Multiclass SMS Message Categorization: Beyond Spam Binary Classification

Fatia Kusuma Dewi, Mgs. M. Rizqi Fadhlurrahman, Mohamad Dwiyan Rahmanianto, Rahmad Mahendra
Faculty of Computer Science, University of Indonesia
{fatia.kusuma, mgs.m31, mohamad.dwiyan}@ui.ac.id, rahmad.mahendra@cs.ui.ac.id

***Abstract*—SMS spam has been growing since mobile phone usage increases. Past researches on SMS spam detection only classified SMS into two categories, spam and not spam. The binary classification of SMS spam prevents the user from seeing the spam messages that they do not really hate, e.g. an advertisement from their favorite product. In this paper, we propose multiclass classification of SMS into: regular, info, ads, and fraud. We use content-based (top-N unigram) as well as non-content based features. The result shows that the best accuracy is achieved by logistic regression that is 97.5 % accuracy with configuration of normalization preprocess and 4096 top-N unigram features.**

Mobile SMS; Spam; Classification; Multiclass

I. Introduction

The growth of the internet and mobile phone usage leads to increase of spam messages in all electronic media. Short Messaging System (SMS), a communication services using a mobile phone, suffered from abundance of spam messages. The growing of spam messages has been assisted by the easiness of sending spam messages to many users at once. Financial Services Authority of Indonesia received 14 thousand reports of fraud SMS in less than one month after the report service opened [1].

There are some factors that contribute to the growing of SMS spams. There are no well-implemented SMS spam filters either at the telecom operator (service provider layer) or SMS application (access layer), so spam messages can be delivered to user's mobile phone unfiltered. Unlike e-mail and other online communication services, SMS does not need the user to be connected to the internet ensuring the spam reach user's mobile phone. The other factor is the availability of cheap bulk SMS sender to send many SMS at once.

Spam messages can disturb mobile phone user by filling up their inbox and overriding important messages from the top of the inbox, distracting the user from reading it. Notification of received SMS spam can bother users too when expecting a relevant message for them. For this reason, spam message could disrupt the productivity of using SMS as communication media. Spam messages are irrelevant or unsolicited messages sent typically to a large number of users. To deal with SMS spam problems, a task called spam detection has been introduced. The method used to detect spam is typically binary classification. An SMS message is categorized into either of two classes: spam or regular message.

[1]media report in https://ekbis.sindonews.com/read/1205678/178/ojk-terima-belasan-ribu-aduan-sms-penipuan-1494950221

One of the spam's types is fraud messages whose the contents can be announcement of winning a fake prize draws or request for money on behalf of people they know like family or friends. Whereas, the other type of spam messages is the advertisement. The reason SMS used as advertisement media is that there are some people who respond to the advertisement SMS [1]. Advertisement may disturb some users, but some users may get benefits of the promotion from products they like. Common SMS spam classification works put the advertisement category into spam, which hides it from users' view [2], [3], [4]. For that reason, we design a new way to categorize spam messages.

We propose task of multi-class SMS categorization. This task is inspired by e-mail classification that split email into several categories [5] which has been already implemented by some email providers. Spam messages will be split into ads (advertisement) and fraud, so the user may view advertisement and promotion messages easier. In addition, we introduce another type of message, namely info, which contains information like bills and the internet's quota of the user from their service provider. SMS data that were used in our experiment is the messages written in Indonesian language.

II. Related Works

Majority works of SMS spam used English SMS data [2], [3], [4]. Related works in other languages have been conducted in Bahasa [6] and Malay language [7]. Most of prior works use supervised learning approach. There are several popular classifiers used in the experiments, such as Naive Bayes, Decision Tree, k-NN, SVM, and Logistic Regression. Naive Bayes and Support Vector Machine relatively have a good performance [8].

Other SMS filtering technique was implemented based on the concept of Artificial Immune System (AIS). The main goal of the spam immune system is to distinguish legitimate message and spam message. The main part of the AIS Engine is the detector, which is regular expression made by combining information from training process [9]. The proposed technique based on AIS Engine has 2% higher accuracy than Naive Bayes classifier. The experimental results show that the detection rate, false positive rate, and overall accuracy of the proposed technique are 82%, 6%, and 91% respectively.

There is another research in filtering spam SMS which inspired by the immune system. The proposed technique is using Dendritic Cell Algorithm (DCA) [10]. The algorithm combines

978-1-5386-3172-0/17/$31.00

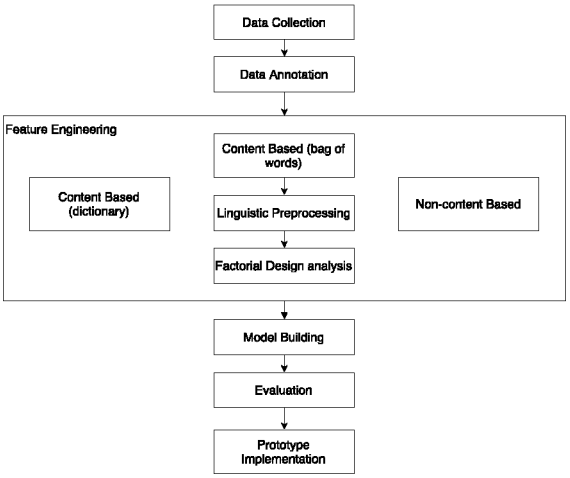

Fig. 1. Methodology Flowchart

NB and SVM classifier to generate the required signal. The experiment showed an excellent result as it achieved 99.95% accuracy.

To best of our knowledge, multi-class classification has not been extensively studied on SMS dataset. So, we reviewed the study of multi-folder categorization of e-mail. For the past years, interest on the multi-folder categorization of e-mail have been placed as the second rank after spam filtering [11]. On multi-folder categorization, most widely used features are e-mail header, e-mail body, terms, phrases, lexical and non-lexical [11]. Some other works on the multi-folder categorization of e-mail include [5], [12].

In their work, [5] had two representation of a document, as a character based N-gram (with N consist of 2 and 3) and as word-based. This work used multilingual personal e-mail dataset from an academic institution. There were 17 different classification categories; one of them was spam. The classification algorithms being used were Naive Bayes, SVM, IBk as instance-based learning, and PART as decision tree learning. The result showed that N-gram representation has slightly better accuracy than word-based representation using 2000 features. Accuracy score was around 95%. This work also concluded that if the feature's amount tends to be small (e.g. less than 300), word-based representation has relatively poor accuracy (between 60% to 75%).

[12] used Parallel SVM model based on iterative MapReduce for improving the computation speed when working on multi-folder e-mail categorization. They used only bag-of-words feature.

III. Methodology

Our study presented in this paper follows methodology as described in Figure 1. First, we collected and annotated the SMS data. Then, the texts of content are tokenized, normalized, and/or stemmed. The features are engineered from content of messages as well as non-content related information, e.g. messages' sender. We built the model using several machine learning algorithms. We evaluate the model using factorial design analysis and feature selection scenario. Finally, we implemented the prototype system for predicting category of a SMS message.

TABLE I
Terms Handled by String Replacement

Rule	Pattern Example
`Dot_Comma`	"." and ","
`Phone_number`	"08123456789", "628123456789"
`Service_number`	"*123*5#", "*5883#"
`Money`	"rp 10.000", "50jt", "Rp 10jt", "Rp50000"
`Date`	"04/07/2017", "13/12", "10/2017", "10/96"
`Percentage`	"50%", "100%"
`Number`	"1", "123", "54", "23"

A. Data Collection

A number of 3664 SMS messages were collected from inbox of six students. Those students were selected with voluntary sampling. The data collection process was supported by Android App that crawled the SMS message content and its metadata. Android App provided the output in form of a CSV file with several headers such as "id", "phone_number", "name", "body", and "has_replied".

B. Data Annotation

After the data being collected, every instance of SMS messages was annotated into 4 categories, such as "regular", "info", "ads", and "fraud". To make standardization of annotation, a guideline was created by observing the characteristics of each category. Three annotators annotated the data. Each annotator annotated the SMS message as many as possible in allocated time. In the end, 2542 SMS messages was successfully annotated and used in all phases of experiment. Annotated data consist of 400 "regular" messages, 980 "info" messages, 1123 "ads" messages, and 39 "fraud" messages. The example of each category can be seen on Table II.

C. Tokenization

Tokenization is splitting the whole text into list of tokens. A token is usually sequence of non whitespace characters. Before the texts are tokenized, particular substrings –most of them are sequence of numbers–were replaced with new predefined strings. The regular expression patterns were created for the string replacement. The purpose is to hinder the number related term to be tokenized carelessly. For example, "Rp 10jt" is tokenized to be "Rp" and "10jt" when ignoring the context. So, it is safer for tokenization process to replace first "Rp 10jt" with string `Money`. The string replacement also aims to generalize variation of writing, such as "Rp 10jt" and "10.000.000". The list of terms handled by string replacement process is summarized in Table I.

Each token is transformed into lowercase format.

978-1-5386-3172-0/17/$31.00©

TABLE II
EXAMPLE MESSAGE OF EACH CATEGORIES

Category	Example Messages	Translated Messages
Regular	Mas...saya sudah di depan wisma cornelius terus kemana lagii yaaa?	Hey... i already in the front of wisma cornelius. where do i go from here?
Info	PktInternet 4GB berakhir pd 01-05-2017 20:24.Pastikan pulsa cukup utk prpanjangan agar tdk kena tarif internet perKB.Info *123# atau klik im3.do/mc	Your internet quota of 4GB is expired in 01-05-2017 20:24. Make sure your credit is enough for automatic subscribing to prevent internet charge per KB. Info *123# or click im3.do/mc
Ads	Kini internetan di Singapura lbh praktis & hemat dengan IM3 Ooredoo. Daftarkan paketnya di *122*1#, mulai Rp25rb/hari (sblm PPN). Info: http://im3.do/smRO"	Now using internet in singapore is easier & cheaper with IM3 Ooredoo. Subscribe now at *122*1#, only from Rp25K/day (before tax). Info: http://im3.do/smRO
Fraud	selamat..SIM CARD anda resmi terpilih sebagai pemenang undian M-TRONIK TH2017 PIN pemenang anda 25b1374 untuk info klik: www.infoterpilihmtronik.ga	Congratulation.. your SIM CARD is choosen as the winner of M-TRONIK Y2017 lottery. Your pin is 25b1374 for more information click www.infoterpilihmtronik.ga

D. Preprocessing

As main features for SMS classification come from text content of the messages, we conduct linguistic pre-processing steps, including normalization, stemming, and stopword removal.

- Text Normalization
  Normalization was done by dictionary lookup. A dictionary containing slang words and the translation in standard formal language was manually compiled. The slang words are obtained from the most frequent term in word cloud.
- Stopword Removal and Stemming
  Stemming is a common technique used in information retrieval to obtain the basic form of a word. On the other hand, stopwords removal is the process of removing common words that often appear too frequent in whole dataset. The stopword removal and stemming process were merged as one linguistic technique in this research. The stemming and stopword removal process in our research was done by using the library from https://sastrawi.github.io/. In addition, stopword removal utilized customized dictionary.

E. Feature Engineering

Feature generation is the first part of feature engineering. The goal of this part is to produce some features for classification. We started from brainstorm phase in which we held the discussion to propose candidate of features. We discussed the datatype of proposed features and thought how to obtain the features from data. In devise phase of feature generation, we decided to incorporate two features sets, that are content-based and non content-based features. The first set consists of bag-of-words and and dictionary features. On the other hand, the latter consists of features that extracted from information about SMS sender. All features are represented as binary values.

- Top-N unigram
  Bag-of-word features consists of N most frequent words appearing in SMS data
- Dictionary features group
  A dictionary containing list of discriminative words was compiled by manual observation of word cloud of messages for each category.
  We looked for some words that tend to be more frequent on one specific class. For example, the word "pemenang" ("winner") word is often found in fraud category (SMS whose content is fake announcement about lottery winner); the word "internet" is most used in info category (Internet provider notifies the user periodically for the usage of their internet package).
- Official sender
  This feature is true if the phone number of the sender consist of numbers with length 1-6 or alphabet with space or dash. For example '9431', 'BANK', and '123'.
- Trusted sender
  This feature is true if the sender's phone number is listed in the whitelist of trusted sender. The whitelist was created based on the senders in the data that registered formally in an institution.
- Saved sender
  The value of this feature is determined by whether the sender's phone number is saved on contact list.
- Has replied
  This feature is true if at least one of previous SMS from the same sender has been ever replied.

To obtain more accurate predictive model, we applied the process to select a subset of relevant features for use in model construction. Wrapping approach is applied as features selection mechanism for top-N unigram. Meanwhile, univariate feature selection is applied to dictionary features group and non content-based features.

F. Model Building

We built the model utilizing four different supervised machine learning algorithms, namely Naive Bayes, Decision Tree, Logistic Regression, and k-Nearest Neighbor (after empirical experiment, we find best k for our data is 3). We ran the experiment using 10-fold cross validation setting. For evaluation, several metrics are measured, such as accuracy, confusion matrix, precision, recall, and F-1 measure.

G. Prototype Implementation

To implement the prototype, we built the system from the data annotation phase. A simple web application was created to annotate the data. After the data annotation phase, we still used the same database until the SMS app prototype creation.

978-1-5386-3172-0/17/$31.00

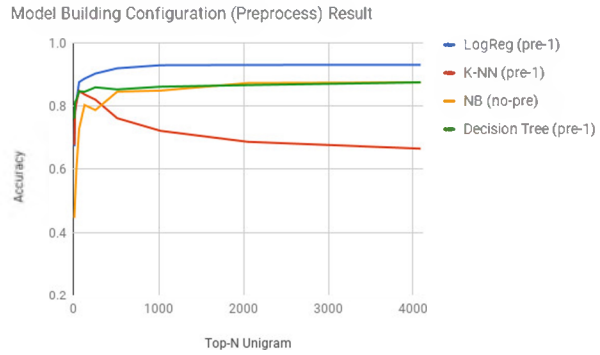

Fig. 2. Model Accuracy Per Algorithm and Preprocessing Technique

1) Infrastructure: We used MongoDB, a NoSQL Database Management System (DBMS), to build the prototype. All raw and processed SMS data were saved in the server.

2) Prediction API: To build the SMS app prototype, an Application Programming Interface (API) was created by using Go Programming Language. The API receives HTTP-POST request with SMS message data. Then, the model predicts the category of the SMS message.

3) Prototype: The SMS app prototype was created by using Ruby on Rails as the framework and PostgreSQL as the DBMS. The SMS app prototype could be accessed in our Heroku server[2].

IV. Results and Analysis

This section presents evaluation and result analysis of conducted experiment.

A. Result and Analysis of Configuration

We want to know which preprocessing techniques are realy needed to boost the classifiers. We tested each preprocessesing scenario into all algorithms in different top-N unigram.

Factorial design analysis [2] was used to determine the best configuration. There are two factors of factorial design being considered, which are linguistic pre-processing technique (LT) and the number of word with the highest frequency (top-N unigram).

The LT factor is the permutation of two types linguistic pre-processing techniques. The LT factor has four levels: no-preprocess (no-pre), normalization only (pre-1), stemming and stopword removal (pre-2), and all pre-processing steps (all).

The word with the highest frequency was obtained by sorting all words in the dataset based frequency. The top-N unigram factor has 10 levels, such as top-8 (2^3) unigrams, top-16 (2^4) unigrams, until top-4096 (2^10) unigrams. The entire experiment to select best configuration could be said as 4 x 10 factorial design. The total experiments are performed 40 times for each classifier.

[2]http://tangkis.herokuapp.com/

TABLE III
Univariate Analysis Custom Feature

Rank	Feature	P-Value
1	is_exist_pesan	3.18E-183
2	is_exist_blogspot	8.76E-139
3	offcial_sender	8.06E-85
4	has_replied	3.77E-83
5	is_exist_silakan	9.14E-75
6	is_exist_sisa	3.16E-73
7	is_exist_pemenang	5.19E-67
8	is_exist_hubung	6.72E-59
9	is_exist_mas	6.44E-49
10	saved_sender	2.62E-48
11	is_exist_internet	1.54E-40
12	is_exist_gratis	1.73E-40
13	trusted_sender	2.45E-38
14	is_exist_berhasil	1.98E-32
15	is_exist_gojek	5.12E-25
16	is_exist_balas	9.97E-24
17	is_exist_beli	8.95E-23
18	is_exist_ketik	9.84E-23
19	is_exist_aplikasi	8.20E-21
20	is_exist_dapatkan	7.13E-13
21	is_exist_trx	5.55E-07
22	is_exist_mbak	8.61E-05
23	is_exist_grab	7.95E-03
24	is_exist_uber	0.19

The accuracy that was examined during factorial design analysis is reported on Figure 2. Best accuracy score for Logistic Regression, 3-NN, and Decision Tree algorithms is achieved by applying pre-processing scenario 1 (pre-1), normalization only. Our result is pretty similar to previous work on SMS spam filtering. Using text normalization and concept generation would performed better for distance and trees classifier [3].

B. Result and Analysis of Feature Selection

Dictionary features group and non-content based features can be customized using feature selection. We did a univariate feature selection using independence test of Chi's Square between feature and output. We eliminated the feature whose p-value is above 0.05 (threshold). The result can be seen on table III. We did not take "is_exist_uber" feature because its p-value is around 0.19.

The chosen custom features were then trained on dataset along side with top-N unigram feature. We compared both model accuracy using top-N unigram features only and using top-N unigram features plus custom features. The accuracy and F1 measure comparison can be seen on table IV. The accuracy using combination of top-N unigram and custom features shows better result for all algorithm compared to accuracy that only using top-N unigram features. We decided to choose to use custom features because it increase the accuracy and didn't decrease the F1 measure.

C. Result and Analysis of Classification Algorithm

Model is built by using the best pre-processing scenario that already explained on subsection A and using custom features that already explained on subsection B. We also include the best top-N unigram Feature for each algorithm. We evaluate

978-1-5386-3172-0/17/$31.00

TABLE IV
ACCURACY WITHOUT AND WITH CUSTOM FEATURE (NOCF AND CF)

Algorithm	NoCF		CF	
	Accuracy	F1 Measure	Accuracy	F1 Measure
Decision Tree	89.7	89.0	93.8	89.0
K-NN	88.1	77.0	91.3	84.0
LogReg	94.3	97.0	97.5	97.0
NB	88.2	86.0	93.7	88.0

TABLE V
PRECISION AND RECALL EACH CATEGORIES LOGISTIC REGRESSION SAG

	Ads	**Info**	**Regular**	**Fraud**
Precision	96.32	98.25	99.11	100.00
Recall	98.55	95.89	99.11	92.86

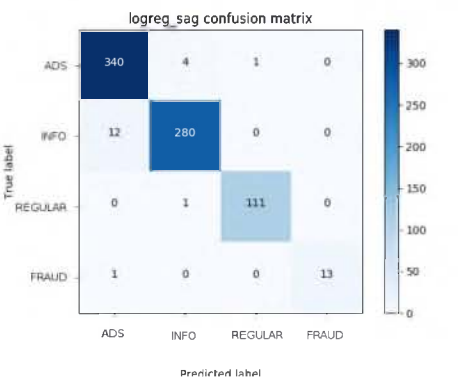

Fig. 3. Confusion Matrix of Logistic Regression SAG Solver

each models by calculating average precision, average recall, and average F1 Measure. The result can be seen on Table VI. Best F1 measure is achieved by logistic regression algorithm, that is 97.47%. Then, the second one is achieved by decision tree, that is 88.78%. Naive bayes has 87.62% accuracy. The least accuracy achieved by K-NN, that is 84.24%.

The confusion matrix of prediction result can be seen on Figure 3. The precision and recall score is reported on Table V. If we look further into precision and recall for each of the categories, both precision and recall have high value (above 95%). However, if we have to choose the importance between precision or recall for each category, it would be different for each category. Info and regular category must have high recall. We assume all those message could be categorized as important message, so we do not want to miss it. Whereas for ads and fraud category, we want to have a high recall. We have an an assumption to begin with that ads and fraud category have a niche group (or no group) of user who will see the messages. So, we do not want more important message categorized on ads and fraud category. However, it is tolerable when both categories is classified on other categories.

If we look at our testing data, the data for fraud category is imbalance relative to other categories. However, the recall score is pretty good. Recall of fraud category is 92.86%. Out of 14 data, there is one data that classified into ads category. There are fraud data that categorized as ads on naive bayes classifier and as info on K-NN classifier. We argue that this misclassification is because of fraud characteristic is more similar to ads and info. There are 5 out of 8 custom features that when the custom features is true for fraud category and they are also true for ads and info category.

V. CONCLUSION

This study conduct a multi-class classification for SMS message. The experiment includes a choice of linguistic technique for preprocess the SMS text. The experiment is using two types of feature, content-based (top-N unigram) and custom features.

All algorithms tested in our experiments can achieve an accuracy above 90% and F1 measure above 88%. The highest accuracy is obtained by logistic regresion algorithm that is

TABLE VI
MODEL EVALUATION USING AVERAGE PRECISION, AVERAGE RECALL, AVERAGE F1 MEASURE, AND ACCURACY

	Precision	**Recall**	**F1**	**Accuracy**
Decision Tree	92.12	88.89	88.79	93.84
K-NN	93.62	80.61	84.24	91.34
LogReg	98.42	96.60	97.47	97.50
NB	87.38	88.09	87.62	93.70

97.5%. We believe that high accuracy (more than 90%) upon all classifiers is supported by discriminative custom features. Applying chi-square's univariate analysis, we select the features whose p-value is below the threshold 0.05. Using the custom features could improve the accuracy, as the features most likely have a correlation to the SMS categories.

Our custom features creation is limited, because it is tailored and derived from observation of our dataset. For future research, we suggest collecting more SMS data in order to produce a generalized custom features. We also suggest to try another classification algorithm to seek a better classification.

REFERENCES

[1] A. Lambert, "Analysis of spam," M.Sc. Thesis, Trinity College Dublin. Department of Computer Science, 2003.

[2] M. V. Aragão, E. P. Frigieri, C. A. Ynoguti, and A. P. Paiva, "Factorial design analysis applied to the performance of sms anti-spam filtering systems," *Expert Syst. Appl.*, vol. 64, no. C, pp. 589–604, Dec. 2016. [Online]. Available: https://doi.org/10.1016/j.eswa.2016.08.038

[3] T. A. Almeida, T. P. Silva, I. Santos, and J. M. G. Hidalgo, "Text normalization and semantic indexing to enhance instant messaging and sms spam filtering," *Knowledge-Based Systems*, vol. 108, pp. 25 – 32, 2016, new Avenues in Knowledge Bases for Natural Language Processing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705116300909

[4] K. Zainal, N. Sulaiman, and M. Jali, "An analysis of various algorithms for text spam classification and clustering using rapidminer and weka," *International Journal of Computer Science and Information Security*, vol. 13, no. 3, p. 66, 2015.

[5] H. Berger, M. Dittenbach, and D. Merkl, "Analyzing the effect of document representation on machine learning approaches in multi-class e-mail filtering," in *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*. IEEE, 2006, pp. 297–300.

[6] I. Shaufiah and I. Asror, "Android short messages filtering for bahasa using multinomial naive bayes," 2006.

978-1-5386-3172-0/17/$31.00©

[7] M. Foozy, C. Feresa, R. Ahmad, and M. F. Abdollah, "A framework for sms spam and phishing detection in malay language: a case study," *International Review on Computers and Software*, vol. 9, no. 7, pp. 1248–1254, 2014.
[8] H. Sajedi, G. Z. Parast, and F. Akbari, "Sms spam filtering using machine learning techniques: A survey," *Machine Learning Research*, vol. 1, no. 1, pp. 1–14, 2016.
[9] T. M. Mahmoud and A. M. Mahfouz, "Sms spam filtering technique based on artificial immune system," *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 1, pp. 589–597, 2016.
[10] A. A. Al-Hasan and E.-S. M. El-Alfy, "Dendritic cell algorithm for mobile phone spam filtering," *Procedia Computer Science*, vol. 52, pp. 244 – 251, 2015, the 6th International Conference on Ambient Systems, Networks and Technologies (ANT-2015), the 5th International Conference on Sustainable Energy Information Technology (SEIT-2015). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050915008674
[11] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi, "Email classification research trends: Review and open issues," *IEEE Access*, 2017.
[12] K. Xu, C. Wen, Q. Yuan, X. He, and J. Tie, "A mapreduce based parallel svm for email classification," *Journal of Networks*, vol. 9, no. 6, pp. 1640–1647, 2014.

978-1-5386-3172-0/17/$31.00