Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 7 No. 1 (2023) 185 - 192 ISSN Media Electronic: 2580-0760

Covid-19 Fake News Detection on Twitter Based on Author Credibility Using Information Gain and KNN Methods

Nanda Ihwani Saputri[1], Yuliant Sibaroni[2], Sri Suryani Prasetiyowati[3]
[1,2,3]Informatika, Fakultas Informatika, Universitas Telkom
[1]nandaihwani@student.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id, [3]srisuryani@telkomuniversity.ac.id

Abstract

Twitter is one of the social media that is used as a tool to share various kinds of information about various kinds of things that are of concern to social media users. One of the information shared is information about COVID-19, which is known that the COVID-19 pandemic is currently spreading throughout the world at a very alarming rate. COVID-19 is an infectious disease caused by SARS-COV-2. The World Health Organization (WHO) claims that the spread of COVID-19 is supported by the spread of false/fake news. So to find out the truth of the news, a COVID-19 fake news detector is needed so that users don't fall for the hoaxes circulating. This study aims to classify COVID-19 news on Twitter based on author credibility. Credibility in question is a person's perception of the validity of information and is a multidimensional concept that is used as a means of receiving information to assess the source of communication. The method used in this research is Information Gain and KNN. KNN (K-Nearest Neighbor) is a supervised learning algorithm that works by classifying a set of data based on classified training data. Information Gain is used to ranking the most influential attributes, and KNN is used to classify data based on learning data taken from the nearest neighbors. The research consists of 6 main stages, namely data collection (crawling data), data preprocessing, feature extraction, feature selection, data split into training data and testing data, KNN stage, and data evaluation stage. The research carried out succeeded in obtaining an accuracy value of 91%, a correlation value between credibility and hoax of 0.115, and a p-value <0.005.

Keywords: twitter; fake news, COVID-19; credibility; KNN; information gain

1. Introduction

The development of communication technology helps humans in sending and receiving information. Virtual communities are starting to form and shift traditional communities. The virtual community that we can meet is social media. Distribution of information in the form of online conversations can be done through social media. The most obvious participation and use of social media can be seen in social media such as Facebook and Twitter [1]. Based on a survey conducted by the Association of Indonesian Internet Service Providers, shows that internet penetration in Indonesia in 2020 will reach 171.17 million people, equivalent to 64.8% of Indonesia's total population [2]. Meanwhile, when a similar survey was conducted in 2017, 143 million internet users were found, equivalent to 54.7% of Indonesia's total population. This shows that there was an increase in the number of internet users from the previous year of 10.1%. This increase in internet access is in line with access to social media, both Facebook and Twitter.

Since its launch in 2006, Twitter has become one of the most popular social media platforms for sharing information, both personal information and a means of interaction in various parts of the world [3]. Dissemination of information on Twitter is done through making tweets. One of the information being disseminated is information about COVID-19, which is known that the COVID-19 pandemic is currently spreading throughout the world at a very alarming rate [4]. The World Health Organization (WHO) even claims that the spread of COVID-19 is supported by the spread of false/fake news [5].

Fake news about COVID-19 seems to spread very quickly on social media [6]. Similar trends have been seen during other epidemics, such as the Ebola, yellow fever, and Zika outbreaks [4]. This is a very worrying development because even a little exposure to fake news can cause public anxiety and distrust [5]. In addition, it is necessary to identify the creator or subject of fake news which will help eradicate a large number of fake news from its origin [7]. Generally, for news spreaders, in addition to the tweets that are made,

Accepted: 17-01-2023 | Received in revised: 03-02-2023 | Published: 06-02-2023

there is social media profile information such as username, location, user description, account creation time, and followers. Which information can be used to obtain fundamental complementary information for background checks as well as a basis for determining the credibility of news spreaders [7]. The credibility in question is a person's perception of the validity of information and is a multidimensional concept that is used as a means of receiving information to assess the source of communication [8].

Previous research presented case studies on credibility based on articles published by "Donald Trump", "Mike Pence", "Barack Obama", and "Hillary Clinton" [7]. Next, it explains 4 case studies related to the credibility of the authors based on published articles. Then, dividing the case into 2 groups, namely republican and democratic. Most of the dataset from "Donald Trump" is evaluated incorrectly by 69%. For the "Mike Pence" dataset it is evaluated at 52%: 48% for true and false news. Meanwhile, the dataset "Barack Obama" and "Hillary Clinton" were considered correct, namely 76% and 73% [7].

Therefore, this study aims to classify news on Twitter tweets, using the Information Gain and KNN methods. This classification is intended to analyze whether the tweet is fake news (hoax) or non-hoax based on the author's credibility.

2. Research Methods

This research was conducted using the Information Gain and KNN methods. The KNN method is used to classify data based on learning data taken from the nearest neighbors [9]. Meanwhile, Information Gain is used to ranking the most influential attributes [10]. This is done because feature selection (Information Gain) is an important part that can optimize classifier performance [10], [11]. It is known that the use of the Information Gain algorithm can reduce the vector dimensions in the dataset [11], [12].

2.1 Research data

The dataset used in this study was obtained from Kaggle, which is tweet data from Twitter, and was collected by Gabriel Preda using the Twitter API combined with Python scripts. Queries are run daily for a set period, to collect a larger sample of tweets. The tweets obtained contain the hashtag #covid19. Tweet data collection started on 25/7/2020 to 29/8/2020, with an initial batch of 17,000 data.

2.2 Research Framework

This study consists of 6 stages, namely data collection (data crawling), data preprocessing, feature extraction, feature selection (Information Gain), data split into training data and testing data, KNN stage, and data evaluation stage (Figure 1).

Table 1. Twitter Data with the Hashtag #COVID19

User name	User followers	text
Time4fisticuffs	9275	@diane3443 @wdunlap @realDonaldTrump Trump never once claimed #COVID19 was a hoax. We all claim that this effort toâ€¦ https://t.co/Jkk8vHWHb3
DIPR-J&K	101009	25 July: Media Bulletin on Novel #CoronaVirusUpdates #COVID19
hr bartender	79956	How #COVID19 Will Change Work in General (and recruiting, specifically) via/ @ProactiveTalent #Recruitingâ€¦ https://t.co/bjZxzGPMbK
Greater Visakhapatnam Municipal Corporation (GVMC)	14357	GVMC sanitation staff carrying out the regular sanitation activities to keep the city clean and prevent the spreadâ€¦ https://t.co/bkrQ5x6BCK
Fergus McPop	1029	Coronavirus Testing Fiasco: St MirrenÂ have pledged to undertake an "urgent review" of their Covid-19 testing procedâ€¦ https://t.co/bfel6gyXIq
Albert Trigg	3956	A review of the recent study (now retracted) which connected #5G with #COVID19
..	..	..
rugby365.com	35049	#ICYMI: New @wallabies coach Dave Rennie will have a few tough decisions to make in the build-up to this year's revâ€¦ https://t.co/trk0GDrii5

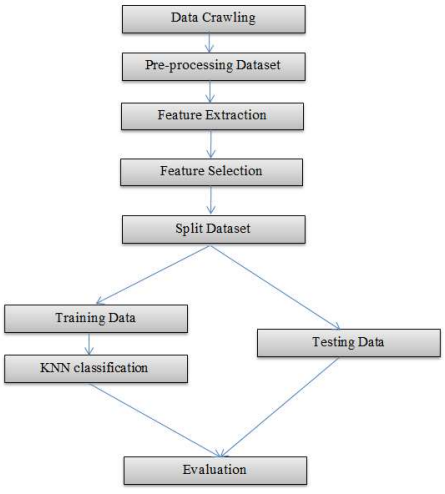

Figure 1. System Flowchart

In the data crawling process, tweets containing the hashtag #COVID19 are collected. Data is obtained from Twitter using the Twitter API combined with

DOI: https://doi.org/10.29207/resti.v7i1.4871