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 proposes 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. The normalization method is important step for cleaning data before using the machine learning method to classify data. The result show that Naïve Bayes to detect Fake news has accuracy 96.08%. Two other more advance methods which are Neural Network and Support Vector Machine achieve the accuracy of 99.90%.

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

I. INTRODUCTION

The huge amounts of information are generated on the social network with various social media's formats. They 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 network. 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 unfortunate 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. 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. A recent event (2017) in Thailand is a good example. 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 make people believe the falsification, sometimes 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 they 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 about the service areas for observing the ceremony 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].

978-1-5386-3555-1/18/$31.00 ©2018 IEEE

B. Machine learning methods

This research uses three methods to classify the believable and unbelievable message from Twitter. They 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 be calculated as follow: [10]

$$P(Cx|T) = \frac{P(T|Cx) \times P(Cx)}{P(T)} \quad (1)$$

- 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. It uses the hyperplane to split two classes data point with the maximum margin [12].

There are four evaluation results: are precision, recall, F-measure and accuracy. They are used to compute four measures: True Positive, True Negative, False Positive and False Negative.

True Positive is the number of messages that is correctly classified by believable messages.

True Negative is the number of messages that is correctly classified by unbelievable messages.

False Positive is the number of messages that is incorrectly classified by believable messages.

False Negative is the number of messages that is incorrectly classified by unbelievable messages.

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

$$Precision = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})} \quad (2)$$

$$Recall = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})} \quad (3)$$

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

$$Accuracy = \frac{\text{True Positive} + \text{True Negative}}{(\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative})} \quad (5)$$

Figure 1 describes the flow of the experiment. The data is collected from Twitter with the selected topics. After the raw data are retrieved, we use the normalization rule to manipulate 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).

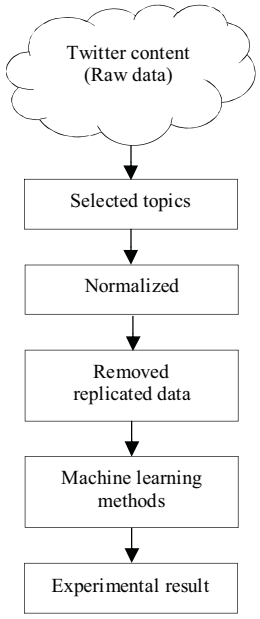

Fig. 1. The overview of work.

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 Octobers 2017 to November 2017.

The twenty two attributes of raw data from Twitter API are selected. They are: Id, Name, IsVerified, ProfileImageUrl, FollowersCount, FriendsCount, FavouritesCount, StatusesCount, Description, Location, TimeZone, UserCreatedDate, 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 (น้ำท่วมกรุงเทพ), rainy (ฝนตก), cataract (น้ำป่า), cracked dam (เขื่อนแตก), cracked dike (ฝายน้ำล้นแตก), incoming storm (พายุเข้า), Kanhoon's storms (พายุขนุน), Long's typhoon (พายุไต้ฝุ่นล้ง), depression (พายุดีเพลสชัน), Rangsit's water release (ปล่อยน้ำท่วมรังสิต), water releasing after the royal ceremony day (ปล่อยน้ำท่วมหลังวันที่ 26 ตุลาคม), earthquakes (แผ่นดินไหว), into the winter (เข้าสู่ฤดูหนาว), cold's disaster (ภัยหนาว) , decreasing temperatures (อุณหภูมิลดลง), climate change (สภาพอากาศแปรปรวน). The second group of topics are the royal cremation ceremony's keywords such as the IXth reign (รัชกาลที่ ๙), the royal ceremony (พระราชพิธีถวายพระเพลิง), the flower for the father (ดอกไม้เพื่อพ่อ), Paklong's market

(ปากคลองตลาด), the exhibition of royal ceremony (นิทรรศการงานพระราชพิธี), to the heaven (สู่ฟ้าเสวยสวรรค์), forever in Thai's mind (สถิตในดวงใจไทยนิรันดร์), to come to the heaven (ส่งเสด็จสู่สวรรคาลัย), free oil fuel (เติมน้ำมันฟรี), Bangchak's free oil (บางจากเติมฟรี), sandalwood's flower (ดอกไม้จันทน์), change the sandalwood's flowers stalks cover from black to white (พันก้านดอกไม้จันทน์จากสีดำเป็นขาว), Chonburi's provincial governor migration (ย้ายผู้ว่าชลบุรี).

The conditional rules for normalized data described as followed.

The first key attribute is Id. It is the Twitter identity 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.

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

Third attribute is Isverified. 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 conditions: 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 divided by 10. Currently, it is in range between 1 and 7, with 1 means that account has 0-9 followers. It is not over than 7 for 1,000,000-9,999,999 followers.

Friendscount, Favoritescount and Statusescount are the number of friends of account, the number of the favorites of account and the status of account respectively. They are calculated similar to Followerscount.

Description is a user's details describing 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'.

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 number of users 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.

HashTags is the tag 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.

RetweetCount is the number of the message was tweet again by one who is not owner. This value is normalized 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 message 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 condition maybe none or one or more of active image's links.

After normalization process, every raw data are set of numbers. The replication data will be removed.

After duplicated data removing, there were 327,784 messages for classified with machine learning process.

The result of experiment with Naïve Bayes, Neural Network, and Support Vector Machine (SVM) with precision, recall, F-measure and accuracy are illustrated in table I.

TABLE I. THE EXPERIMENT'S RESULT.

	Precision	Recall	F-Measure	Accuracy
Naïve Bayes	99.80%	96.10%	97.90%	96.08%
Neural Network	99.80%	99.90%	99.80%	99.90%
SVM	99.80%	99.90%	99.80%	99.90%

The machine learning used in this experiment are Naïve Bayes, Neural Network, and Support Vector Machine (SVM). Twitter messages are classified to two classes: believable and unbelievable. For the rest of measurement: The results from the classification are used to calculate precision, recall, F-Measure, and accuracy.

From the experimental result in the table I, precision are not different with all methods. For the rest of measurement: Precision, Recall, F-Measure and Accuracy, there are difference between Naïve Bayes and Neural Network and SVM. Naïve Bayes gives the lower measures than other two methods. 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]. For example, the report of attempt of IBM's Watson to detect classified news that includes lie content is negative [14]. Motivation of the proliferation the truth and fake news requiring strenuous effort to detection [15]. Human is require to perform fact checking to solve the fake news detection problem [16-17].

From the result of the experiment presented in this paper, 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

[1] Majed AlRubaian, Muhammad Al-Qurishi, Mabrook Al-Rakhami, Sk Md Mizanur Rahman, and Atif Alamri. 2015. A Multistage Credibility Analysis Model for Microblogs. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM '15), Jian Pei, Fabrizio Silvestri, and Jie Tang (Eds.). ACM, New York, NY, USA, 1434-1440. DOI: http://dx.doi.org/10.1145/2808797.2810065

[2] Majed Alrubaian and Muhammad Al-Qurishi. 2016. A Credibility Analysis System for Assessing Information on Twitter, IEEE Transactions on Dependable and Secure Computing, 1-14. DOI : http://dx.doi.org/ 10.1109/TDSC.2016.2602338

[3] Manish Gupta, Peixiang Zhao and Jiawei Han. 2012. Evaluating Event Credibility on Twitter, Proceedings of the 2012 SIAM International Conference on Data Mining, 153-164 , DOI: http://epubs.siam.org/doi/abs/10.1137/ 1.9781611972825.14

[4] Krzysztof Lorek, Jacek Suehiro-Wiciński, Michal Jankowski-Lorek and Amit Gupta. 2015. Automated credibility assessment on twitter, Computer Science, 2015, Vol.16(2), 157-168, DOI: http://dx.doi.org/10.7494/csci. 2015.16.2.157

[5] Ballouli, Rim El, Wassim El-Hajj, Ahmad Ghandour, Shady Elbassuoni, Hazem M. Hajj and Khaled Bashir Shaban. 2017. "CAT: Credibility Analysis of Arabic Content on Twitter." WANLP@EACL (2017).

[6] Alina Campan, Alfredo Cuzzocrea and Traian Marius Truta. 2017. Fighting Fake News Spread in Online Social Networks: Actual Trends and Future Research Directions, IEEE International Conference on Big Data (BIGDATA), 4453-4457

[7] Carlos Castillo, Marcelo Mendoza and Barbara Poblete. 2011. Information credibility on twitter. In Proceedings of the 20th international conference on World wide web (WWW '11). ACM, New York, NY, USA, 675-684. DOI: http://dx.doi.org/10.1145/1963405. 1963500

[8] Muhammad Abdul-Mageed, Mona Diab and Sandra Kübler. 2014. SAMAR: Subjectivity and sentiment analysis for Arabic social media, Computer Speech & Language, Elsevier BV, ISSN: 0885-2308, V.28, Issue: 1, 20-37, DOI: http://dx.doi.org/10.1016/j.csl.2013.03.001

[9] Niall J. Conroy, Victoria L. Rubin and Yimin Chen. 2015. Automatic deception detection: methods for finding fake news. In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community (ASIST '15). American Society for Information Science, Silver Springs, MD, USA, , Article 82 , 4 pages.

[10] Granik M., and Mesyura V. 2017. Fake news detection using Naïve Bayes classifier. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 900-903.

[11] Hertz, J., Palmer, R.G. and Krogh. A.S. 1990 Introduction to the theory of neural computation, Perseus Books. ISBN 0-201-51560-1

[12] Thandar M. and Usanavasin S. 2015 Measuring Opinion Credibility in Twitter. In: Unger H., Meesad P., Boonkrong S. (eds) Recent Advances in Information and Communication Technology 2015. Advances in Intelligent Systems and Computing, vol 361. Springer, Cham

[13] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata and Rob Procter. 2018. Detection and Resolution of Rumours in Social Media: A Survey. ACM Comput. Surv. 51, 2, Article 32 (February 2018), 36 pages. DOI: https://doi.org/10.1145/3161603

[14] W. Vorhies, Using algorithms to detect fake news - The state of the art, 2017, [online] Available: http://www.datasciencecentral.com/profiles/blogs/using-algorithms-to-detect-fake-news-the-state-of-the-art.

[15] Ehsanfar Abbas and Mo Mansouri, 2017. "Incentivizing the dissemination of truth versus fake news in social networks." 2017 12th System of Systems Engineering Conference (SoSE), 1-6.

[16] Hal Berghel. 2017. Alt-News and Post-Truths in the "Fake News" Era. Computer 50, 4 (April 2017), 110-114. DOI: https://doi.org/10.1109/MC.2017.104

[17] Buntain C. and Golbeck J. 2017. Automatically Identifying Fake News in Popular Twitter Threads. 2017 IEEE International Conference on Smart Cloud (SmartCloud), 208-215.