TRACKING COMMUTER TRAIN INTRUSION THROUGH TWITTER CRAWLING

Lya Hulliyyatus Suadaa
Sekolah Tinggi Ilmu Statistik

Abstract

Nowadays, Twitter is a very popular social media, especially in Indonesia. Tweets can be used as a data source to explore information. PT KAI (Kereta Api Indonesia) Commuter Jabodetabek (Jakarta, Bogor, Depok, Tangerang, Bekasi) (PT KCJ) has an official account, Twitter @CommuterLine, to disseminate information related to commuter train. One of important information regularly published in @CommuterLine account are information about commuter train intrusion. PT KCJ uses specific tweet format and certain hashtag to inform people about the intrusion. Intrusion information that is usually published is time of the intrusion, name of the station, train number and train line. #InfoLintas and #InfoLanjut hashtag are used to easier tweet searching. Information extraction processes are adopted to automatically extract commuter train intrusion information from @CommuterLine account Twitter. The statistical analysis about commuter train tweets are visualized in tables and graphs. A prototype system in the form of mobile application is developed to track commuter train intrusion based on the result of the information extraction.

Keywords : *information extraction; commuter train; Twitter crawling; social network mining*

INTRODUCTION

Research Background

Twitter is an online social networking service whose mission is to give everyone the power to create and share ideas and information instantly through the messages called "tweet". Based on Twitter report in 2013, Indonesia was ranked fifth in worldwide user Twitter, with 29 million users, after the US, Brazil, Japan and the UK. Moreover, Jakarta has become the most active Twitter city, exceeding London and Tokyo, which are the second and the third most active cities, respectively [1]. The massive user growth indicates the increasing popularity of Twitter.

As information sources, popularity of Twitter has led to the development of applications and research in various domains such as presidential election [2], public opinion analysis for e-government [3], traffic condition information extraction [4] and tracking flu inflections [5]. Thus, tweets can be used as a data source to explore information.

PT KAI (Kereta Api Indonesia) Commuter Jabodetabek (Jakarta, Bogor, Depok, Tangerang, Bekasi) (PT KCJ) is a subsidiary of PT KAI that manages commuter train of Jabodetabek. Commuter train is one of the important public transportation that is used daily by commuters. PT KCJ uses many media to disseminate information related to train commuter, such as www.krl.co.id website and official Twitter account @CommuterLine. The popularity of Twitter as a data source encourages author to exploit Twitter to analyze and track commuter train intrusion.

Research Question

The research question of this research is how to extract information from Twitter to analyze and track commuter train intrusion.

Objectives

The objectives of this research are the following:

a. Analyze the statistics of train commuter tweets.
b. Analyze and track commuter train intrusion through Twitter.
c. Develop a prototype system to track commuter train intrusion through Twitter crawling.

REQUIREMENTS

User Requirement

Observation has been done on the timeline of the official Twitter account @Commuterline. The results of the observation are the following:

1. There are frequent commuter train intrusions.
2. When commuter train intrusion occurs, @Commuterline would tweet information about the intrusion.
3. The number of followers of @Commuterline account is about 276.000.
4. There is a specific format related to the intrusion tweets

using hashtag #InfoLintas and #InfoLanjut.

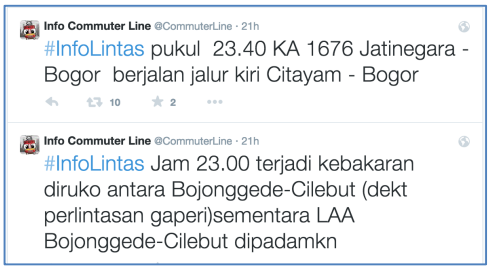

Figure 1. @Commuterline Account Timeline

Besides, a short survey has been done to 10 commuter train customers. The questions are:

1. Where do you frequently get information about the intrusion from?
   a. news b. social media c. others
2. Where do you frequently access the information from?
   a. computer b. smartphone/tablet c. others
3. Are you following account Twitter @Commuterline?

The results of this survey are the following:

1. 60% customers get information about the intrusion from social media, 20% from news, and 20% from others.
2. 70% customers use smartphone/tablet to get information about the intrusion, 10% use computer, and 20% use others.
3. 60% customers follow account Twitter @Commuterline.

Based on the observation and survey results, majority users access the social media, especially account Twitter @Commuterline, from smartphone/tablet to get information about commuter train intrusion. Users need to access their social media first to get the information.

System Requirement

System requirement based on user needs are the following:

1. Application to extract real time commuter train intrusion information from Twitter.
2. Mobile application to access commuter train intrusion information and the pattern with statistics of the intrusion.

METHODOLOGY

The methodology used in this research follows information extraction process shown in Figure 2 [6].

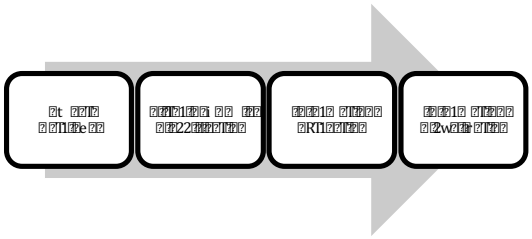

Figure 2. Information Extraction Processes

Tweet Retrieval

In this process, author collects data related to commuter train from the official Twitter account @CommuterLine and specific hashtag. Hashtag is a word or phrase preceded by a hash or pound sign (#) and is used

to identify messages on a specific topic. A certain hashtag can ease the searching process. There are some hashtags that are usually used in Twitter to categorize tweet related commuter train, as follows:

- #InfoCommuter

  This hashtag usually is used in tweets that inform commuters about general information related to commuter train, such as train schedule, train intrusion, and lost item found in train.

- #InfoLintas

  This hashtag is usually used in tweets that inform commuters about commuter train traffic, such as train intrusion, train cancellation, and train line change.

- #InfoLanjut

  This hashtag is usually used in tweets that inform commuters about the continuity or progress of commuter train intrusion.

- #RekanCommuters

  This hashtag is usually used in tweets that welcome or reply commuters.

Twitter provides Twitter APIs to facilitate user to use Twitter functions. There are two kinds of APIs, Twitter REST API and Twitter Streaming API. The REST API provides access to read and write Twitter data while the Streaming API is designed to capture Twitter data from a continuous stream of data in real time.

Filtering/Topic Classification

In this step, author classifies tweets into two categories: tweets that related to the commuter train intrusion and tweets that are related to intrusion solution. The objective is to build a corpus for rule development. Approach used to define model or rule for intrusion Twitter classification is a statistical-based approach. Author used Weka (Waikato Environment for Knowledge Analysis) application as a tool. Weka is a software containing collection of machine learning algorithms for data mining tasks developed by Waikato University [7]. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Figure 3 shows the process flow of model classification development used in this research.

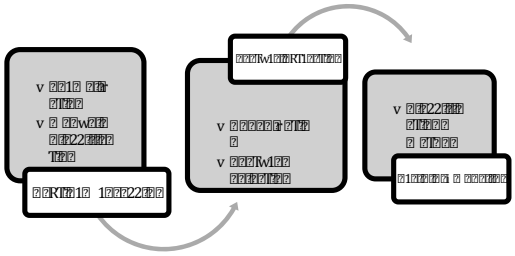

Figure 3. Process Flow Of Model Classification Development

a. Text Preprocessing

   Twitter data contain of unstructured text that needs to be cleaned before processing. Text preprocessing is useful to

ease the data processing in the next step. The processes as follow:

- Formalization
  Tweets usually use an informal language style. In this step, tweets are formalized to a formal language style and some abbreviations are converted.
- Manual Classification
  In this step, author builds corpus as text database for learning process when building a model. The corpus is developed by manually classifying the tweets.

b. Feature Extraction
Author uses the preprocessing feature in Weka application to extract the feature from the corpus. In this step, the feature is extracted from the corpus, as follows:

- Tokenization
  Tokenization is the process of splitting a sentence into words, phrases, symbols, or other meaningful elements. Tokens are separated by whitespace or punctuation characters. In this process, tweets are transformed into bag of words with the StringToWordVector feature provided by Weka.
- Feature selection
  Feature selection is useful to reduce words in every sentence and simplify model development. In this process, author uses stopwords and TF/IDF to select words that are important in the model development. TF/IDF stands for Term Frequency, Inverse Document Frequency, that scores the importance of words (or "terms") in a document based on how frequently they appear across multiple documents or sentences. If a word appears frequently in a document, it is important. However, if a word appears in many documents, so it is not a unique identifier.

c. Training/Modeling
Bag of words and the label are used as training data to develop a model. A classification method is applied to obtain rules to classify the tweets that are related to the train commuter intrusion and those that are not related. Author uses the classification feature in Weka application to build a model.

Information Extraction

Information extraction are conducted by rule-based approach. The rules are built in the previous modeling process. The information extracted from tweets that are related to the commuter train intrusion are the following:

a. Date and Time
   Date and time are extracted from Twitter API. If time details are available from a tweet, the time that is extracted from every tweet is prioritized.
b. Location
   The name of stations are extracted from every tweet and matched with the station dictionary.
c. Train Identity
   Train number and train line are extracted from every tweet and matched with the dictionary.
d. Intrusion Detail
   Intrusion detail is extracted from every tweet.

Information Visualization

The statistical analysis about commuter train tweets are visualized in tables and graphs. A prototype system in the form of a mobile application is developed to track commuter train intrusion based on the result of the information extraction.

ANALYSIS AND DESIGN

Tweet Statistics

Author collects data related commuter train from the official Twitter account @CommuterLine through Twitter API. A simple program developed by using Python has been used to crawl the data. Because of the limitation of Twitter API, the author can only crawl the last 3249 tweets from @CommuterLine. Table 1 shows the summary of @CommuterLine tweets based on Not Reply and Reply type. Not Reply means that the purpose of the tweets is not to reply other tweets and Reply means that the purpose of the tweets is to reply others.

Table 1. @CommuterLine Tweets Summary

No	Hashtag	Type		Total
		Not Reply	Reply	
1	#GoogleDoodle	1	0	**1**
2	#InfoCommuter	95	2	**97**
3	#infokrl	13	0	**13**
4	#InfoLanjut	4	0	**4**
5	#InfoLintas	110	0	**110**
6	#IniAksiKu	1	0	**1**
7	#instafankcj	0	1	**1**
8	#JadwalKRL	0	607	**607**
9	#Rekan	1	0	**1**
10	#RekanCommuters	20	0	**20**
11	#sekilasinfo	117	0	**117**
12	#SERRRPRIZE	2	0	**2**
13	#TarifProgresif	0	2	**2**
14	(blank)	12	2261	**2273**
Total		**376**	**2873**	**3249**

To reduce the bias of information extracted from the tweets, we filter the Not Reply tweets. Figure 4 shows proportion of hashtag used by @CommuterLine from 376 tweets.

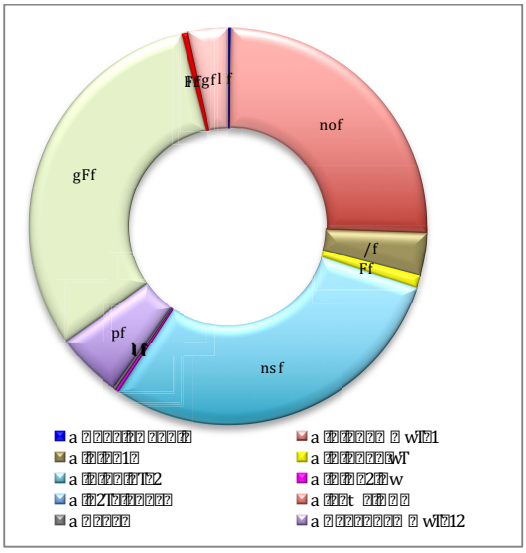

Figure 4. Proportion of Usage Hashtag in @CommuterLine Tweets

Tweets that are related to intrusion are tweets using #InfoLintas and #InfoLanjut hashtags. The #InfoLintas hashtag is usually used in tweets that inform commuters about commuter train traffic, such as train intrusion, train cancellation, and train line

change. The #InfoLanjut hashtag is usually used in tweets that inform commuters about the continuity or progress of commuter train intrusion.

Figure 5 shows the word cloud of tweets with hashtags #InfoLintas and #InfoLanjut. Words that most frequently appear are “InfoLintas”, “KA”, “perjalanan”, “hanya”, “sampai”, “Manggarai”. Manggarai is most related station when train intrusion occurs.

Figure 5. The word Cloud of #InfoLintas and #InfoLanjut Tweets

There are two types of #InfoLintas and #InfoLanjut tweets, tweets that mention train intrusion classified as intrusion tweets and tweets that mention progress or continuity of intrusion classified as solution tweets. Figure 6 and Figure 7 show samples of intrusion tweets and solution tweets, respectively.

Figure 6. Samples of Intrusion Tweets

Figure 7. Samples of Solution Tweets

Figure 8 shows the proportion of tweets type in #InfoLintas and #InfoLanjut tweets. There are 29% tweets about intrusion that mention details of intrusion and 71% tweets about solution that represent the progress of intrusion.

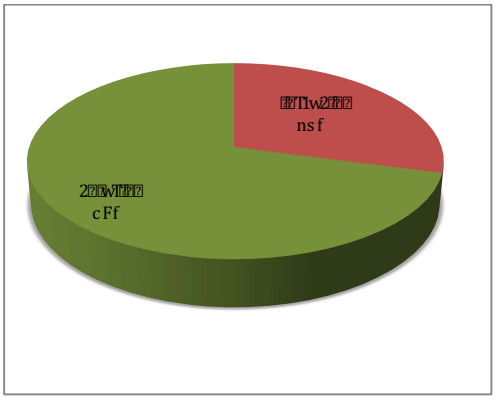

Figure 8. Proportion of Tweets Type in #InfoLintas and #InfoLanjut Tweets

Figure 9 shows word cloud of intrusion tweets. Words that most frequently appear are “InfoLintas”, “KA”, “mengalami”, “gangguan”, and “rangkaian”.

Figure 9. Word Cloud of Intrusion Tweets

The number of commuter intrusion by type and day in a month is shown in Table 2. The intrusion frequently occurs in Tuesday, Wednesday, Thursday, and Friday. The type of the intrusion is mostly unspecified. The rail and signal intrusion frequently occurs in a month.

Table 2. Intrusion by Type and Day in a Month

Intrusion Type	Sun	Mon	Tues	Wed	Thurs	Fri	Sat	Total
accident				2		1		**3**
electrici		1	1				1	**3**
ty								
environment					1	1		**2**
machine				1				**1**
rail		1	2	1	1			**5**
signal	1			1	1	2		**5**
train part					1	1	1	**3**
unspecified	1		3	2	3	1	1	**11**
Total	**2**	**2**	**6**	**7**	**7**	**6**	**3**	**33**

The number of commuter intrusion by time is shown in Table 3. The intrusion frequently occurs from 4:01 until 10:00 and from 16:01 until 23:00.

Table 3. Intrusion By Time

Time		Number of Intrusion	Total
4:01-10:00	4:01-5:00	1	14 (42.42%)
	5:01-6:00	2	
	6:01-7:00	3	
	7:01-8:00	1	
	8:01-9:00	3	
	9:01-10:00	4	
10:01-16:00	10:01-11:00	0	5 (15.16%)
	11:01-12:00	0	
	12:01-13:00	1	
	13:01-14:00	1	
	14:01-15:00	1	
	15:01-16:00	2	
16:01-23:00	16:01-17:00	1	14 (42.42%)
	17:01-18:00	6	
	18:01-19:00	1	

Time		Number of Intrusion	Total
	19:01-20:00	1	
	20:01-21:00	4	
	21:01-22:00	0	
	22:01-23:00	1	
Total		**33**	

Table 4 shows the percentage of commuter intrusion by station. The intrusion frequently occurs in Manggarai and Pasar Minggu station, which are two big stations in Jakarta.

Table 4. Percentage of Intrusion by Station

Station Name	Percentage of Intrusion
Cilebut	5%
Cilejit	3%
Citayam	5%
Depok	5%
Duri	8%
Grogol	3%
Jatinegara	3%
Kampung Bandan	5%
Karet	3%
Kemayoran	3%
Maja	3%
Manggarai	**13%**
Parung Panjang	5%
Parung Panjang	3%
Pasar Minggu	**13%**
Rawa Buaya	3%
Rawa Buaya - Pesing	3%
Serpong	3%
Sudimara	3%
Tanah Abang	3%
Tanah Abang - Palmerah	3%
Tanjung Barat	3%
Tebet	3%
Tenjo	3%
UI	3%

Rule Development

The following is the procedure to derive the rules:

1. Text Preprocessing
   Text preprocessing is done manually to ensure the quality of corpus development as follow:
   - Formalization
     Tweets usually uses an informal language style. In this step, tweets are formalized to a formal language style and some abbreviations are converted to their long form.
   - Manual Classification
     In this step, author builds a corpus as text database for learning process to build a model. The corpus is developed by manually classifying the tweets.
2. Feature Extraction
   Author uses the preprocessing feature provided by Weka application to extract the features from the corpus. The procedure to extract the features are as follow:
   - Tokenization

Author uses StringToWordVector feature in Weka to transform tweets to bag of words.

- Feature Selection
  Author uses stopword and TF/IDF to reduce words that are not relevant to build a rule.

Figure 10 shows preprocessing feature used in Weka application.

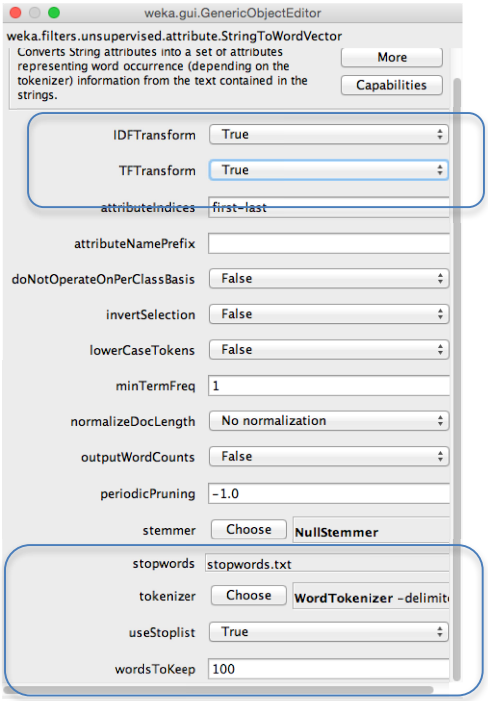

Figure 10. Preprocessing Feature in Weka Application

3. Text Classification
   Author uses classification feature in Weka application to build a model. J48 method is used to build the decision tree of the tweets. A decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. Dependent variable in this process is types of tweet containing two classes, "intrusion" and "solution". The models evaluated by k-fold cross validation, where k is 10. The k-fold cross-validation is a technique to evaluate predictive models by randomly partitioned the samples into k equal size subsamples, where k-1 subsamples are used as training data and a single subsamples is used to evaluate the models. Figure 10 shows the result of text classification.

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

gangguan <= 0
|   rel <= 0
|   |   jam <= 0: solution (59.0/1.0)
|   |   jam > 0
|   |   |   kembali <= 0
|   |   |   |   perjalanan <= 0
|   |   |   |   |   dilalui <= 0: intrusion (6.0)
|   |   |   |   |   dilalui > 0: solution (2.0)
|   |   |   |   perjalanan > 0: solution (4.0/1.0)
|   |   |   kembali > 0: solution (16.0)
|   rel > 0: intrusion (5.0/1.0)
gangguan > 0: intrusion (21.0)

Number of Leaves  :     7

Size of the tree :      13


Time taken to build model: 0.02 seconds
```

Figure 11. The Result of Text Classification

The decision tree based on the result in Figure 11 is shown in Figure 12. The tree is interpreted as follows:

a. If tweet has "gangguan" word, it is classified into intrusion class.

b. If tweet has not "gangguan" word but has "rel" word, it is classified into solution class.

c. If tweet has not "gangguan", "rel", and "jam" words, it is classified into solution class.

d. If tweet has not "gangguan" and "rel" words but has "jam" and "kembali" word, it is classified into solution class.

e. If tweet has not "gangguan", "rel" and "kembali" words but has "jam" and "perjalanan" words, it is classified into solution class.

f. If tweet has not "gangguan", "rel", "kembali" "perjalanan", and "dilalui" words but has "jam" word, it is classified into intrusion class.

g. If tweet has not "gangguan", "rel", "kembali", and "perjalanan" words but has "jam" and "dilalui" words, it is classified into solution class.

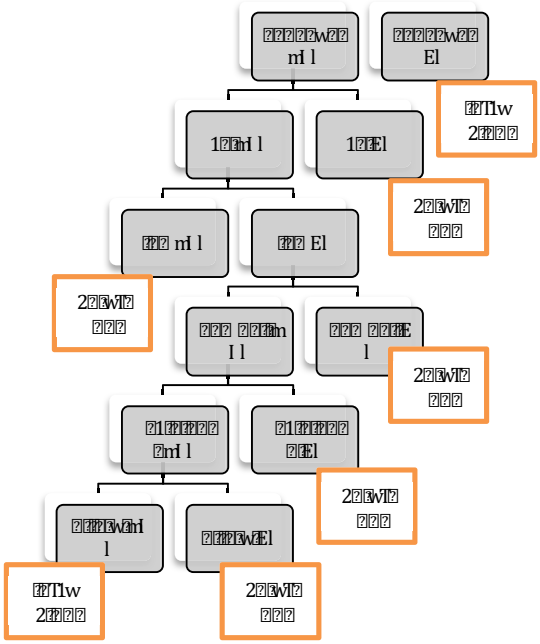

Figure 12. Decision Tree of Intrusion Classification

Figure 13 and 14 depict the accuracy of the result. Figure 13 shows that percentage of correctly classified tweets is 90.2655%. The detail of misclassified tweets are shown in confusion matrix in Figure 14.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         102               90.2655 %
Incorrectly Classified Instances        11                9.7345 %
```

Figure 13. Accuracy of The Result of Text Classification

```
=== Confusion Matrix ===

  a  b   <-- classified as
 24  9 |  a = intrusion
  2 78 |  b = solution
```

Figure 14. Confusion Matrix of The Result of Text Classification

System Design

The rule from analysis step is used in the mobile application to provide commuter train intrusion information, called Si GangKremut (Sistem Gangguan Kereta Komuter). The use case of the application is shown in Figure 15.

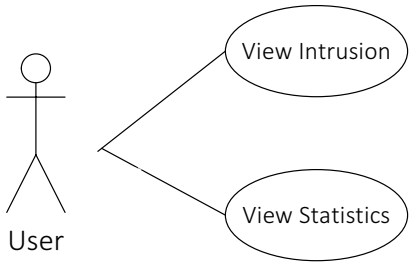

Figure 15. The Use Case Diagram of Si GangKremut

The features of the application are the following:

1. View commuter intrusion
2. View statistics of commuter intrusion

Figure 16 and 17 show the sequence diagram of the application representing process of viewing commuter intrusion and viewing intrusion statistics.

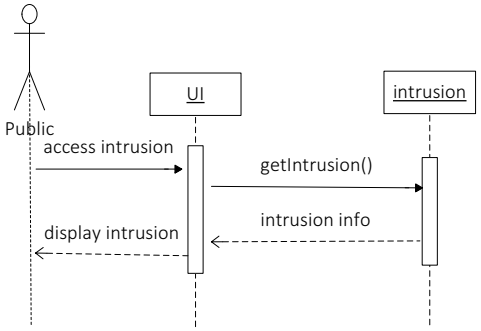

Figure 16. The Sequence Diagram of Commuter Intrusion View

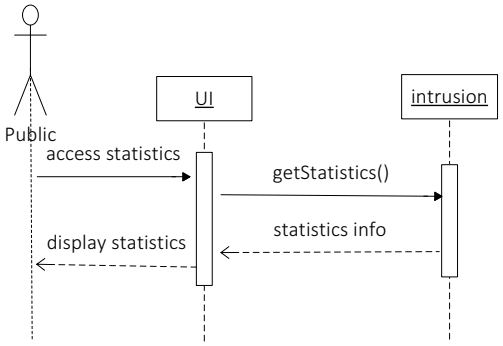

Figure 17. The Sequence Diagram of Commuter Intrusion Statistics View

Figure 18 shows the class diagram representing the data and methods used by the application.

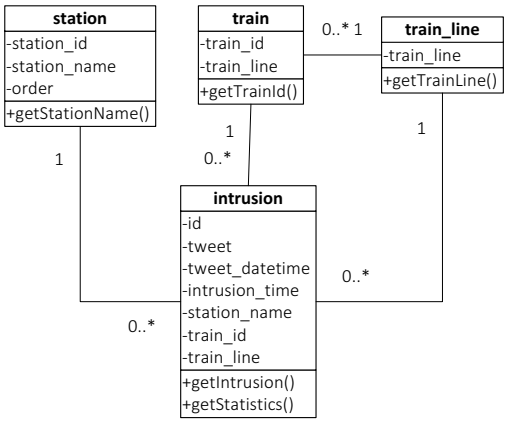

Figure 18. Class Diagram of Si GangKremut

PROTOTYPE SYSTEM

Prototype Design

The rules derived from training/modeling step is used in the crawler and information extraction application. The architecture of the applications is shown in Figure 19.

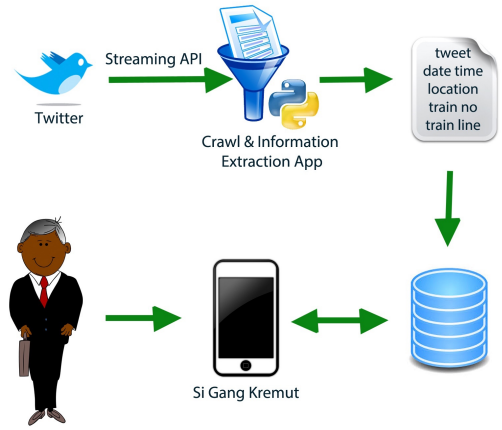

Figure 19. Application Architecture

Crawler and information extraction application has been built by Python to retrieve tweet and extract the information into database. Intrusion data forms a database accessed by mobile application through Javascript Object Notation (JSON) format. The mobile application has been built by Phonegap framework. PhoneGap is an open source solution for building cross-platform mobile applications with standards-based Web technologies such as HTML, JavaScript, and CSS. The features of the system are the following:

1. Display commuter intrusion with the details:
   a. Tweet
   b. Date and time of the intrusion
   c. Name of the station
   d. Train ID
   e. Train Line
2. Display statistics of commuter intrusion

The user interface of mobile application is shown in Figure 20 and Figure 21.

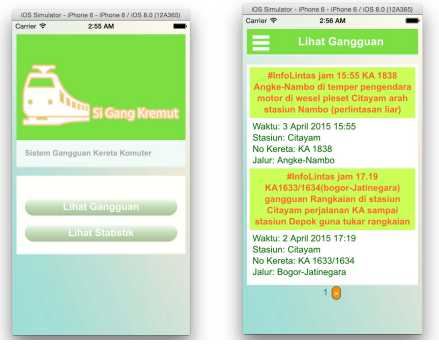

Figure 20. User Interface of Home Menu

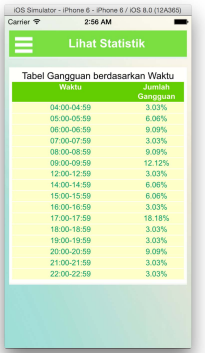

Figure 21. User Interface of View Intrusion and Statistics Menu

Prototype Testing

Testing scenario is used to ensure the functionality of the applications. Table 5 and 6 describe the test scenario of each use case.

1. Use case: View Intrusion

Table 5. Test Scenario of View Intrusion Use Case

ID	Test Case	Data Testing	Expected Result	Statue
S1.1	New intrusion success to be displayed	Tweet: #InfoLintas jam 15:55 KA 1838 Angke-Nambo di temper pengendara motor di wesel pleset Citayam arah stasiun Nambo (perlintasan liar)	Intrusion page with detail new intrusion	Pass

ID	Test Case	Data Testing	Expected Result	Statue
S1.2	New intrusion fail to be displayed	Tweet: Jam 15:55 KA 1838 Angke-Nambo di temper pengendara motor di wesel pleset Citayam arah stasiun Nambo (perlintasan liar)	New intrusion fail to be displayed because tweet doesn't have #InfoLintas hashtag	Pass
S1.3	New intrusion without time detail success to display time detail based on tweet time	Tweet: KA 1838 Angke-Nambo di temper pengendara motor di wesel pleset Citayam arah stasiun Nambo (perlintasan liar)	Intrusion page with detail time intrusion using tweet time	Pass
S1.4	Previous and next intrusion success to be displayed when user click paging button	-	Previous and next intrusion	Pass

2. Use case : View Statistics

Table 6. Test Scenario of View Statistics Use Case

ID	Test Case	Data Testing	Expected Result	Statue
S2.1	Percentage of new intrusion success to be increased in correct time	Tweet: #InfoLintas jam 15:55 KA 1838 Angke-Nambo di temper pengendara motor di wesel pleset Citayam arah stasiun Nambo (perlintasan liar)	Statistics page with increasing percentage in time: 15.00-15.59	Pass

Testing of information extraction application has been done with training data. The results show in Table 7.

Table 7. Accuracy of Information Extraction Application

No	Information	Accuracy
1.	Intrusion time	100%
2.	Station name	60.61%
3.	Train id	63.64%
4.	Train line	63.64%

Based on Table 7, information about station name, train id and train line are below 70% because of OOV (Out of Vocabulary) issue and OOR (Out of Rules) issue.

CONCLUSION

In this study, information extraction by rule based approach are adopted to automatically extract commuter train information from @Commuterline account twitter. The rule is derived from sample tweets by using classification methods. A decision tree as a result of the classification is customized to a rule for information extraction process and used in the crawler and information extraction application. Based on the testing, the prototype can automatically extract commuter train intrusion information. However, information about station name, train id and train line are below 70% because of OOV (Out of Vocabulary) issue and OOR (Out of Rules) issue.

REFERENCES

[1] Alz , "Twitter to open Indonesia office in Jakarta", The Jakarta Post [Online], http://www.thejakartapost.com/news/2014/08/29/Twitter-open-indonesia-office-jakarta.html , 2014. (Accessed: 8 March 2015).

[2] F. Nooralahzadeh, V. Arunachalam, C. Chiru, "2012 Presidential Elections on Twitter -- An Analysis of How the US and French Election were Reflected in Tweets", 19th International Conference on Control Systems and Computer Science (CSCS), 2014.

[3] A.W. Wijayanto, "Desain Sistem Terintegrasi Analisis Persepsi Publik pada Media Sosial Berbasis Internet of Thing untuk Pendukung e-Government Studi Kasus : Badan Pusat Statistik", Konferensi dan Temu Nasional Teknologi Informasi dan Komunikasi (TIK) untuk Indonesia, 2014.

[4] S. K. Endarnoto, S. Pradipta, A. S. Nugroho, J. Purnama, "Traffic Condition Information Extraction & Visualization from Social Media Twitter for Android Mobile Application", International Conference on Electrical Engineering and Informatics, 2011.

[5] A. Lamb, M. J. Paul, M. Dredze, "Separating Fact from Fear: Tracking Flu Infections on Twitter", Proceedings of NAACL-HLT 2013, pages 789–795, 2013.

[6] R. Hanifah, S. H. Supangkat, A. Purwarianti, "Twitter Information Extraction for Smart City, Case Study: Traffic Congestion of Bandung", International Conference on ICT For Smart Society (ICISS), 2014.

[7] http://www.cs.waikato.ac.nz/ml/weka/ (Accessed: 15 March 2015).

[8] M. A. Russell, Mining the Social Web, Second Edition, O'Reilly Media, Inc, USA, 2014.