Detecting Fake News with Machine Learning Method

Supanya Aphiwongsophon  
Department of Computer Engineering  
Faculty of Engineering  
Chulalongkorn University  
Bangkok, Thailand  
supanya.a@student.chula.ac.th

Prabhas Chongstitvatana  
Department of Computer Engineering  
Faculty of Engineering  
Chulalongkorn University  
Bangkok, Thailand  
prabhas@chula.ac.th

Abstract—Fake news has immense impact in our modern society. Detecting Fake news is an important step. This work purposes the use of machine learning techniques to detect fake news. Three popular methods are used in the experiments: Naïve Bayes, Neural Network and Support Vector Machine (SVM). The normalization method is important step for cleansing data before using the machine learning method to classify data. The result show that Naïve Bayes to detect the fake news has accuracy 96.08%. Two other more advance methods which are Neural Network and Support Vector Machine (SVM) achieve the accuracy of 99.90%.

Keywords—fake news; social network; Naïve Bayes; Neural network; Support Vector Machine (SVM)

I. INTRODUCTION

The huge amounts of information are generated on the social networking with various social media’s formats. There have provided very big volumes of posts that explosive increasing of the social media data on the web. When some event has occurred, many people discuss it on the web through the social networking. They search or retrieve and discuss the news events as the routine of daily life. However, very large volume of news or posts made users face the problem of information overloading during searching and retrieving. Unreliable sources of information expose people to a dose of fake news, hoaxes, rumors, conspiracy theories and misleading news.

Some type of news such as bad events from nature phenomenal or climate are unpredictable. When the unexpected events happen there are also fake news that are broadcasted that create confusion due to the nature of the events. Whom known the real fact from the event while the most people believe the forward news from their credible friends or relatives. The fake news comes from the misinformation, misunderstanding or the unbelievable contents which the creditability source. These are difficult to detect whether to believe or not when they receive the news information. Thailand is located in a tropical terrain. The rain is almost throughout the year therefore causes massive flooding in Thailand. Thai Meteorological department present the information of weather forecast, hydrological information and local climate. They have broadcasted the forecasting information to notify the public beforehand and protect their properties. However, the unpredictable natural phenomena’s news such as rain, floods, forest fires, earthquakes, storms, cold and hot weather which could be rapidly spread worldwide with misleading misunderstandings.

Examples in flood conditions may be the rumors such that the reservoir is broken. Flooding in places that have not actually occurred. It is rumored that the water does not flood, but actually the area is flooded. These rumors cause damage in the preparation of the actual disaster.

In this paper we propose the methods to detect fake news. The simple method is Naïve Bayes and the complex method are Neural Network and Support Vector Machine (SVM).

II. LITERATURE REVIEW

A. Definition of fake news

The creditability of information was defined by many words such as trustworthiness, believability, reliability, accuracy, fairness, objectivity, and other with the same concepts and definitions. [1]

There are several research use the machine learning approach to calculate the creditability of tweet’s message [2-5].

Fake news is the contents that claim people to believe with the falsification, sometime it is the sensitive messages. When the messages were received, they will rapidly dispersed it to other. The dissemination of fake news in today’s digital world has effected beyond a specific group. Mixing both believable and unbelievable information on social media has made the confusion of truth. That is the truth will be hardly classified. However, the appearance of fake news causes great threat on the safety of people’s lives and property. There are misinformation (the distributer believes there are true) or disinformation (the distributer knows it is not fact but he intentionally hoax) in fake news proliferation. [6, 7]

For example, In the Royal Cremation Ceremony of His Majesty King Bhumibol Adulyadej of Thailand, many fraud with the serving areas for observing and misinformation were broadcasted. The agenda of the event both the main ceremony
and each province are ambiguous news. There caused confusion to the people who wanted to join the event.

Many research use the sentiment analysis [8] and emotion classification to identify the fake news but it depend on the language’s content.[9]

B. Machine learning methods

This research use three methods to classify the believable and unbelievable message from Twitter, there are Naïve Bayes, Neural network, and Support Vector Machine (SVM).

- Naïve Bayes is the well-known classification method. We define the collected tweet data T and class of data (Cx) which x are believable and unbelievable. The probability of tweet data T in the class Cx can calculate as follow: [10]

\[
P(Cx|T) = \frac{P(T|Cx) \times P(Cx)}{P(T)}
\]

- Neural network is the mathematical model for the information computation process with the connectionism and the parallel distributed processing. This concept comes from the bioelectric network simulation in the neural system.[11]

- Support Vector Machine (SVM) is the classification method of supervised learning. There use the hyperplane to splits two data class’s point with the maximum margin. [12]

There are four evaluation result are precision, recall, F-measure and accuracy which computation from True Positive, True Negative, False Positive, False Negative.

True Positive is the number of messages that correct classify by believable messages.

True Negative is the number of messages that correct classify by unbelievable messages.

False Positive is the number of messages that incorrect classify by believable messages.

False Negative is the number of messages that incorrect classify by unbelievable messages.

The precision, recall, F-measure, and accuracy of classifier are calculate by formula as equation (2) to (5).

\[
Precision = \frac{True \ Positive}{True \ Positive + False \ Positive}
\] (2)

\[
Recall = \frac{True \ Positive}{True \ Positive + False \ Negative}
\] (3)

\[
F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall}
\] (4)

\[
Accuracy = \frac{True \ Positive + True \ Negative}{True \ Positive + True \ Negative + False \ Positive + False \ Negative}
\] (5)

In the figure 1 we collected data from Twitter with the selected topics. After the raw data are retrieved, we using the normalization rule to manipulated them. Next, the process of replication data removing was used. The last process of this work is the machine learning method for data classification that we use Naïve Bayes, Neural network and Support Vector Machine (SVM).

![Fig. 1. The overview of work.](image)

III. EXPERIMENTAL

The purpose of this work is to understand the characteristic of fake news through the features analysis on Twitter’s news that was posted between October 2017 to November 2017.

The twenty two attributes of raw data from Twitter API are Id, Name, IsVerified, ProfileImageUrl, FollowersCount, FriendsCount, FavouritesCount, StatusesCount, Description, Location, TimeZone, TweetCreatedDate, Status, Url, Mentions, Number of Mentions, HashTags, Number of HashTags, RetweetCount, TweetCreatedDate, MessageText and MessageImage.

The data was collected from 948,373 messages, thereafter had been categorized under Twitter API with the topics; There are many topics with the nature phenomena’s keywords such as floods (น้ำท่วม), Bangkok’s floods (น้ำท่วมกรุงเทพฯ), rainy (ฝนตก), cataract (น้ำตาล), cracked dam (เขื่อนแตก), cracked dike (ฝ่ายน้ำลักผัน), incoming storm (พายุเข้า), Kanhoon’s storms (พายุกันฮูน), Long’s typhoon (พายุลัง), depression (พายุไม่เข้า), Rangsit’s water release (ปล่อยน้ำท่วมรังสิต), water releasing after the royal ceremony day (ปล่อยน้ำร่วมงานถวาย rein ๙ พุทธมี), earthquakes (แผ่นดินไหว), into the winter (เข้าฤดูหนาว), cold’s disaster (ภัยหนาว), decreasing temperatures (อุณหภูมิต่ำ), climate change (สภาพภูมิอากาศแปรปรวน) and the royal cremation ceremony’s keywords such as the IXth reign (รัชกาลที่ ๙), the
royal ceremony (พระราชพิธีถวายพระเพลิง), the flower for the father (ดอกไม้จันทน์), Paklong’s market (ปักหลังตลาด), the exhibition of
royal ceremony (นิทรรศการพระราชพิธี), to the heaven (สู่ฟ้าเสวย
สวรรค์), forever in Thai’s mind (ถิ่นใจไทยยิ่งใหญ่บริสุทธิ์), to come to
the heaven (สู่ฟ้าเสวยสวรรค์), free oil fuel (เติมบั้มฟรี), Bangchak’s free oil (บางชากังวลฟรี), sandalwood’s flower
(ดอกไม้จันทน์), change the sandalwood’s flowers stalks cover
from black to white (พันก้านดอกไม้รินหมู่จากสีดำเป็นขาว), Chonburi’s
provincial governor migration (อำเภอจังหวัดชลบุรี).

The conditional rules for normalized data described as
followed.
The first key attribute is Id, which is the Twitter identical
number which consists of two characters attributes that has 9-
10 digits for the older account and 18 digits for the newer
account, so we separate them into two groups normalized.

The second attribute is Name, which is the name of
Twitter’s user which consists of five Condition rules; Thai
characters only, English characters only, the mix of Thai and
English characters, the symbols only, and the other languages
characters.

Third attribute is Iverified, it is the attribute that user
verify themselves with the Twitter. There are two conditions;
true is already verified and false is not verified.

Next attribute is ProfileImageUrl, which is the address
of URL link of the user’s image. There are two condition rules
of ProfileImageUrl; no image or the image with .jpg, .png
or/and other image’s formats.

FollowersCount is the number of followers that follow
each account. The number of followers is normalized from the
number to the digit number of followers that calculate by
divided with 10. Currently, it is in range between 1 and 7, that
start from 1 with means that account has 0-9 followers. It is not
over than 7 for 1,000,000-9,999,999 followers.

Friendcount, Favoritescount and Statusescount are the
number of friends of account, the number of the favorites of
account and the status of account respectively. There are the
number features same as the Followerscount, so that are use the
same condition.

Description is a user’s details that they describe
themselves. The condition of normalization this attribute is the
same as twitter’s name.

Location is where user posts the message. TimeZone is
the zone of time that the accounts were created. There are seven
different zones of location and time zone such as Thailand,
South East Asia, Asia, Australia/New Zealand, Europe/Russia,
US/Canada/Alaska/Hawaii and Africa. However, some user
were not declared their location and time zone or not specified
in seven zones list.

UserCreateDate is the date with user created their
account. The oldest of Twitter user account were 12 years since
Twitter was created in March 2006 and launched in July of the
same year. Therefore we separated it with 0.5 year such as 0.5,
1, 1.5, 2, … 11, 11.5, and 12.

Status is the attribute of user’s situation. The status will
contain two alternative values which contain ‘value’ or ‘none’
is no value.

Url describes the location of message that may or may
not link to other real destination message. However, Url may
contain no value or it can link to the real destination url with
one or more.

Mentions is the detail of users were cited in the message.
It was defined in the message after @ and maybe none or more
than one.

Number of Mentions is the quantity of @ that appear in
the message. This value came from Twitter API.

Hashtags is the detail that user want to describe or
specify topic of the message. It was defined in the message after
# . Hashtags may have none or more than one.

Number of HashTags is the quantity of # in the message.
It came from Twitter API same as number of mentions.

RetweetCount is the number of the message was tweet
again by one who is not owner. This condition is normalized the
number of retweet’s count to the digit number which the range
between 1 and 7 same as the number of followers.

TweetCreatedDate is the date and time with the message
is created. The value was normalized with the period of time
such as 6.01-12.00 A.M., 0.01-6.00 P.M., 6.01-12.00 P.M., and
0.01-6.00 A.M.

MessageText is the detail of the messages sent to each
other via Twitter. This message maybe the owner’s tweet or it
is retweet from the others. The condition of this value are the
owner of message or retweet.

MessageImage is the linking location between the url of
images and its related message. This value came from Twitter API.

Number of Mentions is the number of @ that appear in
the message. This value came from Twitter API.

The machine learning processes of this experiment are
Naïve Bayes, Neural network, and Support Vector Machine (SVM) with precision,
recall, F-measure and accuracy are illustrated in table I.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Naïve Bayes</th>
<th>Neural Network</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>99.80%</td>
<td>99.80%</td>
<td>99.80%</td>
</tr>
<tr>
<td>Recall</td>
<td>96.10%</td>
<td>99.90%</td>
<td>99.90%</td>
</tr>
<tr>
<td>F-Measure</td>
<td>97.90%</td>
<td>99.80%</td>
<td>99.80%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>96.08%</td>
<td>99.90%</td>
<td>99.90%</td>
</tr>
</tbody>
</table>

The experimental result in the table I, the results of
precision are not difference with all processing methods but the
results of recall, F-measure, and accuracy there are separate to
two group with Naïve Bayes vs. Neural network and Support Vector Machine (SVM). The result of recall, F-measure, and accuracy with Naïve Bayes are 96.10%, 97.90%, and 96.08% respectively. Neural network and Support Vector Machine are equivalently results with recall, F-measure, and accuracy are 99.90%, 99.80%, and 99.90% respectively.

IV. CONCLUSION
Fake news is the difficult problem because it is the rumors which it is too hard to identify the fact in contents. [13] The lie’s content was used into the corpus that has effected to the failure of IBM’s Watson prototype testing in the late 2016. [14] Motivation of the proliferation the truth and fake news requiring strenuous effort to detection. [15] It will be use fact checking to solve the fake news detection problem. [16-17]

Fake news can be accurately identified using machine learning methods. In the experiment, selected data collected from Twitter are profiled with twenty two attributes. From this information, all the machine learning methods: Naïve Bayes, Neural Network, Support vector machine, are very good at detecting Fake news with high confidence. Of course, it may not represent the whole spectrum of News in the real-world. However, there is enough evidence that Fake news is not too difficult to detect, at least in some selected domain. It is also difficult to say with confidence how much the result of this experiment can be applied to real-world news. We hope to broaden the scope of our data collection and try to apply our method in a more general way in the future.

ACKNOWLEDGMENT
This research was supported by “The 100th Anniversary Chulalongkorn University Fund for Doctoral Scholarship” and The 90th Anniversary of Chulalongkorn University Fund (Ratchadaphiseksomphot Endowment Fund).

REFERENCES