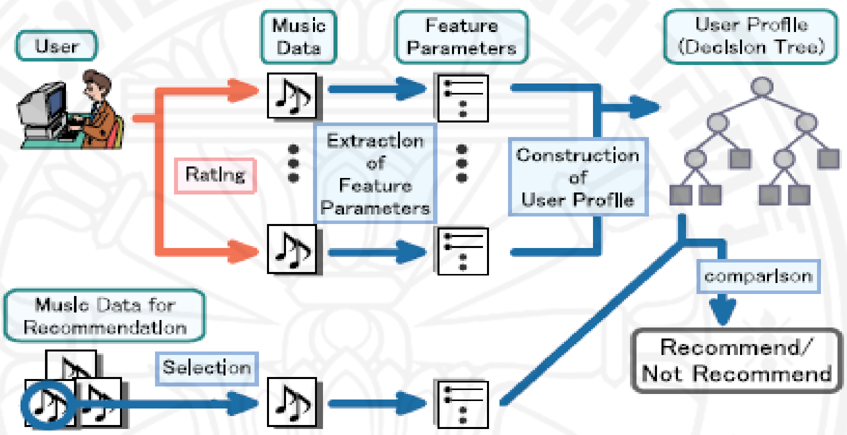
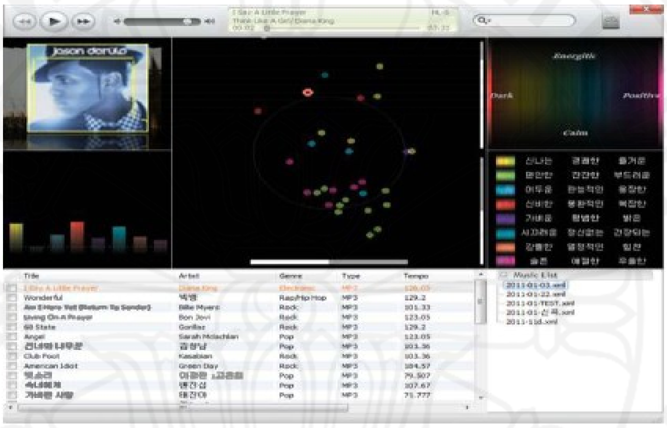


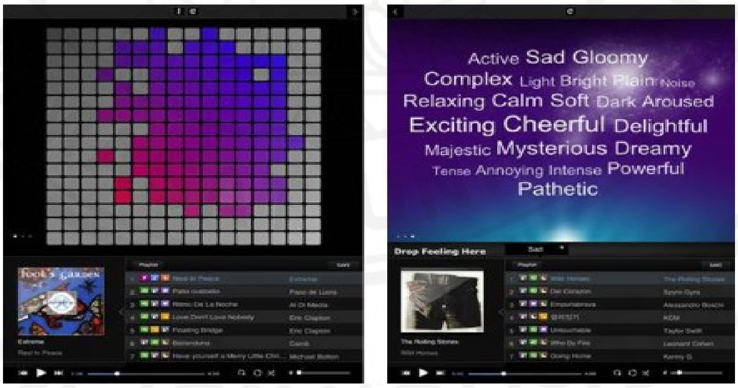
Figure 2.7: Overview of the Iwahama's filtering system

Li et al. [34] proposed a new algorithm for a recommendation system using hidden Markov model (HMM). They download data from GroupLens. It contains 100,000 data form 943 users. Then, transforming data into "genre string" format. Next, computing the probabilities of each "genre string" by using Forward algorithm. To cluster "genre string", they use K-Means with Hamming distance. After, clustering movies, the system can recommend movie to users. Moreover, they compare the experimental result with precision and recall rates of vector space model based (VSM) and HMM based algorithm. The experimental results show that average precision of HMM is higher than VSM based algorithm and HMM based algorithm can avoid small number of similar key words problem by utilizing probabilities distribution.

Kim J. et al. [35] present probability based music mood model and implement this model to create music recommendation system. They create three platform of music recommendation players, for PC, mobile devices, and web. In this work, they used a hybrid approach, which is a combination of categorical and dimensional music model. In addition, categorical model is classified emotion into basic classes such as anger, fear, sadness, happiness and disgust. While, dimensional model can present a model of mood in a circle on two-dimensional bipolar space. It can show mood and relationship between moods. Moreover, they gathered valence-arousal values from 446 music clip and tag the mood by ten persons. For measuring the similarity of mood tags, they used cosine distance and classified mood tags into eight categories by k-means clustering.



(a) Music Recommendation player for PC



(b) Music recommendation app. For mobile device

Figure 2.8: Snapshot of Kim J. et al music recommendation system

Chapter 3
Methodology

This chapter describes methodology in this thesis. It can be divided into four main parts: Section 3.1 and 3.2 describes data collection and annotation respectively. The building of the emotion classifier is shown in section 3.3. Finally, implementation of prototype music recommender system is shown in section 3.4.

3.1 Data Collection

In our research, we focus on Thai language. The crawlers of Thai YouTube comments are constructed using Java. The crawlers collect data from YouTube website. First, we manually collect video clips that contain more than 100 comments. Then, we gather comments by putting Video Id to the YouTube API. Next, we collect all information of each video clip such as name, ID, usernames, comments, comment IDs and timestamp in Java Script Object Notation (JSON) format. Our dataset is comments on Thai video clips, collected from YouTube. From the total of 85 video clips, there are approximately 12,000 comments collected.

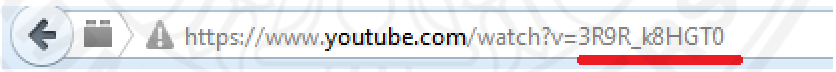


Figure 3.1: An example of Video Id

3.2 Comments Annotation

All comments in our corpus are manually annotated with one of nine tags: Anger, Disgust, Fear, Happiness, Sadness, Surprise, Emotion, Non-Emotion and Junk, by using an annotation tool called "Sansarn Corpus Tagging tool" provided by National Electronics and Computer Technology Center (NECTEC).

NECTEC is an agency under the office of science and technology. Ministry of Science and Technology located at science park, Pathum Thani. The main objectives of NECTEC are research and research support by providing funding support for research, to both government sectors and electronics industries.

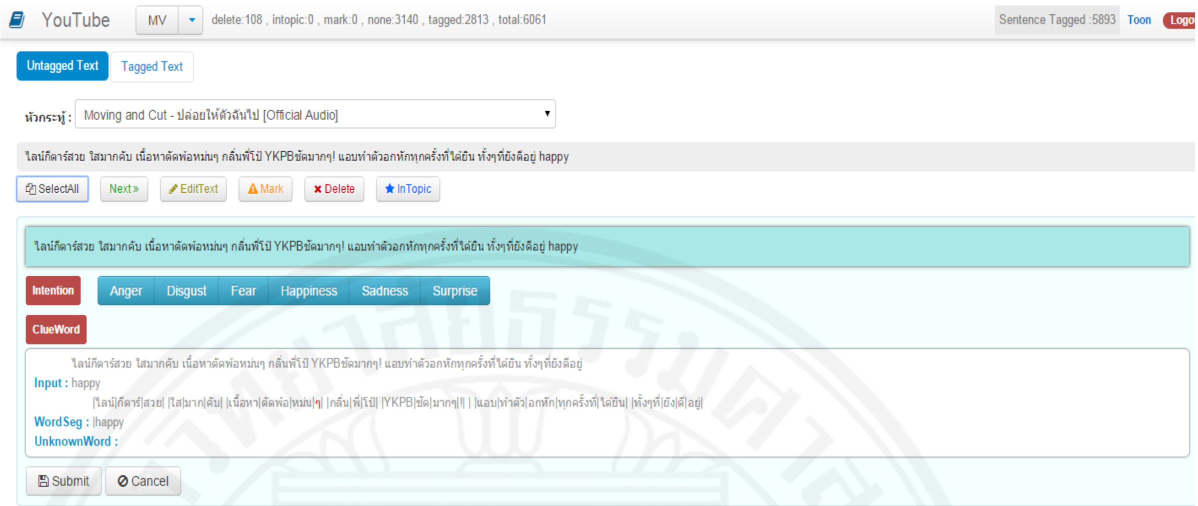


Figure 3.2: Our corpus annotation tool

3.3 Building Emotion Classifier

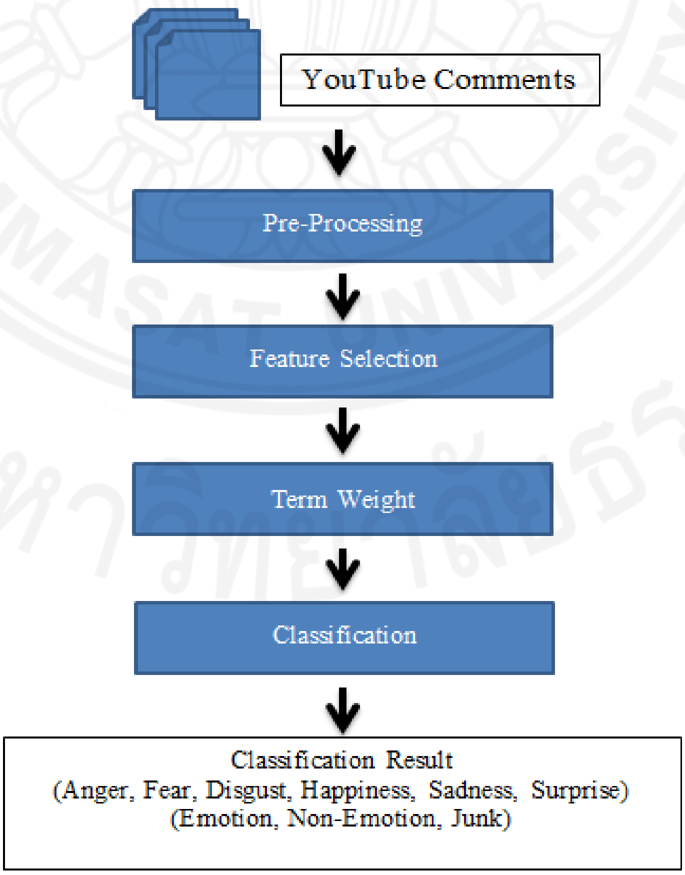


Figure 3.3: Our emotion classification framework

We utilize a two step models. The first one is the emotion filtering model, which filters comments into three groups: Emotion, Non-Emotion and Junk. The second one is the emotion classification model, which will classify emotion comments into six emotions: Anger, Disgust, Fear, Happiness, Sadness and Surprise. To build classifier models, it takes five steps as follow.

3.3.1 Pre-Processing

The collected comments are processed with the following four pre-processing procedures. First, segment each comment into a set of word tokens. Then, remove Thai and English stop words. After that, remove HTML tags and then stem words. In this step, we used "LexTo: Thai Lexeme Tokenizer" and NECTEC LEXiTRON dictionary from NECTEC for segmentation.

3.3.2 Feature Selection

A frequency threshold is set for keyword identification. That is, a word is a keyword if such term appears more than the minimum threshold. Otherwise, it will be removed from the feature set.

3.3.3 Term Weight

This task uses two term features in the classification process. There are:

1. Term Frequency (TF) is a value of how often a term occurs in documents. In our baseline, we use the raw frequency, which is the simplest choice.
2. Term Frequency - Inverse Document Frequency (TFIDF) is a statistic value that shows the important a term in documents. For our weight adjustment, we compute term frequency $tf(t, d)$ equals $f(t, d)$. That is, a frequency of term t in a document d. The inverse document frequency $idf(t,D)$ shows if the term is common or rare of all documents. D is the total number of documents and it is divided by the number of documents, which contain the term. The "add-onesmoothing" of the equation $1+|\{d \in D: t \in d\}|$ solves a division-by-zero problem.

3.3.4 Classification

The thesis has two classification processes. There are the emotion filtering and the emotion classification. Three machine learning are utilized in the classification process. These are:

1. Support Vector Machine
2. Multinomial Naïve Bayes
3. Decision Tree

3.4 Implementing The Prototype Recommender System

Our prototype recommender system is based on content based filtering, which will gives recommendation based on content in the system to users. Figure 3.4 represents the overview of our proposed system.

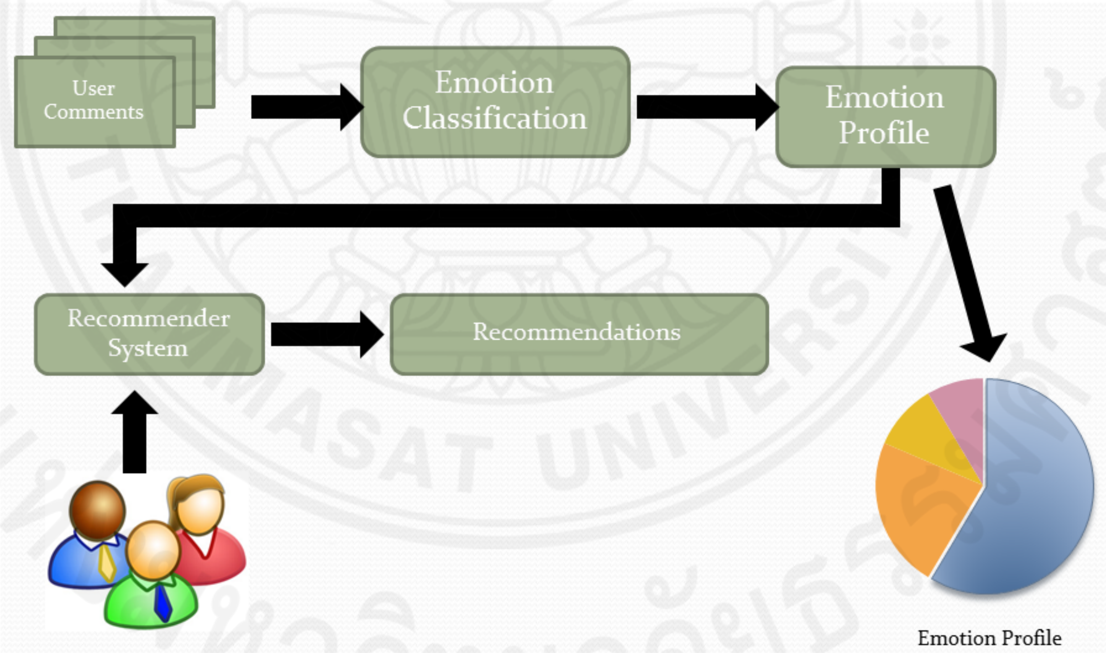


Figure 3.4: Overview of our proposed system

3.4.1 Creating Song Emotion Profile

To create a song emotion profile, we will use user comments as input. Then, we will classify it by using two classifiers. First, emotion filtering, will filter any comments that have no emotion and not related to the music video clips. Finally, emotion classification will classify emotion comments into six emotions.

3.4.2 Measuring Similarity

After we collected song emotion profiles, we need to calculate similarity score for every song. We use feature in song emotion profile to calculate the similarity, which is six emotions score in the song.

Cosine similarity is a comparison of similarity between two documents. Each document is represented by n-dimensional vector that measures the cosine angle between them. The cosine similarity ranges from zero to one. If two vectors have the same angle, the cosine similarity achieve the maximum of one. It means that both songs are exactly the same; otherwise it will be less than one for any other orientation.

The cosine of two vectors can be derived by using the Euclidean dot product formula:

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos\theta$$

Given two vectors of attributes, A and B, the cosine similarity, cos(θ), is represented using a dot product and magnitude as

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum\limits_{i=1}^{n} A_i \times B_i}{\sqrt{\sum\limits_{i=1}^{n} (A_i)^2} \times \sqrt{\sum\limits_{i=1}^{n} (B_i)^2}}$$

3.4.3 Creating Application

The prototype system is a web application. It based on PHP language. The web application consists of three modules, which are Collector module, Similarity Calculator module and Visualization module. The description of the three modules are given below. Figure 3.5 represents our system architecture.

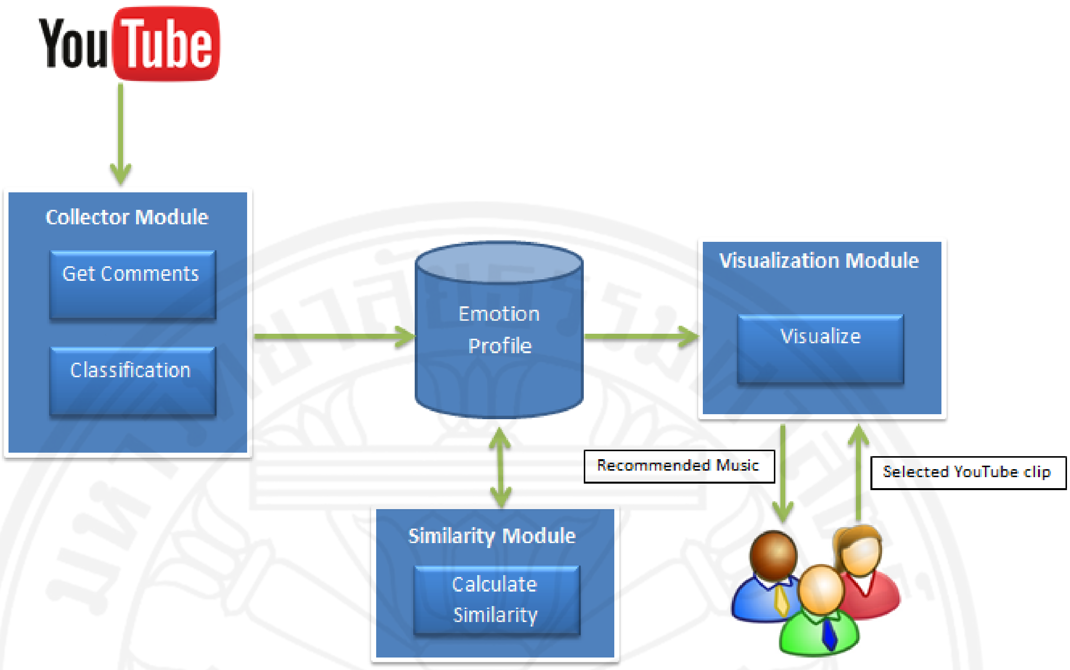


Figure 3.5: System architecture

Collector module: We use this module to collect YouTube comments. First, we manually select YouTube music videos that we desired. Next, putting YouTube ID to the module. It will automatically collect user comments. Then, it classifies all comments that we had collected, and stores it into database.

Similarity Calculator module: After we collect data, we need to calculate similarity for each music video by using Cosine similarity. When calculation is completed, the system selects top five similar score items for each music video and stores this information into the database.

Visualization module: This module presents all song to the user via web application by using Codeigniter framework. The website gives information for each music video to users. For example, emotion profile, emotion comment, and recommended songs that similar to the song that being listened.

Chapter 4
Experimental Result and Discussion

In this chapter, we present our experimental results of emotion classification. Section 4.1 describes the experimental setup in this research. Section 4.2 describes a dataset used in this research. The experimental results show in section 4.3 divided into two parts. The first part is oversampling minority classes and investigating KNN parameter by using SMOTE technique. Second, we summarize original experimental results with results from SMOTE technique. Moreover, prototype recommender system of this thesis is discussed. Finally, the discussion is indicated.

4.1 Experimental setup

A classifier is evaluated by a confusion matrix as shown in Table 4.1. The columns show the predicted classes and the rows show the actual classes. In the confusion matrix, TN is the number of negative samples correctly classified (True Negatives), FP is the number of negative samples incorrectly classified as positive (False Positives), FN is the number of positive samples incorrectly classified as negative (False Negatives) and TP is the number of positive samples correctly classified (True Positives).

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Table 4.1: Confusion matrix

The Overall accuracy is defined in Equation 1.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{1}$$

The Precision is defined in Equation 2.

$$Presicision = \frac{TP}{TP + FP} \tag{2}$$

The Recall is defined in Equation 3.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The F Measure is defined in Equation 4.

$$F\ Measure = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

4.2 Dataset Characteristics

We used two datasets of Thai comments collected from Thai YouTube video clips. From the total of collected 85 video clips, we randomly selected equally from each data set by using YouTube API. The emotion filtering dataset contains 9,571 samples, which is filtering comments into three groups. While the emotion classification dataset contains 5,345 samples which are classified comments into six emotion groups. The data described six types of emotion. Table 4.2 presents an example data for every class. Both dataset contains 7,722 features (word type). The dataset details are shown in Table 4.3 and 4.4

Text	**Class**
แกนี่มันโง่จริงๆ ควาย! Translation: You idiot really stupid.	Anger
เนื้อหาไม่มีไรเลี้ยพูดถึงแต่เรื่องอย่างนั้นจะอ้วกหว่ะ Translation: This content has nothing it talking about …. Disgusting!	Disgust
มึงนี่น่ากลัวหว่ะ Translation: You look horrible.	Fear
โอ๊ย ขำตกเก้าอี้เลยอ่ะ เอาซะอยากไปหามากินเลย ถถถถ Translation: I falling off the chair laughing it make me hungry ha ha ha	Happiness
น้ำตาไหล สะเทือนใจ T_T Translation: I'm cry it hurt T_T	Sadness
เอาเวลา35วิกูคืนมา Translation: give my 35 second back!	Surprise
This class contain emotion comments but don't specific which emotion class.	Emotion
ผมชอบเพลงนี้ที่สุดเลย Translation: I really like this song.	Non Emotion
ใครที่อยากมีเงินใช้เป็นของตัวเอง5000-1000บาทขึ้นไปต่อสัปดาห์พร้อมทั้งมีโปรโมชั่นให้เที่ยวต่างประเทศฟรีบ่อยๆ Translation: Who want money 5000-10000 Thai baht/week with free travelling promotion contract me.	Unrelated

Table 4.2: An example of comments for each class

	Number of emotion comments			
	Emotion	Unrelated	Non Emotion	Total
Emotion Filtering	5,345	1,735	2,491	9,571

Table 4.3: Emotion filtering dataset

For emotion filtering dataset, the majority class is emotion class that contains 5,345 samples. While the minority class are unrelated and non-emotion class contain 1,735 and 2,491 samples respectively.

	Number of emotion comments						
	Anger	Fear	Disgust	Happiness	Sadness	Surprise	Total
Emotion Classification	381	34	37	2,000	2,000	893	5,345

Table 4.4: Emotion classification dataset

For the emotion classification dataset, the majority classes are happiness and sadness contained 2,000 samples. The minority classes are fear and disgust contained 34 and 37 samples respectively.

4.3 Experimental Results

To confirm the filtering and classification performance. We perform two sets of experiments. First, we investigating the results of SMOTE technique. Second, we summarize the experimental results. Next, we apply three standard classifiers, support vector machine, multinomial naïve Bayes and decision tree with default setting by using Weka 3.6 an open source machine learning tool. The results are based on splitting data 80:20

4.3.1 Oversampling Minority Classes and Investigating KNN parameter by using SMOTE technique

The comment of disgust, fear, angry and surprise class are misclassified because amount of comment are fewer than majority class. However, the minority classes are more likely misclassified than the majority class due to machine learning algorithms. In addition, they usually occur in many applications including the YouTube comment corpus. An observation on these experiments is the minority classes are often useful. In our training data, the number of happiness and sadness classes are larger than other class. Since our training data is unbalance, results from classifier may have bias. There are many solutions for this problem. The sampling based approaches are the one of unbalanced data problem solution. Therefore, we selected oversampling method.

NV. Chawla et al. [20] proposed the SMOTE (Synthetic Minority Over-sampling Technique). It was presented to solve, the class imbalance problem. It is one of well-known method according to its modesty and efficiency. Improving classification performance, first, we are initially varying nearest neighbor parameter in both datasets to see effect that we have varied. The experimental results by varying nearest neighbor (KNN) parameter are presented in Figure 4.1 and 4.2. Second, we are oversampling minority class sample by 100 samples for each iteration until it equal to the majority class. For the first iteration, we equalize minority samples in both datasets. So, in the next iteration, we increase minority class sample by 100 samples for each iteration. We set the nearest neighbor (KNN) parameter = 500 for our experiment. The experimental results are shown in Figure 4.3 and 4.4.

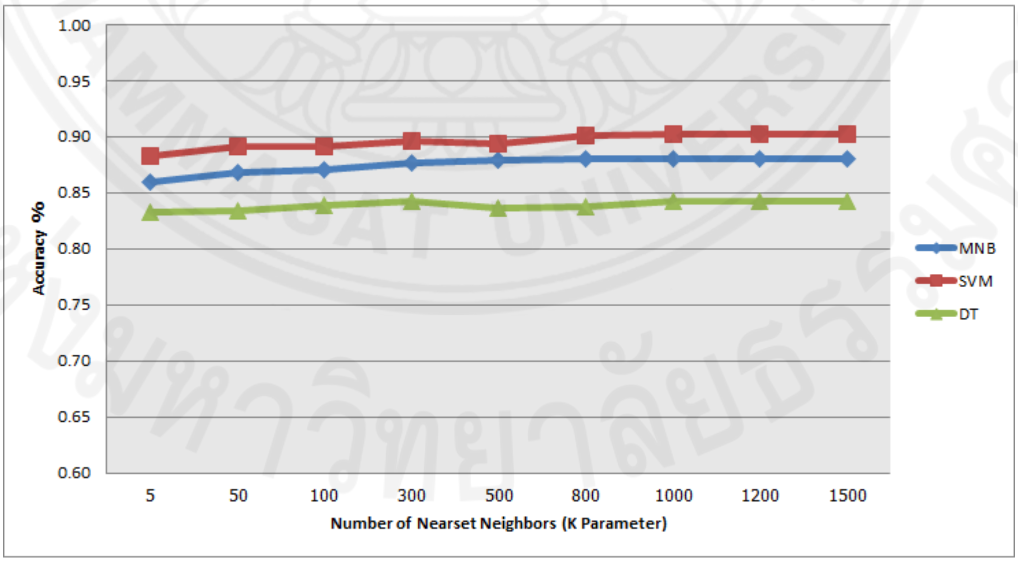


Figure 4.1: Experimental results by varying nearest neighbor parameter in the emotion classification dataset

In Figure 4.1, we varied the nearest neighbor parameter in Emotion Classification dataset with the balance dataset. We found that K = 1,000 gives the highest accuracy with 90.27%for SVM classifier. It is stable when we increased the parameter K.

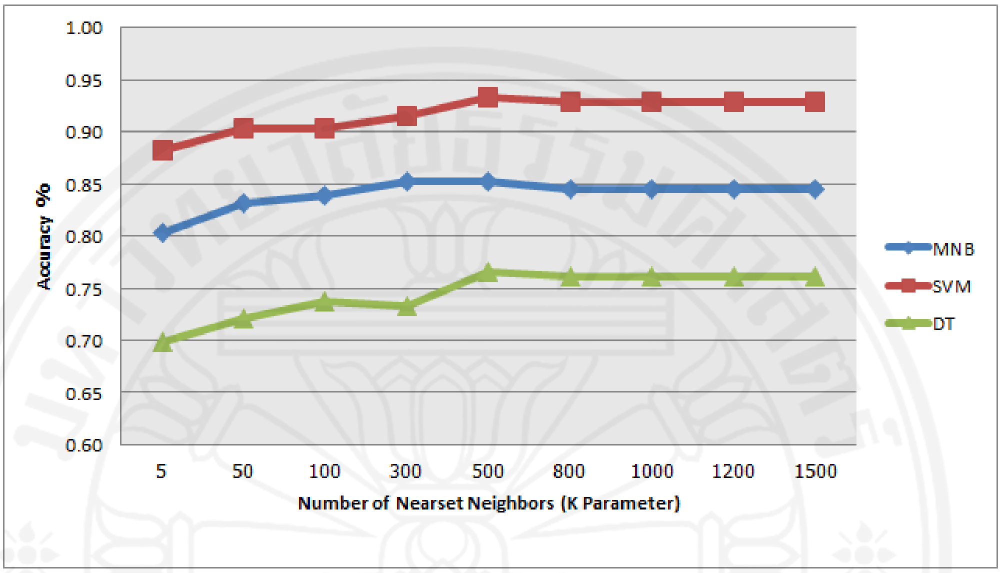


Figure 4.2: Experimental results by varying nearest neighbor parameter in the emotion filtering dataset

In Figure 4.2, we do experiment by varying nearest neighbor parameter in the emotion filtering dataset with the balance dataset. We were found at K=500 it gives the highest accuracy with 93.30%for SVM classifier. Then it slightly drops when we increased parameter K.

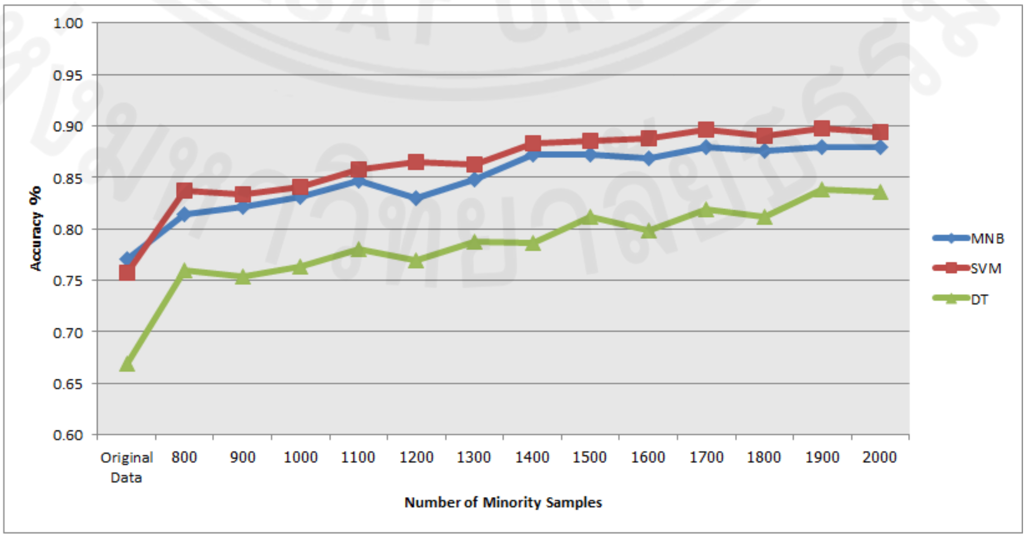


Figure 4.3: The experiment result from the emotion classification dataset

Figure 4.3 summarizes the results of classification with the emotion classification dataset. SVM is the most accurate, followed by multinomial naïve Bayes and decision tree respectively. Finally, we found that SVM at 89.44% with balanced sample size of 2,000 gives the best result.

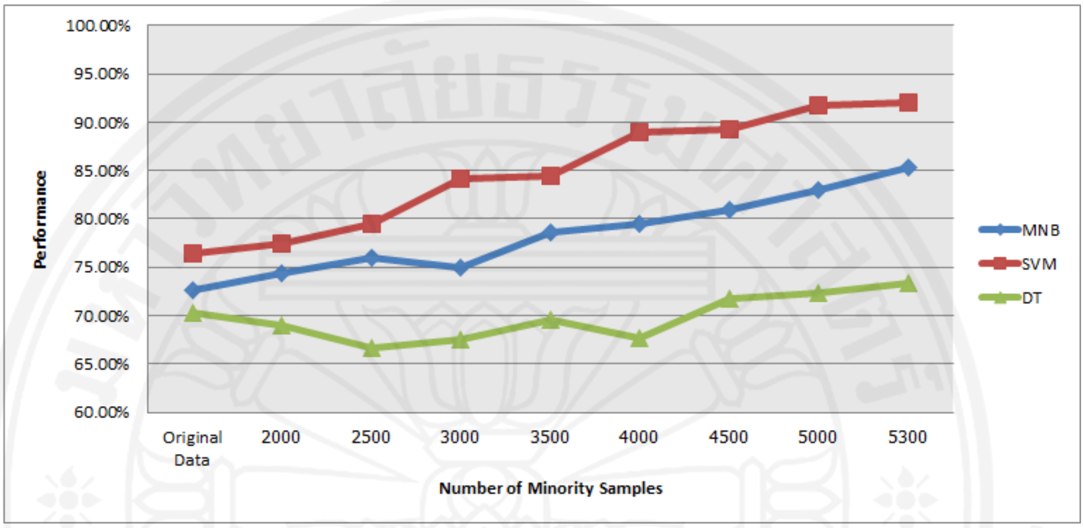


Figure 4.4: Experiment results from the emotion filtering dataset

Figure 4.4 summarizes the results of classification with the emotion filtering dataset. SVM is the most accurate, followed by multinomial naïve Bayes and decision tree respectively. Finally, we found that SVM at 93.30% with balanced sample size of 1,500 is the best accuracy among all methods.

4.3.2. Summarizing Experimental Results

Classifier	Emotion Filtering	Emotion Classification
SVM	74.25	75.68
MNB	74.51	77.08
DT	70.36	66.88

Table 4.5: Original classification accuracy

	Unrelated	Non Emotion	Emotion	Total
Unrelated	**185**	23	132	340
Non Emotion	11	**344**	162	517
Emotion	22	143	**893**	**1058**

Table 4.6: Original confusion matrix of emotion filtering dataset

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Emotion	0.844	0.343	0.752	0.844	0.796	0.758
Unrelated	0.544	0.021	0.849	0.544	0.663	0.773
Non Emotion	0.665	0.119	0.675	0.665	0.67	0.793
Weighted Avg.	0.743	0.225	0.748	0.743	0.738	0.77

Table 4.7: Original Classification performance detail in emotion filtering dataset

In Table 4.6 and Table 4.7, they provide the confusion matrix and classification performance detail for each measurement in emotion filtering dataset with SVM classifier.

	Surprise	Disgust	Anger	Fear	Hap	Sad	Total
Surprise	**83**	0	3	0	58	28	172
Disgust	1	**2**	1	0	2	3	9
Anger	10	0	**36**	0	14	7	67
Fear	0	0	0	**2**	2	4	8
Hap	23	0	3	0	**367**	10	403
Sad	9	0	7	0	49	**315**	380

Table 4.8: Original confusion matrix of emotion classification dataset

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.483	0.048	0.659	0.483	0.557	0.772	Surprise
0.222	0	1	0.222	0.364	0.729	Disgust
0.537	0.014	0.72	0.537	0.615	0.829	Anger
0.25	0	1	0.25	0.4	0.806	Fear
0.848	0.197	0.746	0.484	0.794	0.833	Happiness
0.829	0.119	0.793	0.829	0.811	0.882	Sadness
0.753	0.131	0.751	0.753	0.744	0.839	Weighted Avg.

Table 4.9: Original classification performance detail in emotion classification dataset

In Table 4.8 and Table 4.9, the confusion matrix and classification performance detail are provided for each measurement in emotion classification dataset with SVM classifier.

	Emotion Filtering Accuracy (%)		Emotion Classification Accuracy (%)	
	Original	SMOTE	Original	SMOTE
SVM	76.41	**93.30**	75.68	**89.44**
NBM	72.58	85.32	77.08	87.95
DT	70.36	76.58	66.88	83.64

Table 4.10: Summarize the classification performance

Table 4.10 presents the predicted accuracy for the three classifiers in both dataset. SVM on SMOTE technique performs better than other classifier. Comparing the result in filtering task, SVM using SMOTE yielded accuracy with 93.30% compared to 76.41% with original result. The result improves by 16.9%. Moreover, in classification task, the result of SVM observed from SMOTE was 89.44% compared to original result, which is 75.68%, there is 13.76% improvement.

	Non Emotion	Unrelated	Emotion	Total
Non Emotion	**982**	29	30	1,041
Unrelated	32	**965**	45	1,042
Emotion	85	120	**850**	1,055

Table 4.11: Confusion matrix of emotion filtering dataset

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Emotion	0.891	0.041	0.912	0.891	0.901	0.926
Junk	0.954	0.003	0.995	0.954	0.974	0.983
Non Emotion	0.953	0.054	0.89	0.953	0.92	0.957
Weighted Avg.	0.933	0.032	0.935	0.933	0.933	0.957

Table 4.12: Classification performance detail in emotion filtering dataset

Table 4.11 and Table 4.12, provide the confusion matrix and classification performance detail for each measurement in the emotion filtering dataset with SVM classifier.

	Surprise	Disgust	Anger	Fear	Hap	Sad	Total
Surprise	**350**	0	9	1	30	13	403
Disgust	3	**410**	3	0	6	0	422
Anger	6	2	**389**	0	4	1	402
Fear	0	0	0	**392**	1	0	393
Hap	55	4	5	1	**309**	29	403
Sad	41	2	11	2	26	**309**	391

Table 4.13: Confusion matrix of emotion classification dataset

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Surprise	0.868	0.052	0.769	0.868	0.816	0.94
Disgust	0.972	0.004	0.981	0.972	0.976	0.993
Anger	0.968	0.014	0.933	0.968	0.95	0.989
Fear	0.997	0.002	0.99	0.997	0.994	0.999
Happiness	0.767	0.33	0.822	0.767	0.793	0.926
Sadness	0.79	0.021	0.878	0.79	0.832	0.933
Weighted Avg.	0.894	0.021	0.896	0.894	0.894	0.964

Table 4.14: Classification performance detail in emotion classification dataset

Table 4.13 and Table 4.14, the confusion matrix and classification performance detail are provided for each measurement in emotion classification dataset with SVM classifier.

4.4 Prototype Recommender System

Our prototype system with filtering and emotion classifiers is implemented in PHP language with Codeigniter web framework. The process flow of the system is as follows. Firstly, the users go home page, which show all music video clips in the system as shown in figure 4.5 and 4.6. Then, user selects a music video clip.

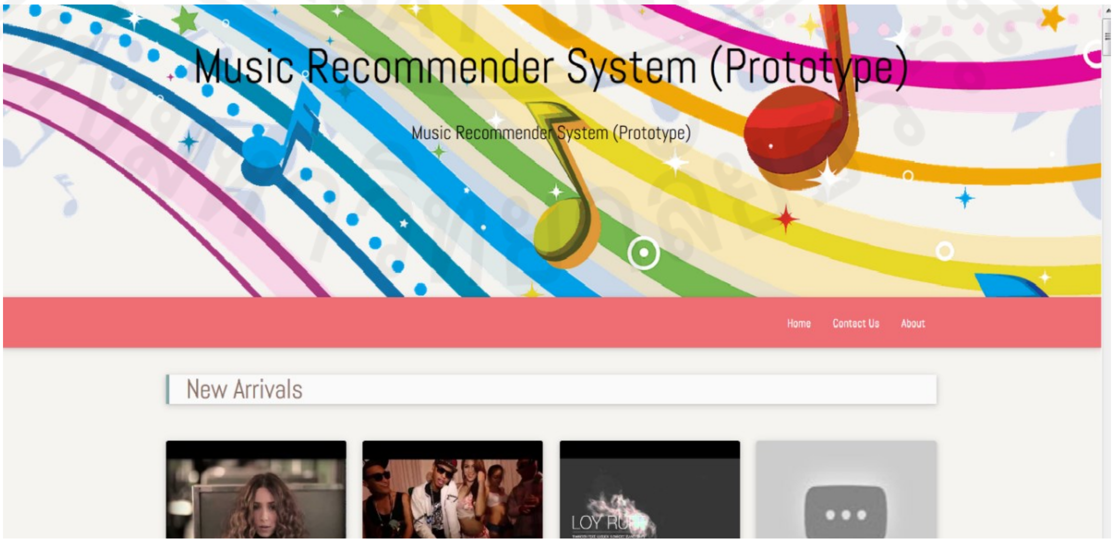


Figure 4.5: Prototype system home page (1)

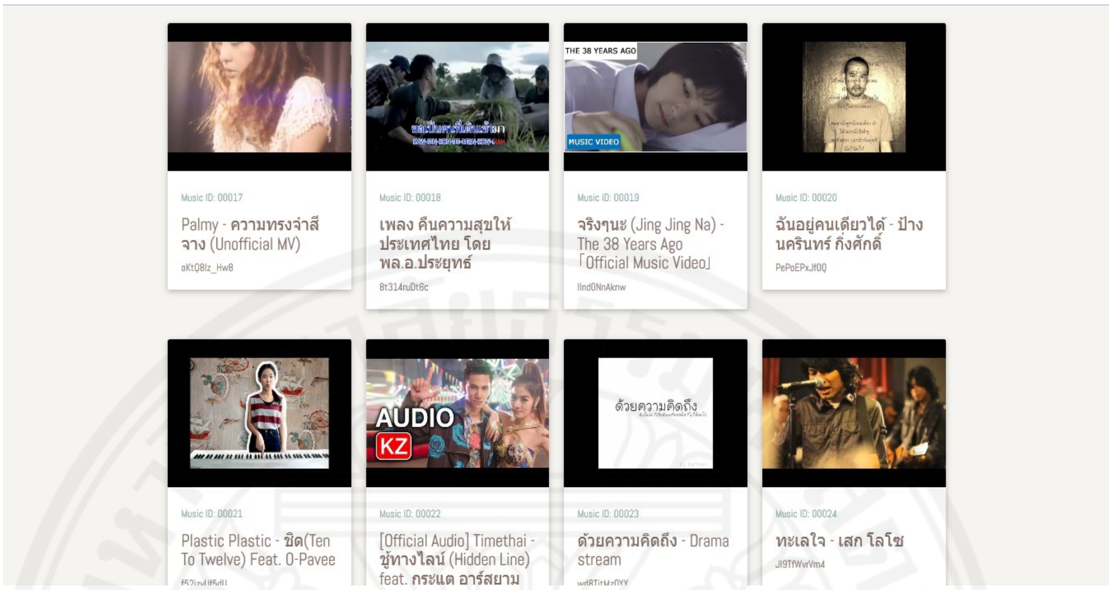


Figure 4.6: Prototype system home page (2)

When users selected the music that user want to listen. Users will go to the next page, which is music pages. In this page, users will get a music video clip, emotion graph, comments which separate into six emotions and music recommendation. As shown in figure 4.7 and figure 4.8.

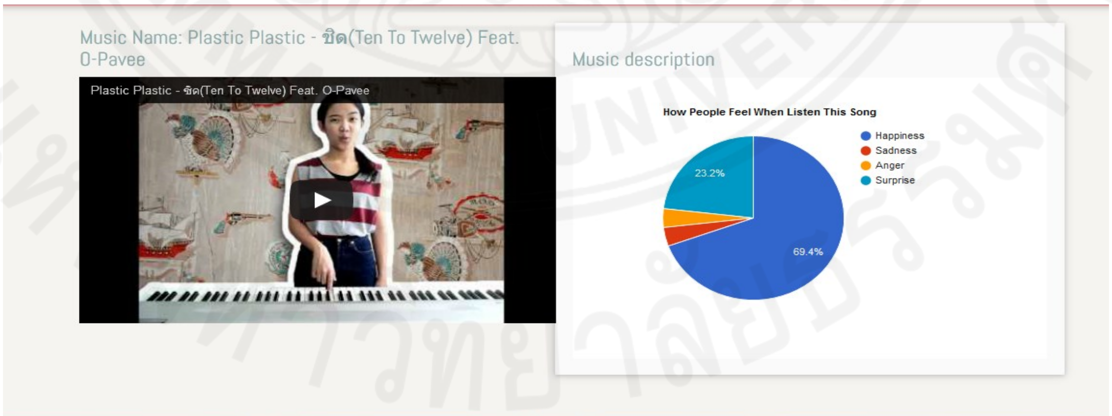


Figure 4.7: Music page

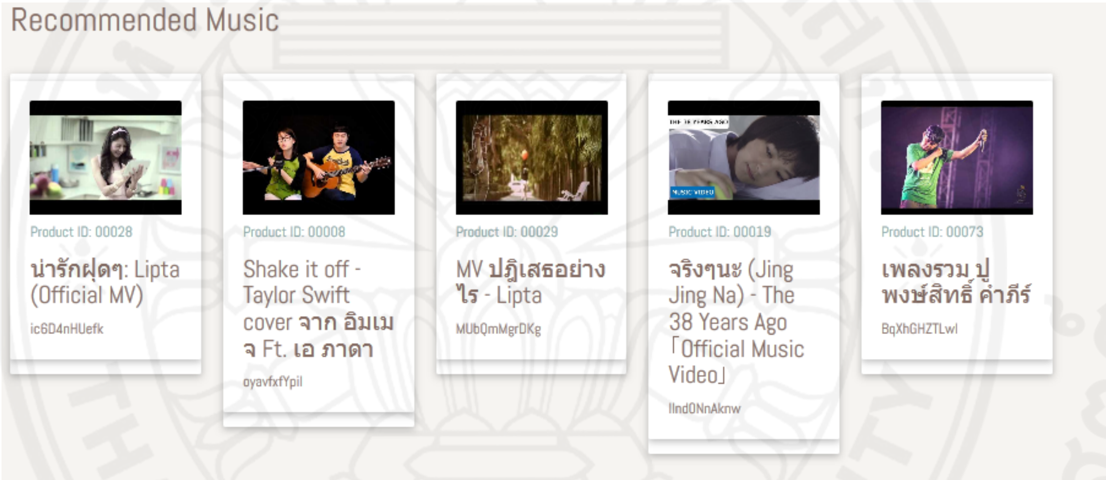


Figure 4.8: Music page recommendation

When users scroll down to the bottom of the page, user will see recommend music which selected top five similar emotion profile music to users as shown in figure 4.8.

4.5 Discussion

When the rapid growth of social media content in Thai Internet users, the contents are valuable for catching the feeling of their users. The emotion classification on Thai YouTube comment can automatically be classified to be the emotion in comment. In this experiment, we build emotion classification on YouTube comments. It consist of two classification models i.e., emotion filtering and emotion classification. The classification results by Support Vector Machine got the accuracy 93.30% on filtering task and 89.43% on classification task.

The second experiment presents a solution for unbalanced dataset problem. When the happiness and sadness classes are larger than the other classes; fear and disgust class, the majority class suppressed the minority class. So, we improve the performance of the unbalanced training data for the emotion classification, we proposed a sampling based algorithm called SMOTE. With this preprocess, the minority class performs well performance compared to the first experimental result. The method improves average precision from 0.751 to 0.896, average recall from to 0.753 to 0.894 and average f-measure from 0.744 to 0.894. While, TP rate in happiness and sadness classes drop from 0.848 to 0.797 and 0.829 to 0.79 respectively, but the rest parameter still good performance.

Chapter 5
Conclusions

This research focuses on the emotion classification on Thai YouTube comments. In this research, we proposed classifying emotion in Thai YouTube comments. We have classified into six basic emotions i.e., anger, disgust, fear, happiness, sadness and surprise. We have applied three machine learning algorithms; multinomial naïve Bayes, decision tree and support vector machine. The original experimental results show that SVM got accuracy with 76.41% on filtering task and 75.68% on classification task. Unfortunately in the first progress, we faced with unbalanced dataset problem. In Classification tasks, it's generally more important to correctly classify the minority class instances. However, in classification problems with imbalanced data, the minority classes are more likely to misclassify than the majority class due to machine learning algorithms design principles. To improve the performance, we use SMOTE: Synthetic Minority Over-sampling Technique. The experiment shows that we can solve the imbalanced data problem and obtain good result. In addition, comparing the result SVM using SMOTE yielded accuracy with 93.30% compared to 76.41% with original result. The result improves by 16.9% on Emotion Filtering dataset. Furthermore in emotion classification dataset the result of SVM observed from SMOTE was 89.44% compared to original result, which is 75.68%, there was 13.76% improvement. Moreover, varying the nearest neighbor parameter helps improving the accuracy. For the best number for the nearest neighbor parameter, we did a number of experiments with different number of parameter should give appropriate value. For our experiment, we found that at K=500 and 1,000 achieved the highest accuracy with SVM classifier in both datasets.

In addition, an equal count of each six basic emotion comments is very necessary. It affects the accuracy of the classifiers. Furthermore, ambiguous or irrelevant comments affect accuracy of the system. We need to eliminate any comments that are ambiguous or irrelevant. It is hard to classify ambiguous comments for example, surprise comment.

"ฟังแล้วแบบ...มันคงง่ายกว่านี้ถ้าเราเป็นเพศเดียวกัน ...เฮ้ย ขนลุกเลย...เนื้อเพลง จั๊กจี้ >,,, <"

"Listen that… it easier if we are same sex… hey! I got goose bumps. It's tickle lyric >,,,<"

From above comment the listener express their feeling that may feel afraid or they may really happy with this song.

As future work, we need to consider the way to improving emotion classification performance e.g., Comparing feature selection e.g., PCA and LSI. Adding more dictionary, common word and stop word. We need to clean a mistake data in our corpus and getting more data. Furthermore, we plan to apply our work to different genre e.g., advertisement, movie and news.

References

[1] YouTube. YouTubeStatistic, https://www.youtube.com/yt/press/statistics.html

[2] Orimaye S. O., Alhashmi S. M. & Eu-gene S. (2012). Sentiment analysis amidst ambiguities in YouTube comments on Yoruba language (nollywood) movies. WWW '12 Companion Proceedings of the 21st international conference companion on World Wide Web, 583-584

[3] Marquez F. B., Mendoza M. & Poblete B. (2013).Combining Strengths, Emotions and Polarities for Boosting Twitter Sentiment Analysis.WISDOM '13 Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining Article No. 2

[4] Strapparava C. & Mihalcea R. (2008).Learning to identify emotions in text. SAC '08 Proceedings of the 2008 ACM symposium on Applied computing, 1556-1560

[5] Nagarsekar U., Mhapsekar, A., Kulkarni, P. & Kalbande D.R. (2013). "Emotion detection from "the SMS of the internet"", 2013 IEEE Recent Advances in Intelligent Computational Systems (RAICS)

[6] Roberts K., Roach M. A., Johnson J., Guthrie J. & Harabagiu S. M. (2012).EmpaTweet: Annotating and Detecting Emotions on Twitter. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)

[7] Ahmed A. A. Esmin., de Oliveira R.L. & Matwin S. (2012).Hierarchical Classification Approach to Emotion Recognition in Twitter. 2012 11th International Conference on Machine Learning and Applications (ICMLA), 381-385

[8] Burget R., Karasek J. & Smekal Z. (2011).Recognition of Emotions in Czech Newspaper Headlines. RADIOENGINEERING, 20(1).

[9] Gao W., Li S., Lee S. Y. M., Zhou G. & Huang C. (2013). Joint Learning on Sentiment and Emotion Classification. CIKM '13 Proceedings of the 22nd ACM

international conference on Conference on information & knowledge management, 1505-1508

[10] Wang W, Chen L. & Thirunarayan K. (2012).Harnessing Twitter 'Big Data' for Automatic Emotion Identification. Sheth A.P., 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, 587-592

[11] Lin, K. H. Y., Yang, C., & Chen, H. H. (2008). Emotion classification of online news articles from the reader's perspective. In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 1, 220-226.

[12] Jia, Y., Chen, Z., & Yu, S. (2009). Reader emotion classification of news headlines. International Conference on Natural Language Processing and Knowledge Engineering NLP-KE 2009, 1-6.

[13] Lertsuksakda R., Netisopakul P. & Pasupa K. (2014). Thai sentiment terms construction using the Hourglass of Emotions, 2014 6th International Conference on Knowledge and Smart Technology (KST), 46-50

[14] Haruechaiyasak, C., Kongyoung, S., & Dailey, M. (2008). A comparative study on thai word segmentation approaches. In Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology ECTI-CON 2008. 1, 125-128.

[15] Inrak, P. & Sinthupinyo, S. (2010). Applying latent semantic analysis to classify emotions in Thai text", 2010 2nd International Conference on Computer Engineering and Technology (ICCET), 6 , 450-454

[16] Chirawichitchai N. (2014).Emotion classification of Thai text based using term weighting and machine learning techniques. 11th International Joint Conference on Computer Science and Software Engineering (JCSSE), 91 – 96

[17] Haruechaiyasak, C., Jitkrittum, W., Sangkeettrakarn, C., & Damrongrat, C. (2008, December). Implementing news article category browsing based on text

categorization technique. In Web Intelligence and Intelligent Agent Technology WI-IAT'08, 3, 143-146.

[18] Ramyachitra D. & Manikandan P. (2014).IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS: A REVIEW. International Journal of Computing and Business Research (IJCBR), 5(4)

[19] Provost F. (2000).Machine Learning from Imbalanced Data Sets 101. Proc. Learning from Imbalanced Data Sets: Papers from the Am. Assoc. for Artificial Intelligence Workshop, 2000 (Technical Report WS-00-05)

[20] Chawla N. V. (2005).Data Mining for Imbalanced Datasets: An Overview Data Mining and Knowledge Discovery Handbook, 853-867

[21] Lokanayaki K. & Malathi A. (2013).Data Preprocessing for Liver Dataset Using SMOTE. International Journal of Advanced Research in Computer Science and Software Engineering, 3(11)

[22] Rivera W. A., Goel A. & Kincaid J. P. (2014).OUPS: a combined approach using SMOTE and Propensity Score Matching. 13th International Conference on Machine Learning and Applications, 424 – 427

[23] Chen Y. (2008).Learning Classifiers from Imbalanced, Only Positive and Unlabeled Data Sets. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 213-220

[24] Li J., Li H. & Yu JL. (2010).Application of Random-SMOTE on Imbalanced Data Mining. Fourth International Conference on Business Intelligence and Financial Engineering (BIFE), 130 – 133

[25] Cheawsakunwattana S. & Sinthupinyo S. (2014).Decision Tree Classification of Imbalanced Data Sets Using Different Weight On Synthesized Data. The Tenth National Conference on Computing and Information Technology NCCIT2014

[26] Li Y., Sun G. & Zhu Y. (2010).Data imbalance problem in text classification. Third International Symposium on Information Processing (ISIP), 301 - 305

[27] Wang J., Xu M., Wang H. & Zhang J. (2006).Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding. 8th International Conference on Signal Processing

[28] Naseriparsa M. & Kashani MMR. (2014).Combination of PCA with SMOTE Resampling to Boost the Prediction Rate in Lung Cancer Dataset. International Journal of Computer Applications, 77(3), 33-38

[29] Ponchob P. & Nitsuwat S. (2010). Recommender System for Notebook Purchasing using Content-Based Filtering. The 6TH National Conference on Computing and Information Technology NCCIT2010.

[30] Tammuangjai P. & Utakrit N. (2012). A Recommender Model for Credit Product using Collaborative Filtering: A Case Study Government Savings Bank. The Eighth National Conference on Computing and Information Technology NCCIT 2012.

[31] Tewari A.S. Kumar, A. & Barman, A.G. (2014). Book recommendation system based on combine features of content based filtering, collaborative filtering and association rule mining. 2014 IEEE International Advance Computing Conference (IACC), 500-503

[32] Fan J., Pan W. & Jiang L. (2014). An improved collaborative filtering algorithm combining content-based algorithm and user activity. 2014 International Conference on Big Data and Smart Computing (BIGCOMP), 88-91

[33] Iwahama, K., Hijikata, Y. & Nishida S. (2014). Content-based Filtering System for Music Data. International Symposium on Applications and the Internet Workshops, 2004. SAINT 2004 Workshops, 480-487

[34] Li H., Cai F. & Liao Z. (2012). Content-Based Filtering Recommendation Algorithm Using HMM. Fourth International Conference on Computational and Information Sciences (ICCIS), 275-277

[35] Kim J., Lee S. & Yoo W. (2013). Implementation and analysis of mood-based music recommendation system. 15th International Conference on Advanced Communication Technology (ICACT), 740-743

[36] Janssens O., Slembrouck M., Verstockt S., Van Hoecke S. & Van de Walle R.(2013).Real-time emotion classification of tweets. 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 1430 – 1431

[37] Arafat J., Halimu C., Habib M. A. & Hossain R. (2013).Emotion Detection and Event Prediction System.Global Journal of Computer Science and Technology Network, Web & Security, 13(13)

[38] Patil T. & Patil S. (2013).Automatic Generation of Emotions for Social Networking Websites using Text Mining. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 1-6

[39] Li J., Xu Y., Xiong H. & Wang Y. (2010).Chinese Text Emotion Classification Based on Emotion Dictionary. 2010 IEEE 2nd Symposium on Web Society (SWS), 170-174

[40] Wang C., Quan C. & Ren F. (2010). Maximum entropy based emotion classification of Chinese blog sentences. 2010 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), 1-7

[41] Jia Y., Chen Z. & Yu S. (2009).Reader Emotion Classification of News Headlines.International Conference on Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009, 1-6

[42] Lin K.H., Changhua Y. & Chen H. H. (2008).Emotion Classification of Online News Articles from the Reader's Perspective. 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 220-226

[43] Pang B., Lee L. & Vaithyanathan S. (2002). Thumbs up?: sentiment classification using machine learning techniques. EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 10, 79-86

[44] Wakade S., Shekar C, Liszka K. J. & Chan C. C.(2012). Text Mining for Sentiment Analysis of Twitter Data. In International Conference on Information and Knowledge Engineering (IKE'12), 109-114

[45] M. Tim Jones, (2012). Recommender systems, Part 1: Introduction to approaches and algorithms. Retrieved from December, 12, 2013, http://www.ibm.com/developerworks/library/os-recommender1/

[46] Surangkana V. (2015). Thailand Internet User Profile 2015. https://www.etda.or.th/documents-for-download.html

Appendices

Appendix A
List of Publication

1. Sarakit P., Theeramunkong T., Haruechaiyasak C. and Manabu O., Classifying Emotion in Thai YouTube Comments, The International Conference on Information and Communication Technology for Embedded System (IC-ICTES 2015), March 2015.
2. Sarakit P, Theeramunkong T and Haruechaiyasak C., Improving Emotion Classification in Imbalanced YouTube Dataset Using SMOTE Algorithm, The 2015 International Conference on Advanced Informatics: Concepts, Theory and Application (ICAICTA2015), 19-22 August 2015