University of Kerbala
College of Computer Science & Information Technology
Computer Science Department

Detecting Vandalism in Crowdsourcing Models

A Thesis

Submitted to the Council of the College of Computer Science & Information Technology / University of Kerbala in Partial Fulfillment of the Requirements for the Master Degree in Computer Science

Written by

Azha Talal Mohammed Ali

Supervised by

Lecturer Dr. Huda Fadhil Hallawi

Asst. Prof. Dr. Noor D. Al-Shakarchy

2023 A.D. 1444 A.H.

Supervisor Certification

I certify that the thesis entitled (**Detecting Vandalism in Crowdsourcing Models**) was prepared under my supervision at the department of Computer Science/College of Computer Science & Information Technology/ University of Kerbala as partial fulfillment of the requirements of the degree of Master in Computer Science.

Signature:

Supervisor Name: Lecturer Dr. Huda Hallawi

Date: 25 / 7 / 2023

Signature:

Supervisor Name: Asst. Prof. Dr. Noor D. Al-Shakarchy

Date: 25 / 7 / 2023

The Head of the Department Certification

In view of the available recommendations, I forward the thesis entitled "Detecting Vandalism in Crowdsourcing Models" for debate by the examination committee.

Signature:

Assist. Prof. Dr. Muhannad Kamil Abdulhameed

Head of Computer Science Department

Date: 25 / 7 / 2023

Certification of the Examination Committee

We hereby certify that we have studied the dissertation entitled **(Detecting Vandalism in Crowdsourcing Models)** presented by the student **(Azha Talal Mohammed Ali)** and examined him/her in its content and what is related to it, and that, in our opinion, it is adequate with (**Excellent**) standing as a thesis for the degree of Master in Computer Science.

Signature: Moh
Name: Mohammed Abdullah Nasir
Title: Prof. Dr
Date: 24/ 7 / 2023
(Chairman)

Signature:
Name: Mohsin Hasan Hussein
Title: Asst. Prof. Dr.
Date:24/ 7 / 2023
(Member)

Signature:
Name: Akeel AbdulKarim Farhan
Title: Asst. Prof. Dr.
Date: 24/ 7 / 2023
(Member)

Signature:
Name: Huda Hallawi
Title: Lecturer Dr.
Date: 24/ 7 / 2023
(Member and Supervisor)

Signature:
Name: Noor D. Al-Shakarchy
Title: Asst. Prof. Dr.
Date:24/ 7 / 2023
(Member and Supervisor)

Approved by the Dean of the College of Computer Science & Information Technology, University of Kerbala.

Signature:
Title : Assist. Prof. Dr.Ahmed Abdulhadi Ahmed
Date: 2 / 8 / 2023
(Dean of Collage of Computer Science & Information Technology)

Dedication

To..

My Mother, My Father Who support & encourage me..

My Husband, Hasan who stand beside me in each step of our life..

My Sons, Ali Rida

Mohammed

Mahdi

For charge me with the energy of innocent love

Abha my sister, Mustafa & Amor.. My brothers

For being present in every situation

Azha T.

Acknowledgement

In the name of Allah, all praise to Allah Most Gracious, Most Merciful. Alhamdulillah, I have finished this research.

I wish to express my gratitude to my supervisor, Dr. Huda Hallawi, for her assistance.

Also, I would like to express my deepest sincere gratitude to my supervisor, Asst. Prof. Dr. Noor D. Al-Shakarchy, for her valuable advice and guidance. Without her assistance, it would not have been possible for me to complete this thesis. I will never forget her encouraging words during the most difficult times. To her, I shall remain thankful.

I would like to take this opportunity to express my gratitude to Assist.Prof. Dr.Ahmed Abdulhadi Ahmed, the dean of the College of IT / University of Kerbala, Assist. Prof. Dr. Mohsin Hasan Hussein, the Deputy Dean for Scientific Affairs, and Assist. Prof. Dr. Muhannad Kamil Abdulhameed, the Head of Computer Science Department, for their cooperation in showing me the proper direction in order to finish this research.

Also, I wish to thank my friends for their assistance and for being there every time.

The Researcher.

Abstract

Crowdsourcing is a common term for collaborative environments that include a group of workers participating in the content of these platforms to create knowledge. The most important and well-known crowdsourcing platforms are Amazon's Mechanical Turk, Kaggle, and Wikipedia. The significant expansion of the encyclopedia made it one of the most visited sites for getting information or using it as a knowledge base for many applications, such as chatbots. This issue comes with another aspect through which users can sabotage the content of Wikipedia, which is known as vandalism.

Vandalism is any attempt to modify the article negatively affecting its quality, which is considered one of the five layers of the model constructed for cyber threats and is known as cyber vandalism. Several automatic detection techniques and related features have been developed to address this issue.

This thesis introduces a deep learning model with a new and light architecture to detect vandalism in Wikipedia articles. The proposed model employs a one-dimensional convolutional neural network architecture (1D-CNN) that can determine the type of modification in Wikipedia articles based on two main stages: the feature extraction stage and the vandalism detection stage. Features are extracted from edits and their associated metadata, as well as new features (reviewers' trust), and then only the salient features are adopted to make a decision about the article; regular or vandalism can contribute to improving the accuracy of prediction. The experiments were conducted on a benchmark dataset, the

PAN-WVC-2010 corpus, taken from a vandalism detection competition hosted at the CLEF conference. The proposed system, with the new features added, has achieved an accuracy of 96%.

Declaration Associated with this Thesis

1- A. T. Mohammed Ali, H. F. Alwan and N. D. Al-Shakarchy, "A Survey on Detecting Vandalism in Crowdsourcing Models," *2022 International Conference on Data Science and Intelligent Computing (ICDSIC)*, Karbala, Iraq, 2022, pp. 25-30, doi: 10.1109/ICDSIC56987.2022.10076011.

2- CyberVandalism Detection in Wikipedia Using Light Architecture of 1D-CNN, Submitted to Karbala International Journal of Modern Science.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Description
1D CNN	One Dimension Convolution Neural Network
AI	Artificial Intelligence
ANN	Artificial Neural Networks
AUC	Area Under the ROC Curve
BOW	Back of Word
CLEF	Conference and Labs of the Evaluation Forum
CM	Configuration management
CM	Confusion Matrix
CRF	Conditional Random Field
CSV	Comma Separated Value
CV	Cross Validation
DCNN	Deep Convolution Neural Networks
DL	Deep Learning
DNN	Deep Neural Network
DoS	Denial of Service
FC	Fully Connected
ICT	Information Communication Technology
LMT	Logistic Model Tree
LNSMOTE	Local Neighbor-Synthetic Minority Oversampling Technique
ML	Machin Learning
MLP	Perceptron Network with Multiple Layers

NLP	Natural Language Processing
NN	Neural Network
PAN-WVC-10	Wikipedia Vandalism Corpus 2010
PAN-WVC-11	Wikipedia Vandalism Corpus 2011
PCFG	Probabilistic Context Free Grammars
POS	Part-of-Speech
ReLU	Rectified Linear Unit (ReLU)
ROC	Receiver Operating Characteristic
SDA	Stacked Denoising Autoencoders
SMOTE	Synthetic Minority Oversampling Technique
SQL	Structured Query Language
UGC	User Generated Content
VDM	Vandalism Detection Model
VDS	Vandalism Detection System
XML	Extensible Markup Language

CHAPTER ONE

INTRODUCTION

1.1 Overview

Modern information systems strongly depend on knowledge bases. For example, web engines use them to enhance query results, Chatbots use them to answer factual inquiries, and fake news detectors use them to verify information. Crowdsourcing is used widely in large-scale knowledge collection [1-3].

Crowdsourcing refers to the practice of obtaining ideas, services, or content from a large and undefined group of people, usually through an online platform. It involves leveraging the collective intelligence and creativity of a diverse group of individuals to solve problems, create content, or generate new ideas. Crowdsourcing can take many different forms, such as asking for product feedback, conducting market research, or soliciting user-generated content like photos or videos. The practice of crowdsourcing has become increasingly popular in recent years due to advances in technology that make it easier to connect and collaborate with large groups of people [4-7].

Crowdsourcing can be an effective way to generate new ideas and engage with a wider community, but it also requires careful management and oversight to ensure that the resulting content or ideas are of high quality and value [4-7].

According to [5] Wikipedia is a free international encyclopedia that was founded as a collaborative model. Without restrictions, any encyclopedia article could be edited anonymously and virtually, and modifications are instantaneously published. This adaptability has assisted its fast expansion in recent years, the dark side of this benefit is that it gets vandalized [6-11] and as a result, reduces the article's quality [8]. The quality of Wikipedia material is important for both humans and machines, as well as the integrity of all human knowledge [9].

Vandalism in crowdsourcing, such as on Wikipedia, refers to the act of intentionally introducing incorrect or misleading information or content into a crowdsourcing platform, disrupting the accuracy and reliability of the information being collected. This can include the submission of false data, spam, inappropriate content, or any other action that deliberately undermines the integrity of the crowdsourcing project. Vandalism can be a major problem for crowdsourcing efforts, as it can lead to inaccurate or incomplete data, which can in turn affect the usefulness and reliability of the information being collected. To prevent vandalism, crowdsourcing platforms often have measures in place, such as community moderation or algorithms that flag suspicious activity, to help detect and remove problematic content. [14].

Wikipedia has a mechanism that allows people to view older versions of articles regardless of what happens to the current edition. Since Wikipedia's foundation, this revision system has maintained track of all changes to all articles. The above allows readers to check the current status of an article at any moment, as well as restore a page to an earlier version if it has been modified maliciously [10] .

Researchers and the Wikipedia community considered vandalism as an open problem studied in online collaborative projects [10,15,16].

The challenge of detecting vandalism is strongly connected to computer security issues such as intrusion detection and spam filtering from emails and weblogs. It's a special type of online defacing [13].

Classification algorithms may be used to detect and identify vandalism [16] of vandalism in Wikipedia. That may be done in three ways: based on the author's reputation, based on the analysis of metadata modifications, and based on the characteristics of text [9,21,22] , all of which are significant in detecting vandal edits in Wikipedia [19].

1.2 Problem Statement

One of the biggest threats to Wikipedia is vandalism. Anyone can edit a Wikipedia page, which means that it is vulnerable to malicious edits, misinformation, and hoaxes, which, in turn, leads to failures in the dependent systems.

As a result, vandalism in such an encyclopedia is seen as a task that needs more attention, not in the detection of vandalism only but also in guaranteeing speed and efficiency in detection.

1.3 The Aim of the Thesis

The aims of this thesis can be illustrated as:

1. Detect vandalized articles in Wikipedia models.
2. Study new characteristics that can be implemented to detect vandalism on Wikipedia’s articles.
3. Design new architecture for vandalism detection model (VDM) using deep learning techniques.

1.4 Objectives of the Thesis

1- Produce an effective Vandalism Detection system detection by employing deep learning techniques with a new light architecture of 1DCNN for Wikipedia articles, which can predict the status of Wikipedia articles based on examining multi-features.

2- Extract new features that can improve the prediction.

1.5 Related Works

Through this section, the most important previous research is exposed regarding one or more research variables.

In general, the subject of vandalism has received wide attention from researchers due to its importance in maintaining the integrity and reliability of the webs and services supplied via the Internet , as well as its relationship to cybersecurity . In crowdsourcing models - like Wikipedia - vandalism is regarded as one of the most complex tasks, as each group of academics focuses on an aspect of this task and the ways to address it. This complexity comes from the fact that any user, who can add any information in any way, can modify the encyclopedia.

Many types of research are concerned with crowdsourcing models and the problem of detecting vandalism in crowdsourcing models such as the following researches:

- Tran & Christen (2013) [27] a new set of text-based functionalities was suggested using Wikipedia data dump; the authors examined the variations between modifications designated as vandalism corrections and revisions connected with vandalism. In addition, the authors employ a multi-language technique to investigate the similarity between bots' results and humans in identifying vandalism and demonstrate that the training classifier on one language may perform well when used with another.

- Götze (2014) [28] explores the influence of resampling the training dataset (PAN-WVC-10 and PAN- WVC-11 in) on classification performance using Random Forest and then considers the Wikipedia vandalism detection challenge as a one-class classification issue by assessing classifiers from the literature. The author introduces the concept of category-relative classifiers and Investigates the vandalized-content that may be similar in different articles, also the category matching of these article, by constructing basic vandalism datasets using the whole revision history of Wikipedia articles and assessing the derived article-relative classifiers. To get a greater classification performance, multiple article-relative classifiers combined.

- Tran et al. (2015) [29] suggests a context-aware concept that considers word dependencies by focusing on a specific type of hidden vandalism in which vandals make sophisticated text modifications that alter the meaning of a sentence without noticeable vandalism markers, to identify sentences that are prone to vandalized. This technique's tag class of words in sentences changed in each revision using a part-of-speech (POS) tagger and for modeling dependencies between tags a conditional random field (CRF) is used. On the same data sets, created a collection of text features and by using a random forest algorithm, vandalism articles were identified. The findings demonstrate that context-aware strategies can supplement existing feature engineering-based approaches as a new tool for anti-vandalism Wikipedia.

- Susuri et al. (2016) [30] recommends identifying vandalism in a monolingual situation using a variety of classifiers (Decision Tree (DT) ,Gradient Tree Boosting (GTB) , Nearest Neighbor (NN) and Random Forest (RF)) . Evaluating their effectiveness when applied across languages using two data sets: the largely unexplored hourly count of views of each Wikipedia article, and the widely utilized edit history of articles. These findings indicate that vandals' view and edit behavior is consistent across languages. As a result of this finding, vandalism models may be developed in one language and then used in another.

- Shulhan & Widyantoro (2016) [31] proposes a method for recognizing vandalism on English Wikipedia employing machine learning algorithms by building a Cascaded Random Forest classifier on a resampled (PAN-WVC-10) English dataset using (LNSMOTE). These two strategies were then compared to Random Forest as a classifier and the Synthetic Minority Oversampling Technique (SMOTE) as a resampler. The results of classifiers tested on the (PAN-WVC-11) English dataset revealed that -in both classifiers- resampling the dataset with LNSMOTE increased the true-positive rate greater than SMOTE .

- Susuri et al. (2017) [32] looked at two Wikipedia language editions: Simple English and Albanian, and found no differences in the results of these classifiers aside from the fact that they had to under-sample the non-vandalism observations to match the number of vandalism observations in the dataset. The findings show that vandals' views and edit behavior are comparable across languages. As a result of this finding, vandalism models were trained in one language and then applied to others.

- Heindorf (2019) [33] designs unique vandalism detectors based on machine learning to save manual review time. To that end, a large-scale vandalism dataset developed, a high predictive-low bias vandalism detector against specific editors, and evaluated in a variety of parameters, trying to compare them to the state-of-art indicated by the Wikimedia Foundation's Objective Modification Evaluation Service and Wiki-data Abuse Filter. This machine learning approach allocates a vandalism mark to each edit made instantly. Letting prompt action against vandalism in multiple procedures of operation: edits with top marks automatically changed back; medium marks edits manually reviewed based on their marks; low marks edits not reviewed at all; and computed 47 features to discover vandalism, trying to consider both content-context data into consideration. For detecting vandalism, bagging and random forests are used as well as multiple-instance learning.

- Martinez-Rico et al. (2019) [6] states that three levels of new characteristics are de-rived from various approaches employed to investigate the usability of leading technology such as deep learning for vandalism detection. While the first set of features was gained by developing a vocabulary of vandals using current semantic-similarity ties in word embedding and DNN, the next set of characteristics, specifically stacked denoising autoencoders (SDA), was gained by using DL techniques to minimize the number of parameters of a BOW model derived from a set of Wikipedia edits. The third set employs graph-based ranking techniques to build a list of vandalism phrases from a Wikipedia vandalism corpus. The above sets of novel features were tested independently and in combination to determine their complementarity; thus, enhancing the state-of-the-art outcomes.

- Santos et al.(2019) [34] developed a methodology integrating (LMT with K-Means) plus taxonomy exploration, using a dataset from Wikipedia's vandalism edition, which comprises data regarding vandalized Wikipedia article edits. The framework provided in this research is a first step in detecting vandalism, and it gives pertinent information about the issues (features) of the action that resulted in vandalism. More research is required to determine how our technology would improve interactions in an online community. When classifying data depicting vandalism, the trials reveal that the ML model has a precision of 78.1% and a recall of 63.8%. The aforementioned model was utilized in this study to detect norm violations and provide the perpetrator with information about the aspects of their activity that make it a violation.

Table 1.1: Samples of the results obtained, corpora and classifiers utilized for training from literature.

No.	Author – Year	Corpus	Classifier	Classification Result			
				Recall	**F1**	**Precision**	**AUC**
1	Tran & Christen 2013 [39]	Dump History Wikipedia	Gradient Tree Boosting	87%	87%	87%	-
2	Götze 2014[28]	PAN-WVC-10	Random Forest	60%	-	60%	-
		PAN-WVC-11		39%	-	92%	-
3	Tran et al.2015[29]	PAN-WVC-11	Random Forest	81%	-	81%	-
4	Susuri et al.2016[30]	Dump History Wikipedia	Decision Tree - Gradient Tree Boosting Nearest Neighbor - Random Forest	83% - 83% 69% - 84%	-	84% - 83% 69% -83%	-
5	Heindorf 2019[33]	WDVC-2015	Random Forest	97%	-	40%	-
6	Martinez-Rico et al. 2019[6]	PAN-WVC-10	Random Forest	96%	79%	85%	-
7	Santos et al. 2021[34]	PAN-WVC-10	Logistic Model Trees + K-Means	63.8%	-	78.1%	-

1.6 Thesis Organization

After Chapter 1, which presents an introduction to the entire thesis, and literature review. The rest of the thesis is structured as the following:
Chapter 2: presents a comprehensive description of the main principles, which are used later in this thesis.
Chapter 3: clarifies the main steps of designing a Vandalism Detection Model using deep neural networks.
Chapter 4: shows the employment of the proposed system and experimental results of this implementation as well as each step.
Chapter 5: conclusions of this thesis and gives suggestions for future work are shown in this chapter.

CHAPTER TWO

THEORITACAL BACKGROUND

2.1 Overview

Since the establishment of Wikipedia in 2001, more than 56 million articles have been generated by millions of contributors . People use the most popular online encyclopedia for a wide range of purposes including research, background knowledge, learning about areas of interest for school or job, and fact-checking[40-41]. Wikipedia has a distinctive feature in which people can freely edit and publish articles. Despite the benefits of this feature, which contribute to the vast expansion of the encyclopedia, it is the main cause of vandalism[6].

According to the Merriam-Webster dictionary, "vandalism" is defined as the intentional or malicious damage or defacement of private or public property. Within Wikipedia, vandalism occurs in the form of a deliberate edit with the intent of providing incorrect information or hiding information by deleting content, abusive language, advertisements, and/or irrelevant text. Identifying and correcting the defaced article can distract the editor from developing or extending new articles or other vital activities, so the reader will receive no or incorrect information due to the deletion[35,18].

The user community is responsible for the first attempts of identifying vandalism manually by reporting the presence of vandalism, these attempts results in the creation of several bots. These bots analyze newly created revisions, implement hand-crafted rule sets, and identify instances of vandalism. The use of a wide variety of statistical and machine-learning techniques has contributed to the extremely complicated nature of these approaches as time has passed[17].

The community that acts as a knowledge base would greatly benefit from an automated solution to assist in reviewing such things as vandalism [38,44].

Deep Learning is regarded as one of the most recent advances in machine learning. It assisted researchers in developing solutions that could only be imagined a decade ago, due to a lack of data management and data processing capacity [41-42]. Convolutional Neural Networks (CNNs) have grown as the de facto paradigm for a wide range of Computer Vision and ML tasks [43-44]. When the NN's depth is deep enough, the CNN is considered universal. This implies that it may be used to approximate any continuous function to arbitrary accuracy [45] . The key advantage of CNNs seems to be that they combine the feature extraction and classification operations into a single machine-learning body that can be adjusted concurrently to enhance classification performance. This eliminates the need for hand-crafted elements or other post-processing [46-47].

Recent research has shown that, with a suitable systematic approach, compact 1D CNNs can outperform all old and conventional techniques. [43] . The primary benefits of the 1D CNN classifier are its low-complexity architecture and practical, cost-effective real-time hardware implementation[48-49] . The Deep Learning Approach has been successful in many topics [50] close to or even similar to the topic of detecting vandalism in Wikipedia, such as Wikipedia Vandal [51] ,Fake News detection [52], Anomaly Detection [53] , one-class classification problems [54], and Cyberbullying [55].

2.2 Introduction to Crowdsourcing

With Web 2.0, users can create their own content, use it in different ways, be a part of the culture, and connect with other users [3].

Online communities are fantastic places to look for new ideas and genius, and they are one of the most amazing ways to re-engineer relationships between users and organizations [56] .Crowdsourcing is a term that refers to an online, distributed problem-solving and production model with a growing number of interested business parties. It is based on the open-world approach, in which peers connect and work without being organized according to a managerial or hierarchical structure. Thousands of people add their own unique perspectives to a body of knowledge, which is the basis of our information and knowledge environment [3].

Jeff Howe flamed it through his article entitled "The Rise of Crowdsourcing" in 2006 for Wired Magazine. Howe defines crowdsourcing as "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call" [57].

In 2012, Estellés-Arolas and González-Ladrón-de-Guevara did a literature review to find out what made crowdsourcing different from other Internet projects. They found that crowdsourcing is defined in different ways.

These definitions contain eight key features that each crowdsourcing endeavor must have, according to the authors. The following are the components:

1. There is a distinct group of people.
2. There is a task with a clear objective.

3. The crowd's remuneration is unmistakable.
4. The crowdsourcer is identified explicitly.
5. The payment that the crowdsourcer will receive is established.
6. It's a participatory online process that's been allocated to you.
7. It employs a variable-length open call.
8. It takes advantage of the Internet.

"Crowdsourcing is a sort of participatory online activity in which an individual, an institution, a nonprofit organization, or a firm proposes the voluntary undertaking of a work to a group of individuals of different knowledge, heterogeneity, and number, via a flexible open call" [59].

The completion of a task of varying complexity and modularity to which the crowd is expected to contribute by contributing their time, money, knowledge, and/or experience always results in mutual gain.

"The user will be satisfied with a specific type of need, such as financial, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and use to their advantage what the user has brought to the venture, which will take different forms depending on the type of activity undertaken" [58].

From the aforementioned components, the general approach to crowdsourcing is illustrated as follows in Figure (2.1) [59].

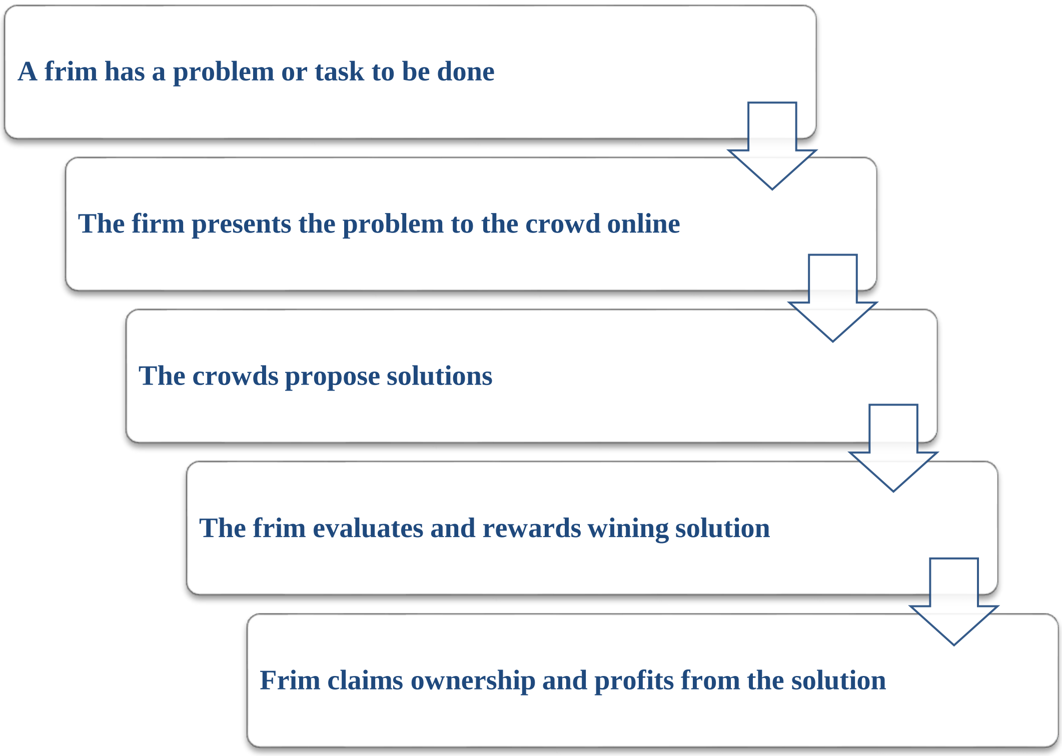

A frim has a problem or task to be done
The firm presents the problem to the crowd online
The crowds propose solutions
The frim evaluates and rewards wining solution
Frim claims ownership and profits from the solution

Figure 2.1 : The general approach of crowdsourcing [59].

2.3 Crowdsourcing Typologies

Estellés-Arolas and González-Ladrón-de-Guevara proposed an integrated typology divided into five categories [58] :

1- Crowdcasting is a method of distributing information to a large number of crowdsourcing projects that are similar to contests, in which a problem or task is presented to the crowd, and the person who solves it first or improves it is paid (i.e., the "incentive").

2- Collaborative crowdsourcing: Crowdsourcing efforts involve communication among members of a crowd while the initiative's creator remains on the sidelines. Two types of subtypes differ in their final goal:

- Brainstorming with a large group of people Massive online brainstorming sessions in which a variety of ideas are presented and the community can support them with comments and votes (e.g., IdeaJam)
- Crowdfunding: Customers address the doubts and difficulties of other customers in this instance, so they don't need to contact professional customer care (i.e., get satisfaction).

3- Crowdcontent: In these crowdsourcing jobs, people pool their labor and knowledge to create or find various sorts of content, but not in a competitive manner. There are three types of subtypes:

- Crowdsourcing: initiatives in which the audience should produce material, such as when translating short amounts of text or labeling photographs individually (i.e., Amazon Mechanical Turk).
- Crowdsourcing is a method of gathering information from a large number of people. Crowdsourcing projects are where the general public searches the internet for content for any reason (i.e., peer-to-peer patent review).
- Analyzing crowds. Initiatives in which the public searches for multimedia documents like films or photographs rather than the Internet (e.g., Stardust@home)

4- Crowdfunding is a fourth option. An individual, organization, or enterprise seeks cash from the crowd in exchange for a reward in a crowdfunding campaign (i.e., Kickstarter).

5- Public opinion. The goal in this scenario is to learn about user opinions on a specific problem or product through votes, comments, tags, or even the sale of shares (i.e., ModCloth, Intrade)[58].

2.4 Crowdsourcing Models

Crowdsourcing models can be created on either public or private platforms. Several firms have created such applications during the last year's popularity. Examples include Mechanical Turk, book review websites, YouTube, Flickr, Quora, Wikipedia, and Demand Media [60].

Wikipedia ranks among the most popular websites on the internet (and, with nearly half a trillion page views each year), is a remarkable source of data today) [61]. Wikipedia is a crowdsourcing model that has all the characteristics of crowdsourcing platforms.

These features helped in the widespread and significant growth of Wikipedia, which includes being free because anybody can use, copy, share, and alter Wikipedia articles, including for commercial reasons, as long as the end result is similarly licensed under the same terms. Collaborative since Wikipedia materials are developed via the cooperation of thousands of users; anybody, even if not registered, can edit Wikipedia and contribute to discussions regarding content and policy. Wikipedia is multilingual, as it has versions in 240 languages and is still growing. Global-scale because of its ten years of existence [62].

The English Wikipedia is considered the most comprehensive and well-organized. Wikipedia has almost 4 million articles about various topics. Furthermore, the form of the article requires that the headline be a name or a description of the issue, according to the Wikipedia approach.

And it covers a wide range of topics, including arts, geography, politics, technology, athletics, and games. Wikipedia's features as a database for knowledge extraction are not limited to scale but also include a rich layout, sense clarification based on URLs, concise link texts, and well-organized content [37].

The five Wikipedia pillars commonly characterize Wikipedia’s essential concepts [63] .

These five pillars are defined as follows by Wikipedia's Five Pillars [64]:

1. **"Wikipedia is an encyclopedia":** It is not appropriate for advertisement, news, or primary research activities.

2. **"Wikipedia is written from a neutral point of view":** Individual opinions and experiences must be ignored, and the majority's various viewpoints must be presented.

3. **"Wikipedia is free content that anyone can use, edit, and distribute."** Contributions do not require a user account.

4. **"Wikipedia's editors should treat each other with respect and civility":** Users are not permitted to use vulgarity or abuse.

5. **"Wikipedia has no firm rules":** There are no set rules for modifying an article. As a result of this lack of information, new users frequently make mistakes [28,64].

Additional principles, however, often take priority in practice, expanding on the pillar "Wikipedia is free content that anyone may use, modify, and share" to include various sorts of "openness" and "freedom."

Furthermore, the community has prioritized references and attribution for knowledge as a means of checking for plagiarism and copyright laws, as well as a defense mechanism against criticism of the integrity of Wikimedia content [63].

The revision history that is recorded for each article on Wikipedia is a unique feature that permits keeping track of every modification ever made to the encyclopedia. At the same time, the revision history is what causes Wikipedia to have a lot of information, which makes it hard to understand it all [65].

2.5 Cyber Security

All big data-type applications require high-level software and Information Communication Technology (ICT) skills; the combination of the pieces of information obtained by data mining necessitates high-level software and ICT skills, and all these applications demand a high level of cyber protection [14].

Cybersecurity refers to a wide range of policies, methods, technologies, and procedures that work to secure the confidentiality, integrity, and availability of computational resources, networks, software, and data against threats. With AI-enhanced evasive cyber attackers, cyberspace is evolving into a place where there is no trust since any device or program might potentially be used for malicious purposes [66] . It can be broadly categorized into two types of crimes: ones that explicitly harm computer systems or devices, such as botnets, viruses, or DOS attacks, and those that are made possible by computer networks or devices but have a primary target, such as identity fraud, phishing scams, information warfare, or cyberstalking.

The term "cybercrimes" refers to a wide spectrum of prohibited behavior, which may include "cyber vandalism," which is the act of damaging or destroying data as opposed to stealing or abusing it [15].

2.6 Vandalism

Reliance is increasing day by day on crowds to gather knowledge, and this increase gives the possibility for many users to modify and change the content, allowing some modifiers to sabotage and destroy the knowledge base [67]. Vandalism is considered one of the five layers of the model built for cyber threats known as "cyber vandalism." Instead of misusing or stealing data, cyber vandalism involves damaging or destroying it [68] .

According to Wikipedia, vandalism is defined as "an intentional addition, deletion, or alteration of content that directly threatens the encyclopedia's integrity" [33] . Because Wikipedia is a society that may reflect our community and culture, the frequency, quality, and intent behind all of these vandalisms might have cultural consequences [69] . Vandalism demonstrated heterogeneous characteristics [70] , and evidence that this issue cannot be handled solely through human inspection: as the information in Wikipedia grows by the day, the effort required to manually review them will outgrow the community's resources. Reviewing millions of articles each month places significant pressure on the knowledge base community [67].

Therefore, many approaches have emerged, whether machine learning or bots, to identify and detect the problem of vandalism, and this makes solving the problem faster as well as saving time and effort for volunteers to focus their efforts on adding new high-quality content instead of wasting efforts on other issues [67]. Also, Wikipedia has implemented a protection page policy that restricts a page from general editing to avoid further vandalism or edit wars [39].

From all this, we define vandalism as "the attempt to modify the article in a way that negatively affects the article's quality."

2.6.1 Vandalism Types

The dynamic nature of Wikipedia encourages many various types of vandals to make harmful alterations, such as manipulating facts, including vulgarity, or deleting information [29]. Users who regularly vandalize Wikipedia despite warnings are flagged to administrators and prevented from making future modifications [71] .

It is critical to classify and recognize vandalism in order to learn how to deal with it through a variety of methods, such as securing the page, modifying system settings, or blocking it [72] . Figure (2.2) illustrates the classification of vandalism types.

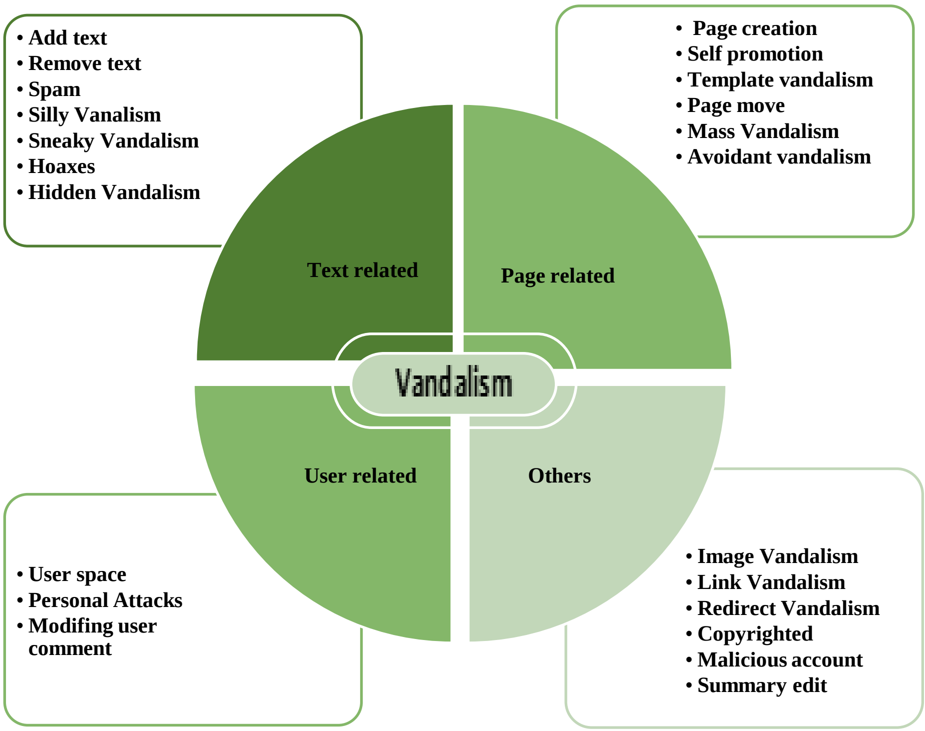


Figure 2.2 : Vandalism Types Classification (Owned).

Returning to Figure (2.2), the section "Others" consists of a set of vandalism types, including those connected to image vandalism in Wikipedia, which is done by uploading inappropriately shocking and explicit images inappropriately on Wikipedia. Whereas "link and redirect vandalism" refers to adding, removing, modifying, and redirecting links in Wikipedia for offending, such as by adding spam links or changing facts, such as by adding external links to irrelevant sites[72].

Creating malicious accounts with intentionally offensive or disruptive nicknames is considered vandalism, whether the account is active or not. Edit summary: Vandalism is frequently associated with malicious account creation, to leave a mark on the record that cannot be effectively removed. Only a few editors with special capabilities above administrators can change edit summaries. "Copyrighted" means material is repeatedly uploaded, even after being warned, by vandals intentionally uploading or utilizing material on Wikipedia in a way that violates Wikipedia's copyright guidelines [72].

"User-related" vandalism refers to vandalism that occurs on user pages or user talk pages, or that alters user behavior, such as modifying or deleting other users' comments from a conversation. Sometimes changing a user's vote also included were the deletion of recent warnings from the user discussion page and the addition of insults, vulgarity, and the like to user pages or user talk pages [72].

Bullying, harassing, threatening, and making harsh or insulting statements about other users are examples of "personal attacks." Adding pages or creating content for existing pages that only serves to degrade the subject is likewise prohibited [72].

The other part of "page-related" vandalism is composed of "page creation vandalism," which refers to making new pages for the sole purpose of malicious activity, such as advertising sites, personal attack webpages, and pages offering nonsense. Self-promotion vandalism is also about creating a page but about oneself, family members, or businesses and companies with which one is personally connected [72].

Mass vandalism and template vandalism involve having a negative impact on many pages by damaging or disrupting a template or by running a Vandalbot that aims to vandalize or spam a large number of articles [72].

The removal of tags such as {afd} and {copyvio} to conceal deletion candidates or avoid deletion of such content actually increases the likelihood that the article will be removed, and the removal of recent warnings of vandalism or other significant infractions from the user's talk page is considered to be avoidant vandalism [72].

The "text-related" area is an essential part of the vandalism detection technique since understanding the type of vandalism aids in selecting the best system features [72].

Text vandalism is the intentional addition of text that is unrelated to the topic of the article. Text removal is about deleting all or significant sections of an article's content for no apparent reason or replacing all or portions of pages with text that does not belong[72].

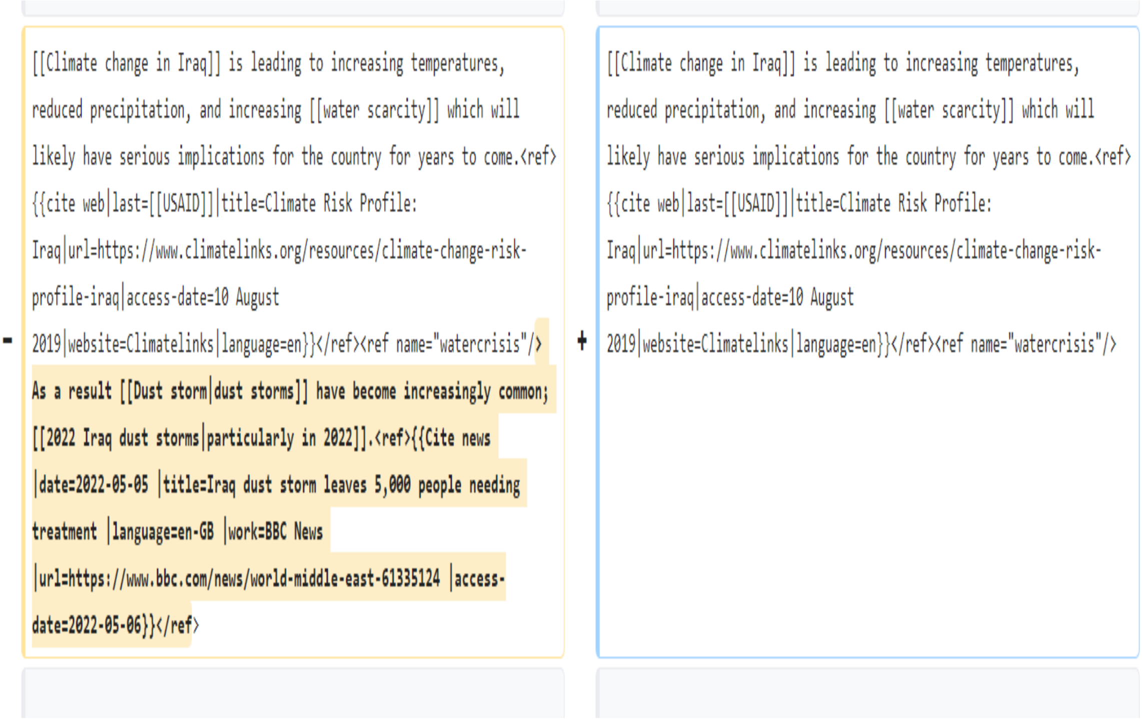
In spam vandalism, the vandal adds text to any page that supports an interest that serves the user, as well as adds external links to the site(s) that promote an interest that benefits the user, even if the site contains material relevant to the article. Silly vandalism Adding nonsense, random characters, and humor to article space intentionally [72].

Sneaky vandalism that is difficult to detect or avoids detection, such as minor bad-faith information modification that goes unnoticed, results in a bad update followed by a good one. Because some watch lists only show the most recent modification to an article, the poor edit may go unreported [72].

Creating hoaxes involves combining believable misinformation into publications and employing fake references. And hidden vandalism is vandalism that is not apparent in the finished rendering of the article but is visible while editing [72].

2.6.2 Examples of Vandalism on Wikipedia

According to Wikipedia, the English-language Iraq page is one of the most vandalized pages within the continent of Asia [73] . Figure (2.3) shows an example of vandalism in which, on the Iraq page, specifically in the Climate Change section, news information was added to the article with the addition of links to news sites as a resource, which contradicts Wikipedia's policy, that "Wikipedia is an encyclopedia": It is not appropriate for advertisement, news, or primary research activities.

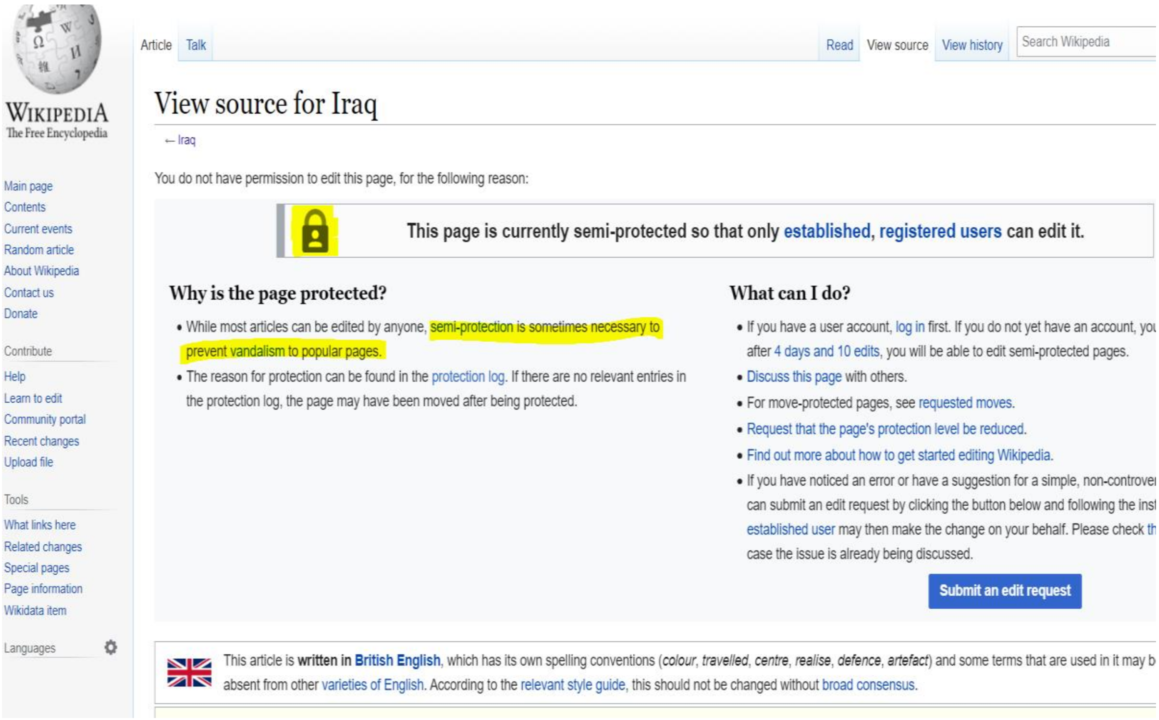


```
- [[Climate change in Iraq]] is leading to increasing temperatures, reduced precipitation, and increasing [[water scarcity]] which will likely have serious implications for the country for years to come.<ref>{{cite web|last=[[USAID]]|title=Climate Risk Profile: Iraq|url=https://www.climatelinks.org/resources/climate-change-risk-profile-iraq|access-date=10 August 2019|website=Climatelinks|language=en}}</ref><ref name="watercrisis"/> As a result [[Dust storm|dust storms]] have become increasingly common; [[2022 Iraq dust storms|particularly in 2022]].<ref>{{Cite news |date=2022-05-05 |title=Iraq dust storm leaves 5,000 people needing treatment |language=en-GB |work=BBC News |url=https://www.bbc.com/news/world-middle-east-61335124 |access-date=2022-05-06}}</ref>
```

```
+ [[Climate change in Iraq]] is leading to increasing temperatures, reduced precipitation, and increasing [[water scarcity]] which will likely have serious implications for the country for years to come.<ref>{{cite web|last=[[USAID]]|title=Climate Risk Profile: Iraq|url=https://www.climatelinks.org/resources/climate-change-risk-profile-iraq|access-date=10 August 2019|website=Climatelinks|language=en}}</ref><ref name="watercrisis"/>
```

Figure 2.3 : Example of vandalism: adding news as information to the encyclopaedia.

In Figure (2.4), Wikipedia provides a "protected page policy," which restricts a page from general editing to avoid further vandalism or edit-wars.



View source for Iraq

← Iraq

You do not have permission to edit this page, for the following reason:

This page is currently semi-protected so that only established, registered users can edit it.

Why is the page protected?

- While most articles can be edited by anyone, semi-protection is sometimes necessary to prevent vandalism to popular pages.
- The reason for protection can be found in the protection log. If there are no relevant entries in the protection log, the page may have been moved after being protected.

What can I do?

- If you have a user account, log in first. If you do not yet have an account, you after 4 days and 10 edits, you will be able to edit semi-protected pages.
- Discuss this page with others.
- For move-protected pages, see requested moves.
- Request that the page's protection level be reduced.
- Find out more about how to get started editing Wikipedia.
- If you have noticed an error or have a suggestion for a simple, non-controver can submit an edit request by clicking the button below and following the inst established user may then make the change on your behalf. Please check th case the issue is already being discussed.

This article is **written in British English**, which has its own spelling conventions (*colour*, *travelled*, *centre*, *realise*, *defence*, *artefact*) and some terms that are used in it may b absent from other varieties of English. According to the relevant style guide, this should not be changed without broad consensus.

Figure 2.4 : Wikipedia protection page policy.

A few instances of vandalism that started in a knowledge base and moved to information systems are as follows:

- Apple’s Siri referred to Angela Merkel, the chancellor, as a pig.
- According to Yahoo's expert panels, Barack Obama is the creator of ISIS.

All of these cases were triggered by vandalism in the underlying information bases and resulted in negative publicity for the company [33].

2.7 Vandalism Detection Process

In Crowdsourcing, vandalism detection has been broadly defined as a binary machine-learning problem with the goal of classifying a pair of related revisions as vandalized or regular based on edit category features

[74] . The general process of detecting vandalism in a collaborative community shown in Figure (2.5).

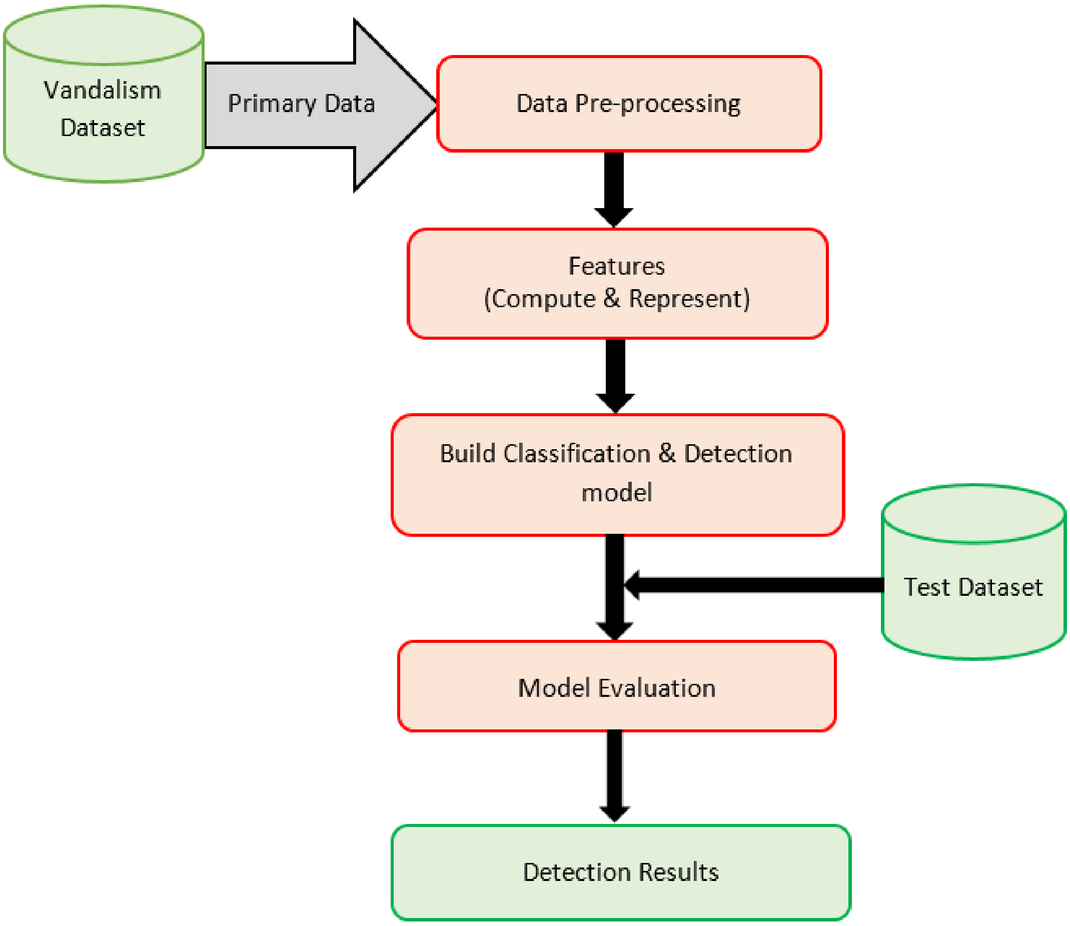


Figure 2.5 : Vandalism Detection Process.

2.7.1 Datasets

Various datasets were often utilized for the evaluation of automatic vandalism detection algorithms; the most essential ones are outlined in this section.

- Webis-WVC-07: The first standardized test collection for comparing vandalism detection algorithms is the Webis Wikipedia

Vandalism Corpus (Webis-WVC-07), used first by [11] . It has 940 revisions, of which 301 have been flagged as vandalism by human reviewers; this corpus is no longer valid because it has been upgraded to the below-listed corpus [75].

- PAN-WVC-10: The corpus contains 32452 revisions on 28468 Wikipedia articles, 2391 of which are vandalism edits. The corpus was annotated using Amazon's Mechanical Turk; 753 people were hired and cast over 150000 votes on the revisions, ensuring that each update was reviewed by at least three annotators [76].

- PAN-WVC-11: The corpus contains 29949 revisions on 24351 Wikipedia articles, 2813 of which are vandalism edits. 9985 English edits, 9990 German edits, and 9974 Spanish edits are included in the corpus. The corpus was crowdsourced using Amazon's Mechanical Turk; each edit was shown to a group of annotators, who were asked to evaluate whether it was vandalism or not, and the annotators' agreement was analyzed to label an edit. This corpus adds to the PAN-WVC-10, which only contains English revisions. To obtain more representative findings, both samples may be used [77].

- Dump History of Wikipedia: All modifications of all articles and other Wikipedia pages are recorded and released as XML or SQL dump files by Wikipedia, which offers them for free to interested users [78].

2.7.2 Features

Three feature categories are considered: textual-language, metadata, and reputation (user) features. Textual features (e.g. DIGIT_RATIO, BIASED_WORDS) are computed by examining the new edit's annotation text, or rather both edit's annotation texts, of a revision, although they may need text processing methods of various sophistications, while language features are frequently simple to compute for individual languages but must be adapted to be transferable [28,79].

Metadata features (e.g., timestamp) are straightforwardly generated from fundamental edit attributes or obtained from the analysis of extra Wikipedia material, such as database processing . The user category (e.g., an individual editor) includes user-related metadata features derived depending on the user's editing behavior. Such values that examine the past conduct of an entity engaged in the edit have a significant computational cost since huge sections of Wikipedia’s history must be analyzed [28,79].

Some research has examined the possibility of adding a category of new features using deep learning techniques and combining them with other categories, such features as state-of-the-art characteristics that a vandalism detection system may use to identify the existence of vandalism on Wikipedia, moreover additional characteristics based on several ranking algorithms applied to "a co-occurrence graph". It is made up of linked terms that appear in the same areas of words as the considered type of article (vandal or regular) [6].

Generally, features can be extracted from: (1) the revision itself; (2) a comparison of the revision against another revision (i.e., a diff); or (3) information generated from earlier or subsequent revisions [17,34] . Taxonomy described space features with examples in Figure (2.6).

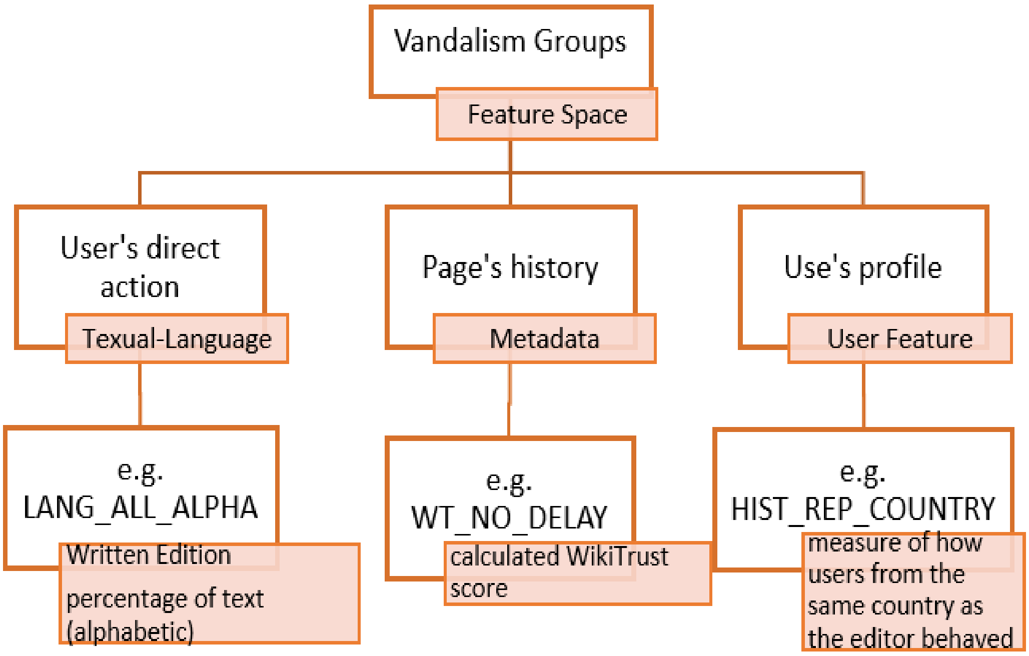


Figure 2.6: Vandalism Feature Space.

2.8 Deep learning

Deep learning (DL) is an expanded area of machine learning (ML) that has been demonstrated to be extremely beneficial in the fields of text, image, and speech recognition. Deep learning techniques are a collection of algorithms that have similarities to the interaction between stimuli and neurons in the human brain. DL offers a wide range of applications in machine vision, language understanding, speech synthesis, image production, and other fields. These methods are simple enough to learn in both supervised and unsupervised environments [80].

DL algorithms are relied essentially on the notion of artificial neural networks ANN. The availability of rich data and suitable processing power has made training such algorithms easy in today's society. As more data is collected, the performance of deep learning models continues to improve. A better representation of this can be seen in Figure (2.7) [80].

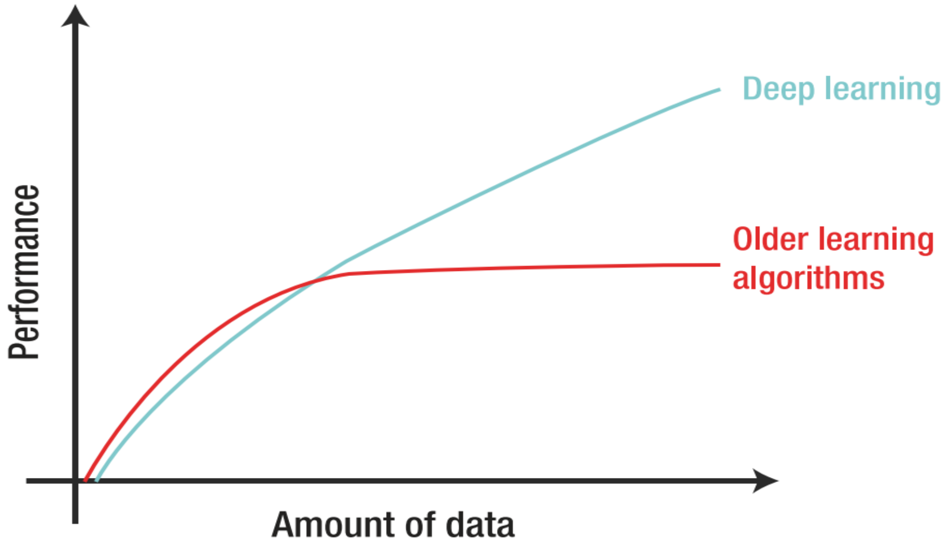


Figure 2.7: Data science approaches are scaled according to the amount of data [83].

The difference between a deep and a shallow network, and why the term "deep learning" gained currency are represented in Figure (2.8) [80].

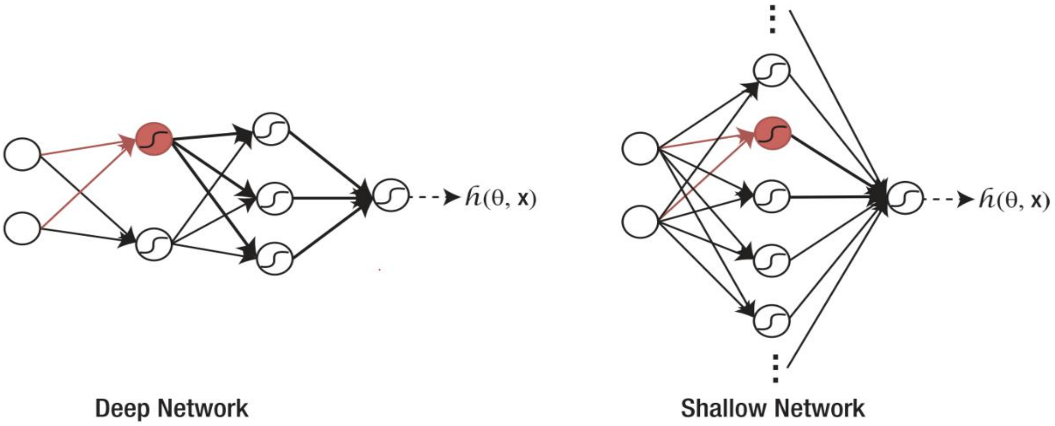


Figure 2.8: Representation of deep and shallow networks [83].

Neurons in a neural network often work together to form a layer. Multiple layers are then joined to form the network. Deep Neural Networks (DNN) are neural networks with hidden layers that are deep and rich. Hidden layers are extra layers added to a network to add additional processing when the task is too demanding for a small network. The number of hidden layers will be in the hundreds. DNN is noted for its creativity and precision. There are various types of DNN, many of which are alerted to function on image data and many of which are text data sources[81] .

Deep learning was introduced to reduce the increased number of parameters in the multilayer perceptron and its complicated architecture made it complex to use[52].

2.9 Deep Convolution Neural Networks (DCNN)

Convolutional Neural Networks (CNNs) are a kind of ML which uses DL and relied on a subset of the feed-forward NN . The CNN concept was motivated by the discovery of a visual mechanism in the brain known as the visual cortex. The visual cortex is made up of many cells, which are known as receptive fields and are responsible for determining light in small, overlapping, and sub-regions of the visual field. These cells represent local filters over the input space, while the more complicated cells present broader receptive fields. CNNs are a potential method for processing multimedia data such as photos, videos, voice, or audio, and can be used for image recognition and categorization.

CNN can be used in the same areas with different dimensions depending on the application, such as employing one-dimensional 1D-

CNN with speech processing, which deals with just one-dimensional (1D) represent temporal axis convolved to produce 1D features (temporal feature). Two- Dimension CNN is used in image processing, with the 2D kernel representing spatial characteristics across the image.

Ultimately, three-dimensional CNN is used in the video. To express spatial features over time, processing and 3D kernels convolve with spatial information over image sequences (video).

CNN performs two roles, one for feature extraction and the other for feature categorization. Each function employs some layers that are used to accomplish a given task. Convolution, nonlinear (activation), pooling, and fully connected (dense) layers in addition to some generalization layers. These layers are stacked for each function as following:

- **Feature extraction function:** CNN's feature extractor composed of one or more convolution layers, each following by a non-linear layer that indicates the activation function used in that layer to differentiate the special signal of usable features on each hidden layer. The features maps of the input sample are retrieved and produced by the (convolution layer CL) . In other words, the CL acts on the input data and the filter kernel coefficients produced during the training phase as if it were a local filter. The CL creates low-level features in the input data as a set of primitive patterns. The following CL can mix these core features and secondary features to create the higher level feature patterns. A trigger function is utilized in both general neural networks and CNNs to distinguish the special signal of valuable characteristics on each hidden layer. Numerous functions, including 'Rectified Linear

Units (ReLUs) and continuous trigger (non-linear) functions,' can be implemented with CNN as activation functions or non-linear layers.

By employing a subsampling (pooling) layer, the CNNs model becomes more robust in the presence of noise and blurring. This layer is responsible for decreasing the resolution of the features of an input data. This reduction in feature resolution is achieved by combining the outputs of a neuron, which are clusters in one layer, into a single neuron in the next layer using one of the pooling functions such as MAX pooling, MIN pooling, or AVERAGE pooling.

- **Classification function:** Lastly, a standard NN with one or more hidden layers can be concatenated, with all of these layers referred to as dense layers or fully connected (FC) layers. The final layer of the model (also the most recent FC layer) represents the output layer that provides classification decision-making [82].

When the depth of the neural network is great enough, a DCNN is universal, which means it may be used to approximate any continuous function to arbitrary accuracy.

The key advantage of CNN over its predecessors is that it automatically recognizes significant features without the need for human assistance, making it the most widely used. CNN has three main benefits: similar representations, sparse interactions, and parameter sharing [47].

2.10 One Dimension Convolution Neural Networks (1D-CNN)

Both feature extraction and classification tasks are fused into one process that may be adjusted to optimize the classification performance.

This is the primary advantage of 1D-CNNs, which can also result in low computational complexity because the sole significant operation is a sequence of 1D convolutions, which are just linear weighted summing of two 1D arrays.

Such a linear operation during the Forward and Back-Propagation operations can effectively be executed in parallel. This is also an adaptive implementation because the CNN topology allows for alterations in the input layer dimension, allowing the output CNN layer's sub-sampling factor to be changed adaptively.

1D CNNs are considered a modified version of the recently developed 2D CNNs. Many studies have shown that 1D CNNs are advantageous and thus preferable to their counterparts in dealing with 1D signals in certain applications for the following reasons [43] :

- The computational complexity of 1D differs significantly from that of their counterparts. That is, when the same design, network, and hyperparameters are used, the computational complexity of a 1D CNN is much lower than others.

- In general, especially in recent studies, most 1D CNN applications have used compact (with 1-2 hidden CNN layers) configurations with networks with 10 K parameters. Networks with shallow architectures are much easier to train and implement.

- 1D-CNN is thought to be a hardware generalization network. This implies that any CPU implementation on a standard computer is feasible and relatively fast for training compact 1D CNNs with few hidden layers (e.g., 2 or less) and neurons (e.g., less than 50).

- Due to their low computational requirements, compact 1D CNNs are well-suited for real-time and low-cost applications[43].

The architecture of a general 1D-CNN as shown in figure(2.9) below [83].

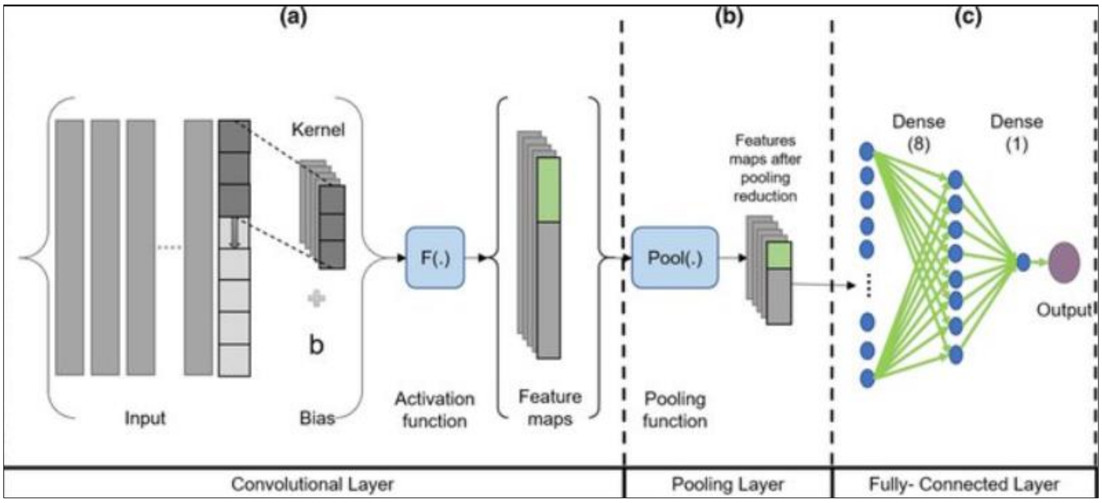


Figure 2.9 : 1D-CNN general architecture[83].

2.11 Missing Value

Missing values are typically attributed to: human mistakes in data processing, machine error owing to malfunctioning equipment, respondents' refusal to answer specific questions, drop-out from studies, and merging unrelated data. The missing values problem is ubiquitous in all domains that deal with data and produces a variety of concerns such as performance degradation, data analysis issues, and biased results due to differences in missing and full numbers.

Furthermore, the severity of missing values is determined in part by the amount of missing data, the pattern of missing data, and the process causing the missingness of the data [84].

Missing values are handled in two ways :

2.11.1 Delete Rows with Missing Values

During performing analysis, any entries with missing values are removed/discarded. Because there is no need to evaluate value, deletion is considered the simplest approach. However, deletion has significant drawbacks because it introduces bias in analysis, especially when missing data is not dispersed randomly[84].

if a row has many missing values then you can choose to drop the entire row[85] .

2.11.2 Interpolation

Missing values can also be imputed using interpolation. Pandas interpolate method can be used to replace the missing values with different interpolation methods like ‘polynomial’, ‘linear’, ‘quadratic’. The default method is ‘linear’ [85].

Interpolation is used to find missing values by using their neighbors. When the average does not match the missing numbers well, then it must go to a different strategy, which is most often interpolation[86].

2.12 Normalization

It is a common practice in the machine learning community to normalize the data beforehand to rely on common assumptions for which classifiers are known to behave well, such as zero-mean and unit-variance. Standard strategies often allow for significantly improved processing with statistical approaches.

Converting the dynamics to [0,1] using

$$\mathbf{I' = (I - I_{min}) / (I_{max} - I_{min})} \qquad \mathbf{(1)}$$

Where $\mathbf{I'}$ = scaled value , and $\mathbf{I_{min}}$ and $\mathbf{I_{max}}$ are the minimum and maximum values in the dataset, respectively.

This is helpful mostly for numerical optimization that rarely behaves well with very large values and relates more to the implementation than the theoretical standpoint [87].

Normalization techniques are essential for accelerating the training and improving the generalization of deep neural networks (DNNs), and have successfully been used in various applications. Normalization defined as a general transformation, which ensures that the transformed data has certain statistical properties[88].

2.13 Performance Measurement

The performance of a trained classifier is measured against test data using the detection performance metrics of precision and recall. From the errors in the confusion matrix, calculate the following measures of a classifier's confusion matrix shown in Figure(2.10):

		Actual	
		P	N
Predicted	P	*TP*	*FP*
	N	*FN*	*TN*

Figure 2.10: confusion matrix.

Where the letter P represents class 1 and the letter N represents class.

As a result, TP indicates the number of correct positive detections, FP indicates the number of incorrect negative detections, FN indicates the number of incorrect negative detections, and TN indicates the number of correct negative detections. Measures of precision and recall of the prediction error can be calculated from the data in the confusion matrix as follows:

Precision = TP / (TP+FP) (2)

Recall=TP/(TP+FN) (3)

Most classification approaches, rather than providing a definitive class label, instead provide a confidence score based on the quality of the input data. The final classification is reached by picking a threshold value that determines the class relative to the level of confidence attained. The prediction error measures for each practicable threshold can be obtained by modifying the threshold for the vandalism confidence values.

By plotting the ordered recall values against the precision values and modifying the classification threshold, one can generate the precision-recall (PR) curve. By examining the region under the curve, we can avoid the difficulty of choosing a threshold in advance.

Also, the area under the curve can be used to make statements about the effectiveness of a classifier in a given detection scenario and the appropriateness of a given threshold [28].

FP-rate = FP / (FP+TN) (4)

TP-rate = TP / (TP+FN) (5)

2.14 Evaluation the Performance

Cross-validation is a technique used for determining a ML model's performance on new data by comparing the performance of multiple ML models on a subset of that data using a single parameter called k. In other terms, the small training set is used to make an educated guess to show how well the model performs when predicting new, unseen data.

K - Folds is a data resampling technique employed to evaluate the generalization capabilities of predictive models and to avoid overfitting to quantify genuine prediction error and adjust model parameters. Cross-validation, belongs to the Monte Carlo method family [89] . Figure (2.11) shows the 10-fold cross-validation procedure for training and testing.

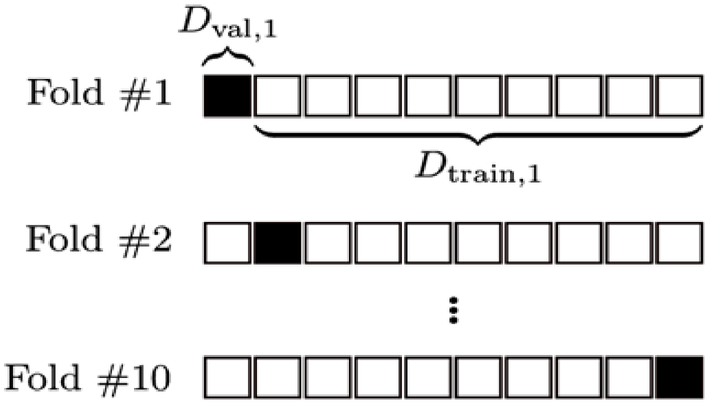


Figure 2.11: Using 10 -Fold Cross-Validation Procedure[89].

CHAPTER THREE

PROPOSED METHOD

3.1 Overview

The proposed Vandalism Detection System (VDS) is a collection of data mining, data analysis, and ML techniques to analyze and classify Wikipedia articles in order to make predictions about Vandalism status. VDS involves extracting multi data from articles then analyzing these data in order to extract features that can be used to make an accurate classification of vandalism status.

The objective of this thesis is to come with essential features that can clearly show the vandalize changes on the text data and metadata of articles from a set of training/ testing articles collected from Wikipedia with regular and vandalized states. A Proposed (VDS) monitors the Wikipedia articles continuously based on detection of multi-features changing, thereby giving the ability to make predictions about Wikipedia article state in future and alarms or remove the undesired change. In this chapter, the proposed system is implemented in two major modes: training and testing modes. Each mode contains several stages and steps. The following sections clarify the details of each stage separately.

3.2 Proposed Vandalism Detection System

The proposed system presented Vandalism Detection of Wikipedia articles based on robust multi-feature type extraction. The proposed model extracts, analyzes as well as classifies the textual and metadata with reviewers trust features of articles in the same model by employing one Dimensional Convolutional Deep Neural Networks (1D CNN) with new light architecture. The proposed system is performed two main stages: the preprocessing stage then vandalism detector stage, each stage contained several steps that performed different functions. The first stage in a proposed system is to do data preprocessing. Missing value in a dataset is a very common phenomenon in reality, yet a big problem in real-life scenarios. Missing Data can also refer to as "Not Available" values. Sometimes many datasets simply arrive with missing data, either because it exists and was not collected or it never existed. The second stage is the vandalism detection, which is performed through the 1DCNN model and consists of two stages: The salient feature extraction step and the classification step.

The general block diagram of the proposed system stages illustrated in figure (3.1) and described in the algorithm (3.1). There are additional processes implemented. The first one is preparing the dataset to be suitable for the Deep Neural Network model; which is implemented before training the proposed models, and the last one is the evaluation process implemented to evaluate the performance of the proposed system based on test data.

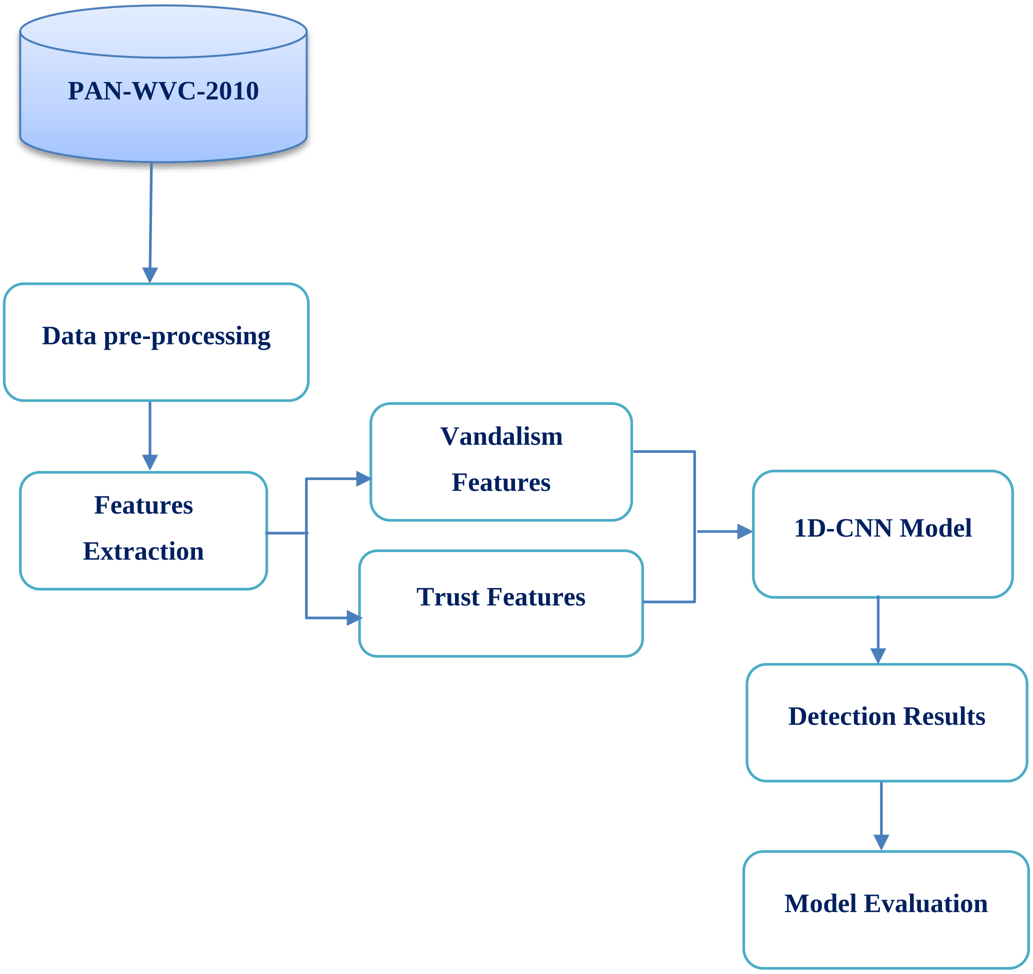


Figure 3.1: The block diagram of the proposed VDS.

Algorithm 3.1: VDS Model Algorithm.

```
Input: RD: Raw Data.
Output: Detect article class (Vandalism or Regular).
Step 1: Data Preprocessing.
        o Eliminate Missing value.
Step 2: Features Extraction:
    - Vandalism features.
    - Trust features.
    - Gathering all features.
Step 3: Detection Stage:
    - Extracting Salient Features step
        - Convolved features.
        - Select the strongest features by (sampling down).
    - Classification step.
End.
```

3.3 Preparing Dataset

The process of preparing raw data so that it is suitable for further processing and analysis. One of the most important steps in preparing the dataset is the process of dividing it and choosing the ratio between the training and testing datasets. This approach allowing finding the model hyper parameter and estimating the generalization performance. The commonly used ratio is 80:20, which means 80% of the data is for training and 20% for testing, the dataset division work of the proposed model is illustrated in figure (3.2).

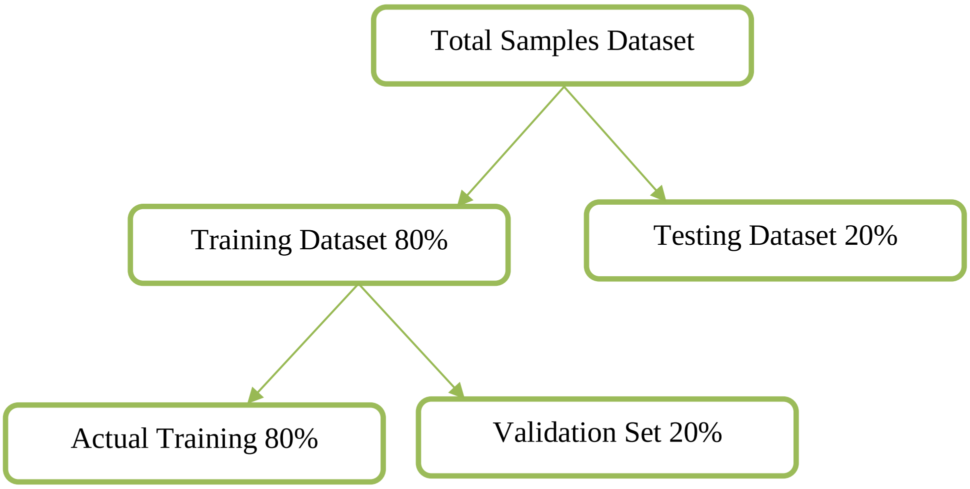


Figure 3.2 : The Dataset preparing work.

3.4 Data Preprocessing

Uncleaned data cannot be processed by most of the machine learning algorithms. The pre-processing stage performs all preparation work on the inputs to provide generalization for the system to become suitable for implementation on a proposed VDM because the actual inputs might have some unsuitable data. This stage consists of two step : handle missing value, and normalization.

3.4.1 Handle Missing Value

The presence of missing values in a dataset can affect the performance of a classifier built using that dataset as a training sample. Many datasets have missing data, either because it exists but was not collected or because it never existed. For example, Suppose during a survey:

- Different users may choose not to share all their required information.
- Human error could cause a failure in recording the values.

During data preparation, several missing values were eliminated, as were certain fields where the annotator's answers were (not sure) removed. Missing values are handled in two ways :

1- **Delete Rows with Missing Values:** one of the quick techniques could used to treat missing values. If any columns have more than half of the values as null, then the entire column can dropped. In the same way, rows can also be dropped if having more column values as null. Before using this method an important issue must be kept in mind that should not lose information. Because if the information deleted is contributing to the output value then should not use this method because this will affect the output.

The proposed system deals with a limited number of features (columns); and dropping any column that has many, missing values may be led to lose this feature. Therefore, cannot use this method with columns. On the other hand, the proposed system was implemented with a huge dataset. So deleting some rows will not make much difference and output results do not depend on the deleted data.

2- **Interpolation** : After dropping the rows that have more than half of the values as null, other missing values in the dataset are imputed using the ‘interpolation’ technique to estimate unknown data points between two known data points based on different interpolation methods like ‘polynomial’, ‘linear’, ‘quadratic’. Polynomial Interpolation is a more precise approach which used in the proposed system because linear interpolation has some chances of introducing error.

3.4.1 Normalization

Neural Networks models deal with small weight values during inputs processed. The learning process can be disrupted or slow down due to large integer values inputs. Normalization is the process of scaling data to a standard range in order to remove differences in the magnitude of values. This is important because data with different scales can lead to issues in training machine-learning models.

3.5 Features Extraction

There are two types of features that can be extracted and adopted with this stage:

3.5.1 Traditional Features – Vandalism Features

Consisted of textual and meta-data features. Where text features include all the text characteristics of an article. The textual features are taken from the text files (revisions of articles) that contained in the dataset; on the other hand, the metadata is a data that offers information about the article’s edits. These features are extracted from the CSV files, which are also contained in the dataset. The details for this type of the applied features are described in Table (3.1) .

Table 3.1 : Features for vandalism.

Features for vandalism		
Features Type	**Feature**	**Description**
Textual	Size increment	Article, absolute size increment, i.e., |new|-|old|
	Size ratio	Calculate size ratio Calculate digit ratio Calculate upper to lower case ratio Calculate upper to all ratio Calculate non alphanumeric ratio
	Longest character sequence	Vandalism often consists of long strings of the same character, such as (aaggggghhhhhh!!!!!, sssoooo huge).
	Character distribution	Character distribution of the inserted text with respect to the expectations useful for detecting nonsense.

	Character diversity	Expression giving the length of the inserted text as a percentage of the total length of characters.
	Word categories ratio	Vulgarism, Biased and Pronoun ratio are defined and listed.
	Average term frequency	Average frequency of inserted terms relative to the old revision text
	Longest word	Longest inserted word (links are not considered)
Meta-Data	Is anonymous	The editor is anonymous or not.
	Comment character distribution	Character distribution of the comment.
	Comment character diversity	Measure of different characters compared to the length of comment.
	Comment longest character sequence	Long strings of the same character.
	Comment longest word	Longest inserted word in the comment.
	Comment word categories	Vulgarism, Biased and Pronoun in the comment.
	Comment length	Append a length of the comment.

3.5.2 New Features –Trust Features

Furthermore, a new features **"The Annotator Trust"** was extracted and added to meta-data based on the annotators' information which decide whether or not the respective edit is vandalism. Giving trust to the annotator is computed by extracted many related sub-features that can be illustrated in Table (3.2).

Additionally, annotator decides whether or not more than one article is vandalized, and his assessments are correct, with points of trust given for the annotator's age and gender.

Table 3.2 : The Annotator Trust Sub-features.

Sub-Feature	Description
Annotator-id	unique annotator identifier
Age	the age of the annotator
Sex	gender of the annotator
Reading	how often the annotator reads Wikipedia
Editing	how often the annotator edits Wikipedia
Vandalizing	whether or not the annotator has vandalized Wikipedia
Noticing	how often the annotator notices vandalism on Wikipedia
All predictions	How many total predictions for article edits for each annotator.
True predictions	How many correct predictions for article edits for each annotator, whether Vandalism or normal.
False predictions	How many false predictions for article edits for each annotator, whether Vandalism or normal.
Trust-reading	trust is given to annotators who read Wikipedia on a daily or monthly basis, and who will be given trust in their decision
Trust-editing	trust is given to annotators who edit Wikipedia on a daily or weekly basis, and who will be given trust in their decision
Trust-vandalizing	The one who previously vandalized Wikipedia is not given trust in his decision
Trust-noticing	trust is given to annotators who noticing vandalism on Wikipedia daily, weekly, or monthly basis, and who will be given trust in their decision
Trust-percentage	Number of true predictions to total number of predictions

3.6 1D-CNN Model - Vandalism Detection Stage

The most crucial point in the real-time applications is the performance speed, whereas the training and performance time of existing methods have long times. For that, a new light vandalism detection models proposed to detect the Wikipedia articles status based on 1DCNN model with new light architecture.

The proposed 1DCNN model designed with light architecture. This architecture has a relatively small number of parameters compared to other CNN architectures, which makes it computationally efficient and faster to execute on devices with limited resources. At the same time, the light CNN model is easier to train and deploy the model on resource-constrained devices than other CNN architectures due to using a compact network structure, which requires less memory.

The primary goals of the proposed model are to decrease execution time with increased accuracy as well as the system can be implemented with reasonable hardware capability. Figure (3.3) illustrate the model stages.

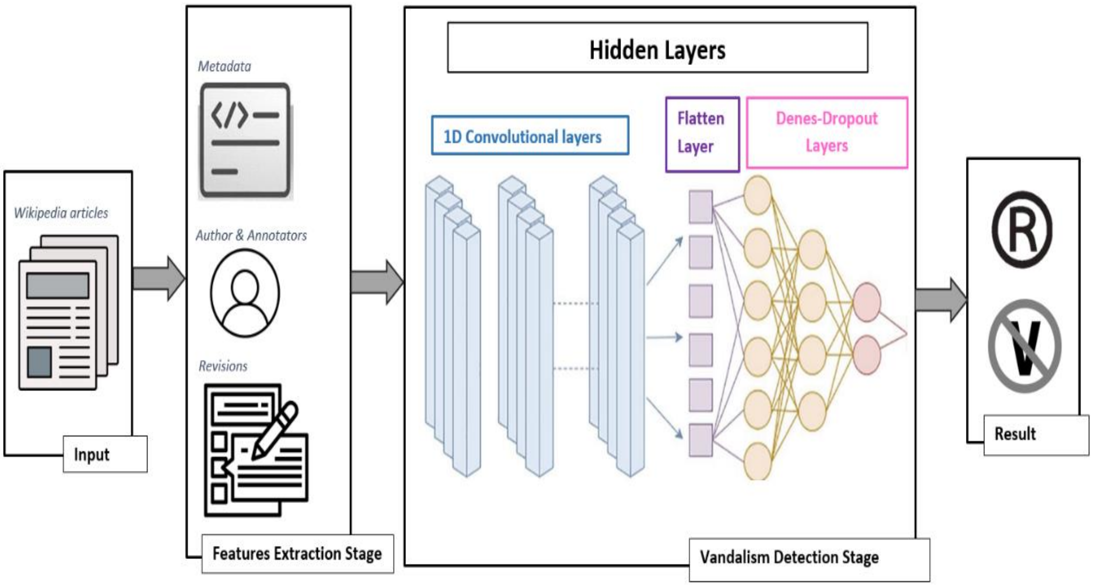


Figure 3.3 : VDS Stages.

The proposed CNN model consists of seven layers - showed in figure (3.4) below - : two of one-dimensional convolutional layers have depth size (64) and kernel size (3). Each one merged with activation (non-linear) layer used non-saturating rectified linear units (ReLU). In order to handle sample noise, ReLU was combined with the CNN layer to collect important features and reject weak ones. ReLU performed this by only keeping the parts of the features that have positive values and throwing away the rest as noise.

Third layer is batch normalization, which can improve training stability, convergence speed, and overall performance. Batch normalization can be normalizes the activations of each layer in the CNN by adjusting and scaling them, which make the model more resilient to variations in the input data distribution and thus assets to reduce the impact of noise.

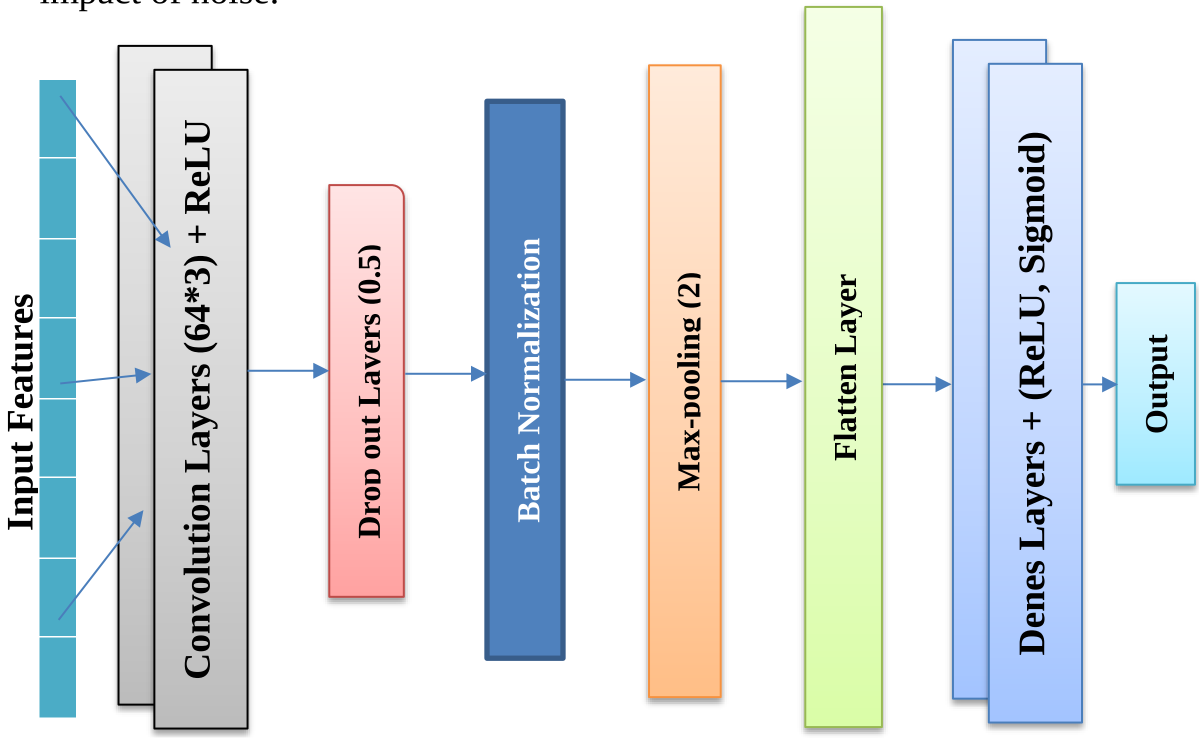


Figure 3.4 : The architecture of the proposed 1D-CNN.

The 1D max-pooling (4th) layer; with pool size (2); creates an effective and efficient feature hierarchy which is a crucial sub-step in feature extraction, and generalization, leading to effective classification tasks. This layer implements down sampling and reduces the computational complexity of subsequent layers. In order to reduce overfitting and improve noise tolerance in the proposed model, a dropout layer was added with 0.5%.

The final two layers are dense layers that have (100 and 1 neurons) and implemented (ReLU and sigmoid activation functions) respectively. High-level representations are learned during these two layers based on the extracted features and final predictions are made by leveraging the fully connected nature to capture global information and relationships, leading to effective classification. During this stage, the important, prominent and influential features in the decision-making process are selected through a set of operations happened between the model layers this step known as (**salient feature extraction**).

The summary representation of the proposed model can be illustrated in table (3.3), which visualize information of layers and their order in the proposed model, output shape and number of parameters (weights) for each layer and finally the total number of parameters (weights) in the proposed model.

Table 3.3: The summary representation of proposed model.

Block no.	Layer type	Output shape	Params no.
1	reshape_1 (Reshape)	(None, 19, 1)	0
	conv1d_1 (Conv1D)	(None, 17, 64)	256
	conv1d_2 (Conv1D)	(None, 15, 64)	12352
2	dropout_1 (Dropout)	(None, 15, 64)	0
	max_pooling1d_1(Max-pooling)	(None, 7, 64)	0
	flatten_1 (Flatten)	(None, 448)	0
3	dense_1 (Dense)	(None, 100)	44900
	dense_2 (Dense)	(None, 1)	101
Total params:		**57,609**	
Trainable params:		**57,609**	
Non-trainable params:		**0**	

3.7 Model Evaluation

The proposed system is evaluated based on the testing dataset. The Log Loss, CM, Acc., Recall, F1-Measure, and Precision are used to evaluate the proposed system. The concept of Cross-Validation employed to make the evaluation process more reliable. Finally, the procedure of training and testing is repeated ten times using 10-folds.

3.8 Cross Validation (K-Folds)

The cross-validation procedure estimates the skill of a ML model on unseen data by evaluating the ML models on specific resampling data set based on a single parameter called k. In other terms, the limited sample used to estimate how the model generally expected to predict un-used data during the training of the model.

The proposed model are given k = 10, for each value of K, it split the dataset into Training (80%) +Validation (10%) = 90% and Testing = 10% and it run the algorithm (3.2), recording the testing performance according to the metric used (adopted accuracy). Finally, the average of the performance computed to represent the result.

Algorithm 3.2: 10-Fold Cross-Validation

```
Input: Total Dataset
Output: final Testing accuracy.
Step 1: Shuffle the dataset randomly
Step 2: Determine the value of K as K is 10, Split the dataset
        into k groups.
Step 2: for each value of K:
    a) Select unique training and testing datasets by take the group as
       testing set and the remaining groups as a training data set:
            KFoldTesting = subset (Data)
            KFoldTraining = subset (Data)
    b) Train and record performance
            KFoldPerformance[i] = Train (KFoldTraining, KFoldTesting)
    c) Retain the evaluation score and discard the model
Step 3: Summarize the performance skill of the model
        TotalPerformance = ComputePerformance(KFoldPerformance)
End
```

3.9 Summary

This chapter presents the proposed system for the thesis, which is a Model for detecting vandalism in Wikipedia articles. The first step is to prepare the data by dividing it in a ratio of 80:20 for training and testing, respectively.

VDM relies on a set of features, including traditional features of vandalism that were used in previous research and features that were extracted based on the information of annotators included in the dataset (PAN-WVC-2010). Before extracting the features, the dataset needs to be pre-processed to obtain reliable features for improving the classification process. The missing data in the dataset has been processed by deleting rows and interpolating.

After merging all the features, they are entered into the 1DCNN model, which consists of 7 layers, during which the influential features are selected, noise and negative values are eliminated, and finally a decision is made on whether the article is vandalized or not.

After obtaining the classification results, which are shown in the next chapter, the system was evaluated to see its performance on the selected dataset.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Overview

The effectiveness of the proposed system illustrated in the previous chapter is tested with different parameters values and the results of implementation will be analyzed in this chapter. It is worth to mentioning that the evaluation of the proposed system stages is implemented separately. A public dataset is applied as a case study to determine the behavior of this model. The datasets that are used with the proposed system, as well as, the system requirements are presented in the next sections. Other sections describe the results obtained from each stage of the proposed system.

4.2 System Requirement

First of all, to perform machine learning and deep learning on any dataset, the software/program requires a computer system powerful enough to handle the computing power necessary. So the proposed system is implementing using the following:

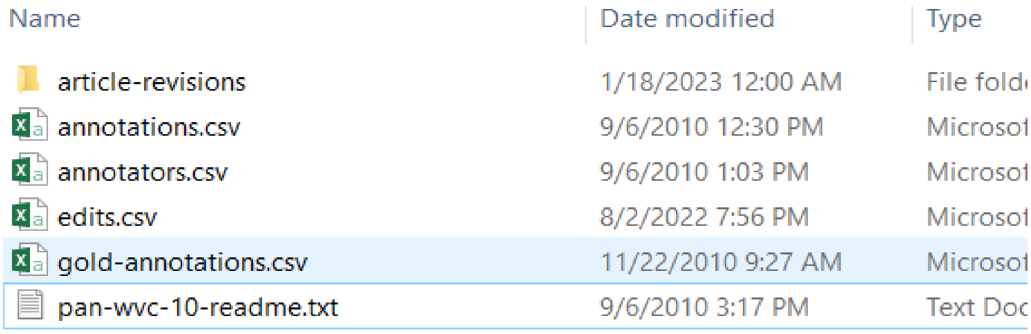
- **Hardware:**

1. Central Processing Unit (CPU): Intel(R) Core (TM) i7-4510U CPU @ 2.60 GHz.
2. RAM: 8 GB.
3. Hard Disk: 1 TB + 500 GB.
4. Graphics Processing Unit (GPU): NVIDIA GeForce 840M

- **Operating System:** Windows 10 Pro-64-bit.
- **Programming Language:** Python

4.3 Dataset

PAN Wikipedia Vandalism Corpus 2010, PAN-WVC-10 is a corpus containing edits from the edit logs of the Wikipedia encyclopedia which humans have annotated with regard to whether they are regular edits or rather vandalism edits. Figure (4.1) show the folder of dataset contents.



Name	Date modified	Type
article-revisions	1/18/2023 12:00 AM	File fold
annotations.csv	9/6/2010 12:30 PM	Microsof
annotators.csv	9/6/2010 1:03 PM	Microsof
edits.csv	8/2/2022 7:56 PM	Microsof
gold-annotations.csv	11/22/2010 9:27 AM	Microsof
pan-wvc-10-readme.txt	9/6/2010 3:17 PM	Text Doc

Figure 4.1 : Dataset's Folder Contents.

4.3.1 Article-revisions

Contains the revisions of Wikipedia articles which are either source, or result of an edit operation.

Directories "part1 - part65" each part directory contains up to 1000 text files, each of which is the revision of a Wikipedia article. The name of a file is its revision identifier, which is referenced in the edits.csv file. The contents of each text file are the plain wiki-text of the revision.

4.3.2 Other Files

The dataset content of four csv files shows in Table (4.1) below :

Table 4.1: Show in details the CSV files that represent dataset meta-data.

No.	File Name	Description	Columns		Note
			Name	**Description**	
1	edits.csv	List of edits to be used to train a vandalism detector	Edited	unique edit identifier	(editid) column - primary key of the table.
			Editor	user name / IP of editor	
			Oldrevisionid	identifier of the old, source revision	
			Newrevisionid	identifier of the new, destination revision	
			Diffurl	URL showed the difference between old **-** new revision	
			Edittime	time of edit	
			Editcomment	comment of editor left when submitting the edit	
			Articleid	unique identifier of an edited article	
			Articletitle	the title of the edited article	
2	gold-annotations.csv	Classification of each edit as vandalism or regular edit	Edited	reference to the editid of edits.csv	editid column - primary key of this table.
			Class	Classification of edit (regular or vandalism)	
			Annotators	Num. of annotators who concur with this classification	
			Totalannotators	total number of annotators who reviewed the edit	

3	annotations.csv	Annotations that supplied by the different annotators.	Edited	reference to the editid of edits.csv	The pair (editid,annotatorid) - primary key of this table
			Annotatorid	reference to annotatorid of annotators.csv	
			Class	classification of the referenced edit (regular, vandalism, dunno) dunno : the annotator was uncertain	
			decisiontime	time before the annotator made chose a class	
			Submittime	date of submission annotation	
4	annotators.csv	annotators which supplied the annotations in annotations.csv	Annotatorid	unique annotator identifier	annotatorid - primary key of this table
			age	age of the annotator	
			sex	gender of the annotator	
			Reading	how often the annotator reads Wikipedia	
			editing	how often the annotator edits Wikipedia	
			Vandalizing	whether or not the annotator has vandalized Wikipedia	
			Noticing	how often the annotator notices vandalism on Wikipedia	

4.3 Vandalism Detection Stage

The metrics used in the proposed model are the accuracy, loss and mean square error functions that use in each step of training and validation processes - A learning curve is an important, graphical representation that can provide insight into such learning behavior by plotting generalization performance against the number of training examples-.

The proposed model like all Neural Networks models implements two consequent phases: the learning phase then testing phase. The proposed model's learning phase behavior on specific training and validation datasets in each epoch is represented in plotting figures, which indicate the outcomes of metrics employed in each epoch, as illustrated in figure (4.2), and the experimental findings of the proposed model are reported in table (4.2).

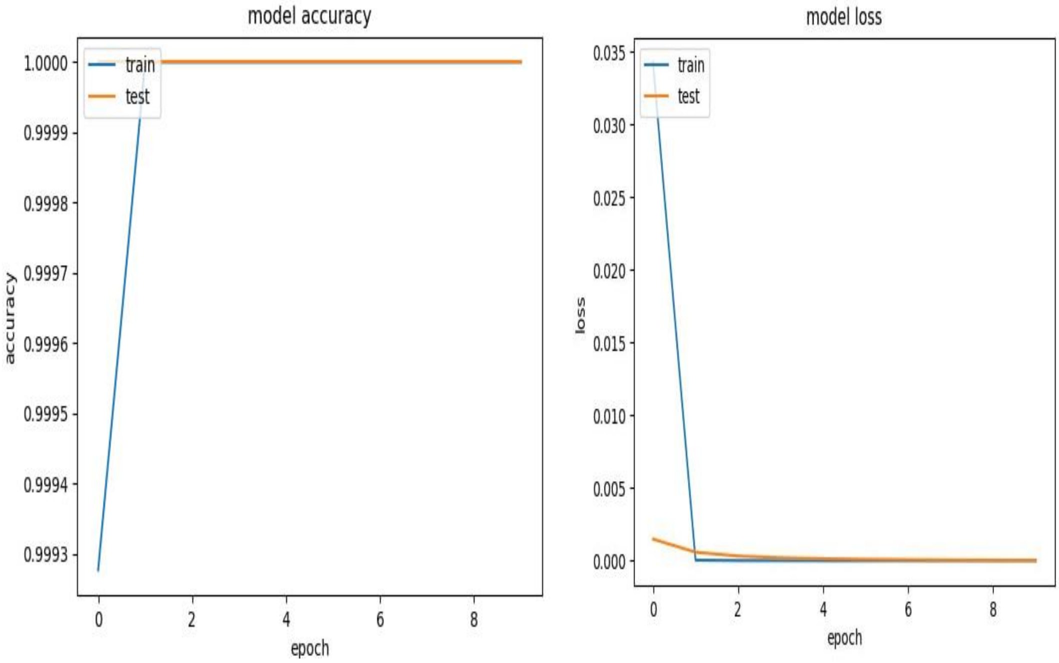


Figure 4.2 : The Accuracy and Loss functions of learning phase.

Table 4.2: The experimental results of the proposed model in training and validation sets.

Model	No. of epochs	Accuracy		Loss Fun.		Mean Squared Error	
		Training data	Testing data	Training data	Testing data	Training data	Testing data
1D CNN	5	1.00	1.00	1.6694e-07	1.5736e-05	1.6694e-07	1.5736e-05
	10	1.00	1.00	1.1934e-07	5.1260e-06	1.1934e-07	5.1260e-06
	20	1.00	1.00	1.1921e-07	1.1921e-06	1.1921e-07	1.1921e-06

The experimental results of the hyper-parameters (batch size and learning rate) table (4.3) represents tuning in the model for batch size choose and table (4.4) for learning rate choose.

Table 4.3: batch size tuning of VDS model.

Batch size	Mode	Loss func.	MSE	Accuracy
16	Training	0.002	0.002	0.97
	Testing	0.002	0.002	0.98
24	Training	0.002	0.002	0.98
	Testing	0.001	0.001	0.99

Table 4.4: Learning Rate tuning of VDS model.

Learning Rate	Mode	Loss func.	MSE	Accuracy
32	Training	0.001	0.001	1.00
	Testing	0.000	0.000	1.00
64	Training	0.000	0.000	1.00
	Testing	0.001	0.001	0.99

The evaluation results of the VDS are presented through the table (4.5) for the confusion matrix and table (4.6) presents the evaluation of the model in term of Performance Metrics.

Table 4.5: The Confusion Matrix CM.

N= 6488		Predictive model	
		Yes	No
Actual Recommended	Yes	6000	0
	No	0	488

Table 4.6: the evaluation measures values of VDS.

Measure	**Value**
Accuracy	1.0000
Precision	1.0000
Recall	1.0000
F1-Score	1.0000

4.4 K-Fold Cross-Validation

Performing K-fold cross-validation is used to further strengthen the network against the over-fitting. This technique employs the smart ruse by performing K rounds of training-validation-testing on, different, non-overlapping, equally-proportioned Training (80%), Validation (10%), and Testing (10%) sets.

The 10-Fold CV implemented with proposed VDS model and the experimental results can be represented in the Table (4.7).

Table 4.7: 10-Fold CV for VDS model.

No of fold	**Accuracy of each fold**	**model accuracy**
#1	*100.00%*	*100.000% (+/-0.000)*
#2	*100.00%*	
#3	*100.00%*	
#4	*100.00%*	
#5	*100.00%*	
#6	*100.00%*	
#7	*100.00%*	
#8	*100.00%*	
#9	*100.00%*	
#10	*100.00%*	

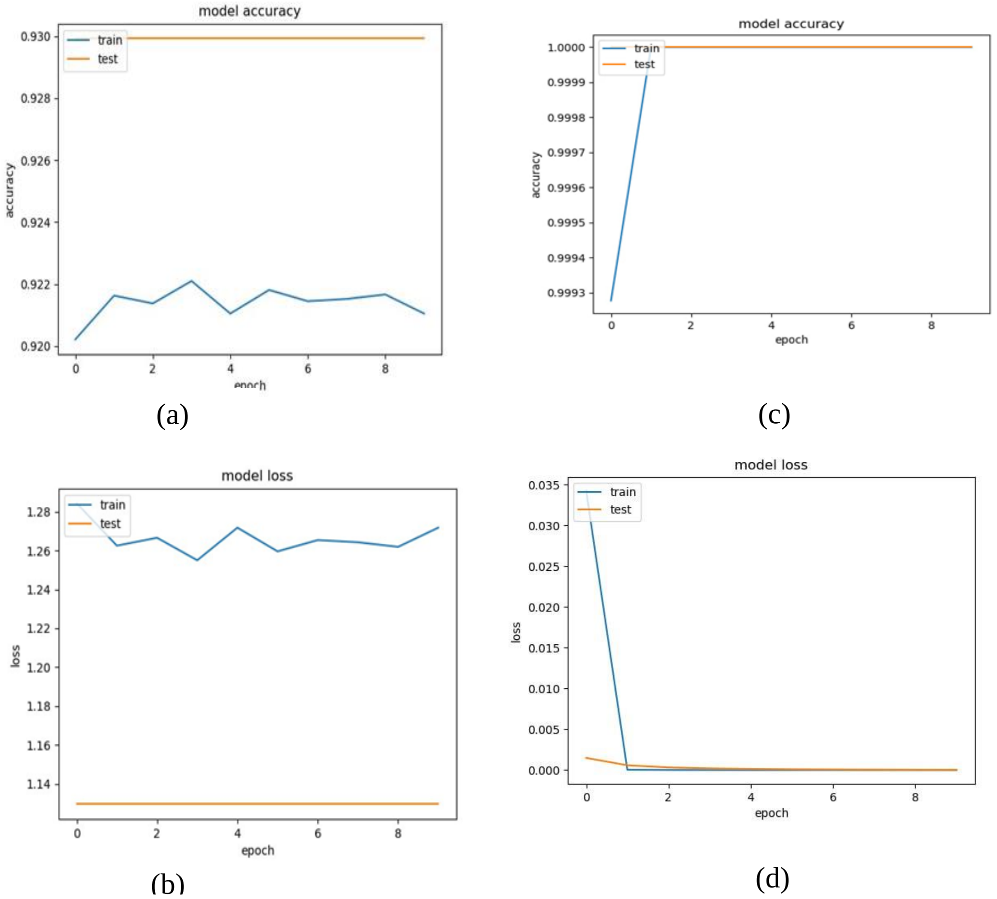
4.5 Comparison Results of Proposed System

Several experiments were conducted on the proposed system , during which different features were used in order to achieve the best accuracy .

In this section, the most two prominent experiences that have passed on the system reviewed to compare the use of traditional features and the new proposed features

- The first experiment: In this experiment, the traditional features that presented in Table (3.1) were used. These features are among the most important features related to vandalism that were used in previous research.

  Where these features showed results of an estimated accuracy of 92% when used with the architecture of the proposed system. The accuracy curve of the first experiment shows in Figure (4.3) below.
- The second experiment: In this experiment, the new proposed features that presented in Table (3.2) were used In addition to the traditional features. The used of the new features that called (trust features) with the traditional (standard) features, made the architecture of the proposed system is directed towards a correct classification of the results with a high accuracy of up to 100%. The accuracy curve of the first experiment shows in Figure (4.3) below.



(a)

(c)

(b)

(d)

Figure 4.3 : The proposed model's accuracy and loss function – first experiment (a) accuracy and (b) loss and second experiment (c) accuracy and (d) loss.

Adding new features (new variables) helps the model to clarify invisible relationships between data. In addition, these features were selected after processing the missing and uncertain data. All these factors help improve the model's accuracy.

In the framework of comparing the results, Table (4.8) presents a comparison between the results of previous research and the proposed model.

Table 4.8: VDS Accuracies for different scenarios of the PAN-WVC-10 Dataset.

No.	***Methods***	***Classifier***	***Recall***	***F1***	***Precision***	***AUC***
1	*Harpalani et al.* [18]	*Logit Boost*	*47%*	*58%*	*73%*	*93%*
2	*Götze* [28]	*Random Forest*	*60%*	-	*60%*	-
3	*Martinez-Rico et al.* [6]	*Random Forest*	*96%*	*79%*	*85%*	-
4	*Santos et al.* [34]	*Logistic Model Trees + K-Means*	*63.8%*	-	*78.1%*	-
5	*Proposed System VDS*	*1D-CNN*	*100***%**	*100%*	*100%*	*100%*

4.6 Imbalanced Dataset Processing using Stratified Cross-Validation

This technique provides train and test indices to split data into train and test sets. Stratified k-fold cross-validation is the same as k-fold cross-validation, but does stratified sampling instead of random sampling. The folds are made by preserving the percentage of samples for each class.

The 10-Fold implementation with the proposed VDS and the experimental results can be represented in the table (4.9).

Where a list of possible accuracy is presented for each fold, the maximum accuracy is 100% and the minimum accuracy is 92%. The final model accuracy is 96%.

Table 4.9: Stratified Validation of 10-Fold for VDS model.

No. of fold	**Accuracy of each fold**	**Model accuracy**
#1	*92%*	***96%***
#2	*96%*	
#3	*98%*	
#4	*100%*	
#5	*96%*	
#6	*96%*	
#7	*98%*	
#8	*94%*	
#9	*94%*	
#10	*100%*	

4.7 Summary

This chapter presents the results of the proposed system, beginning with a presentation of the nature of the dataset and its files.

After that, display the results of different experiments that passed VDM. These results explain the difference in accuracy between using traditional features and using the new features. In addition, a presentation of the evaluation results using K-fold.

Because of the unbalanced classes in the dataset, the system was evaluated using the stratified k-fold technique, which is a method to select a percentage from each class.

The average accuracy is 96%, and it depends on the system's accuracy.

CHAPTER FIVE

CONCLUSIONS AND FUTURE WORKS

5.1 Overview

Vandalism detection is an important issue in sustaining the quality of Wikipedia articles.

Detecting vandalism on Wikipedia articles is an on-going challenge, and there are several limitations associated with vandalism detection methods. Vandalism detection algorithms often rely on patterns, keywords, or statistical anomalies to identify potentially malicious edits. However, these algorithms may struggle to understand the broader context and intent behind an edit as some forms of vandalism may be subtle or context-dependent. Vandalism techniques constantly evolve to bypass detection mechanisms. Vandalism can take the form of subtle changes, biased edits, or misleading information that may be challenging to detect using automated algorithms. Vandalism creators can also adapt their tactics based on the detection methods in place, making it a cat-and-mouse game between vandals and detection systems.

One of the most important challenges in the vandalism problem is that it is related to human behavior. In this thesis, new features related to human behavior were extracted, including the trust of reviewers. This chapter shows the conclusions of the thesis, and some suggestions for future work.

5.2 Conclusions

Some social disorder problems, such as disturbing the peace and trespassing, are often associated with the Vandalism problem. At the same time, reducing and limiting the damage resulting from the occurrence of vandalism is through early detection and rapidly resolved of vandalism. The detection system using deep neural networks can be considered the best approach from other traditional approaches in terms of accuracy and loss functions.

The main conclusions of this research can be summarized below. The proposed model achieves successful detection of the Wikipedia article status to vandalism edits or regular edits.

The light 1D CNN model provides a low complexity which leads to the possibility of applying the model with limited computational capability and requirements and reasonable run time.

The implementation of the ReLU activation function in the nonlinear layer integrated with the convolution layer provides the ability to process the noise associated with inputs. This integrated layer extracts salient features and neglects the weak ones including noise by removing all the weak and noise elements from the sequence and keeping only those carrying a positive value. The proposed vandalism detection model enables to generalize of the editing and viewing features in Wikipedia.

Using stratified cross-validation enables evaluation of the proposed model with balanced classes for each fold. And overcome the problem of unbalanced data.

The model accuracy achieved 96% when employing new (trust features) that are used together with the conventional vandalism features to increase the accuracy and reach to efficient model.

5.3 Future Works

- Vandalism identification may be based on the type of vandalism or the identity of the vandal. It is possible to know the type based on the extracted features, whether they were from the text or from meta-data. and the identification of vandals It can be found by tracking the user's behavior if he is registered in Wikipedia or by tracking his IP using modern technologies.

- The proposed system could be improved to determine the vandalism type in addition to its detection, such as add, delete, change, and so on. Another improvement is determining the location of vandalism from articles. Other datasets that deal with the problem of vandalism can be applied or combining more than one dataset to extract more features and evaluate the model.

- Reducing the use of traditional features and relying on new features to reduce the time required to extract traditional features.

REFERENCES

[1] S. Heindorf, Y. Scholten, G. Engels, and M. Potthast, “Debiasing vandalism detection models at wikidata,” in *The World Wide Web Conference (WWW ’19)*, 2019, pp. 670–680, doi: 10.18420/inf2019_48.

[2] K. Madanagopal and J. Caverlee, “Towards Ongoing Detection of Linguistic Bias on Wikipedia,” in *International World Wide Web Conference Committee*, 2021, pp. 629–631, doi: 10.1145/3442442.3452353.

[3] D. Schall, *Service-oriented crowdsourcing: architecture, protocols and algorithms*. 2012.

[4] C. Grimpe, M. K. Poetz, N. Rietzler, and F. Waldner, “The Impact of Cooperation in Innovation Contests: Poison Pill, Placebo, or Tonic?,” *Acad. Manag. Proc.*, vol. 2021, no. 1, p. 15971, 2021, doi: 10.5465/ambpp.2021.15971abstract.

[5] C. V. Zsi, D. H. Ait, M. V. Rsa, and H. Leo, “Situated , crowdsourced m icrolearning : From micro - content to micro - arguments,” 2007, pp. 1–29, doi: 10.13140/RG.2.1.4252.3281.

[6] J. R. Martinez-Rico, J. Martinez-Romo, and L. Araujo, “Can deep learning techniques improve classification performance of vandalism detection in Wikipedia?,” *Eng. Appl. Artif. Intell.*, vol. 78, no. 0952–1976, pp. 248–259, 2019, doi: 10.1016/j.engappai.2018.11.012.

[7] K. Y. Itakura and C. L. A. Clarke, “Using dynamic markov compression to detect vandalism in the wikipedia,” *Proc. - 32nd Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, SIGIR 2009*, pp. 822–823, 2009, doi: 10.1145/1571941.1572146.

[8] C. Hube, “Methods for Detecting and Mitigating Linguistic Bias in Text Corpora,” Gottfried Wilhelm Leibniz, 2020.

[9] K. Y. Wong, M. Redi, and D. Saez-Trumper, *Wiki-Reliability: A Large Scale Dataset for Content Reliability on Wikipedia*, vol. 1, no. 1. Association for Computing Machinery, 2021.

[10] J. Carter, “ClueBot and vandalism on Wikipedia,” 2010, [Online]. Available: http://www.acm.uiuc.edu/~carter11/ClueBot.pdf.

[11] M. Potthast, B. Stein, and R. Gerling, “Automatic Vandalism Detection in Wikipedia,” *Springer-Verlag Berlin Heidelb. 2008*, pp. 663–668, 2008.

[12] A. Halfaker and R. S. Geiger, “ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia,” *Proc. ACM Human-Computer Interact.*, vol. 4, no. CSCW2, 2020, doi: 10.1145/3415219.

[13] K. Smets, B. Goethals, and B. Verdonk, “Automatic vandalism detection in wikipedia: Towards a machine learning approach,” *AAAI Work. - Tech. Rep.*, vol. WS-08-15, pp. 43–48, 2008.

[14] M. Lehto and P. Neittaanmaki, *Cyber Security: Analytics, Technology and Automation*, vol. 78. 2015.

[15] A. Razzaq, A. Hur, H. F. Ahmad, and M. Masood, “Cyber security: Threats, reasons, challenges, methodologies and state of the art solutions for industrial applications,” *Proc. - 2013 11th Int. Symp. Auton. Decentralized Syst. ISADS 2013*, vol. 5, no. 2, 2013, doi: 10.1109/ISADS.2013.6513420.

[16] Q. T. Truong, G. Touya, and C. de Runz, “OSMWatchman: Learning how to detect vandalized contributions in OSM using a random forest classifier,” *ISPRS Int. J. Geo-Information*, vol. 9, no. 9, 2020, doi: 10.3390/ijgi9090504.

[17] B. T. Adler, U. P. De Valencia, P. Rosso, and A. G. West, “Wikipedia Vandalism Detection : Combining Natural Language , Metadata , and Reputation Features,” in *Lecture Notes in Computer Science: Computational Linguistics and Intelligent Text Processing 6609*, 2011, pp. 277–288, doi: http://dx.doi.org/10.1007/978-3-642-19437-5_23.

[18] M. Harpalani, M. Hart, S. Singh, R. Johnson, and Y. Choi, “Language of vandalism: Improving Wikipedia vandalism detection via stylometric analysis,” *ACL-HLT 2011 - Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol.*, vol. 2, no. 2009, pp. 83–88, 2011.

[19] N. Joshi, F. Spezzano, M. Green, and E. Hill, “Detecting Undisclosed Paid Editing in Wikipedia,” *WWW ’20 Proc. Web Conf. 2020*, pp. 2899–2905, 2020, doi: https://doi.org/10.1145/3366423.3380055.

[20] A. Belani, “Vandalism Detection in Wikipedia: a Bag-of-Words Classifier Approach,” pp. 1–15, 2009, [Online]. Available: http://arxiv.org/abs/1001.0700.

[21] S. M. M. Velasco, “Wikipedia Vandalism Detection Through Machine Learning : Feature Review and New Proposals ∗ Lab Report for PAN at CLEF 2010,” 2010.

[22] A. G. West, S. Kannan, and I. Lee, “Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata,” in *Proceedings of the 3rd European Workshop on System Security, EUROSEC’10*, 2010, vol. 1752050, pp. 22–28, doi: 10.1145/1752046.1752050.

[23] Q. Wu, D. Irani, C. Pu, and L. Ramaswamy, “Elusive vandalism detection in Wikipedia: A text stability-based approach,” *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 1797–1800, 2010, doi: 10.1145/1871437.1871732.

[24] S. C. Chin, W. N. Street, P. Srinivasan, and D. Eichmann, “Detecting wikipedia vandalism with active learning and statistical language models,” *Proc. 4th Work. Inf. Credibility, WICOW ’10*, pp. 3–10, 2010, doi: 10.1145/1772938.1772942.

[25] S. Javanmardi, D. W. McDonald, and C. V. Lopes, “Vandalism detection in Wikipedia: A high-performing, feature-rich model and its reduction through Lasso,” in *WikiSym 2011 Conference Proceedings - 7th Annual International Symposium on Wikis and Open Collaboration*, 2011, pp. 82–90, doi: 10.1145/2038558.2038573.

[26] E. Alfonseca, G. Garrido, J. Y. Delort, and A. Peñas, “WHAD: Wikipedia historical attributes data: Historical structured data extraction and vandalism detection from the Wikipedia edit history,” *Lang. Resour. Eval.*, vol. 47, no. 4, pp. 1163–1190, 2013, doi: 10.1007/s10579-013-9232-5.

[27] K. N. Tran and P. Christen, “Cross-language learning from bots and users to detect vandalism on Wikipedia,” *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 3, pp. 673–685, 2013, doi: 10.1109/TKDE.2014.2339844.

[28] P. C. Götze, “Advanced Vandalism Detection on Wikipedia,” Bauhaus-Universität Weimar, 2014.

[29] K. N. Tran, P. Christen, S. Sanner, and L. Xie, “Context-aware detection of sneaky vandalism on wikipedia across multiple languages,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9077, pp. 380–391, 2015, doi: 10.1007/978-3-319-18038-0_30.

[30] A. Susuri, M. Hamiti, and A. Dika, “Machine learning based detection of vandalism in Wikipedia across languages,” *2016 5th Mediterr. Conf. Embed. Comput. MECO 2016 - Incl. ECyPS 2016, BIOENG.MED 2016, MECO Student Chall. 2016*, pp. 446–451, 2016, doi: 10.1109/MECO.2016.7525689.

[31] M. Shulhan and D. H. Widyantoro, “Detecting vandalism on English Wikipedia using LNSMOTE resampling and Cascaded Random Forest classifier,” in *International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, 2016, pp. 1–6, doi: 10.1109/ICAICTA.2016.7803106.

[32] A. Susuri, M. Hamiti, and A. Dika, “Detection of vandalism in wikipedia using metadata features – Implementation in simple English and Albanian sections,” *Adv. Sci. Technol. Eng. Syst.*, vol. 2, no. 4, pp. 1–7, 2017, doi: 10.25046/aj020401.

[33] S. Heindorf, “Vandalism Detection in Crowdsourced Knowledge Bases,” Paderborn University, 2019.

[34] T. F. dos Santos, N. Osman, and M. Schorlemmer, “Learning for Detecting Norm Violation in Online Communities,” 2021, [Online]. Available: http://arxiv.org/abs/2104.14911.

[35] K. N. Tran and P. Christen, “Cross-language learning from bots and users to detect vandalism on Wikipedia,” *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 3, pp. 673–685, 2015, doi: 10.1109/TKDE.2014.2339844.

[36] M. Steinkasserer, T. Ruprechter, and D. Helic, “Investigating Western Bias in Wikipedia Articles about Terrorist Incidents,” in *The 17th International Symposium on Open Collaboration (Companion)*, 2021, pp. 1–5.

[37] G. A. Al-Sultany and H. J. Aleqabie, “Enriching Tweets for Topic Modeling via Linking to the Wikipedia,” *Int. J. Eng. Technol.*, vol. 7, no. 19, pp. 144–150, 2018, doi: https://doi.org/10.14419/ijet.v7i4.19.

[38] S. Heindorf, M. Potthast, B. Stein, and G. Engels, “Towards Vandalism detection in knowledge bases: Corpus construction and analysis,” in *the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’15)*, 2015, vol. 10, no. 1145, pp. 831–834, doi: 10.1145/2766462.2767804.

[39] F. Spezzano, K. Suyehira, and L. A. Gundala, “Detecting pages to protect in Wikipedia across multiple languages,” *Soc. Netw. Anal. Min.*, vol. 9, no. 1, p. 0, 2019, doi: 10.1007/s13278-019-0555-0.

[40] K.-N. D. Tran, “Detecting Vandalism on Wikipedia across Multiple Languages,” Australian National University, 2015.

[41] E. U. H. Qazi, A. Almorjan, and T. Zia, “A One-Dimensional Convolutional Neural Network (1D-CNN) Based Deep Learning System for Network Intrusion Detection,” *Appl. Sci.*, vol. 12, no. 16, pp. 1–14, 2022, doi: 10.3390/app12167986.

[42] S. Dong, P. Wang, and K. Abbas, “A survey on deep learning and its applications,” *Comput. Sci. Rev.*, vol. 40, no. 00379, pp. 1–22, 2021, doi: 10.1016/j.cosrev.2021.100379.

[43] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, “1D convolutional neural networks and applications: A survey,” *Mech. Syst. Signal Process.*, vol. 151, no. 107398, pp. 1–21, 2021, doi: 10.1016/j.ymssp.2020.107398.

[44] D. Bhatt *et al.*, “Cnn variants for computer vision: History, architecture, application, challenges and future scope,” *Electron.*, vol. 10, no. 20, pp. 1–28, 2021, doi: 10.3390/electronics10202470.

[45] D. X. Zhou, “Universality of deep convolutional neural networks,” *Appl. Comput. Harmon. Anal.*, vol. 48, no. 2, pp. 787–794, 2020, doi: 10.1016/j.acha.2019.06.004.

[46] S. Kiranyaz, M. Zabihi, A. B. Rad, T. Ince, R. Hamila, and M. Gabbouj, “Real-time phonocardiogram anomaly detection by adaptive 1D Convolutional Neural Networks,” *Neurocomputing*, vol. 411, no. 0925–2312, pp. 291–301, 2020, doi: 10.1016/j.neucom.2020.05.063.

[47] L. Alzubaidi *et al.*, “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” *J. Big Data*, vol. 8, no. 1, pp. 1–74, 2021, doi: 10.1186/s40537-021-00444-8.

[48] T. Ince, “Real-time broken rotor bar fault detection and classification by shallow 1D convolutional neural networks,” *Electr. Eng.*, vol. 101, no. 2, pp. 599–608, 2019, doi: 10.1007/s00202-019-00808-7.

[49] S. M. Shahid, S. Ko, and S. Kwon, “Performance Comparison of 1D and 2D Convolutional Neural Networks for Real-Time Classification of Time Series Sensor Data,” in *International Conference on Information Networking*, 2022, vol. 2022-Janua, pp. 507–511, doi: 10.1109/ICOIN53446.2022.9687284.

[50] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, “The Computational Limits of Deep Learning,” *arXiv Prepr. arXiv2007.05558*, pp. 1–33, 2022, [Online]. Available: http://arxiv.org/abs/2007.05558.

[51] S. Yuan, P. Zheng, X. Wu, and Y. Xiang, “Wikipedia Vandal Early Detection: From User Behavior to User Embedding,” in *Machine Learning and Knowledge Discovery in Databases*, vol. 10534 LNAI, 2017, pp. 832–846.

[52] W. H. Bangyal *et al.*, “Detection of Fake News Text Classification on COVID-19 Using Deep Learning Approaches,” *Hindawi Comput. Math. Methods Med.*, vol. 2021, no. 5514220, pp. 1–14, 2021, doi: 10.1155/2021/5514220.

[53] G. Pang, C. Shen, L. Cao, and A. Van Den Hengel, “Deep Learning for Anomaly Detection: A Review,” *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–36, 2020, doi: 10.1145/3439950.

[54] P. Zheng, S. Yuan, X. Wu, J. Li, and A. Lu, “One-class adversarial nets for fraud detection,” in *the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI’19/IAAI’19/EAAI’19)*, 2019, pp. 1286–1293, doi: 10.1609/aaai.v33i01.33011286.

[55] S. Neelakandan *et al.*, “Deep Learning Approaches for Cyberbullying Detection and Classification on Social Media,” *Hindawi Comput. Intell. Neurosci.*, vol. 2022, no. 2163458, pp. 1–13, 2022, doi: https://doi.org/10.1155/2022/2163458.

[56] Daren C. Brabham, *Crowdsourcing*. MIT Press, 2013.

[57] H. Ulbrich, M. Wedel, and H. Dienel, *Internal Crowdsourcing in Companies*. springer, 2021.

[58] F. J. Garrigos-Simon and Y. Narangajavana, *From crowdsourcing to the use of masscapital. the common perspective of the success of apple, facebook, google, lego, tripadvisor, and zara*. 2015.

[59] H. O. Ikediego, M. Ilkan, A. M. Abubakar, and F. Victor Bekun, “Crowd-sourcing (who, why and what),” *Int. J. Crowd Sci.*, vol. 2, no. 1, pp. 27–41, 2018, doi: 10.1108/ijcs-07-2017-0005.

[60] A. Doan, M. J. Franklin, D. Kossmann, and T. Kraska, “Crowdsourcing applications and platforms,” *Proc. VLDB Endow.*, vol. 4, no. 12, pp. 1508–1509, 2011, doi: 10.14778/3402755.3402809.

[61] F. Petroni *et al.*, “Improving Wikipedia Verifiability with AI,” pp. 1–16, 2022, [Online]. Available: http://arxiv.org/abs/2207.06220.

[62] P. Rosso, “Wikipedia Vandalism Detection,” Politécnica de Valencia, 2011.

[63] M. PROFFITT, *Leveraging wikipedia*, ALA editio. chicago: American Library Association Extensive, 2018.

[64] “Wikipedia:Five pillars.” https://en.wikipedia.org/wiki/Wikipedia:Five_pillars#:~:text=Respect your fellow Wikipedians%2C even,on the part of others.

[65] O. Ferschke, “The Quality of Content in Open Online Collaboration Platforms: Approaches to NLP-supported Information Quality Management in Wikipedia,” *Tech. Univ. Darmstadt*, p. 224, 2014.

[66] S. Abaimov and M. Martellini, *Machine Learning for Cyber Agents*. Springer, 2022.

[67] M. June *et al.*, “False Intel Detection In Crowd Source Knowledge Base,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 3, pp. 2158–2164, 2021, doi: 10.30534/ijatcse/2021/921032021.

[68] R. A. and D. G. Gopal and Scope:, Eds., *Cyber Security and Digital Forensics*. Scrivener, 2022.

[69] C. A. Damas and K. Mochetti, “An analysis of homophobia on vandalism at Wikipedia,” *Proc. 2019 Res. Equity Sustain. Particip. Eng. Comput. Technol. RESPECT 2019*, 2019, doi: 10.1109/RESPECT46404.2019.8985876.

[70] S.-C. Chin and W. N. Street, “Enriching Wikipedia Vandalism Taxonomy via Subclass Discovery,” 2011.

[71] B. Suh, G. Convertino, E. H. Chi, and P. Pirolli, “The singularity is not near: Slowing growth of Wikipedia,” *Proc. 5th Int. Symp. Wikis Open Collab. WiKiSym 2009*, vol. 1, no. 650, 2009, doi: 10.1145/1641309.1641322.

[72] “Wikipedia:Vandalism_types.” https://en.wikipedia.org/wiki/Wikipedia:Vandalism_types.

[73] “Most vandalized pages.” https://en.wikipedia.org/wiki/Wikipedia:Most_vandalized_pages.

[74] J. Daxenberger and I. Gurevych, “Automatically classifying edit categories in wikipedia revisions,” *EMNLP 2013 - 2013 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, no. October, pp. 578–589, 2013.

[75] “Webis-WVC-07,” *January 1, 2007*. https://webis.de/data/webis-wvc-07.html.

[76] “PAN-WVC-10,” *2010*, 2010. https://webis.de/data/pan-wvc-10.

[77] “PAN-WVC-11,” *2011*, 2011. https://webis.de/data/pan-wvc-11.html.

[78] “Data dumps.” https://meta.wikimedia.org/wiki/Data_dumps.

[79] B. T. Adler, L. De Alfaro, and I. Pye, “Detecting Wikipedia Vandalism using WikiTrust,” pp. 22–23, 2010, [Online]. Available: http://institute.lanl.gov/isti/issdm/papers/vandalism-adler-report.pdf.

[80] P. Goyal, S. Pandey, and K. Jain, *Deep Learning for Natural Language Processing: Creating Neural Networks with Python*. Apress, 2018.

[81] F. A. Abdulghani and N. A. Z. Abdullah, “A Survey on Arabic Text Classification Using Deep and Machine Learning Algorithms,” *Iraqi J. Sci.*, vol. 63, no. 1, pp. 409–419, 2022, doi: 10.24996/ijs.2022.63.1.37.

[82] N. D. K. Al-shakarchy, “Driver Drowsiness Detection Based on Spatio-Temporal Features with 3D Convolutional Neural Networks,” University of Babylon, 2020.

[83] N. Vakitbilir, A. Hilal, and C. Direkoğlu, “Hybrid deep learning models for multivariate forecasting of global horizontal irradiation,” *Neural Comput. Appl.*, vol. 34, no. 10, pp. 8005–8026, 2022, doi: 10.1007/s00521-022-06907-0.

[84] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, “A survey on missing data in machine learning,” *J. Big Data*, vol. 8, no. 140, pp. 1–37, 2021, doi: 10.1186/s40537-021-00516-9.

[85] N. Tamboli, “All You Need To Know About Different Types Of Missing Data Values And How To Handle It What is a Missing Value ? How is Missing Value Represented In The Dataset ?,” *Data Science Blogathon Introduction*, pp. 1–12, 2023.

[86] Agrawal, “Interpolation - Power of Interpolation in Python to fill Missing Values,” *Data Science Blogathon Introduction*, pp. 1–6, 2021.

[87] N. Audebert, B. Le Saux, and S. Lefevre, “Deep learning for classification of hyperspectral data: A comparative review,” *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, 2019, doi: 10.1109/MGRS.2019.2912563.

[88] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, and L. Shao, “Normalization Techniques in Training DNNs: Methodology, Analysis and Application,” pp. 1–20, 2020, doi: 10.1109/TPAMI.2023.3250241.

[89] D. Berrar, “Cross-validation,” *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1, no. January 2018, pp. 542–545, 2018, doi: 10.1016/B978-0-12-809633-8.20349-X.

الخلاصة

التعهيد الجماعي هو مصطلح شائع يطلق على البيئات التعاونية التي تتضمن مجموعة من المشاركين في صناعة محتوى هذه المنصات وانشاء المعرفة . من أهم وأشهر منصات التعهيد الجماعي هي Amazon's Mechanical Turk و Kaggle و Wikipedia.

تعتبر ويكيبيديا واحدة من أكثر نماذج التعهيد الجماعي حيوية، حيث يساهم المستخدمون في إنشاء المقالات وتحريرها. مما أدى إلى موسوعيتها الامر الذي جعلها واحدة من أكثر المواقع زيارة للحصول على المعلومات أو استخدامها كقاعدة معرفية للعديد من التطبيقات مثل روبوتات الدردشة. كل هذه الخصائص المميزة يمكن للمستخدمين استغلالها سلباً في تغيير محتواها، وهو ما يعرف بالتخريب.

التخريب هو أي محاولة لتعديل المقال بشكل يؤثر سلبًا على جودة المقالة. يعتبر التخريب أحد الطبقات الخمس للنموذج الذي تم إنشاؤه لوصف التهديدات السبرانية ، وهو ما يعرف بالتخريب السيبراني.

بشكل عام، فأن "التخريب السيبراني" هو فعل إتلاف أو إضرار البيانات بدلاً عن سرقتها أو إساءة استخدامها. تم تطوير العديد من تقنيات الكشف التلقائي والميزات ذات الصلة لمعالجة هذه المشكلة.

تقدم هذه الأطروحة نموذجًا للتعلم العميق بهندسة معمارية جديدة لحل مشكلة التخريب في مقالات ويكيبيديا. يستخدم النموذج المقترح شبكة عصبية تلافيفية أحادية البعد (1D CNN) لتحديد نوع التعديلات التي تتم على مقالات ويكيبيديا هل هي تعديلات (عادية أو تخريبية) ، وفي الوقت نفسه ، تم استخراج ميزات جديدة واعتمادها في العمل ، مما ساهم في تحسين دقة النموذج . حيث أجريت التجارب على مجموعة بيانات معيارية ، مجموعة PAN-WVC-2010 من مسابقة الكشف عن التخريب التي استضافت في مؤتمر CLEF. بلغت الدقة التي حققها النظام المقترح بالميزات الجديدة 96٪.

جامعة كربلاء
كلية علوم الحاسوب وتكنولوجيا المعلومات
قسم علوم الحاسوب

كشف التخريب في نماذج التعهيد الجماعي

رسالة ماجستير
مقدمة الى مجلس كلية علوم الحاسوب وتكنولوجيا المعلومات / جامعة كربلاء وهي جزء من متطلبات نيل درجة الماجستير في علوم الحاسوب

كتبت بواسطة
أزهى طلال محمد علي الجعاوي

بإشـراف

م.د. هدى فاضل حلاوي

أ.م.د. نور ضياء الشكرجي

1444 هـ **2023 م**