

Phát hiện tin giả dựa trên nội dung và ngữ cảnh xã hội sử dụng học máy

Lê Thị Xinh^{1,*}, Lê Quang Hùng², Nguyễn Thị Ngọc Bích³,
Nguyễn Thị Kim Phụng², Phạm Trần Thiện²

¹Khoa Sư phạm, Trường Đại học Quy Nhơn, Việt Nam

²Khoa Công nghệ thông tin, Trường Đại học Quy Nhơn, Việt Nam

³Công ty Fujinet, Bình Định, Việt Nam

Ngày nhận bài: 07/04/2023; Ngày sửa bài: 18/06/2023;

Ngày nhận đăng: 06/09/2023; Ngày xuất bản: 28/10/2023

TÓM TẮT

Bài báo này trình bày nghiên cứu về phát hiện tin giả dựa trên nội dung tin và ngữ cảnh xã hội sử dụng học máy. Đầu tiên, chúng tôi phân tích các khái niệm liên quan, các phương pháp phát hiện tin giả. Tiếp theo, chúng tôi mô hình hóa nhiệm vụ này như một bài toán phân lớp nhị phân, biểu diễn nội dung tin và ngữ cảnh xã hội dưới dạng véc-tơ đặc trưng. Sau đó, chúng tôi sử dụng một số thuật toán học máy để xây dựng mô hình phân lớp. Kết quả thực nghiệm với ba thuật toán học máy: Support Vector Machine, Naive Bayes và k -Nearest Neighbors trên bộ dữ liệu FakeNewsNet cho thấy hiệu quả của phương pháp đề xuất.

Từ khóa: Phát hiện tin giả, nội dung tin, ngữ cảnh xã hội, phân lớp, học máy.

*Tác giả liên hệ chính.

Email: lethixinh@qnu.edu.vn

Fake news detection based on news content and social contexts using machine learning

Le Thi Xinh^{1,*}, Le Quang Hung², Nguyen Thi Ngoc Bich³,
Nguyen Thi Kim Phuong², Pham Tran Thien²

¹Faculty of Education, Quy Nhon University, Vietnam

²Faculty of Information Technology, Quy Nhon University, Vietnam

³Fujinet Company, Binh Dinh, Vietnam

Received: 07/04/2023; Revised: 18/06/2023;

Accepted: 06/09/2023; Published: 28/10/2023

ABSTRACT

This paper presents research on detecting fake news based on news content and social context approach using machine learning. First of all, we analyze related concepts, methods of detecting fake news. Next, we model this task as a binary classification problem, representing news content and social context as feature vectors. Then we use machine learning algorithms to build the classification model. Experimental results with three machine learning algorithms: Support Vector Machine, Naive Bayes and k -Nearest Neighbors on the FakeNewsNet dataset show the effectiveness of the proposed method.

Keywords: *Fake news detection, news content, social context, classify, machine learning.*

1. INTRODUCTION

The development of online social media platforms (e.g. Facebook, Twitter, Instagram, etc.) brought about a significant increase in the accessibility of information on the one hand, and accelerated the propagation of fake news on the other hand. As a result, the influence of fake news is growing and threatening the safety of the community.¹ The scope of fake news was most marked during the 2016 US presidential election campaign. The top 20 election fake news received 8,711,000 shares and comments on Facebook, larger than the total of 7,367,000 shares and comments on top 20 election stories from 19 major media outlets.²

Distinguishing true news from fake news is one of the difficult tasks for humans. Psychosocial and media studies show that people's ability to detect deception ranges from 55%–58%.³

There have been several expert-based manual fake news detection tools, platforms and websites (e.g. PolitiFact, Snopes) and community-based (e.g. Fiskkit, VAFC) so far. However, manual fake news detection is not suitable for the large amount of information generated, especially on social media.⁴ Therefore, the research direction fake news detection [automatic] has become a "hot" topic recently.⁵⁻⁹ In which, fake news detection can be classified

*Corresponding author.

Email: lethixinh@qnu.edu.vn

into two approaches namely (i) content-based and (ii) propagation-based.¹⁰⁻¹³

Content-based fake news can be detected by analyzing the news content. Meanwhile, propagation-based fake news detection exploits how news spread on social networks. The "life-cycle" of fake news has three basic stages: (1) content creation, (2) publication,

and (3) propagation as illustrated in Figure 1. Propagation-based approach using social context information is difficult to apply in predicting fake news before the third stage (before fake news is spread on social media). Therefore, it is necessary to detect fake news early to prevent its spread (i.e., when fake news is at the publication stage and it has not yet spread widely).

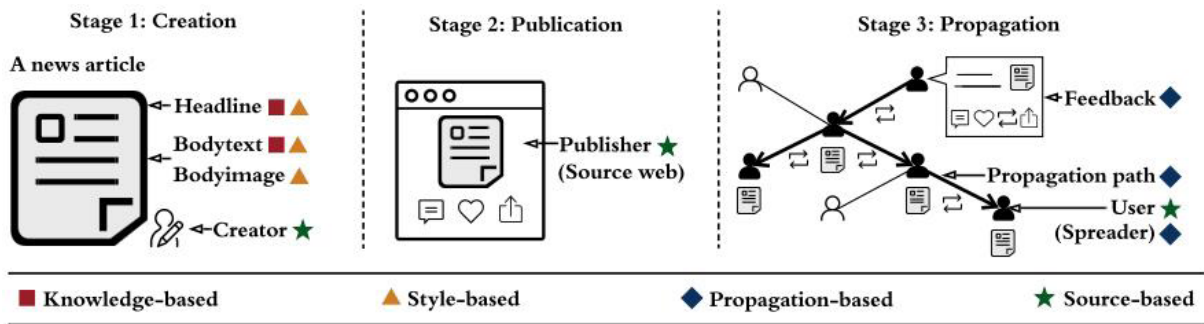


Figure 1. Fake news life cycle and detection methods.²

In this paper, we present a research on detecting fake news according to content-based and social context approach using machine learning. The main contributions of the paper are:

- 1) Analysis of related concepts, methods of detecting fake news.
- 2) Proposal of a method to detect fake news based on news content and social context using machine learning, including: (i) modeling this task as a binary classification problem; (ii) representing content at the lexical and social context level in the form of feature vectors; and (iii) using machine learning algorithms to build classification models.
- 3) Experiment to evaluate the effectiveness of classification models on the FakeNewsNet dataset.

The rest of the paper is organized as follows: Section 2 presents an overview of fake news and fake news detection. Section 3 then presents news content and social context-based fake news detection method using machine learning followed by Section 4 describing experiments. Section 5 wraps up the article with the conclusion.

2. OVERVIEW OF FAKE NEWS AND FAKE NEWS DETECTION

2.1. Definition of fake news

Current studies about fake news detection often involves the following concepts: fake news, false news, satire news, disinformation, misinformation, and rumor. These concepts can be distinguished based on three characteristics: (i) authenticity, (ii) intention, and (iii) whether information is news. Table 1 summarizes related concepts based on these characteristics.¹ For example, disinformation has false authenticity [news or not news] with bad intentions.²

Table 1. Compare related concepts.

Concepts	Authenticity	Intention	News or not news?
Fake news	False	Bad	News
False news	False	-	News
Satire news	-	Not bad	News
Disinformation	False	Bad	-
Misinformation	False	-	-
Rumor	-	-	-

According to Zhou,² “fake news is intentionally false news published by a news outlet”. Typically, news agencies publish news in the form of articles with the following components: title, content, author (including user’s feedback) as illustrated in Figure 1.

2.2. Fake news detection methods

2.2.1. Content-based

Content-based approaches include (i) knowledge-based and (ii) style-based/writing-style. Knowledge-based fake news detection evaluates the veracity of news by comparing knowledge drawn from verified news content with known facts (i.e. true knowledge). Similar to knowledge-based methods, style-based fake news detection also focuses on news content analysis. This process includes two steps called style representation (using language features) and style classification (using machine learning models). While the knowledge-based method mainly evaluates the authenticity of the news, the style-based method can assess the intention of the news.^{2,14-15}

2.2.2. Propagation-based

Propagation-based approach uses social context information to detect fake news, for example, how fake news spreads on social networks, who spreads it, and how spreaders connect with each other.

The news ecosystem on social media provides social contextual information regarding three basic entities: publishers, news [pieces], and users.

Figure 2 and Figure 3 illustrate the spread of news. In Figure 3, p_1 , p_2 and p_3 are the publisher of the news a_1 , ..., a_4 and u_1 , ..., u_6 are the users sharing these news. In addition, users tend to form social links with people with similar interests.¹⁶⁻¹⁷

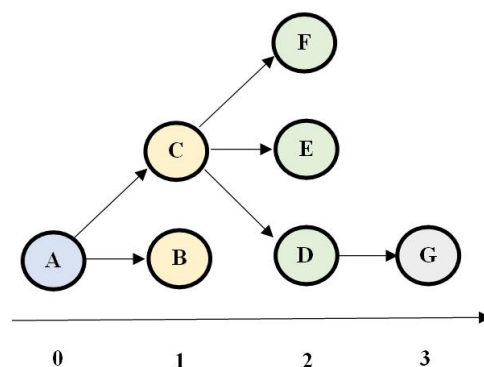


Figure 2. Tree structure-based news propagation.

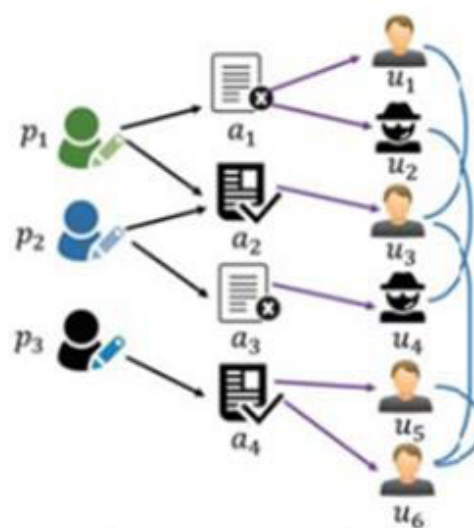


Figure 3. The relationship between publisher and user.

In general, the input to a propagation-based method can be (i) news cascades or (ii) self-defined propagation graphs.

A news layer is a tree structure that represents the direct spread of news on social network (example in Figure 2).² The root node corresponds to the user who first shared the news (i.e. the initiator); other nodes in the layer correspond to users who then forwarded the post after it was posted by the parent node. In a news propagation network (subgraph of a social network), each network corresponds to news, each node in the network represents a user, and an edge between two nodes represents the relationship of the two users. For example, in Figure 3, edge $(p \rightarrow a)$ represents publisher p publishing news a , edge $(a \rightarrow u)$ represents new a spread by user u and edge $(u_1 \rightarrow u_2)$ is social relationship between u_1 and u_2 .¹⁶

Thus, propagation-based fake news detection focuses on categorizing (i) news layers or (ii) self-defined graphs.

3. PROPOSED METHODS

In this section, we present a news content and social context-based fake news detection method using machine learning. First, we model this task as a binary classification problem. Next, we represent the news content at the lexical level according to the BoW model (Bag of Words) as a feature vector and connect to the social context vector. Then we use machine learning algorithms to build the classification model.

3.1. Problem definition

We consider the problem of detecting fake news based on content (part of text), where $A = \{a_1, a_2, \dots, a_n\}$ is the set of n news. Suppose the news to be verified a can be represented as a feature vector $v \in R^k$. The task of verifying the content-based is defined a function f , such that:

$$f: v \xrightarrow{D} y \quad (1)$$

Where $y \in \{0,1\}$ (0 – true news, 1- fake news) is the predicted label of the news and $D = \{(v_i, y_i) | v_i \in R^k, y_i \in \{0,1\}, i = 1..n\}$ is the training dataset. The training dataset D consists of n news, each news $a_i \in D$ is represented by the feature vector v_i with the label y_i .

And news content and social context based fake news detection problem is defined as follows: Let $N = \{n_1, n_2, \dots, n_{|N|}\}$ is a set of news, each of which is labeled as $y_i \in \{0,1\}$, $y = 1$ is the fake news and $y_i = 0$ is the true news. The news n_i is represented by the news content (news body) and side information such as (title, source, author, ...). When n_i is posted on a social network, it is usually interacted with by social network users $U = \{u_1, u_2, \dots, u_{|U|}\}$. Social context includes user interactions such as comments, posts, likes/shares, etc.

$$SC(n_i) = ((u_1, sc_1, t_1), (u_2, sc_2, t_2), \dots, (u_{sc}, sc_{sc}, t_{sc})) \quad (2)$$

Each tuple (u, sc, t) refer to user u 's context sc for news n_i in timet. Here, a user can interact with a post multiple times.

Task of this problem is to find a model M to predict the label $y(n_i) \in \{0,1\}$ for each news based on the news content and social context. Therefore, this task is defined by Equation (3):

$$y(n_i) = M(C(n_i), SC(n_i)) \quad (3)$$

Where $C(n_i)$ is news content and $SC(n_i)$ is the social context of the news. Figure 4 describes the problem of detecting fake news through news content and social context.

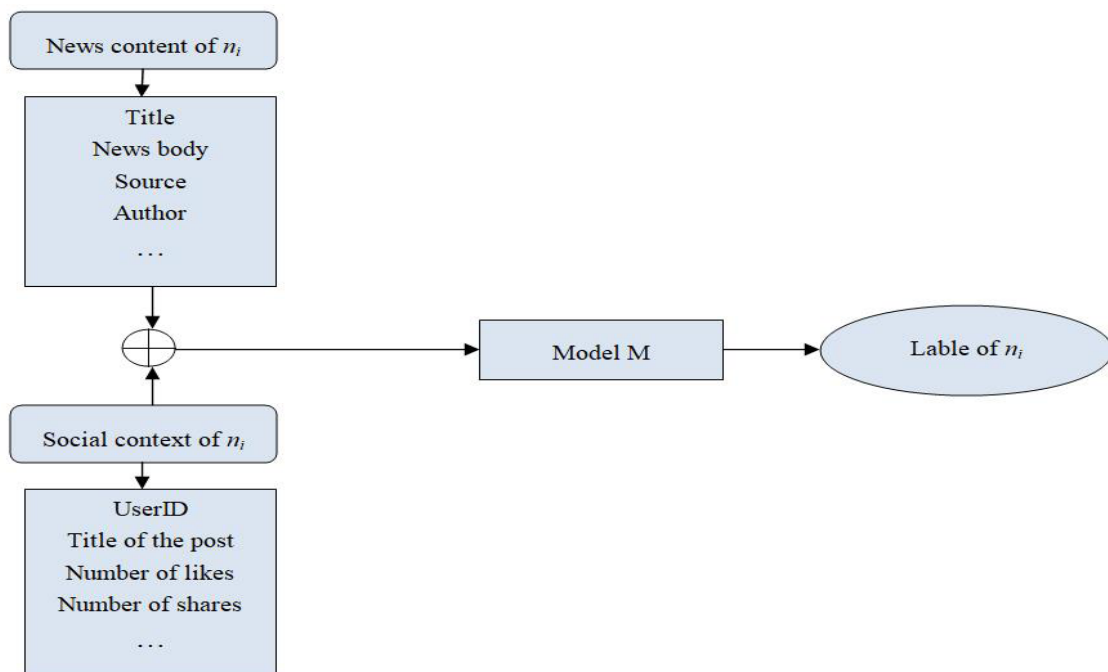


Figure 4. Description of problem.

3.2. News content and social context

3.2.1. News content

News content is the main component that makes up the story/event (news body) and includes the following auxiliary information:

- Source (e.g. <https://dantri.com.vn>, <https://saobiz.vn>).
- Headline is the title that describes the main topic of the news. The title is often named so that it attracts the reader's attention.
- Author
- Publication time

3.2.2. Social context

The social context of a news can be posts, likes, shares, replies, etc. When the features relevant to the news content are insufficient or unavailable, the social context is useful information for authenticating the news. Ancillary information related to the social context is as follows:

- Social network users (user)
- Title is the title or short caption of the post. This title is closely related to the title of the news.
- Score is the rating for a post given by other users, which determines the acceptance or disapproval of the post by other users.
- Number of comments is the number of comments on a post, this characteristic shows the popularity of the post.
- The upvote/downvote ratio estimates the approval/disapproval of other user's posts.
- User credibility: This is a feature that helps determine if users are prone to spreading fake news. For example, if the user's other posts are not trusted, it is likely that the next post will also be unreliable.

3.2.3. Representation of news content and social context

News content description features have four [language] levels: (i) lexicon, (ii) syntax,

(iii) discourse and (iv) semantic. In this step, we represent the content at the lexical level according to the BoW model. Suppose the dataset contains n news $N = \{n_1, n_2, \dots, n_{|N|}\}$ with a total of t words $W = \{w_1, w_2, \dots, w_t\}$. Let x_j^i is the number of words w_j appearing in n_i . Then, the normalized frequency of w_j for the news n_i is calculated according to Equation (4).

$$w_j = \frac{x_j^i}{\sum_{j=1}^t x_j^i} \quad (4)$$

Thus, the new n_i is represented as a feature vector $v_i = \{w_1, w_2, \dots, w_t\}$.

3.3. Classification model

Figure 5 shows an overview of the model that we propose to use.

• First, from the raw dataset (as shown in Figure 6), the data is preprocessed and extracted featuring news content and social context, respectively. The input is news content (identifier of news, publishing source, title of news, main content) and social context features (number of likes, number of shares, user identifier), the output is a vector representation of news content and social context. For each news $n_i \in A$ (set of news), we represent it as a feature vector $v_i \in R^k$. This representation is tailored to each machine learning algorithm. The vector representations are combined to produce a single representation that is passed as input at the next stage. The final output is passed to the classifier.

• Next, we use machine learning algorithms to train the classification model (traditional machine learning algorithms such as Support Vector Machine (SVM), Naive Bayes (NB), k -Nearest Neighbor (k -NN)).

• Finally, we use the classification model to predict whether the input data is true or fake news. The model's prediction results are compared with actual (labeled) data to evaluate the model's effectiveness.

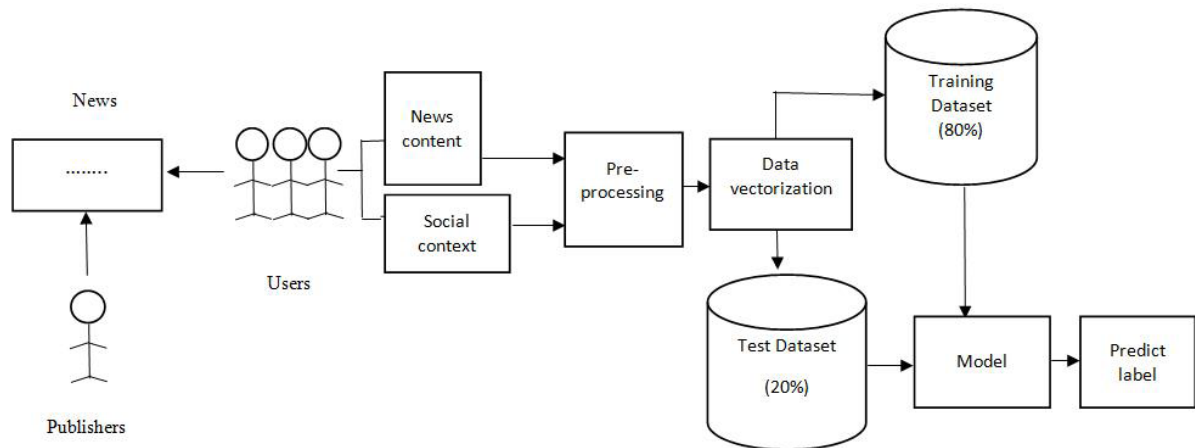


Figure 5. Proposed model.

3.3.1. Naive Bayes

Naive Bayes algorithm uses conditional probability between attributes and class label to determine the class of a data sample to be classified.¹⁸

Let D is the training dataset: $D = \{X_1, X_2, \dots, X_n\}$. Where, each X_1 is represented by a vector containing m attributes $X_1 = \{x_{11}, x_{12}, \dots, x_{1m}\}$. Let C be a set of labels consisting of p classes: $C = \{C_1, C_2, \dots, C_p\}$.

Given the data sample $X = \{X_1, X_2, \dots, X_m\}$, the Naive Bayes classifier will predict X belong to class C_i if: $P(C_i|X) > P(C_j|X)$, ($1 \leq i, j \leq p, i \neq j$). The process of classifying data sample X according to Naive Bayes algorithm is described in Algorithm 1.

Algorithm 1 Naive Bayes classifier algorithm

Input: D : Training dataset; C : Label set; X : New data samples.

Output: Label of X

- 1: for $i = 1$ to p do
 - 2: $P(C_i) = |X_i|/|D|$ //Calculate the probability of occurrence of the class C_i , where $|X_i|$ is the number of data samples belonging to the class C_i .
 - 3: $P(C_i|X) = P(x_1|C_i) \times P(x_2|C_i) \times \dots \times P(x_n|C_i) \times P(C_i)$ //Calculate the classification probability C_i is the label of the data sample X .
 - 4: $label(X) \leftarrow C_i$ if $P(C_i|X)$ is the biggest.
-

3.3.2. Support Vector Machine

The main idea of the SVM algorithm is that given a training set represented in vector space, where each data sample is a point, this method finds a decision hyperplane h that can best divide the points on this space into separate classes. The quality of this hyperplane is determined by the distance of the nearest data point of each class to this plane. The larger the boundary distance, the better the decision plane, and the more accurate the classification. The purpose of the SVM algorithm is to find the maximum boundary distance to give the best classification result.¹⁹

Suppose we need to classify a data sample X into one of two classes $C_1 = -1$ and $C_2 = 1$. The SVM classification algorithm is detailed in Algorithm 2.

Algorithm 2 SVM classifier algorithm

Input: D : Training dataset; C : Label set; X : New data samples; T : bias coefficient; $W = \{w_1, w_2, \dots, w_n\}$: The set of hyperplane coefficients (vector weights).

Output: Label of X

- 1: for $i = 1$ to n do
 - 2: $f(X) \leftarrow \text{sign}(T + \sum w_i \times x_i)$
 - 3: $z = T + \sum w_i \times x_i$
 - 4: if ($z \geq 0$) then
 - 5: $\text{sign}(z) \leftarrow 1$
 - 6: else $\text{sign}(z) \leftarrow -1$
 - 7: $label(X) \leftarrow \text{sign}(z)$
-

The SVM classification algorithm depends on the weight vector parameters W and the bias coefficient T . The goal of SVM is to estimate W and T to maximize the margin between the data classes.

3.3.3. *k*-Nearest Neighbor

k-NN is one of the simplest supervised learning algorithms. *k*-NN algorithm classifies new data points based on *k* nearest data points (*k* - neighbors). The measure used to calculate the distance between two data points can be Euclidean, Manhattan, Minkowski, Cosine.²⁰ Algorithm 3 describes the steps to classify a data point according to the *k*-NN algorithm.

Algorithm 3 *k*-NN classifier algorithm

Input: D : Training dataset; C : Label set; X : New data samples; K : Number of neighbors.

Output: Label of X

1: for $i = 1$ to n do

2: $d(X_i, X)$ //Calculate the distance between X and the data points in the training set.

3: end for

4: Create a set $I = \{X_i\}$, $|I| = K$ // I contains K data point closest to X .

5: for $i = 1$ to p do

6: Count the number of classes C_i that appear in I

7: $label(X) \leftarrow C_i$, if C_i is the most appear class.

4. EXPERIMENT

4.1. Dataset

A significant challenge for automated fake news detection is the availability and quality of datasets. In the experiment, we use two datasets PolitiFact and GossipCop.²¹

Table 2 describes the news type, size and number of label 1 (fake news) statistics of the datasets. PolitiFact dataset has news of the type of article, and it includes 1,056 news, in which 432 news is fake news.

Table 2. Description of experimental data.

Dataset	Type	Size	Number of label 1
PolitiFact	article	1,056	432
GossipCop	article	22,140	5,323

Figure 6 is a snapshot of the data in the GossipCop dataset. The data includes news content and social context. In which, the content features include id (identifier of news), news_url (publishing source), title (title of news), news_body (main content), count_like (number of likes), count_share (number of shares), user_ids (user identifier, each with 18 numbers). For example, the first data line in Figure 6 has the following features:

	A	B	C	D	E	F	G	H
1	id	news_url	title	news_body	count_like	count_share	user_ids	label
2	gossipcop-2493749932	www.daily	Did Miley Cyr	Congratulations mig	12096	5421	2843290759029	fake
3	gossipcop-4580247171	hollywooc	Paris Jackson & Cara Delevingne E		746	1167	9928955082671	fake
4	gossipcop-941805037	variety.coi	Celebrities Joi Elaine L. Chao, a vet		1269	529	8533593535328	fake
5	gossipcop-2547891536	www.daily	Cindy Crawfo	In a move that left r	870	577	9888219051961	fake
6	gossipcop-5476631226	variety.coi	Full List of 201	Good morning. (Wa	756	921	9557927936324	fake
7	gossipcop-5189580095	www.tow	Here's What F	Federal prosecutors	2932	147	8902530052993	fake
8	gossipcop-9588339534	www.foxn	Biggest celebr	As he prepared last	4564	125	6832263807425	fake
9	gossipcop-8753274298	www.eonl	Caitlyn Jenner	Taiwan scrambled	4511	1000	1026891446081	fake
10	gossipcop-8105333868	www.inqu	Taylor Swift R	President Xi Jinping	1966	2476	8189285335694	fake
11	gossipcop-2803748870	www.huff	For The Love	First it was the icy sr	629	1185	8160302481900	fake
12	gossipcop-7312096991	www.msn	Miley Cyrus, L	Abdul Ali Shamsi ha	11188	917	9301835055138	fake
13	gossipcop-5328748354	yournews	Miley Cyrus C	SEOUL, South Korea	8573	93	9142087416078	fake
14	gossipcop-9878194459	www.mar	Selena Gome	Photographs recent	4529	1741	1016502265060	fake
15	gossipcop-9521617242	www.eonl	Critics' Choice	President Obama de	538	500	9386812614086	fake

Figure 6. Part of the data in the GossipCop dataset.

D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
news_body	count_like	count_share	user_ids	label											
Congratulations might be in order for Miley Cyrus and Liam Hemsworth, according to new reports. Insiders told Life & Style magazine the pair secretly web in a 'hippie-style' ceremony at their Malibu, California mansion. 'They've actually done the deed for real this time! Miley and Liam recently had a secret wedding at her Malibu mansion,' the source said. 'Neighbors had no clue it was a wedding. Kids were running around everywhere. It was a hippie-style wedding.' The source continued: 'She wore a white, flowey dress and all her animals were roaming around, it was crazy! The food was vegetarian and organic.' As for their vows, the anonymous informant said: 'They expressed their love, how they're each other's best friend and want to grow old together. 'There wasn't a dry eye in the house.' Life & Style reported that, rather than invite someone to perform at their wedding, Miley did the honors herself. She sang a rendition of godmother Dolly Parton's song, Islands in the Stream. Wayne Coyne of the Flaming Lips also performed a few songs. Miley's mother and father, Billy Ray and Trish Cyrus, were both present along with siblings Noah, Trace, Brandi, Braison and Christopher. Liam's famous family attended as well. According to the report, both Chris and Luke Hemsworth looked on as Miley and Liam said their 'I Dos'. Neither Miley nor Liam have shared photos from their rumoured wedding day, but this won't be their last opportunity. According to the outlet, the couple plan to have another wedding ceremony in Australia so that more of Liam's family can be involved. As for their post wedding plans, the source said Miley and her alleged hubby have their sights set on a baby. 'It's been Miley's dream to have a baby,' they said. Here's what you need to know.															

Figure 7. Main content of the news.

user_ids	label														
2843290759029268	482843327445599682562843354125902970892843599187928801282843853851513323532844563888334520322846448782	673182732846565589635645452847053919657943042847053919574016012847624255537233922847660898277212172847697668062576642	8476976949062041628476977544233779228476977298030182428476977865518694428476977707391795228477191375304294428490294320	183705628491893997530316928559226671688499328563542187297996928563551692430131328573132410128793729595453355579392029	6639632710459392296651276626825216700137152143699968700501642412560384700599457767890944700925213031141376757598980373	045248757614017808437248757810204448743425757855454886293504818454872299163649818532208990371841818622382524997632818	6236447378718728243280593356922908402379266425364498402779880674304008403055567368724488403126410421043208403203834378	772488403213781506703368403254493896007688403467926716989458403591157549998108403591401196994578403664596750622748403	9688158698291284041337065005056084044084752711680084044429973291008084046255013536563384184926864633856084967804105588	736092585119156055654492742672865323008092753101221245747392758887557652889792758887555136307292758912119076044992760	6510406029314928442019529162752928532613056352256928653424752758784953201937574342658953336408650993664953692192064884	737953999445384888321954307670344916992101003977889145241610101841509706547201010184410115727361101019144461775257610	1022901390899200010102624286707097661010334767701561344101039380914479923210104184440162590721011127764261396481101372	3578834014208101372360157134029110137238136856698891016000562893123584101643170513209344010364442963486638081060594707	447853058106059544648835891210606795022404362241060683397649952770106070857359077376110607222723731865610608121262002

Figure 8. User identifier.

- id: gossipcop-2493749932
- news_url:
www.dailymail.co.uk/tvshowbiz/article-5874213/Did-Miley-CyrusLiam-Hemsworth-secretly-married.html
- title: Did Miley Cyrus and Liam Hemsworth secretly get married?
- news_body: (Figure 7)
- count_like: 12096
- count_share: 5421
- user_ids: (Figure 8)

4.2. Experimental setup

We use three machine learning algorithms, including: SVM, NB, and *k*-NN to train

classification models on two different datasets. From the input datasets, we preprocess the data by removing stop word and special symbols, then vectorize the data matching each algorithm at the lexical level.

The training data and the test data were split in a ratio of 8:2, using a 5-fold cross-validation method.

To evaluate the classification models, we use the confusion matrix as shown in Table 3, where:

- TP (true positive): Number of news predicted to be fake news and actually fake news;
- FN (false negative): Number of news that are predicted to be fake news when in fact they are true;

- FP (false positive): Number of news that are predicted to be true when they are actually fake;
- TN (true negative): Number of news predicted to be true and in fact true.

Table 3. Confusion matrix performance.

Actual ↓ Prediction →	Fake news	True news
Fake news	TP	FP
True news	FN	TN

P (Precision), R (Recall) and F_1 are calculated as follows:

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{7}$$

We implement machine learning algorithms and evaluate classification models based on the open source tool Scikit-learn.²² We use the following classification models:

- SVM classification model: *SVC (kernel = linear)*
- Naive Bayes classification model: *GaussianNB()*
- *k*-NN classification model: *KNeighborsClassifier()*

4.3. Results and discussion

Table 4, Table 5 and Table 6 present the

experimental results of the classification models on the datasets that are news content only (PolitiFact (C) and GossipCop (C)) and social context only, respectively (PolitiFact (SC) and GossipCop (SC)), and combine both news content and social context (PolitiFact (C+SC) and GossipCop (C+SC)). Experimental data show that all three classification models (SVM, NB and *k*-NN) achieve the measure of F_1 above 75%. It can be seen from the experimental results that when using a dataset combining news content and social context, almost all three models give better classification results. Specifically, when applying the *k*-NN algorithm on the PolitiFact (C+SC), the F_1 measure is 7.4% higher when running on the PolitiFact (C) and 7.7% higher when running on the PolitiFact (SC). In another case, when applying the SVM algorithm on the GossipCop (C+SC) dataset, the F_1 measure is 2.8% higher when running on the GossipCop (C) dataset and 2.6% higher when running on the GossipCop(SC) dataset.

Figure 9 and Figure 10 show the comparison of the F_1 measure between the classifiers on PolitiFact(C), PolitiFact(SC) and PolitiFact(C+SC) datasets; and between classifiers on GossipCop(C), GossipCop(SC) and GossipCop (C+SC) datasets. Experimental results show that most algorithms applied to datasets that combine news content and social context give better results of measuring F_1 when applied on datasets with only news content or social only social context.

Table 4. Experimental results on datasets with only news content.

Dataset→	PolitiFact (C)			GossipCop (C)		
Metric→	P	R	F_1	P	R	F_1
<i>k</i> -NN	0.609	0.995	0.755	0.827	0.879	0.852
NB	0.821	0.862	0.841	0.87	0.655	0.751
SVM	0.795	0.922	0.853	0.876	0.904	0.890

Table 5. Experimental results on datasets with only social context.

Dataset→	PolitiFact (SC)			GossipCop (SC)		
Metric→	P	R	F ₁	P	R	F ₁
k-NN	0.603	1.0	0.752	0.831	0.894	0.861
NB	0.831	0.817	0.824	0.875	0.645	0.743
SVM	0.793	0.849	0.820	0.876	0.910	0.892

Table 6. Experimental results on datasets combining news content and social context.

Dataset→	PolitiFact (C+SC)			GossipCop (C+SC)		
Metric→	P	R	F ₁	P	R	F ₁
k-NN	0.887	0.779	0.829	0.798	0.856	0.826
NB	0.874	0.832	0.852	0.831	0.868	0.849
SVM	0.815	0.917	0.863	0.883	0.956	0.918

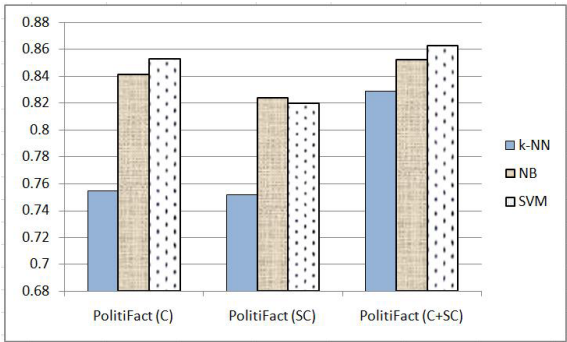


Figure 9. Compare the F_1 measure on the PolitiFact(C), PolitiFact(SC) and PolitiFact(C+SC) datasets.

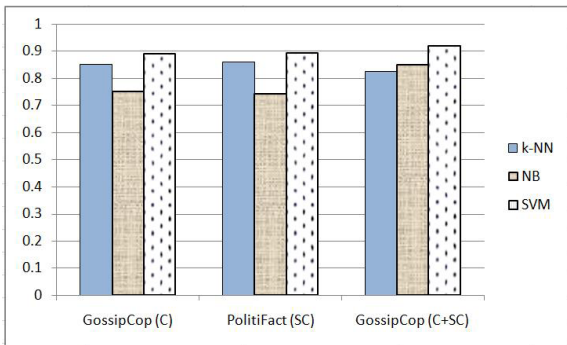


Figure 10. Compare the F_1 measure on the GossipCop(C), GossipCop(SC) and GossipCop(C+SC) datasets.

5. CONCLUSION

In this paper, we have presented a research on detecting fake news based on news content and social context approach using machine learning.

We have analyzed related concepts, methods of detecting fake news. We have modeled this task as a binary classification problem, representing content and social contexts as feature vectors. Then we used machine learning algorithms to build the classification model. Experimental results with three machine learning algorithms (SVM, NB and k -NN) on two different datasets show the effectiveness of the proposed method.

In the future, we plan to extend this study towards content analysis in terms of natural language processing at the syntactic and semantic levels, and build a Vietnamese dataset for research on fake news detection problem.

Acknowledgement

This research is conducted within the framework of science and technology projects at institutional level of Quy Nhon University under the project code T2022.762.18.

REFERENCE

1. R. Oshikawa, J. Qian, W. Y. Wang. *A survey on natural language processing for fake news detection*, The 12th Language Resources and Evaluation Conference, France, 2020.
2. X. Zhou, R. Zafarani. *A survey of fake news: fundamental theories, detection methods*,

- and opportunities, *ACM Computing Surveys (CSUR)*, **2020**, 83(5), 1-40.
3. V. L. Rubin. *On deception and deception detection: content analysis of computer mediated stated beliefs*, The American Society for Information Science and Technology, USA, 2010.
 4. R. Zafarani, M. A. Abbasi, H. Liu. *Social media mining: an introduction*, Cambridge University Press, United Kingdom, 2014.
 5. A. Choudhary, A. Arora. Linguistic feature based learning model for fake news detection and classification, *Expert Systems with Applications*, **2021**, 169, 114-171.
 6. Y. Bang, E. Ishii, S. Cahyawijaya, Z. Ji, and P. Fung. *Model generalization on covid-19 fake news detection*, Communications in Computer and Information Science, United States, **2021**, 128-140.
 7. C. Song, K. Shu, B. Wu. Temporally evolving graph neural network for fake news detection, *Information Processing & Management*, **2021**, 102712.
 8. J. A. Nasir, O. S. Khan, I. Varlamis. Fake news detection: a hybrid cnn-rnn based deep learning approach, *International Journal of Information Management Data Insights*, **2021**, 1(1), 100007.
 9. R. K. Kaliyar, A. Goswami, P. Narang. Deepfake: Improving fake news detection using tensor decomposition-based deep neural network, *The Journal of Supercomputing*, **2021**, 77, 1015-1037.
 10. V. P. Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea. *Automatic detection of fake news*, The 27th International Conference
 11. A. Jain, A. Kasbe. *Fake news detection*, 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), India, 2018.
 12. J. C. S. Reis, A. Correia, F. Murai, A. Veloso, F. Benevenuto. Supervised learning for fake news detection, *IEEE Intelligent Systems*, **2019**, 34(2), 76-81.
 13. N. K. Conroy, V. L. Rubin, Y. Chen. *Automatic deception detection: Methods for finding fake news*, The Association for Information Science and Technology, USA, 2015.
 14. O. Ngada, B. Haskins. *Fake news detection using content-based features and machine learning*, 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDSE), Australia, 2020.
 15. H. E. Wynne, Z. Z. Wint. *Content based fake news detection using n-gram models*, The 21st International Conference on Information Integration and Web-based Applications & Services, Indonesia, 2019.
 16. K. Shu, S. Wang, H. Liu. *Beyond news contents: the role of social context for fake news detection*, The twelfth ACM international conference on web search and data mining, Australia, 2019.
 17. P. K. Verma, P. Agrawal. PropFND: propagation based fake news detection, *Applications of Artificial Intelligence and Machine Learning: Select Proceedings of ICAAAIML 2021*, Springer, Singapore, 2022, 557-569.
 18. J. Han, J. Pei, M. Kamber. *Data mining: concepts and techniques*, Elsevier, India, 2011.
 19. V. Vapnik, I. Guyon, T. Hastie. Support-vector machines, *Machine Learning*, **1995**, 20, 273-297.
 20. T. Hastie, R. Tibshirani. Discriminant adaptive nearest neighbor classification, *IEEE Transactions on Pattern Analysis Andmachine Intelligence*, **1996**, 18(6), 607-616.
 21. K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu. Fakenewsnet: a data repository with news content, social context, and spatio temporal information for studying fake news on social media, *Big Data*, **2020**, 8(3), 171-188.
 22. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau. Scikit-learn: machine learning in python, *The Journal of Machine Learning Research*, **2011**, 12, 2825-2830.