A Framework for Hoax News Detection and Analyzer used Rule-based Methods

SY. Yuliani[1]
Information Security and Networking Research
Group(InFORSNET), Faculty of Information
Communication Technology, Universiti Teknikal Malaysia
Melaka, Melaka, Malaysia, Widyatma University

Mohd Faizal Bin Abdollah[2]
Information Security and Networking Research
Faculty of Information Communication Technology
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

Shahrin Sahib[3]
Information Security and Networking Research
Group(InFORSNET), Faculty of Information
Communication Technology, Universiti Teknikal Malaysia
Melaka, Melaka, Malaysia

Yunus Supriadi Wijaya[4]
Informatics Department
Telkom University
Bandung, Indonesia

Abstract*—Currently, the era where social media can present various facilities can answer the needs of the community for information and utilization for socio-economic interests. But the other impact of the presence of social media opens an ample space for the existence of information or hoax news about an event that is troubling the public. The hoax also provides cynical provocation, which is inciting hatred, anger, incitement to many people, directly influencing behavior so that it responds as desired by the hoax makers. Fake news is playing an increasingly dominant role in spreading Misinformation by influencing people's Perceptions or knowledge to distort their awareness and decision-making. A framework is develope dataset collection of hoax gathered using web crawlers from several websites, using classification techniques. This hoax news will be categorized into several detection parameters including, page URL, title hoax news, publish date, author, and content. Matching each word hoax using the similarity algorithm to produce the accuracy of the hoax news uses the rule-based detection method. Experiments were carried out on eleven thousand-hoax news used as training datasets and testing data sets; this data set for validation using similarity algorithms, to produce the highest accuracy of hoax text similarity. In this study, each hoax news will label into four categories, namely, Fact, Hoax, Information, Unknown. Contributions propose Automatic detection of hoax news, Automatic Multilanguage Detection, and a collection of datasets that we gather ourselves and validation that results in four categories of hoax news that have measured in terms of text similarity using similarity techniques. Further research can be continued by adding objects hate speech, black campaign, blockchain technique to ward off hoaxes, or can produce algorithms that produce better text accuracy.

***Keywords*—Component; hoax; news; framework; web crawling; detection; multilanguage; unsupervised algorithm; similarity algorithm**

I. Introduction

Social media are very supportive and act as an interface to spread information, facilitate communication as well as news. News defined as new information or information about something that is happening, presented in print, broadcast, internet, or word of mouth to third people or many people. News must account for its truth, actual, and informative; news must be current and interesting [1], which can significantly affect many people [2][3]. Some news, especially on social media, uses false news as a political weapon [4],[5]. Social media for news consumption is a double-edged sword. On the, there is no need to pay a lot of money to get information, because access to information is easily available for now, and very quickly with seconds information can be disseminated, this makes people search for and consume news from social media. Even though fake news and real news are straightforward to distinguish [6], Hoaxes are impossible to predict, and when they distributed, it's hard to stop. Headlines from hoax news often use fewer words and nouns. A hoax is a news article that is intentional and untrue and can mislead readers. The public still cannot distinguish between true and false news, this happens as a result of low public literacy of messages on social media, [5] Hoax is defined as a series of information that is intentionally misled, but sold as truth. Other words closely related to Hoax are Fake News, Allcott [6] "Hoax News Defined as a deliberate and verifiable news article that can mislead readers", and other definitions from some of the latest research. [7],[8],[9],[10],[11], there are many more studies related to hoax news detection.

In general, hoaxes, identification can be identified with four criteria. First, hoax information usually has the characteristics of chain letters by including sentences like "Spread this to everyone you know; otherwise, something unpleasant will happen." Second, hoax information usually does not include the date of the event or does not have a real-time or can be verified, for example, "yesterday" or "issued by..." statements that do not show clarity. Thirdly, hoax information usually does not have an expiry date on the information alert, even though the actual presence of that date will also prove nothing, but can have a prolonged disquieting effect. Fourth, no identifiable organization is cited as a source of information or includes the organization but is usually not linked to data.

One of the characteristics of hoax news consists of disinformation and Misinformation; these two words are different but interrelated, According to Bernd [12], Misinformation and disinformation are strictly related to information, Misinformation is inaccurate information and disinformation is false information that spreads on social media networks [13]. Data from Kemominfo there are around 800,000 sites in Indonesia that have indicated as hoax news spreaders, social media has been by individuals for personal and group benefits by spreading harmful content that causes unrest and mutual suspicion in the community [14]. With the difference between the number of facts and hoaxes, "many current news verification mechanisms already exist, and a large volume of information, then a new automatic hoax news detection approach is needed" [15].

The last stage is validation testing the detection results by matching the similarity of hoax news text, title, or content hoax news. An explanation of the related hoax detection framework will explain in the next section.

II. Related Work

The concept of news hoax is often associated with, fake news, lies, disinformation, misinformation, deception, misleading, rumors, fraud, some related work can be found at, for disinformation [16][17], for misinformation [12], for fake news [18][19],for deception[20] ,for lies [5], for rumors [21], for fraud [22][23], for misleading [24], etc. Problems related to this topic have been seen concerning classification, detection, filtering. Likewise, most published works have hoax news detection as hoax classification problems or not hoaxes. The following is some research work related to news hoaxes.

Arjun [25] proposes an automatic hoax framework technique that classifies hoax news into several classes, true, mostly true, half true, almost untrue, false, and shorts with models based on Convolutional Neural Networks (CNN) and Bi-directional Long Short Term Memory (Bi-LSTM). Representations obtained from these two models included in the Multi-layer Perceptron Model (MLP) the final classification.

Ishak et al. [26] proposed a framework of a text-based deception detection system using the Levenshtein Distance algorithm. Then identifying the potential deception of the email content by comparing it with a database of deception, the required component consists of three main components: preprocessing of text, detection of deception, and detection of a new deception. For the Pre-processing text stage, collect emails that will test for validity as genuine or fraudulent emails.

Shu et al. [9] propose the Tri-Relationship hoax detection framework, called TriFN, with data objects taken from social media. This technique explores the correlation of publisher bias, news establishments, and relevant user involvement simultaneously, and proposes Tri-Relationship. Shu et al. provide two comprehensive real-world fake news datasets to facilitate hoax news research.

Tacchini et al.[27], have an automated online As a contribution that is by showing that Facebook posts can be classified with high accuracy as a hoax or not a trick based on users who "like" them. System presents two classification techniques, one based on logistic regression, the other based on new adaptations crowdsourcing Boolean algorithm. The dataset consists of 15,500. Facebook posts and 909,236 users, the research results obtained classification accuracy exceeding 99% even when the training set contained less than 1% of the job. Shows the power of the technique for which they purposed the system worked even to users who like hoaxes and Fact posts these results indicate that the diffusion pattern mapping information can be a useful component for automatic hoax detection system.

Veronica et al. [28] make a framework for the automatic identification of fake content in online news. And has two contributions. The First contribution introduces two new data sets for the detection of hoax news, covering seven different news centers. It further explains the collection, announcement, and validation processes in detail and presents some exploratory analysis about identifying linguistic differences in fake and legitimate news content. Another controversy, conducting a series of learning experiments to build an accurate phony news detector provides a comparative analysis of automatic and manual identification of fake news. Support Vector Machine (SVM) classifier and five-fold cross-validation, with accuracy, precision, recall, and measures averaged over the five iterations, using classification techniques for hoax news detection.

Orissa Rasywir [29] make a framework for the conduct hoax news classification in Indonesian using a statistical approach based on machine language, with application based on text categorization, where the proposed system consists of preprocessing, feature extraction, feature selection and execution of classification models. The machine learning algorithm technique chosen in the hoax news classification system is Naïve Bayes (NV), Support Vector Machine (SVM). The results of the conclusions Research results in The best experimental results achieved with naïve bays algorithms with unigram features where feature selection uses union operations between information gain and mutual information.

Sethi [30] proposes a prototype framework or social argumentation to verify the validity of alternative facts intended to help curb the spread of false news, a prototype system used social argumentation to verify the validity of proposed alternative facts and help with detection of hoax news, uses the principles of fundamental argumentation in a graph-theoretic framework which also combines the semantic and web.

III. Methodology

A. Hoax News Methodology

Hoax detection methodology used in this study into three stages: stage one is PreProcessing, stage two is a process, and phase three is The Post Process. Fig. 1 below is a chart of the hoax news detection methodology research stages.

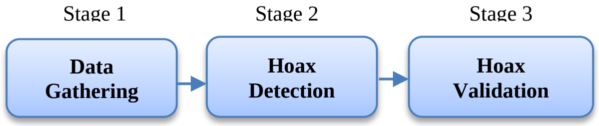

Fig. 1. Hoax News Methodology

Hoax news detection methodology consists of three stages, one stage named Pre-Processing, the preprocess stage, namely the stage of data gathering, by crawling from the web, and then a web register containing hoax news data, after the data is collected, by crawling than the data will then be stored in a database, and will classify as data hoaxes and not hoaxes.

The methodology here is the Process stage at this stage will be the second stage known as hoax news detection, hoax news detection is done by taking extracted data and searching data by extracting features to determine many words, and similarity in words, which will then proceed to the next method, feature selection, and labeling the data, which is divided into four news categories, Hoax, Fact, Information, Unknown.

The third phase is the Post Process phase; this phase validation the similarity of text extraction from the detection of hoax text data security using the similarity algorithm and produces the results of the validation done in the form of a percentage.

B. *Hoax News Dataset Collection Method*

The Hoax News Dataset collection is done in several stages, namely, Registering website & Social Media, Automatic Parse, Web Crawlers, Pre-processing, Labeling Categorizing Classification, Validity. For phase one, which collects data, data collection done by asking for the URL address, next is the Automatic Parse process, and finally crawling the deceptive data, which will then move into the dataset. Fig. 2 is for Stage one:

1) Registering website: The initial stages of collecting hoax data are by taking hoax data from several websites that contain hoax news that has gone through the analysis stage of checking hoax news facts that have verified the web used is taken from the web for English-language hoaxes including: www.hoaxslayer.net, www.sophos.com, www.truthorfiction.com, www.symantec.com, and for Indonesian-language hoaxes produced from several government websites and online media websites that have been trusted including stophoax.id, turnbackhoax.id, indeks.kompas.com, and cekfakta.com.

2) Automatic parse: Automatic parsing is a way to break up a series of hoax words that will produce a description tree that will be used in the next compilation stage, namely, syntactic analysis [31],[32] related research about parse[33]. Automatic Parse consists of three stages and below is the Fig. 3 of the automatic parsing stage.

a) Planning and Defining Object: Planning and Defining Object is achieved by writing code cleaner for the web crawlers through scrapping and storing various data from various news article websites or blog posts from multiple sites, each with a different template and layout. Contains the title of the article, another contains the website's title, and the title of the article in <span id = "title">, then collects a number of "types" of data, news reviews, news articles, author from various website, and which stores this data type as an object that can be read and written to a database.

SELECT WEBSITE

(insert into table)

DEFINE OBJECTS

1) Title
2) Text
3) Category
4) Author
5) Date Published
6) Analysis

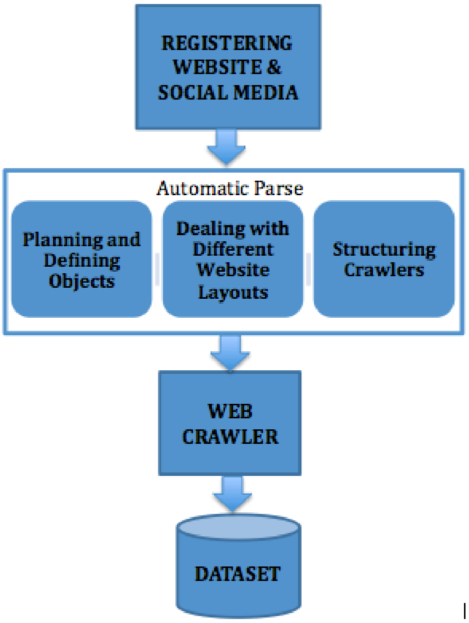
REGISTERING WEBSITE & SOCIAL MEDIA
Automatic Parse
Planning and Defining Objects
Dealing with Different Website Layouts
Structuring Crawlers
WEB CRAWLER
DATASET

Fig. 2. Hoax News Dataset Collection Method.

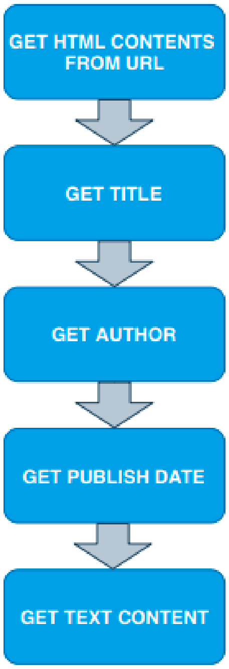

Fig. 3. Hoax News Automatic Parsing Method.

b) Dealing with Different Website Layout: Dealing with Different Website Layouts is extracting relevant and useful data from various websites, without having upfront knowledge about the structure of the site itself. Identify the title and main content of a page and retrieve the URL, string, or object. Website classes store h1 string tags that indicate where titles can found. Necessary template extensions can handle missing fields, collect various types of data, crawl only through certain parts of the website, and store more complicated information about the page.

https://www.hoax-slayer.net/hiv-food-and-drink-contamination-hoaxes-continue/

Inspect Html Element

Defining HTML Title position

```
<h1 class="post-title single-post-title">

HOAX -‘Breast Larvae Infestation From Undergarments’

</h1>
```

c) Structuring Crawlers: Structuring Crawlers is to create a type of website layout that is flexible and can be modified, incorporating this method into a well-structured and expandable website crawler that can collect links and find data automatically.

Get Html Page Source

file_get_contents (php function)

Get Title Html Positions

```
@$first_title = explode(‘<h1 class="post-title single-post-title">’, $response );

@$second_title = explode(’ </h1>‘, $first_title[1]);

@$title = $second_title[0];
```

7) Web crawlers: Web Crawler is a program that parses the hypertext structure of the web [34], starting with an initial address called a seed and secretly visiting the web address on a web page [35]. Related research about crawler [36],[37],[38], Hoax data collected from several websites that contain hoaxes, following the stages in collecting hoax data using crawler techniques:

- Stage one makes a website name or URL that contains hoax news that will be crawled. Starting with the copy of the start page URL, and the total page to the last page, then suffix or case and system status finish or begin.

- Stage two is to check the language to use, Indonesian or English.

- Stage three does Description; the Description contains a summary of the web.

- Stage four, namely the Page URL is a series of characters according to a specific standard format, for example, characters / or =

- Step five stemming is the process of changing the affected words into essential words, and this is needed if the hoax news is in Indonesian, we give a sign then we give the indication "YES" if the news hoax is in English then no stemming is needed "NO."

- Stage Six Header Page: Header Page explains about pages to help search engines find content.

- Element HTML Tag: Category, Title, Author, Date Publisher, Text, Web Text, Analysis, Rating

C. *Pre-processing*

The second stage is the Pre-processing Stage. In-Text mining applications, textual data representation dramatically influences the accuracy of the results [39]. The preprocessing phase consists of the process of data cleaning and feature selection. In this study, preprocessing was carried out in the following Fig. 4.

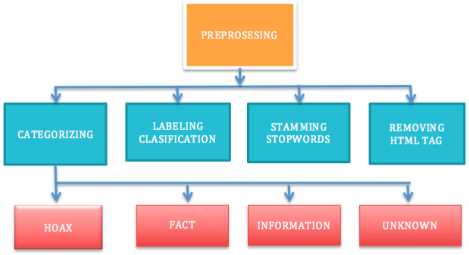

Fig. 4. Preprocessing Phase.

1) Removing html tag: Removing HTML Tags is designed to strip HTML tags from the text. It will also strip embedded JavaScript code, style information (style sheets), as well as a system inside PHP tags (<php > <php > < >), one time the command, will replace the sequence of new (multiple) lines and the next will be a list of allowed tags.

2) Case folding: Case Folding is the stage of changing all character letters to lowercase. Next, do the character removal by removing all the characters that considered not needed, namely all aspects other than letters (a-z).

3) Tokenizing: Tokenizing is the process of getting words from documents, which are the embryo of attributes. The next process. Synonym Normalization is a stage of normalizing words that have the same meaning to reduce the dimensions of the data and get quality attributes. Hardalov defines stop words as the most common and functional words in a language [40].

4) Stop words: Stopword Removal removes words that do not affect class classifications, such as conjunctions, clothing words, pronouns, etc. The process by filtering each word with the database that has provided. [github.com/masdevid/ID-Stopwords]

5) Stemming: Stemming is a process that finds the essential word of a word. By eliminating all affixes (affixes) useful consisting of prefixes, insertions (infixes), suffixes (suffixes), and confides (the combination of prefix and suffix) in words derivative. Stemming is used to replace the form of a word that becomes the basic word of the word that fits the structure of good Indonesian morphology right.

D. Labeling Categorizing Classification

Data labeling is one of the processes for labeling hoax data categories, some related research [41], [42]. Hoax news labeling is a stage by labeling hoax news data, labeling is done, on datasets that have not to be labeled, or unknown data. An anonymous data process will carry out several stages, including labeling news information, then determine features selection and selected features using unsupervised algorithms, and produce news classification in the form of labels into categories, namely, Fact, Hoax, Information. The following Fig. 5 shows the process of data labeling stages.

E. Multilanguage Detection

Multilanguage identification: Language identification is the task of automatically detecting the language in hoax news data — some of the same research for multilanguage detection [43]. Vu, uses multilanguage by introducing statistical rule-based methods to create rules for detecting in various languages [44]. Bouarara [45], uses multilanguage to detect email with learning-based methods automatically and ranks the steps to generate spam emails [47]. Contributing to our research by using the technique of collecting hoax news from the web in several different countries or languages, and produce site data effects are news hoax news from several languages including, Indonesian, English and Malaysian. The following steps are carried out for automatic language detection, randomly select documents for each List.

F. Validity

Validation is to detect hoax news using a similar method. Most existing approaches consider the hoax news problem as a classification problem that predicts whether a news article is a hoax or not. [9]. Problems encountered in text mining are large amounts of data, high dimensions, data and structures, constantly changing data, and data noise [46]. Overcoming unstructured data, we need to evaluate the accuracy of words. A similarity algorithm is used in this evaluation to detect hoax news [47]. The similarity algorithm calculates the similarity between two strings; the complexity of this algorithm is O, where N is the length of the longest string. The parameter used is. The first string. The second string. The third is to calculate the similarity in percent's. The similar text will calculate the similarity in percent, by dividing the results of similar text by the average length of the given string times 100. Finding the longest general substring first, and then doing this for the prefix and suffix, recursively calculate the number of matching characters. The length of all common substrings found is added. Example:

Example #1 similar_text() argument swapping example

```
<?php

$sim = similar_text('bafoobar', 'barfoo', $perc);

echo "similarity: $sim ($perc %)\n";

$sim = similar_text('barfoo', 'bafoobar', $perc);

echo "similarity: $sim ($perc %)\n";
```

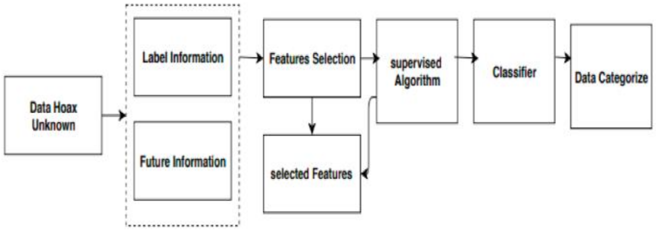

Fig. 5. Hoax News Labeling Categorizing.

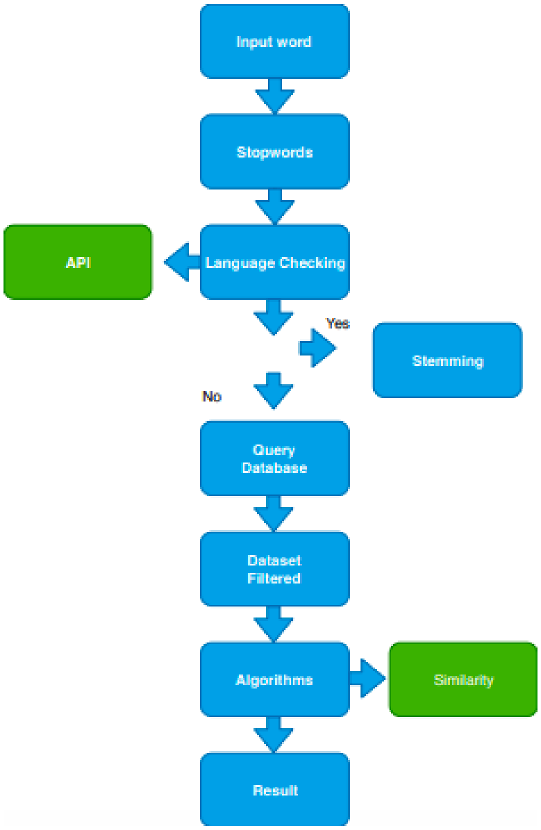

Fig. 6. Hoax New Validity Method.

This example shows that swapping the first and second argument may yield different results as shown in Fig. 6.

Stages of Validation of hoax news similarity detection:

- Word Input: Input by comparing some of the words contained in the hoax news dataset provided in the hoax detection engine. Document reading: reading text documents

- Stop word: Optimization by removing all words that classified as stop words. Stop words are common words that usually appear in large numbers and are considered to have no meaning. Stop words are generally used in information retrieval, stop words for English include, among others, the, while for Indonesian, among others are, at, too.

- Language checking: Check the language using Multilanguage using an application that is already available in the API.

- Stemming: return various kinds of word formations to the representation of essential words, stemming method requires input in the form of words contained in a document, by producing output in the form of crucial words. Basic word search for words that influence Indonesian.
- Database Query: Search for words contained in the hoax news database.
- Algorithm: The selection is carried out one by one testing of four algorithms namely similarity algorithm, Levenshtein algorithm, Smith-Waterman algorithm, Damerau Levenshtein algorithm.
- Result: Shown Results of the similarity of words in percent and the amount of time needed to process in seconds.

G. *Hoax News Detection and Analyzer Framework*

Generally, the hoax news detection framework represented as follows in Fig. 7:

Detection Hoax framework has three stages where the first stage is called the stage Preprocess, and this preprocess stage we named the stage of gathering hoax news from several web sites, the scene consists of a web Registering by entering a web URL. The Automatic Parse step is taking the clean text from unnecessary HTML marks. There are three stages for automatic parsing i.e., Planning and defining objects, then dealing with different website layouts and structure crawling. And the third stage in gathering data is web crawlers.

The second stage is "Process"; at this stage, the process is called the Process Data Hoax Detection Generator. This process has three stages, namely Preprocessing, Labeling, and Categorization; the Pre-processing stage consists of Removing HTML Tags, then stemming, then stop word stages. The output of this process will produce four categories of hoax news, namely Hoax, Fact, Information, and Unknown.

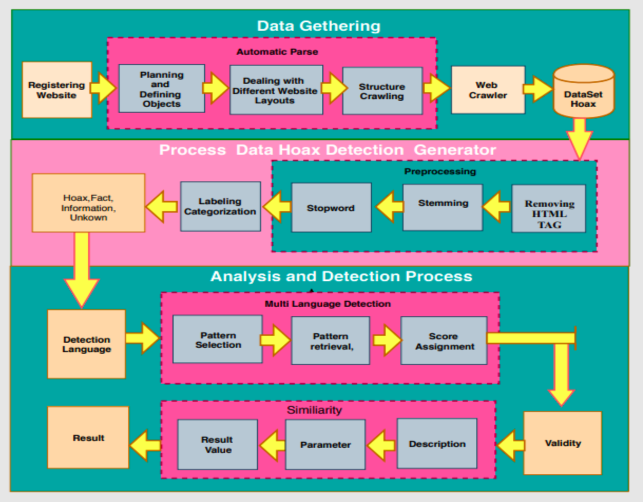

Fig. 7. Hoax News Detection Framework.

The third stage is the Post-processing process; this stage is named the Analysis and Detection Process. This stage produces two methods, namely Multi Detection Language and Validity and Result. The Multilanguage detection process consists of three processes, namely pattern selection, pattern retrieval, and score assignment. The Validation process is the Similarity text technique, which has three methods, namely Description, finding Parameters, and producing Result Value similarity of data hoax words in percent.

IV. Result

Result of research of the hoax news detection: The contribution of the hoax news detection framework. Stage of gathering the hoax news dataset. Collecting data is carried out by the crawling method, which results in twelve contributions thousand hoax or corpus data sets. The second contribution is hoax news detection through the preprocessing stage. Analysis Using the text of news headlines or news content. Producing a 12000 accurate and reliable dataset can set more versions of several different parts of the world can then strength by multilingualism. Detection of results in terms of accuracy and processing time will be well validated. Additional research can be done. Hoax data labeling with classification techniques using unsupervised algorithms results in the participation of four categories of hoax news, namely, hoax, fact, information, and unknown.

The third contribution is multilanguage detection; at this stage, the API; this is a language detection web service. It accepts the text and returns the results with the detected language code and score. Validation detection of hoax, this validation produces hoax news accuracy using the similarity text algorithm method, and results in a percentage of the accuracy of text similarity, and calculates the length of time it takes to detect hoax news in seconds. The strengths of developing a proposed hoax news detection framework are hoax datasets, hoax data labels, multilanguage, and validation of detection results.

V. Conclusion

Social media is one of the media that makes it very easy for us to get information, find references in the scientific field; it is also straightforward for us to do business and get news. Still, of the many good things produced by social media, there are also things bad stuff, i.e., the amount of false or fake news information known as a hoax. Conclusions of the research on creating a text-based hoax detection framework. This framework is needed to reduce the dissemination of hoax news in the community. This study's findings may well detect hoax news. The analysis is done using the news title text or news content. Producing a 12000 accurate and reliable dataset can set more versions of several different parts of the world can then strengthened by multilingualism. Detection of results in terms of accuracy and processing time will be well validated. Additional research can be done

ACKNOWLEDGEMENTS

This work been supported under Universiti Teknikal Malaysia Melaka research grant Gluar/CSM/2016/FTMK-CACT/100013. The Author would like to thanks University Teknikal Malaysia Melaka, CyberSecurity Malaysia and all member of INFORSNET research group for their incredible support in this project.

REFERENCES

[1] J. Fiske, Understanding Popular Culture. Paperback, 1989.

[2] Fang, "Writing Style Differences in Newspaper , Radio , and Television News Writing Style Differences in Newspaper , Radio , and Television News Irving Fang," no. 2, 1991.

[3] S. Y. Yuliani, S. Sahib, M. F. Abdollah, M. N. Al-mhiqani, and A. R. Atmadja, "Review Study of Hoax Email Characteristic," vol. 7, pp. 778–782, 2018.

[4] M. N. Al-Mhiqani, R. Ahmad, W. Yassin, A. Hassan, Z. Z. Abidin, N. S. Ali, and K. H. Abdulkareem, "Cyber-Security Incidents: A Review Cases in Cyber-Physical Systems," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 1, pp. 499–508, 2018.

[5] H. Berghel, "Lies, Damn lies, and fake news," Computer (Long. Beach. Calif)., vol. 50, no. 2, pp. 80–85, 2017.

[6] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," J. Econ. Perspect., vol. 31, no. 2, pp. 211–236, 2017.

[7] E. Mustafaraj, "The Fake News Spreading Plague : Was it Preventable ? The Fake News Spreading Plague : Was it Preventable ?," 2017.

[8] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic Deception Detection: Methods for Finding Fake News."

[9] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media : A Data Mining Perspective," no. i, 2016.

[10] M. Potthast, "A Stylometric Inquiry into Hyperpartisan and Fake News," pp. 231–240, 2018.

[11] V. L. Rubin, N. J. Conroy, Y. Chen, and S. Cornwell, "Fake News or Truth ? Using Satirical Cues to Detect Potentially Misleading News." pp. 7–17, 2016.

[12] B. C. Stahl, "On the Difference or Equality of Information ,Misinformation , and Disinformation : A Critical Research Perspective," vol. 9, 2006.

[13] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media : A Data Mining Perspective," vol. 19, no. 1, pp. 22–36, 2016.

[14] B. Fitnah, J. Ajak, K. Milenial, B. Hoax, M. Pembangunan, H. Bentuk, L. Teror, P. Masyarakat, T. B. Informasi, J. M. Terprovokasi, K. Hitam, P. B. Politik, M. C. Menjadi, and A. M. Hoax, "Heboh HOAX Nasional."

[15] Y. Chen and V. L. Rubin, "Towards News Verification : Deception Detection Methods for News Discourse Towards News Verification : Deception Detection Methods for News Discourse," vol. 2015, 2015.

[16] S. Kumar, R. West, and J. Leskovec, "Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes," Www, pp. 591–602, 2016.

[17] N. A. Karlova and K. E. Fisher, "' Plz RT ' : A Social Diffusion Model of Misinformation and Disinformation for Understanding Human Information Behaviour," vol. 2012, pp. 1–17.

[18] Ö. Özgöbek and J. A. Gulla, "Towards an Understanding of Fake News," 2017.

[19] J. Thorne, M. Chen, G. Myrianthous, J. Pu, X. Wang, and A. Vlachos, "Fake News Detection using Stacked Ensemble of Classifiers," pp. 80–83, 2017.

[20] V. L. Rubin, Y. Chen, and N. J. Conroy, "Deception Detection for News: Three Types of Fake News," Proc. Assoc. Inf. Sci. Technol., vol. 52, no. 1, pp. 1–4, 2015.

[21] S. Zannettou, M. Sirivianos, J. Blackburn, and N. Kourtellis, "The Web of False Information : Rumors , Fake News , Hoaxes , Clickbait , and Various Other Shenanigans," pp. 1–26.

[22] S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, "Automatic detection of online recruitment frauds: Characteristics, methods, and a public datase," Futur. Internet, vol. 9, no. 1, pp. 1–19, 2017.

[23] L. Akoglu and C. Faloutsos, "Opinion Fraud Detection in Online Reviews by Network Effects," pp. 2–11, 2012.

[24] Y. Chen, N. J. Conroy, and V. L. Rubin, "Misleading Online Content : Recognizing Clickbait as ' False News ,'" pp. 15–19, 2014.

[25] A. Roy, K. Basak, A. Ekbal, and P. Bhattacharyya, "A Deep Ensemble Framework for Fake News Detection and Classification," 2017.

[26] A. Ishak, Y. Y. Chen, and S. Yong, "Distance-based Hoax Detection System," pp. 215–220, 2012.

[27] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some Like it Hoax: Automated Fake News Detection in Social Networks," 2017.

[28] B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic Detection of Fake News," no. August, 2017.

[29] E. Rasywir, A. Purwarianti, and K. Kunci, "Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin," J. Cybermatika, vol. 3, no. 2, 2015.

[30] R. J. Sethi, "Spotting Fake News : A Social Argumentation Framework for Scrutinizing Alternative Facts."

[31] I. Retrieval, Introduction to Information Retrieval. 2008.

[32] D. O. One and P. Description, "Web Crawling Models."

[33] J. Nilsson and J. Hall, "Parsing Formal Languages using Natural Language Parsing Techniques," no. October, pp. 49–60, 2009.

[34] C. Castillo and M. Marin, "Effective Web Crawling," no. November 2004.

[35] B. J. Holmes, Crawling the Web with Java 6. 2005.

[36] A. Heydon and M. Najork, "Mercator : A scalable , extensible Web crawler," vol. 2, pp. 219–229, 1999.

[37] B. Pinkerton, "WebCrawler : Finding What People Want," 2000.

[38] C. Olston, "Web Crawling."

[39] F. A. Ozbay and B. Alatas, "A Novel Approach for Detection of Fake News on Social Media Using Metaheuristic Optimization Algorithms," pp. 62–67, 2019.

[40] M. Hardalov, I. Koychev, and P. Nakov, "In Search of Credible News," no. November 2017, 2016.

[41] "Ranking, Labeling, and Summarizing Short Text in Social Media," no. May, 2013.

[42] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," Proc. 21st Int. Conf. World Wide Web - WWW '12, p. 191, 2012.

[43] M. Lui, J. H. Lau, and T. Baldwin, "Automatic Detection and Language Identification of Multilingual Documents," vol. 2, pp. 27–40, 2014.

[44] M. T. Vu, Q. A. Tran, F. Jiang, and V. Q. Tran, "Multilingual Rules for Spam Detection," vol. 1, no. July 2009, pp. 107–122, 2014.

[45] H. A. Bouarara, T. Moulay, R. M. Hamou, A. Amine, and T. Moulay, "A Novel Bio-Inspired Approach for Multilingual Spam Filtering," vol. 11, no. September, 2015.

[46] A. Huang, "Similarity Measures for Text Document Clustering," no. April, 2008.

[47] M. Vuković, K. Pripužić, and H. Belani, "An intelligent automatic hoax detection system," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5711 LNAI, no. PART 1, pp. 318–325, 2009.