Journal of Theoretical and Applied Information Technology

30th April 2022. Vol.100. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CRITICAL EVALUATION ON SPAM CONTENT DETECTION IN SOCIAL MEDIA

[1] ANTONIUS RACHMAT CHRISMANTO, [2]ANNY KARTIKA SARI, [3] YOHANES SUYANTO

[1] Faculty of Information Technology, Universitas Kristen Duta Wacana, Indonesia
[1,2,3] Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Indonesia
E-mail: [1]anton@ti.ukdw.ac.id, [2]a_kartikasari@ugm.ac.id, [3]yanto@ugm.ac.id

ABSTRACT

The spam content detection problem is still challenging due to its complexity, feature extraction process, language, context-aware detection capabilities, performance, and evaluation method. Spam content detection is different from spammers' detection and thus requires a different approach. This paper aimed to conduct a comprehensive literature review for "spam content detection" to identify the various approaches taken and generate up to date issues, especially in the social media case study. Literature data are collected from 2015 to 2021 based on seven journal repository databases and filtered into 69 main articles. This research compared the latest approaches and methods to see the gaps between these studies. Discussions on the approach, research media, dataset, feature extraction & selection, the language, context-based or not, the algorithm, performance, future research direction, and challenges were carried out. Additionally, this paper also discussed spam content on Indonesian social media and provided comprehensive suggestions for possible implementation, further research direction, and a possible new approach. This article can be used to develop new approaches, methods, and models in detecting spam content on social media.

Keywords: *Spam Content Detection; Social Media; Literature Survey; Systematic Review; Future Research And Challenge*

1. INTRODUCTION

Since the Internet has become a platform on web 2.0, users can read and create their content. Internet users are now the creators of their content for various purposes: conveying messages, doing business, doing hobbies, and sharing thoughts/ideas in texts, images, or videos. Technological developments on the Internet platform, starting from Short Message Service (SMS), Multimedia Messaging Service (MMS), email, links, and social media, allow users to quickly and easily spread the contents. Contents on the Internet, the web, and especially social media platforms can contain positive or negative content. Along with Internet development, negative content is becoming more and more. These negative contents take many forms, known as "spam." The word "spam" became known from the 1970s on the BBC television show called "Monty Python's Flying Circus," which refers to the name of a food that is repeatedly mentioned so that it becomes something that is no longer desirable anymore [1]. Spam emerged in the Internet context in 1978 when Gary Thuerek sent spam emails to users on the ARPANET [2]. Spam can be defined as irrelevant, unwanted, and usually occurs repeatedly in large numbers in various forms that interfere with users' information flow processing and even obscure it. Spam can take many forms, such as spam emails, spam advertisements, click spam, spam links, spam news, and other spam content [3]. The term web spam first appeared in 1996, according to Convey in The Boston Herald newspaper [4]. In the context of social media, spam can appear in user posts and comments

Spam can come from two sources: spam that originates from its users (known as spammers) and spam that originates based on its content. Spammers are active users who can intentionally generate spam content regularly, using different content according to the target domain [5]. Spammers use accounts that frequently change so that the other parties cannot detect them. This account even comes from a particular organization which could even threaten national security [6]. Spam contents can be in the form of spam images [7], spam videos [8], spam application [9], spam text (SMS / short message ([10], [11], [12], [13], and [14]), email ([15], [16], [17], and [18]), advertisements, spam links, posts/articles ([19], [20], [21], [22], [23], and [24]), spam comments on social media: Twitter ([25], [26], [27], and [28], [29]), YouTube ([30],

[31], [32], [33], [34], and [35]), Facebook ([36], [24]), Instagram ([37], [38], [39], [40], [41], [42], [43], [44], [45], and [46]). In this paper we focus on the text spam content because its popularity and more researched by researchers.

For example, several manuals, semi-manual, and automatic techniques can be used to detect spam comments on social media. Using the manual techniques, we can see whether it is specific to the post, whether there is a suspicious URL, whether the user uses an actual name/real email or not, or uses several different emails [47]. Using semi-manual techniques, we can use post/comments filter, captcha, HTML tags remover, IP address blacklists filter, and providing comment restrictions [48]. In addition, spam detection can also be done automatically using several content-based filtering methods, link-based filtering, and user behavior such as clicks, gestures, sessions, and others [49]. Automatic spam detection can be done using machine learning (ML) algorithms (ex: Naive Bayes (NB) [50] and Support Vector Machine (SVM) [51]). It also can be done using deep learning (DL) algorithms (ex: Long Short-term Memory (LSTM) [52] and Transformer [53]).

Based on this background, spam detection is a problem that has existed for a long time and is still being faced and fought by various parties. We conduct a systematic review for the specific topic: spam content detection on social media, especially spam comments. A systematic review can synthesize research findings systematically, transparently, and reproducible [54]. This article tries to find various literature that is comprehensive, relevant, up-to-date, and, finally, contributes to 1). a survey on research development comprehensively related to spam content detection, 2). discussion from various points of view in several categories, from case studies, media/area used, the dataset used, pre-processing and features extraction, language used, context spam detection, and methods/algorithms used, including machine learning and deep learning approach, 3). analyzing various methods used based on trends, 4) providing findings and input in research gaps and challenges, and 5) drawing conclusions and suggestions for steps that can be taken to deal with spam content cases in social media, especially in Indonesia.

This article is organized as follows: Introduction section contains introducing spam, spam history, forms of spam, introduction to spam on social media, and paper's contribution. The Related Works section talks about related research. The Method/Algorithm section talks about the research methodology. The Results & Discussion section discusses the detailed analysis of spam content cases according to the main problem, dataset, language, preprocessing, feature, method, the possibilities for its application, the research gaps, challenges, and future research that can still be worked out. Finally, The Conclusion section summarizes this article and plans for the following steps in this research area.

2. RELATED WORKS

Research on content spam detection surveys was conducted in several articles in [55], [56], [57], [58] and [59]. Previous research states that spam content detection is not easy and needs steps to handle the complexity, such as pre-processing, manual and automatic features, and machine learning techniques. The results also depend on the language and the training datasets.

Article [55] collected the literature from 2005-2015. It stated that spam content detection could be divided into content-based, source-based, and hybrid-based. Based on the content, it usually appears on the text, so it is necessary to take specific features so that the detection method can be more accurate. Meanwhile, we can see spam from the user's origin based on the source. This technique utilizes the original IP address and can also be seen from their activity logs on certain social media. Some of the features that can be used to detect spam comments are unusual content in comments, the percentage of comments containing specific spam words, and how redundant the comments are [55].

Wu et al. [57] conducted a survey analysis of the literature regarding Twitter spam detection. This article divides Twitter spam detection techniques based on syntax analysis, feature analysis, and blacklist methods. Based on the syntax analysis, detection is carried out based on whether the Tweet contains a link/not, contains a specific spam keyword, or contains a suspicious username. The post/comment was then retrieved using TF-IDF or the sparse graph. Based on Twit's features, spam detection can be done through account statistical information, Twitter statistics, and whether a Twit contains campaign content/not. It can also use the Twitter social graph feature to see the relationship between Twitter accounts. The article [56] is also in line with Wu surveying Twitter spam detection based on the account, content, social graph, and hybrid. This article also discusses the features that can be used for Twitter spam detection from these techniques. Meanwhile, Talha and Kara [56] focus

on a survey of Twitter spam accounts based on an account-based detection method that is not in our scope. They also mention a little about content-based spam detection using manual features, graph-based using node relationship features that are not lightweight, and hybrid-based. Unfortunately, they only focus on the Twitter spam problem.

Poonkodi & Sukumaran [58] discuss a survey on feature selection techniques and spam detection techniques on social media using ML. First, they categorized spammers into four types: spammers, phishers, fake users, and promoters. They surveyed feature selection categorized into a filter-based, wrapper, and embedded. Finally, based on detection algorithm, they use linear classification-based algorithms (Logistic Regression, NB, SVM, k-Nearest Neighbor, Decision Tree, Random Forest), clustering algorithms (K-Means), and hierarchical clustering. They also surveyed performance analysis methods using accuracy, FP rate, FN rate, and the Mathews Correlation Coefficient. They suggest using DL mathod to overcome the issues in ML algorithm.

The last article in [59] is the most comprehensive survey compared to the other. This article discusses social spam, spamming process, spam taxonomy, features, ML and DL methods, and challenges in this research area. They discuss more general social spam cases, not only social media text content. They also focus more on sentiment and sarcasm detection in online social networks (OSN).

Based on the previous and related research above, the difference between this evaluation survey and previous related studies are 1) in the period of data collection (2015 to early 2021), the focus of detection of spam content on social media, discussion of case studies, datasets used in the research, pre-processing, and features used, language, context, and proposed methods. This article focuses on the spam content in text, especially in spam comments, while articles [55], [56], [57], and [58] conducted a spam survey on social media in a general form. The motivation of this article is to find effective methods to detect spam comments based on the context and the usage of emoji features in detection techniques. The discussion in this article also focuses on two main techniques, machine learning and deep learning methods, and also about the pre-processing techniques, while [57] and [58] discuss only the machine learning technique. Finally, this article's findings give insight into the further research direction and a suggestion about the possibility of context-based spam comments technique. Finally, Table 1 gives the different comparations about these survey research in spam detection, including our article.

Table 1: Summary of Existing Survey in Social Media Spam Detection

Ref. Article	Topics covered
[55]	Spam detection survey in general based on manual features technique survey
[56]	Twitter spam account, content-based, and hybrid technique survey
[57]	Twitter spam detection using ML and graph survey
[58]	Spam detection in social media using ML survey
[59]	Spam detection on an online social network in general (multimedia), spammer account, social spamming, ML/DL method, features, performance, and research challenges survey
Ours	Spam detection on social media in text format, dataset, social media source, features generation & selection, language, emoji features, ML & DL, performance, future research challenge, study case, and context-based survey

3. RESEARCH METHOD

In this survey article, we use the systematic review method. The steps in this systematic review are design, conduct, analysis, structuring, and writing review [54]. In the analysis step, we identify, screen, and check the eligibility articles included in the review process [60]. First, we define the research topic: “spam content detection” in the design step. Then we conduct, analyze, and structure a literature review by collecting the research papers from the reference library. The reference library is taken from various data sources, such as scientific journal articles, conference proceedings articles, and books collected from computer science and information technology. Reference databases were collected from seven primary scientific databases: Scopus, Science Direct, IEEE Explore, Springerlink, Arxiv, SinTA-accredited Indonesian journals, and Google Scholar with the first keyword filter: "spam detection," and then followed by the second keyword filter based on each database as displayed on Figure 1 and Figure 2 describes the systematic literature review steps that we conduct.

In the first stage, filtering was carried out on the seven main source libraries and produced 13121 data. From these results, the second stage was carried out with the filtering keywords: "social

media" or "comment" or "Instagram" or "Twitter" or "Facebook" or "Youtube." The articles were then subjected to exclusion and inclusion based on journal level, subject, source of type, language, year, and abstract skimming to see the relevance. The final result was 69 articles. The relevancy means that various techniques are a spam content detection topic. The final results are a detailed review for each article to be discussed in more depth. Figure 1 and Figure 2 display the literature collection method, while Table 2 describes the literature filtering steps.

Table 2: Articles Data Based on Systematic Review (2015 – 2021)

Sources	Keywords	Count of docs	Details	Total
Scopus www.scopus.com	Filter1: "Spam Detection"	479	Conference: 294, Article: 185	12
	Filter2: "Social Media" Or "Twitter" Or "Youtube" Or "Facebook" Or "Instagram" Or "Comments."	309	Article: 121, Conference: 188	
IEEE Explore www.ieeexplore.com	Filter1: "Spam Detection"	291	Conference 251, Journal: 35, Early Access: 3, Magazine 2	23
	Filter2: "Social Media" Or "Twitter" Or "Instagram" Or "Youtube" Or "Facebook" Or "Comment."	186	Conference: 162, Journal: 24	
Science Direct www.sciencedirect.com	Filter1: "Spam Detection"	482	Review Article: 29, Research Article: 453	-
	Filter2: "Social Media, Twitter, Youtube, Facebook, Instagram, Comments."	39	Review Article: 8 Research Article: 22, Book Chapter: 7, Mini-Review: 1, Other: 1	7
SpringerLink https://link.springer.com	Filter1: "Spam Detection"	458	Article: 458, Chapter: 348, Conference: 274, Work Entry: 15	-
	Filter2: "Social Media, Twitter, Youtube, Facebook, Instagram, Comments"	12	Article: 10, Chapter: 2, Conference: 2	3
Arxiv (https://arxiv.org/)	Filter1: "Spam Detection"	66	-	7
Garuda DIKTI http://garuda.ristkbrin.go.id	Filter1: "Spam"	45	Full Pdf Only	6
Google Scholar http://scholar.google.com	Filter1: "Spam Detection"	11300	-	11
	Filter2: "Spam Detection", "Social Media", "Twitter", "Youtube", "Facebook", "Instagram", "Comments."	221	-	
TOTAL 69				

From 69 collected and validated literature data, the detection of spam content can be divided into several parts, as can be seen in Figure 3, that is:

1. Spam detection based on content (text, sound, image, or video). Text-based content can appear in the form of posts/status on social media such as Twitter ([61], [57], [62], [63]) and Facebook ([64], [36], [65], [66]). It can also appear in comments on Instagram ([38], [37], [44], [40]), Facebook ([36], [24]), and Youtube ([33], [32]). It appears in article/document [55], email ([66], [67]), message / SMS ([13], [12]). Text spam detection is more researched than image spam ([68], [7], [69]), voice spam ([70], [71]), or video spam ([72], [8]) because text processing is much easier.
2. Spam detection based on sources and creators. Spam sources come from social media, articles, news, telecommunications providers, and specific websites. Spreaders of spam content can come from human sources (regular users), robots (user bots) ([5], [73], [74]), or a specific organization that deliberately spreads it.
3. Spam detection based on the targets. It can be grouped into two parts: based on the platform: online [75], and based on the possible risks: role-based user and temporal-based user. Role-based users can be seen based on their age, while temporal-based users can be seen based on the event and time of spreading spam content. Based on online and offline platforms, the most frequent targets are public figures, for example, artists/actors and politicians ([38], [45], [40], [44]).
4. Detection based on the method. It can be divided into two parts: ML and DL.

Popular ML algorithms used in spam detection include NB [33], Complement NB [43], SVM KNN [45], [51], XGBoost [40], Gradient Boosting [76], Decision Tree, and AdaBoost [77]. Meanwhile, popular DL methods include CNN [51], RNN, LSTM [10], Bi-LSTM [25], GRU, and most recently, Transformer [78] such as BERT [53].

There is an increase in research on spam detection topics from year to year, especially in 2019-2020, as shown in Figure 4. This figure proves that much research can be developed on this topic due to the widespread use of technology, the increasing number of Internet users, social media, and the development of spam content detection methods [77]. In addition to Figure 4, this paper also displays the bar chart based on the language used in Figure 5. English is the most widely used language in spam text content detection, which means the opportunity is still open for other languages. In Figure 6, we can see the distribution statistics of articles based on social media platforms used in spam content detection research. In Figure 7, we can see that previous research is not context-based spam detection ("no" is 68%). In Figure 8, we can see that previous research ignores the emoticons features ("excluding emoticons" is 93%). The number of research that is not context-based and not used the emoticon feature is still massive, so further research is still needed on these two things.

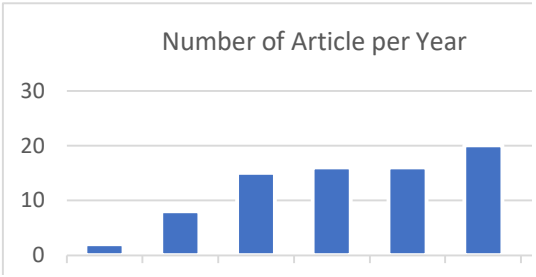

Figure 4: Number of Spam Content Detection Research per Year

This article will discuss the methods/algorithms used, the comparison between methods, the performance, the differences in the methods used, the challenges that still exist, and the possibilities for development, adjustment, and implementation that can still be used done. Additional discussion was also given on handling the detection of spam content in a unique case study of Instagram based on the situation in Indonesia.

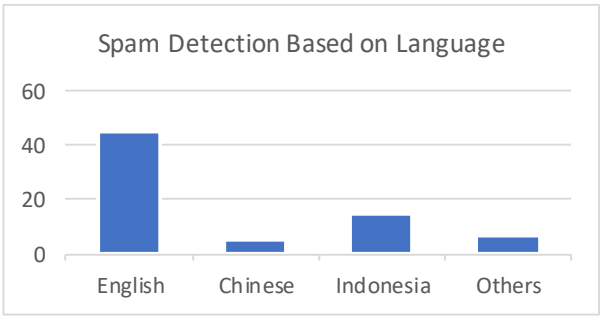

Figure 5: Spam Content Detection Research Based on The Language

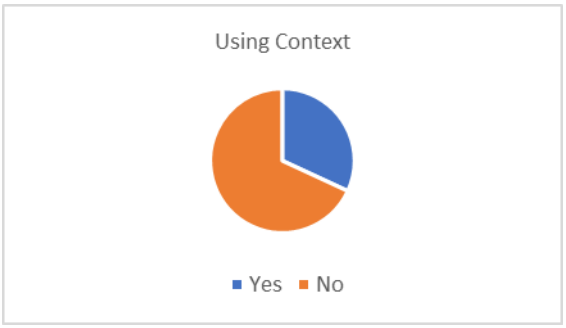

Figure 7: Percentage of Context-based Detection on Spam Content Detection Research

Figure 8: Percentage of Emoji Feature Usage in Spam Content Detection Research

4. RESULT AND DISCUSSION

This section aims to answer some research questions: 1). What are the most widely used means for spreading spam content? In what languages are the most spam content researched? What datasets are used? How is the use of pre-processing methods? What is feature generation & selection used? Moreover, what is the trend of using automatic methods (machine learning and deep learning)? Finally, we analyze spam content detection in the case of social media, the model implementation, research gap, and give some future research and direction.

4.1. The Spam Content Problem

Spam content is a crucial problem to solve. Based on the literature collected, detecting spam content problems in text forms occurs in SMS/message, email, news/articles, posts, and comments made on social media. We will be detailed the spam content problems in the following paragraph.

SMS is one of the most popular text-based messages before being replaced by instant messaging. SMS popularity is proven by the number of SMSes sent per month increasing to 7700% from 2008 to 2018. The price of sending SMS is getting cheaper, and it is more effective than email because 75-80% of emails are not read [10], while SMS will always be opened, especially at night before the sleeping time [12]. However, the more popular SMS is, the more spam SMS content has emerged. According to [12], 68% of mobile phone users experience spam SMS. Spam messages via email and instant messaging such as Whatsapp/Telegram are also increasing. This message usually asks the user to click a link or forward a message to others to become a chain message in exchange for a particular gift [11]. Now, some emails contain messages that are not true that can disrupt the business flow and cause consumers to lose trust in a product because of these spam messages [13].

Websites and social media platforms are also the newest means of spreading spam information. Such spam information spreads quickly in fake news, which causes misleading information ([38], [24]). Apart from spam news/articles on the website, there are also fake/spam accounts on social media such as Twitter [79] and spam content on Twitter in English [62], Turkish [26], and Bangla [78]. Spam comments also appear on Instagram ([37], [43], [80], [38], [81], [39]), or spam comments on Youtube [32]. As displayed in Table 3, 30 articles use case studies of spam posts and comments in SMS, email, article, web, forum, online store, and social media. Social media that are mostly used in case studies are Twitter and Instagram.

Table 3: List of Research-Based on The Problem, Dataset, and Language

The Problem	Dataset	Language	Year	Ref.
SMS and Email Spam	SMS UC Irvine Repository	English	2019	[12]
	SMS UC Irvine Repository	English	2019	[10]
	SMS Spam Collection dari [82]	English	2019	[11]
	SMS UC Irvine Repository	English	2019	[29]
	SMS National University Corpus	English	2020	[13]
	Ling Email Spam	English	2019	[17]
Article / News spam	News dataset (AG's news and Sogou news) [83]	English	2015	[84]
	News article with category of Economy, Health, Technology, Sports, and Politics	Indonesian	2017	[19]
	SMTnews, output from machine translation systems from news document	English	2018	[85]
	From Li Ronghao, Fudan University consists of 8.316 documents	Chinese	2018	[21]
	Chinese news (6000 data)	Chiinese	2018	[23]
	Blog dataset	Inggris	2019	[76]
	News reports	Azerbaijani	2019	[86]
	Newsgroup article	English	2020	[87]
	News dataset	Bangla	2020	[78]
Post/Status spam in Social Media	American public figure's Twitter post (10 million)	English	2016	[88]
	Facebook spam post (1795067 data)	English	2017	[36]
	Twitter post (50.000 spam and 2.632 not spam)	English	2018	[25]
	Twitter post-Honeypot and Spam Post Detection (SPD)	English	2018	[62]
	Twitter post (164.549 Twit written by 74 users)	Turkish	2019	[26]
	Twitter post with 6.000 Twit	English	2019	[27]
	Twitter dataset	English	2019	[89]
	Twitter post and account	English	2020	[28]
	Twitter post contains 58.159 tweets and SMS dataset 5.574 data	English	2020	[29]
Spam comments in Social Media	Spam comments from Instagram (24.602 data)	Indonesian	2016	[40]
	Spam comments from Instagram	English	2017	[37]
	Spam comments from Instagram (14.500 data)	Indonesian	2017	[39]
	Spam comments from Instagram (14.500 data)	Indonesian	2017	[80]
	Komentar spam	English	2018	[31]

	dari YouTube (13000 data)			
	Komentar spam dari YouTube (10000 data)	English	2018	[30]
	Komentar spam dari Instagram (17000 data)	Indonesian	2018	[90]
	Komentar spam dari Instagram (17000 data)	Indonesian	2018	[41]
	Komentar spam pada YouTube (1956 data)	English	2019	[33]
	Spam comments from Instagram (2.600 data)	Indonesian	2019	[43]
	Spam comments from Instagram (14500 data)	Indonesian	2019	[91]
	Spam comments from Instagram (1400 data)	Indonesian	2019	[44]
	Spam comments from YouTube (350 data)	English	2019	[32]
	Spam comments from YouTube (400000 data)	English	2020	[35]
	Spam comments from YouTube	English	2020	[34]
	Spam comments from Instagram (14.500 data)	Indonesian	2020	[45]
Others	Comments form Instagram online store (2.810 data)	Indonesian	2017	[42]
	Review from TripAdvisor	English	2017	[20]
	Comments form online store (300 data)	Indonesian	2020	[46]
	Spam post in the forum	Indian	2019	[92]
	Review from airline web (14460 data)	English	2020	[93]

4.2. Dataset and Language

Based on the dataset used in research, it is interesting to discuss that many researchers collect the datasets by themselves, so it can be said that there are still rarely standard datasets to be used in detection comparison. For example, many Twitter datasets are collected by the researchers using the Twitter API or scraping technique, which certainly takes a very long time ([25], [26], [27], [62]). These researchers also collected Instagram and Youtube datasets by themselves ([38], [43], [33]). The dataset language is mainly in English, but there are some in the other language, such as Indonesian [94], Bangla [78], and Turkish [26]. The lack of standard datasets in a particular language can also make it is challenging to compare.

4.3. Pre-processing, Features Extraction, and Selection

Text in general and social media are remarkably unstructured and need to be converted into a structured form. Pre-processing is the first step that is usually done before the detection process. This pre-processing step is not always carried out nor explained in some research ([12], [25], [28], [37], [40], [57]). Standard pre-processing stages used in the literature are text tokenization, data cleaning, normalization, stemming, lemmatization, and stopwords removal. Tokenization means converting long text into short tokens, which are usually words that are separated by a specific separator. These article ([27], [39], [45], [44], [76], [82], and [83]) used tokenization method. Data cleaning is the stage of cleaning meaningless parts of the text or will not affect computer training data later [10]. Data cleaning steps are removing numbers, punctuation marks, special characters, or emoticons/emojis. ([29], [84]). Stemming is the process of taking root words/essential words to reduce the number of tokens that can be generated. According to [95], stemming refers to a rough heuristic process that cuts off the ends of words in the hope of achieving the goal of finding the root word correctly, including removing derivational affixes. Some stemming algorithms are very dependent on the language used. Lemmatization is similar to stemming, but there are slight differences. Lematisation usually refers to doing something right using vocabulary and morphological analysis of words, usually aimed at removing only inflectional ends and restoring the word's base or dictionary form, known as the lemma ([21], [19]). Stopwords removal is the process of removing tokens that are included in the stopwords data, that are useless words (these researches use the stopwords removal step [10], [29], [19], [43], [39], [45], [44]). Article [20] states that the pre-processing method that has the most effect on performance is the stemming and stop words removal processes.

Normalization changes the tokens in a particular language according to the correct spelling and writing. The normalization method is a unique technique that is still a big challenge in terms of method, speed, and language ([43], [46]). To normalize words, we can use the spelling correction technique. Spelling correction, referred to as typo correction, is done first using a dictionary-based. The development of methods used in spelling correction is as follows: dictionary-based, based on the minimum edit distance search, based on

similarity key (phonetic), rule-based, based on probability, based on word embedding, and using DL, such as encoder-decoder architecture (seq2seq) [96].

Feature extraction is a method for retrieving all features/traits that can be training data in intelligent system algorithms. According to [97], feature extraction methods categorizes 1). based on filtering using word frequency, mutual information, and information gain, 2). based on fusion/combination, such as the weighted KNN, 3). mapping-based, such as using LSI and PCA, 4). cluster-based, such as using the chi-square method and concept indexing, and 5). DL, such as autoencoder, restricted Boltzmann machine, deep belief network, CNN, and RNN.

The feature extraction can be divided into hand-engineered features or automatic features extraction. Hand-engineered features are feature extraction methods that have been predetermined by researchers based on the researcher's knowledge. These features are usually still taken automatically using a program. Some examples of hand-engineered Twitter features are user profile features, account information, and pairwise engagement [57] because researchers think these features are perfect for detecting spam posts. Article [28] also uses quite a lot of hand-engineered features on Twitter, while [37] and [40] use hand-engineered features on Instagram. Automatic feature extraction is an automatic feature retrieval method based on existing text content. This method is often used in DL-based techniques because DL can automatically retrieve token features from the extracted word embedding. Word embedding is generated using Word2Vec (these researches use Word2Vec for features generation [12], [21], [11], [25], [26] [42]), Word2Vec-SM [21], GloVe ([29], [83]), ELMO [98], or Transformer based ([76], [85]). Some tools used in feature extraction are RapidMiner [39], Weka [99], NLTK [11], and Sci-kit Learn from Python. Table 4 summarizes the various pre-processing methods, features extraction, and selection uses. In Table 4, the 'V' sign means if the process is done, while it is empty if it is not done. Column MF = Manual Features, T = Tokenization, S/T = Stemming / Lemmatization, SW = Stopwords removal, BOW = Bag of Words, TFIDF = Term Frequency Inverse Document Frequency, WE = Word Embedding, and E = Emoji. From Table 4, we can see that most of the researches use standard preprocessing steps. However, the latest trend is to use automatic features (embedding) or a combination of features generation and selection. From Table 4, a research gap can also be seen in that most researchers do not use emoji as a feature. Even though in the case of social media, the use of emojis should not be ignored.

Table 4: The Pre-Processing Method, Features Extraction and Selection in Spam Content Detection

REF	MF	T	S / L	SW	BOW	TFIDF	WE	E
[12]	V						V	
[10]	V	V		V		V		
[11]	V	V		V			V	
[29]	V	V		V	V	V	V	
[13]	V	V		V	V			
[17]	V	V	V	V		V		
[84]		V					V	
[19]	V	V	V	V		V		
[85]	V	V		V			V	
[21]	V	V	V	V		V		
[23]		V	V	V	V			
[100]	V	V			V		V	
[86]	V	V		V	V	V		
[88]		V		V			V	V
[101]		V					V	V
[36]	V	V			V			
[25]	V						V	
[62]	V						V	
[26]	V	V		V			V	
[27]	V	V		V				
[102]		V					V	
[28]	V							
[29]		V		V		V	V	
[40]	V							
[37]	V							
[39]	V	V	V	V		V		
[80]	V	V	V	V		V		
[33]	V	V	V					
[31]	V	V			V	V	V	
[103]		V			V	V		
[91]		V	V	V		V		
[41]		V	V	V		V		
[43]	V	V	V	V		V		
[92]		V		V		V		
[44]	V	V	V	V		V		
[32]	V	V	V	V	V			
[35]		V			V			
[45]	V	V	V	V		V		
[42]	V	V	V	V		V	V	
[20]	V	V	V	V		V		
[46]	V	V		V		V		

4.4. Method / Algorithm of Content Spam Detection and The Performances

In NLP and text mining, the steps for processing text data for various tasks can be performed as follows ([104], [105]) data collection, pre-processing, features generation & selection, mining task algorithm, evaluation & analysis, and the last is implementation. The mining task (classification, detection, prediction, categorization) is the core part, how the system can perform its duties

according to its purpose. Evaluation can be seen from the method/algorithm performance in learning and detection. Several performance matrices used in detecting spam content are accuracy, recall, precision, F-measure (F1), loss, Area Under Curve (AUC), and system speed [77]. An example of concrete implementations in the detection of Instagram spam content is implemented using the web services and browser extensions (ex: Firefox/Chrome) because the researchers cannot modify Instagram ([41], [90], [91]). There is also an implementation of spam content detection on mobile application-based short messages using DL in [11] and Chrome-based browser extension on news spam detection [24]. Figure 9 describes the general concept of processing mining tasks for text, and in the next section, we will discuss methods/algorithms in ML and DL, especially spam content detection.

4.4.1. Machine Learning (ML) Based Method

ML methods require excellent pre-processing and feature extraction capabilities so that the system can learn well. ML methods are considered shallow learning techniques, whereas DL is referred to as DL with updates to traditional ML methods [106]. In ML techniques, the mining task algorithm used depends on the case. There are three main mining tasks: classification/ detection, grouping/ clustering, and prediction. The widely used algorithms in spam content detection are NB and SVM ([12], [10], [17], [19], [25], [26], [28], [46], [42], and [84]). NB is a simple method and is derived from the Bayes probability theory. NB has two variants: the multinomial and multivariate Bernoulli model. Article [107] compares five multinomial and multivariate NB variants in text classification: multinomial Bernoulli, multinomial Bernoulli term frequency attributes, multinomial Bernoulli boolean attributes, multivariate Gauss, and flexible NB based on their Receiver Operating Characteristics curve. The result shows that multinomial Bernoulli with boolean attributes is the best, with an average ROC of 97%. SVM is a supervised learning algorithm that learns from data and finds the best hyperplane based on specific functions to separate classes maximally in vector space [42]. SVM uses several parameters that need to be tuned so that the results are better than without tuning.

NB is one of the most widely used and easy-to-use ML algorithms but has lower accuracy than SVM. In spam content detection, performance comparisons between NB and SVM were performed in ([10], [17], [19], [25], [26], [43], [39], [40], [84]). According to [10], SVM achieved an accuracy of 97.81%, beating NB, which was only 80.54%. In contrast, [17] researches that NB defeated SVM in the case study of spam messages. In that article, multinomial NB achieved higher precision and recall levels than SVM when used in news article classification combined with TF-IDF vector-weighting pre-processing techniques. Ban et al. use SVM in Twitter to classify spam. NB, SVM, and several other ML algorithms are compared to deep learning techniques. Based on the results also in [25], it was reported that DL was indeed the most superior, SVM was superior to NB, where NB was in the worst position. The winner of the experiment was Random Forest (RF) [25]. Article [39] states that SVM also has the highest accuracy, recall, precision, and F1 than RF and NB. In detecting spam comments on Instagram Indonesian-language, SVM was also superior to NB. At the same time, in [45], SVM was defeated by KNN and KNN Distance Weight, which reached 91% accuracy in the same case study. Finally, in an article by Qiao et al. (2018), SVM is also higher in performance than NB, but it is still inferior to DL techniques. It can be said that ML methods are good, but they are still inferior to DL.

Article [77] compares some ML algorithms in text classification, namely SVM, Random Forest, AdaBoost, Decision Tree (C4.5), NB, KNN, and Logistic Regression. They tested on 71 datasets that contain classification, detection, or sentiment analysis cases. It reports that SVM is in third place, where RF occupied the second position, and the best is Gradient Boosting (GB) [77]. Extreme Machine Learning is also compared with DL using the Scikit-Learn Python library. Judging from the average performance, GB is the fastest algorithm, but SVM and KNN are the most efficient. The accuracy of DL and ELM is quite good, but the execution time is not good. The DL method is not always best based on these 71 datasets. The study also featured the 11 best classification learning algorithms in Area Under Curve: GB, RF, SVM, ELM, C45, SRC, Logistic Regression, AdaBoost, KNN, NB, and finally, DL. GB ranks first in terms of the best testing efficiency, followed by DL and SVM [77].

Some ML methods are used in text classification and spam detection based on the ML method. The performance of ML methods relies not only on the algorithm but also on the pre-processing and feature generation step as a pipeline. There is also a chance to combine ML algorithms for better performance.

4.4.2. Deep Learning (DL) Method

DL methods can be considered a subset of ML characteristics, with additional characteristics: 1). DL can use massive data, but the increasing rate in its accuracy is proportional to the amount of training data [12], 2). DL uses a concept that mimics the human brain's thinking and comes from the Artificial Neural Network technique, 3). DL requires hardware resources in high specifications than ML, 4). DL does it automatically in feature engineering, while ML still requires complicated manual features [97], 5). DL can be very time-consuming compared to ML, and 6) DL is more challenging to interpret in terms of ease of interpretation because of its automatic learning. DL is still a state-of-the-art method for mining tasks, including detecting spam content based on existing research.

DL is trending techniques right now. Article [108] first introduced DL and continues to be developed until now. Depending on data quality [96], it can automatically take features, has more complex input parameters, has more layers of neurons, and the resulting output can be more diverse. Based on these changes in characteristics, DL requires enormous computational resources. Several studies have shown that the training process using DL requires high hardware specifications ([11], [25], [77], [109]).

Some DL algorithm that widely used in classification & text spam detection are Convolutional Neural Network (CNN) ([12], [29], [42]), Recurrent Neural Network (RNN) [10] and its variants, such as Long Short-Term Memory (LSTM) ([12], [29], [42], [13], [83]), Gated Recurrent Unit (GRU), attention-based LSTM [110], and Bi-LSTM [25], and also traditional Transformer [26] & Bidirectional Encoder Representations from Transformer (BERT) [76]. CNN is a DL model that was first used in image detection. In-text processing, CNN can automatically extract vector text information with these steps [12]: creating word metrics, identifying hidden features from the dataset, and text classification. During its development, the CNN architecture underwent many modifications to improve its performance, such as combined with the attention mechanism and LSTM [111]. The architecture of CNN that is used in text classification can be seen in Figure 10.

RNN is a DL architecture used to process sequential data based on the time-step. RNN is used in machine translation, text summarization, or sentiment classification in text processing. RNN can handle sentence sequences with token (word) processing. However, RNN also has weaknesses, such as 1). processing cannot be done in parallel because the process is done sequentially, 2). the possibility of a vanishing gradient problem, the information from sequences will be lost if the gradient calculation gets smaller and smaller, and 3). It has a long training process [112].

LSTM tries to solve the RNN vanishing gradient problem [113]. LSTM uses various gates in its architecture, including input gates, forget gates, and output gates. Using LSTM, a vanishing gradient does not occur, and the system can forget less critical information. One of the variations of LSTM is Bi-LSTM [114], an LSTM that uses two LSTMs, one taking input in a forward direction and the other taking input in a reverse direction. Bi-LSTM effectively increases the amount of information available to the neural network, increasing the context available for processing (for example, knowing what words immediately follow and precede the words in a sentence). Bi-LSTM takes advantage of LSTM because it can handle forward and backward sequential input.

The previously discussed architectures still have weaknesses, such as 1). they cannot be parallel, 2). there is still the possibility to experience a vanishing gradient, and 3). the training process is slower than the LSTM [115]. Based on these problems, Google Brain created a new DL called Transformer. The Transformer relies only on the attention mechanism [116]. With this architecture, training will be faster, no longer experiencing vanishing gradients, and the process can be done in parallel. Transformer achieves state-of-the-art for Neural Machine Translation processes [115]. The Transformer used in text classification does not use encoder and decoder parts like in NMT, but just the encoder part. The output of the encoder part then passes to a neural network and finally to the softmax function. The Transformer (encoder part) architecture used in text classification can be seen in Figure 11. Based on Table 5, Transformers outperforms ML methods and some DL methods.

DL for classification/ detection/ categorization requires tuning, proper optimization functions, and regularizer functions on the final layer. The activation function used in binary classification tasks usually uses the sigmoid function. For multi-classes, it can use the softmax or ReLU function. The optimization function improves the objective function capability (error function) by maximizing or minimizing the objective function. The

optimization functions commonly used are SGD, RMSprop, Adam, AdaDelta, AdaGrad, Adamax, and Nadam. Article [117] tried to examine Adam, SGD, and Adadelta to classify semantic similarities between two sentences and found that Adam is the best and fastest function. The regularizer function prevents overfitting or underfitting from learning against the dataset.

Based on the comparison of ML and DL methods in Table 5, the average performance of DL methods (accuracy, F1, AUC, ROC, precision, and recall) has better results than ML methods. These results cause many DL methods to be chosen, used, and further developed by researchers today. DL becomes a more promising model and method but still requires performance tuning and adjustment of model implementation in the actual field (production).

Table 5: Methods/Algorithms and Performance on Spam Content Detection in Social Media

Algorithm / Methods	Result / Performance	Dataset	Category	Ref
Gradient boosting, Random forest, Extra trees, MLP, SVM, SVM + MLP	Extra Trees (AUC: 0.986) Random Forest AUC: (0.986) Gradient Boosting (AUC: 0.988) MaxEnt (AUC: 0.93) MLP (AUC: 0.96) SVM (AUC: 0.93)	Twitter spam post	ML	[62]
Transformer Encoder compared with NB, SVM, and Random Forest	Transformer encoder resulting F1-score of 89.3% The accuracy result is 89.4%	Twitter in the Turkish language	DL, Transformers	[26]
Fuzzy Logic	The accuracy result is 74%	Dataset spam article in social media	Fuzzy Logic	[27]
SVM, K Nearest Neighbours, NB, and Random Forest, and ANN	The accuracy of SVM is 0.95 The accuracy of KNN is 0.95 The accuracy of NB is 0.93 Accuracy of RF is 0.96 Accuracy ANN is 0.97	Tweet spam	ML, DL	[28]
DL: CNN, LSTM, BiLSTM, Transformers BERT	BERT Acc: 98%, F1 score 98%, BiLSTM 96%	SMS Spam and Twitter	DL Transformers	[29]
ML methods: NB, SVM, and XGBoost	The usage of fastText increasing model and accuracy, fastening features extraction SVM and XGBoost F1-scores of 0.9601 and 0.9512	Dataset post and public figure in Indonesian's Instagram	ML	[40]
Random Forest	Random Forest, k-fold 10. Accuracy is 96%	Instagram spam post	ML	[37]
NB and SVM	Accuracy is 78.5% with SVM without stemming, and data is unbalanced.	Dataset from artist and actor Indonesia that have follower more than 10 billion on Instagram	ML	[39]
NB	Accuracy is 77% in the balanced class because the dataset is small. In the imbalanced dataset, the accuracy 72%	Dataset from artist and actor Indonesia	ML	[80]
Complementary NB and SVM	Accuracy CNB is 93% with a small dataset, and for the SVM is 87% accuracy SVM is suitable for	Dataset from artist and actor Indonesia	ML	[43]

	the balanced dataset.			
NB	Result: F1-measure = 0.83, recall = 0.98 and precision = 0.72	Instagram comments by scrapping technique	ML	[44]
Distance Weighted KNN compared to KNN	Accuracy of DW-KNN: 0.918 Accuracy of KNN: 0.84	Dataset from artist & actor Indonesia	ML	[45]
SVM and CNN	Accuracy CNN is better than SVM: 84.23%	Dataset Instagram online shop	ML, DL	[42]
NB	Accuracy 80%, precision 0,76 and recall 0,94	Dataset Instagram online shop	ML	[46]

4.5. Spam Content Detection Analysis in Social Media

Social media is a means for users to carry out various kinds of social media's positive impacts, including the ease of making friends, and connections occur virtually with new friends or those who have not met for a long time due to distance. Another good impact is that social media can also be used to promote activities, promote merchandise, and others. User groups that frequently use social media are user groups of public figures (for example, political figures, community leaders, artists, actors, and many more). These public figures become better known, closer / connected to their fans (followers) using social media. All activities of public figures can be easily recognized and followed by fans. For example, in Indonesia, public figures with many fans/followers are artists/actors (known as artists). Artists in Indonesia usually use the social media Facebook. YouTube, Twitter, Instagram, and Tiktok [118]. Instagram is one of the most popular social media used by public figures because of its effectiveness [91]. The more famous an artist is, the more followers he has, and the more spam comments will be in it [31]. Spam comment can disrupts information flow on a particular post/status [38].

Handling spam comments on social media is not easy. To deal with spam comments on social media, we can do it manually, for example filtering specific keywords that indicate spam. However, not all social media have this feature. On Instagram, for example, there is a feature to filter certain words [42] automatically. Still, it can only use English, while [119] uses some [120]. The second way is to make social media private so that other users cannot search, find, or ask for friends. However, this method is impossible in public figures because private accounts are impossible for public figures.

Spam content on social media is so massive that it can be challenging to handle. According to [42], there are several challenging problems such as informal language, it contains emoticons/emojis, there are a lot of abbreviations and typos, and it contains code-mix data. Spam content is very dependent on its status/post (post - comments pair problem), varied length of comments but short (1-3 sentences @ five words), and the structure of posts and comments is reply-response. There is no hierarchy, only mentions using the character '@.'

Detecting spam content on social media depends on its language. For example, spam detection on social media based on the Indonesian language is still rare [19]. The pre-processing stage is essential and dramatically influences spam content detection, especially ML methods. Article [20] stated that the stemming and stopwords removal processes influenced the final mining task. Stemming and stopword removal, of course, really depend on the language used. Apart from stemming and stopword removal, the normalization of text according to the language used to become valid words (has the correct meaning and writing) is also determined by the language. However, the Levenshtein algorithm's normalization method in [121] has not used test data from social media datasets with a very high error rate. In addition to the Levenshtein method, research [122] used the Longest Common Subsequence (LCS) method combined with a dictionary. Finally, the article [123] uses the dataset from Twitter and Facebook in the Indonesian language. This research used the word embedding skip-gram and CBOW method, followed by a Levenshtein and Jaro-Winkler distance score. From [123], it was found that the accuracy of improvement reached 79%. This research result is excellent considering that text on social media contains severe "damage," such as incorrect spelling, abbreviations, typos, using code-mix (mixed language), and many symbols as said in [42].

Instagram has different characteristics from Twitter. Twitter uses the concept of tweet post, reply, and retweet. According to [57], spam content on Twitter usually comes from Twitter posts containing: URL / link that contains irrelevant things, specific irrelevant keywords, a large number

of other usernames, or irrelevant text content. Unlike Twitter, spam content on Instagram pairs between post captions accompanied by photos/images and comments. Most of the spam content on Instagram comes from spam comments, not posts. In the case of the artist, the more famous an artist is, the higher the spam content of comments on each post. The more well-known an artist is, the more followers he has, and the more spam comments will be in it. Comment spam disrupts information flow from comments on a particular post/status [38].

From the literature collected in Table 4, a few researchers use context-based spam detection and scarce uses emoji/emoticon feature, even though the emojis are much found in Instagram posts and comment data. There is also very little research that has been done until the implementation stage ([41], [90], [91]). Article [41] tries to implement the Firefox browser extension that users can use directly but is still limited to desktop browsers and cannot be applied to Instagram mobile using the REST web service [90].

From the various studies that we have been collected, no one has considered/implemented spam content detection on social media, which is treated as a case of data pair detection between posts and spam comments (post-spam comments pair). From this perspective, whether a comment is considered spam depends on the posting caption. DL methods have been widely used to perform sentence-pair problem classification tasks to detect two texts/sentences Semantic textual similarity / semantic sentences relatedness (STS). STS can be used in sentence pairs between question and answer or the relatedness between two sentences ("equivalent," "similar," "specific," "no alignment," "related," "opposite"). Article [117] uses LSTM to achieve 75% accuracy, while research in [111] uses coupled LSTM with better accuracy of 85%.

To implement spam comment detection on social media based on the STS post-comments pair concept, we consider CNN, LSTM, LSTM-Attention, Bi-LSTM Attention, or Transformer based can be used. Based on our knowledge, this approach can be used to detect spam comments that capture the context of the post on social media. This approach can use the alignment features of posting and comments and get the context.

Article [124] created CNN architecture that enhanced with additional attention to detect similarity in sentence pair modeling. In the convolution section, attention is added that takes information from the word embedding in the previous layer so that attentive context and matching vectors can be obtained based on the first and second sentences as their partners. Architecture in [114] has six layers: word embedding, multi-window attention, attentive convolution, max pooling, and similarity measurement, and is continued to fully connected layer and output layer. The results were applied to question-answering English datasets and achieved the highest accuracy of 0.88 on the Quora dataset.

Article [112] said that LSTM is used in accomplishing STS tasks using the coupled LSTM model. Two LSTM parallels are used to retrieve sentence pair information. That research is used two different coupled LSTM architectures: loosely coupled (LC) and tightly coupled (TC) LSTM, and finally made a stack of four LC and TC LSTMs. As a result, the four-stacked TC-LSTM achieved the best accuracy of 85.1. Bi-LSTM is used as part of sentence encoding to detect STS cases combined with weighting to optimize pair-loss in learning sample data, namely SentPWNet (Sentence Pair Weighted Network) [125]. SentPWNet provides two contributions: using local weights to measure the level of information from sentence pairs and learning from complex sentence pair data iteratively until later it converges. The result is that the SentPWNet method combined with Bi-LSTM as a sentence encoder has the highest accuracy compared to regular LSTM and ordinary CNN. Of course, this result is auspicious to be developed further. To measure the similarity between two pairs of sentences, we can use the method in [116] that the Sentence Discrepancy Prediction (SDP) ensemble method reaches the highest macro-F1 0.89. When combined with Transformer, it reaches an F-1 of 0.9.

4.7. Future Research and Challenge

Based on the analysis and discussion that has been carried out in the previous section, there are still many challenges that have been faced and will still be faced in the field of spam content detection on social media. As the final part of this paper, several challenges and future research direction could be addressed, such as:

a. The need for open datasets and gold standard datasets. Judging from the research data previously discussed, it is still rare to find a dataset for system learning to detect spam content for various languages. Based on statistics, most of the datasets are in English and Chinese. The

statistic is undoubted because English and Chinese are the first and second international languages. However, other under-resourced languages are still lacking. Even if there is, it is usually collected manually by researchers, and then it is not shared with the public. If many standard datasets can be used in this field, it can be used as a comparison criterion for a suitable detection method's performance. There are also not many unique datasets available for social media.

b. Normalization, and context-based feature selection research challenges. Unlike algorithms, text normalization requires datasets and dictionaries with specific languages. Not to mention the dictionary for abbreviations that are very context-dependent on the language used. The next difficulty faced on social media is the high rate of writing errors, whether intentional or unintentional, the number of non-standard abbreviations, and slang languages closely related to the times and context.

c. The need for context-based emoji/emoticon features in the following research. Emojis and symbols are widely used on social media. Many users express their posts or comments by using emoticons/emojis only. Most of the researchers throw away and ignore these emoticons from existing studies. We can imagine so much text content discarded/not used if it turns out that more content is written only with symbols and emojis. Several studies started creating vector representations of emoji/emoticons based on the dataset taken from social media Twitter/Facebook and represented in word embedding as research in [88] and [101]. The concept of emoji vectors is essential to be used as a spam content detection feature, especially on social media. The context for detecting spam content is also crucial for successful detection. Much research is still needed regarding using the emoji feature in detecting spam content on social media. Finally, the hypothesis that posting captions & spam comments on social media are considered a problem with semantic sentence pair classification is also promising to continue to be developed to improve spam detection performance.

d. The generalizability of spam content detection methods**.** This problem is general in ML/DL: how an ML system that is already good for specific datasets can also apply to other datasets, especially if the reality is that the data on social media is vast and continuously increasing and changing based on context, time, and trends. For example, several ways can be done by learning supervised learning models and semi-supervised or even unsupervised learning. For example, research using the transfer learning method, such as the BERT architecture, is a promising solution, as seen in articles in [87] and [126]. Learning must be able to be carried out on data stream learning (system learning against data that continues to increase) continuously as in research by [127] and [128] so that the detection system can continue to learn new data.

e. Real implementation model. The challenges in actual implementation and use directly by users are a problem in itself. Many studies only stop testing algorithms with specific datasets and show high results but rarely show that the model can be applied in the field. The development of an intermediary interface is mandatory because the detected system is usually a system originating from other parties such as Twitter, Facebook, Instagram, and Youtube. Several studies have used a browser extension to implement the model with web services as the intermediary. The development of an intermediary system and its implementation are difficulties researchers must consider and continue exploring.

This article also has some drawbacks and limitations. It only focuses on text spam content in supervised learning. There is no specific discussion about the dataset because of the difficulty of searching public standard databases regarding text spam content in multiple languages. The actual implementation case cannot be described clearly because of the limited information from the relevant research source, especially in the Indonesian social media case.

5. CONCLUSION

This paper has tried to collect various research references from 2015 to 2021 with specific scientific methods that can be justified. We collect

the works of literature to seek research topics on social media's up-to-date spam content detection. The discussion is carried out comprehensively in terms of the main problems, the dataset, the pre-processing stages, whether context-based/not, whether emojis/emoticons feature usage/not, the algorithm for detection, the performance of methods, and finally, the analysis & discussion. From the development of existing detection techniques, we discuss that spam content detection can be seen as a post-spam-comments-pair problem to detect it more accurately using deep learning methods, especially hybrid BiLSTM-CNN Attention-based architecture. This approach is a promised new approach to detect relatedness between post and comment and spam comments in the future. This paper provides insights regarding detecting spam content on social media, which can be improved and specially applied in the real world. The development of the algorithmic method used has also been discussed to look for gaps and improve their performance. Finally, the challenges of spam content detection were also discussed, especially a new approach using posting-comments pair to detect spam comments based on posting context. For further research, we will investigate more about the sentence-pair classification model as promised model to apply in the spam content detection on social media, especially to get the context-based spam more accurate.

REFERENCES:

[1] P. Sawers, "The origin of the word 'spam' | THE GOOD WORD," *The Good Word*, 2010. http://www.thegoodword.co.uk/2010/09/20/the-origin-of-the-word-spam/ (accessed Feb. 26, 2021).

[2] InternetSociety, "What Is Spam | Internet Society," *Internet Society*, 2014. https://www.internetsociety.org/resources/doc/2014/what-is-spam/ (accessed Feb. 26, 2021).

[3] S. Geerthik, "Survey on Internet Spam : Classification and Analysis," *Int. J. Comput. Appl. Technol.*, vol. 4, no. June, pp. 384–391, 2013.

[4] E. Convey, "Porn sneaks way back on web," *The Boston Herald*, 1996. https://scholar.google.com/scholar?cluster=12315291433618083 15&hl=en&as_sdt=2005&sciodt=0,5 (accessed Feb. 26, 2021).

[5] N. Saidani, K. Adi, and M. S. Allili, "A semantic-based classification approach for an enhanced spam detection," *Comput. Secur.*, vol. 94, p. 101716, 2020, doi: 10.1016/j.cose.2020.101716.

[6] Z. Guo, L. Tang, T. Guo, K. Yu, M. Alazab, and A. Shalaginov, "Deep Graph neural network-based spammer detection under the perspective of heterogeneous cyberspace," *Futur. Gener. Comput. Syst.*, vol. 117, pp. 205–218, 2021, doi: 10.1016/j.future.2020.11.028.

[7] S. Srinivasan *et al.*, "Deep Convolutional Neural Network Based Image Spam Classification," *Proc. - 2020 6th Conf. Data Sci. Mach. Learn. Appl. CDMA 2020*, pp. 112–117, 2020, doi: 10.1109/CDMA47397.2020.00025.

[8] S. Kanodia, R. Sasheendran, and V. Pathari, "A Novel Approach for Youtube Video Spam Detection using Markov Decision Process," *2018 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2018*, pp. 60–66, 2018, doi: 10.1109/ICACCI.2018.8554405.

[9] J. Coopersmith, *SPAM: A Shadow History of the Internet by Finn Brunton*, Reprint ed., vol. 55, no. 4. The MIT Press, 2014.

[10] A. Chandra and S. K. Khatri, "Spam SMS Filtering using Recurrent Neural Network and Long Short Term Memory," *2019 4th Int. Conf. Inf. Syst. Comput. Networks, ISCON 2019*, pp. 118–122, 2019, doi: 10.1109/ISCON47742.2019.9036269.

[11] A. Zainab, D. Syed, and D. Al-Thani, "Deployment of deep learning models to mobile devices for spam classification," in *Proceedings - 2019 IEEE 1st International Conference on Cognitive Machine Intelligence, CogMI 2019*, 2019, no. CogMI, pp. 112–117, doi: 10.1109/CogMI48466.2019.00024.

[12] P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS Spam," *Futur. Gener. Comput. Syst.*, vol. 102, pp. 524–533, 2020, doi: 10.1016/j.future.2019.09.001.

[13] D. Ganesh Kumar, M. Kameswara Rao, and K. Premnath, "A recurrent neural network model for spam message detection," in *Proceedings of the 5th International Conference on Communication and Electronics Systems, ICCES 2020*, 2020, pp. 1042–1045, doi: 10.1109/ICCES48766.2020.09137940.

[14] T. Xia and X. Chen, "A discrete hidden

Markov model for SMS spam detection," *Appl. Sci.*, vol. 10, no. 14, 2020, doi: 10.3390/app10145011.

[15] S. Bajaj, N. Garg, and S. K. Singh, "A Novel User-based Spam Review Detection," *Procedia Comput. Sci.*, vol. 122, pp. 1009–1015, 2017, doi: 10.1016/j.procs.2017.11.467.

[16] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," *Heliyon*, vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[17] F. Z. Ruskanda, "Study on the Effect of Preprocessing Methods for Spam Email Detection," *Indones. J. Comput.*, vol. 4, no. 1, p. 109, 2019, doi: 10.21108/indojc.2019.4.1.284.

[18] A. Al-Ajeli, R. Alubady, and E. S. Al-Shamery, "Improving spam email detection using hybrid feature selection and sequential minimal optimisation," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 19, no. 1, p. 535, 2020, doi: 10.11591/ijeecs.v19.i1.pp535-542.

[19] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli, and Rudy, "News Article Text Classification in Indonesian Language," *Procedia Comput. Sci.*, vol. 116, pp. 137–143, 2017, doi: 10.1016/j.procs.2017.10.039.

[20] W. Etaiwi and G. Naymat, "The Impact of applying Different Preprocessing Steps on Review Spam Detection," *Procedia Comput. Sci.*, vol. 113, pp. 273–279, 2017, doi: 10.1016/j.procs.2017.08.368.

[21] W. Tian, J. Li, and H. Li, "A Method of Feature Selection Based on Word2Vec in Text Categorization," *Chinese Control Conf. CCC*, vol. 2018-July, pp. 9452–9455, 2018, doi: 10.23919/ChiCC.2018.8483345.

[22] Xuemei Bai, "Text classsification based on LSTM and attention," in *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, 2018, pp. 29–32, doi: 10.1109/ICDIM.2018.8847061.

[23] C. He and Y. Shi, "Research on Chinese spam comments detection based on Chinese characteristics," *2018 IEEE 4th Int. Conf. Comput. Commun. ICCC 2018*, pp. 2608–2612, 2018, doi: 10.1109/CompComm.2018.8781051.

[24] S. R. Sahoo and B. B. Gupta, "Multiple features based approach for automatic fake news detection on social networks using deep learning," *Appl. Soft Comput.*, vol. 100, p. 106983, 2021, doi: 10.1016/j.asoc.2020.106983.

[25] X. Ban, C. Chen, S. Liu, Y. Wang, and J. Zhang, "Deep-learnt features for Twitter spam detection," *2018 Int. Symp. Secur. Priv. Soc. Networks Big Data, Soc. 2018*, pp. 22–26, 2018, doi: 10.1109/SocialSec.2018.8760377.

[26] A. E. Yüksel, Y. A. Türkmen, A. Özgür, and A. B. Altınel, "Turkish tweet classification with transformer encoder," in *International Conference Recent Advances in Natural Language Processing, RANLP*, 2019, vol. 2019-Septe, pp. 1380–1387, doi: 10.26615/978-954-452-056-4_158.

[27] P. Tehlan, R. Madaan, and K. K. Bhatia, "A spam detection mechamism in social media using soft computing," *Proc. 2019 6th Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2019*, pp. 950–955, 2019.

[28] P. Mishra, "Correlated Feature Selection for Tweet Spam Classification," *arXiv*, 2019, [Online]. Available: http://arxiv.org/abs/1911.05495.

[29] R. Ghanem and H. Erbay, "Context-dependent model for spam detection on social networks," *SN Appl. Sci.*, vol. 2, no. 9, pp. 1–8, 2020, doi: 10.1007/s42452-020-03374-x.

[30] A. O. Abdullah, M. A. Ali, M. Karabatak, and A. Sengur, "A comparative analysis of common YouTube comment spam filtering techniques," in *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*, Mar. 2018, pp. 1–5, doi: 10.1109/ISDFS.2018.8355315.

[31] S. Aiyar and N. P. Shetty, "N-Gram Assisted Youtube Spam Comment Detection," *Procedia Comput. Sci.*, vol. 132, pp. 174–182, 2018, doi: 10.1016/j.procs.2018.05.181.

[32] N. Alias, C. F. M. Foozy, and S. N. Ramli, "Video spam comment features selection using machine learning techniques," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 15, no. 2, pp. 1046–1053, 2019, doi: 10.11591/ijeecs.v15.i2.pp1046-1053.

[33] N. M. Samsudin, C. F. B. Mohd Foozy, N. Alias, P. Shamala, N. F. Othman, and W. I. S. Wan Din, "Youtube spam detection

framework using naïve bayes and logistic regression," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, pp. 1508–1517, 2019, doi: 10.11591/ijeecs.v14.i3.pp1508-1517.

[34] P. S. Khodake, "Analysis and Detection of Spam Comments on Social Networking Platforms like YouTube using Machine Learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 8, no. 9, pp. 1280–1282, 2020, doi: 10.22214/ijraset.2020.31677.

[35] R. Abinaya, E. Bertilla Niveda, and P. Naveen, "Spam detection on social media platforms," *2020 7th Int. Conf. Smart Struct. Syst. ICSSS 2020*, pp. 31–33, 2020, doi: 10.1109/ICSSS49621.2020.9201948.

[36] J. Kim, D. Seo, H. Kim, and P. Kang, "Facebook Spam Post Filtering based on Instagram-based Transfer Learning and Meta Information of Posts," *J. Korean Inst. Ind. Eng.*, vol. 43, no. 3, pp. 192–202, 2017, doi: 10.7232/jkiie.2017.43.3.192.

[37] W. Zhang and H.-M. Sun, "Instagram Spam Detection," in *2017 IEEE 22nd Pacific Rim International Symposium on Dependable Computing (PRDC)*, Jan. 2017, pp. 227–228, doi: 10.1109/PRDC.2017.43.

[38] A. Chrismanto and Y. Lukito, "Klasifikasi Komentar Spam Pada Instagram Berbahasa Indonesia Menggunakan K-NN," in *Seminar Nasional Teknologi Informasi Kesehatan (SNATIK)*, 2017, pp. 298–306.

[39] A. Chrismanto and Y. Lukito, "Identifikasi Komentar Spam Pada Instagram," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 8, no. 3, p. 219, 2017, doi: 10.24843/lkjiti.2017.v08.i03.p08.

[40] A. A. Septiandri and O. Wibisono, "Detecting spam comments on Indonesia's Instagram posts," *J. Phys. Conf. Ser.*, vol. 801, no. 012069, pp. 1–7, 2017, doi: 10.1088/1742-6596/755/1/011001.

[41] A. Chrismanto, W. Raharjo, and Y. Lukito, "Design and Development of REST-Based Instagram Spam Detector for Indonesian Language," *Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018*, pp. 345–350, Sep. 2018, doi: 10.1109/ISEMANTIC.2018.8549725.

[42] F. Prabowo and A. Purwarianti, "Instagram online shop's comment classification using statistical approach," in *Proceedings - 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2017*, 2018, pp. 282–287, doi: 10.1109/ICITISEE.2017.8285512.

[43] N. A. Haqimi, N. Rokhman, and S. Priyanta, "Detection Of Spam Comments On Instagram Using Complementary Naïve Bayes," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 13, no. 3, p. 263, Jul. 2019, doi: 10.22146/ijccs.47046.

[44] B. Priyoko and A. Yaqin, "Implementation of naive bayes algorithm for spam comments classification on Instagram," in *2019 International Conference on Information and Communications Technology, ICOIACT 2019*, 2019, pp. 508–513, doi: 10.1109/ICOIACT46704.2019.8938575.

[45] A. Chrismanto, Y. Lukito, and A. Susilo, "Implementasi Distance Weighted K-Nearest Neighbor Untuk Klasifikasi Spam dan Non-Spam Pada Komentar Instagram," *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 2, p. 236, 2020, doi: 10.26418/jp.v6i2.39996.

[46] R. L. Musyarofah, E. U. Utami, and S. R. Raharjo, "Analisis Komentar Potensial pada Social Commerce Instagram Menggunakan TF-IDF," *J. Eksplora Inform.*, vol. 9, no. 2, pp. 130–139, 2020, doi: 10.30864/eksplora.v9i2.360.

[47] K. Hines, "How to Identify and Control Blog Comment Spam on WordPress," *Neil Patel*, 2021. https://neilpatel.com/blog/control-blog-comment-spam/ (accessed Feb. 27, 2021).

[48] G. Mishne, D. Carmel, and R. Lempel, "Blocking blog spam with language model disagreement," in *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web, AIRWeb 2005 - Held in Conjunction with the 14th International World Wide Web Conference*, 2005, no. May, pp. 1–6.

[49] N. Spirin and J. Han, "Survey on Web Spam Detection: Principles and Algorithms," *SIGKDD Explor. Newsl.*, vol. 13, no. 2, pp. 50–64, 2012, doi: 10.1145/2207243.2207252.

[50] A. Khatun, M. H. Matin, A. Miah, and R. Miah, "Comparative Study on Text Classification," *Int. J. Eng. Sci. Invent.*, vol. 9, no. 9, pp. 21–33, 2020, doi: 10.35629/6734-0909012133.

[51] G. M. Shahariar, S. Biswas, F. Omar, F. M. Shah, and S. Binte Hassan, "Spam Review Detection Using Deep Learning," *2019 IEEE 10th Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEMCON 2019*, pp. 27–33, 2019, doi: 10.1109/IEMCON.2019.8936148.

[52] S. Saumya and J. P. Singh, "Spam review detection using LSTM autoencoder: an unsupervised approach," *Electron. Commer. Res.*, vol. 20, no. 0123456789, 2020, doi: 10.1007/s10660-020-09413-4.

[53] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[54] H. Snyder, "Literature review as a research methodology: An overview and guidelines," *J. Bus. Res.*, vol. 104, no. August, pp. 333–339, 2019, doi: 10.1016/j.jbusres.2019.07.039.

[55] M. Alsaleh, A. Alarifi, F. Al-Quayed, and A. Al-Salman, "Combating comment spam with machine learning approaches," *Proc. - 2015 IEEE 14th Int. Conf. Mach. Learn. Appl. ICMLA 2015*, pp. 295–300, 2016, doi: 10.1109/ICMLA.2015.192.

[56] A. Talha and R. Kara, "A Survey of Spam Detection Methods on Twitter," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 3, pp. 29–38, 2017, doi: 10.14569/ijacsa.2017.080305.

[57] T. Wu, S. Wen, Y. Xiang, and W. Zhou, "Twitter spam detection: Survey of new approaches and comparative study," *Comput. Secur.*, vol. 76, pp. 265–284, Jul. 2018, doi: 10.1016/j.cose.2017.11.013.

[58] T. Poonkodi and S. Sukumaran, "A SURVEY ON FEATURE SELECTION AND MACHINE LEARNING ALGORITHMS FOR SPAM DETECTION," *J. Inf. Comput. Sci.*, vol. 13, no. 10, pp. 59–68, 2018.

[59] S. Rao, A. K. Verma, and T. Bhatia, "A review on social spam detection: Challenges, open issues, and future directions," *Expert Syst. Appl.*, vol. 186, no. March, p. 115742, 2021, doi: 10.1016/j.eswa.2021.115742.

[60] D. Moher *et al.*, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *PLoS Med.*, vol. 6, no. 7, 2009, doi: 10.1371/journal.pmed.1000097.

[61] R. Krithiga and D. R. E. Ilavarasan, "Machine learning techniques for spammer identification: State of the art and analysis," *J. Crit. Rev.*, vol. 7, no. 1, pp. 446–448, 2020, doi: 10.31838/jcr.07.01.87.

[62] I. Inuwa-Dutse, M. Liptrott, and I. Korkontzelos, "Detection of spam-posting accounts on Twitter," *Neurocomputing*, vol. 315, pp. 496–511, Nov. 2018, doi: 10.1016/j.neucom.2018.07.044.

[63] D. Gunawan, R. F. Rahmat, A. Putra, and M. F. Pasha, "Filtering Spam Text Messages by Using Twitter-LDA Algorithm," *2018 IEEE Int. Conf. Commun. Networks Satell. Comnetsat 2018 - Proc.*, pp. 1–6, 2018, doi: 10.1109/COMNETSAT.2018.8684085.

[64] M. Chakraborty, S. Pal, R. Pramanik, and C. Ravindranath Chowdary, "Recent developments in social spam detection and combating techniques: A survey," *Inf. Process. Manag.*, vol. 52, no. 6, pp. 1053–1073, Nov. 2016, doi: 10.1016/j.ipm.2016.04.009.

[65] M. K. Sohrabi and F. Karimi, "A Feature Selection Approach to Detect Spam in the Facebook Social Network," *Arab. J. Sci. Eng.*, vol. 43, no. 2, pp. 949–958, Feb. 2018, doi: 10.1007/s13369-017-2855-x.

[66] K. Chowdhury, "Spam Identification on Facebook , Twitter and Email using Machine Learning," *Cent. Eur. Res. J.*, vol. 6, no. 1, pp. 18–26, 2020.

[67] T. Verma and N. S. Gill, "Email Spams via Text Mining using Machine Learning Techniques," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 4, pp. 2535–2539, 2020, doi: 10.35940/ijitee.d1915.029420.

[68] S. Khawandi, F. Abdallah, and A. Ismail, "A Survey On Image Spam Detection Techniques," in *3rd International Conference on Computer Science and Information Technology (COMIT 2019)*, 2019, no. January, pp. 13–27, doi: 10.5121/csit.2019.90102.

[69] C. Y. Su, D. F. Shen, and G. S. Lin, "An image spam detection method," *2017 IEEE Int. Conf. Consum. Electron. - Taiwan, ICCE-TW 2017*, pp. 71–72, 2017, doi: 10.1109/ICCE-China.2017.7991000.

[70] G. Vennila and M. S. K. Manikandan,

"Detection of Human and Computer Voice Spammers Using Hidden Markov Model in Voice over Internet Protocol Network," *Procedia Comput. Sci.*, vol. 115, pp. 588–595, 2017, doi: 10.1016/j.procs.2017.09.169.

[71] M. Swarnkar and N. Hubballi, " SpamDetector : Detecting spam callers in Voice over Internet Protocol with graph anomalies ," *Secur. Priv.*, vol. 2, no. 1, p. e54, 2019, doi: 10.1002/spy2.54.

[72] Y. Yusof and O. H. Sadoon, "Detecting Video Spammers in Youtube Social Media," in *ICOCI Kuala Lumpur. Universiti Utara Malaysia*, 2017, pp. 228–234, [Online]. Available: http://www.uum.edu.my.

[73] G. Lingam, R. R. Rout, D. V. L. N. Somayajulu, and S. K. Ghosh, "Particle Swarm Optimization on Deep Reinforcement Learning for Detecting Social Spam Bots and Spam-Influential Users in Twitter Network," *IEEE Syst. J.*, pp. 1–12, 2020, doi: 10.1109/JSYST.2020.3034416.

[74] M. Orabi, D. Mouheb, Z. Al Aghbari, and I. Kamel, "Detection of Bots in Social Media: A Systematic Review," *Inf. Process. Manag.*, vol. 57, no. 4, p. 102250, 2020, doi: 10.1016/j.ipm.2020.102250.

[75] H. Gupta, M. S. Jamal, S. Madisetty, and M. S. Desarkar, "A framework for real-time spam detection in Twitter," *2018 10th Int. Conf. Commun. Syst. Networks, COMSNETS 2018*, vol. 2018-Janua, no. January, pp. 380–383, 2018, doi: 10.1109/COMSNETS.2018.8328222.

[76] M. Li, B. Wu, and Y. Wang, "Comment Spam Detection via Effective Features Combination," 2019, doi: 10.1109/ICC.2019.8761340.

[77] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, "An up-to-date comparison of state-of-the-art classification algorithms," *Expert Syst. Appl.*, vol. 82, pp. 128–150, 2017, doi: 10.1016/j.eswa.2017.04.003.

[78] T. Alam, A. Khan, and F. Alam, "Bangla Text Classification using Transformers," *arXiv*, Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.04446.

[79] M. McCord and M. Chuah, "Spam Detection on Twitter Using Traditional Classifiers," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, 2011, pp. 175–186.

[80] A. Chrismanto and Y. Lukito, "Deteksi Komentar Spam Bahasa Indonesia Pada Instagram Menggunakan Naive Bayes," *J. Ultim.*, vol. 9, no. 1, pp. 50–58, 2017, doi: 10.31937/ti.v9i1.564.

[81] I. Wicaksono, "Sistem identifikasi komentar negatif pada instagram," Surakarta, 2020. [Online]. Available: http://eprints.ums.ac.id/82841/7/Naskah Publikasi - Upload Perpus.pdf.

[82] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the study of SMS spam filtering: New collection and results," *DocEng 2011 - Proc. 2011 ACM Symp. Doc. Eng.*, no. January, pp. 259–262, 2011, doi: 10.1145/2034691.2034742.

[83] X. Zhang, J. Z. Zhao, and Y. LeCun, "Character-level Convolutional Networks for TextClassification," in *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015, pp. 649–657, [Online]. Available: https://dl.acm.org/doi/10.5555/2969239.2969312.

[84] C. Qiao *et al.*, "A new method of region embedding for text classification," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018, no. 2012, pp. 1–12.

[85] L. Zhang and D. Moldovan, "Rule-based vs. Neural Net Approaches to Semantic Textual Similarity," *Proc. First Work. Linguist. Resour. Nat. Lang. Process.*, pp. 12–17, 2018, [Online]. Available: https://www.aclweb.org/anthology/W18-3803.

[86] U. Suleymanov, B. K. Kalejahi, E. Amrahov, and R. Badirkhanli, "Text Classification for Azerbaijani Language," *Comput. Syst. Sci. Eng.*, vol. 35, no. 6, pp. 467–475, 2018, doi: 10.32604/csse.2020.35.467.

[87] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv*, vol. 21, pp. 1–67, 2019.

[88] F. Barbieri, F. Ronzano, and H. Saggion, "What does this emoji mean? A vector space skip-gram model for twitter emojis,"

in *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 2016, pp. 3967–3972.

[89] G. Jain, M. Sharma, and B. Agarwal, "Spam detection in social media using convolutional and long short term memory neural network," *Ann. Math. Artif. Intell.*, vol. 85, no. 1, pp. 21–44, 2019, doi: 10.1007/s10472-018-9612-z.

[90] A. R. Chrismanto, W. Sudiarto, and Y. Lukito, "Integration of REST-Based Web Service and Browser Extension for Instagram Spam Detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 12, 2018, doi: 10.14569/IJACSA.2018.091253.

[91] A. Chrismanto, W. Raharjo, and Y. Lukito, "Firefox Extension untuk Klasifikasi Komentar Spam pada Instagram Berbasis REST Services," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 2, p. 146, 2019, doi: 10.26418/jp.v5i2.33010.

[92] P. Ratadiya and R. Moorthy, "Spam filtering on forums: A synthetic oversampling based approach for imbalanced data classification," *arXiv*, Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.04826.

[93] M. A. Ullah, S. M. Marium, S. A. Begum, and N. S. Dipa, "An algorithm and method for sentiment analysis using the text and emoticon," *ICT Express*, no. xxxx, pp. 10–13, 2020, doi: 10.1016/j.icte.2020.07.003.

[94] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," *arXiv*, Sep. 2020, [Online]. Available: https://www.aclweb.org/anthology/2020.aacl-main.85.

[95] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*, 1st editio. Cambridge: Cambridge University Press, 2008.

[96] S. Ahmadi, "Attention-based Encoder-Decoder Networks for Spelling and Grammatical Error Correction," *arXiv*, Sep. 2018, [Online]. Available: http://arxiv.org/abs/1810.00660.

[97] H. Liang, X. Sun, Y. Sun, and Y. Gao, "Text feature extraction based on deep learning: a review," *Eurasip J. Wirel. Commun. Netw.*, vol. 2017, no. 1, pp. 1–12, 2017, doi: 10.1186/s13638-017-0993-1.

[98] M. E. Peters, M. Neumann, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings ofNAACL-HLT 2018*, 2018, pp. 2227–2237, [Online]. Available: http://allennlp.org/elmo.

[99] S. Sharmin and Z. Zaman, "Spam detection in social media employing machine learning tool for text mining," in *Proceedings - 13th International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2017*, 2018, pp. 137–142, doi: 10.1109/SITIS.2017.32.

[100] M. Li, B. Wu, and Y. Wang, "Comment Spam Detection via Effective Features Combination," *IEEE Int. Conf. Commun.*, vol. 2019-May, 2019, doi: 10.1109/ICC.2019.8761340.

[101] B. Eisner, I. Augenstein, T. Rockt, and S. Riedel, "emoji2vec : Learning Emoji Representations from their Description," 2016.

[102] G. Jain, M. Sharma, and B. Agarwal, "Spam detection in social media using convolutional and long short term memory neural network," *Ann. Math. Artif. Intell.*, vol. 85, no. 1, pp. 21–44, Jan. 2019, doi: 10.1007/s10472-018-9612-z.

[103] A. O. Abdullah, M. A. Ali, M. Karabatak, and A. Sengur, "A comparative analysis of common YouTube comment spam filtering techniques," *6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding*, vol. 2018-Janua, pp. 1–5, 2018, doi: 10.1109/ISDFS.2018.8355315.

[104] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*, 1st ed. Cambridge University Press, 2007.

[105] S. M. Weiss, N. Indurkhya, T. Zhang, and F. J. Damerau, *Text mining: Predictive methods for analyzing unstructured information*. Springer New York, 2005.

[106] Suyanto;, K. N. Ramadhani, and Mandala; Satria, *Big Data, Deep Learning Modernisasi Machine Learning untuk Big*, 1st ed. Bandung: Penerbit Informatika, 2019.

[107] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with Naive Bayes - Which Naive Bayes?," 2006.

[108] Geoffrey E. Hinton;, S. Osindero;, and Y.-W. Teh;, "A fast learning algorithm for

deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006, doi: 10.1162/neco.2006.18.7.1527.

[109] R. Chowdhury, K. G. Das, B. Saha, and S. K. Bandyopadhyay, "A Method Based on NLP for Twitter Spam detection," *Preprints*, no. July, 2020, doi: 10.20944/preprints202007.0648.v1.

[110] X. Sun and W. Lu, "Understanding Attention for Text Classification," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, no. 1999, pp. 3418–3428, doi: 10.18653/v1/2020.acl-main.312.

[111] Q. Fu, C. Wang, and X. Han, "A CNN-LSTM network with attention approach for learning universal sentence representation in embedded system," *Microprocess. Microsyst.*, vol. 74, 2020, doi: 10.1016/j.micpro.2020.103051.

[112] P. Liu, X. Qiu, and H. Xuanjing, "Recurrent neural network for text classification with multi-task learning," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2016-Janua, pp. 2873–2879, 2016.

[113] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[114] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[115] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[116] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.

[117] A. Besti, R. Ilyas, F. Kasyidi, and E. C. Djamal, "Semantic classification of scientific sentence pair using recurrent neural network," *Int. Conf. Electr. Eng. Comput. Sci. Informatics*, vol. 2020-Octob, pp. 150–155, 2020, doi: 10.23919/EECSI50503.2020.9251897.

[118] Databooks, "Ini Media Sosial Paling Populer Sepanjang April 2020," *Databooks*, 2020. https://databoks.katadata.co.id/datapublish/2020/05/25/ini-media-sosial-paling-populer-sepanjang-april-2020 (accessed Nov. 04, 2020).

[119] Instagram, "How do I filter out and hide comments I don't want to appear on my posts on Instagram? | Instagram Help Centre," *Instagram Help*, 2020. https://www.facebook.com/help/instagram/700284123459336 (accessed Mar. 06, 2021).

[120] R. Burns, "188 Spam Words to Avoid: How to Stay Out of Spam Filters," *Active Campaign*, 2019. https://www.activecampaign.com/blog/spam-words (accessed Mar. 06, 2021).

[121] V. Mawardi, N. Susanto, and D. Naga, "Spelling Correction for Text Documents in Bahasa Indonesia Using Finite State Automata and Levinshtein Distance Method," *MATEC Web Conf.*, vol. 164, p. 01047, Apr. 2018, doi: 10.1051/matecconf/201816401047.

[122] D. Gunawan, Z. Saniyah, and A. Hizriadi, "Normalization of abbreviation and acronym on microtext in Bahasa Indonesia by using dictionary-based and longest common subsequence (LCS)," *Procedia Comput. Sci.*, vol. 161, pp. 553–559, 2019, doi: 10.1016/j.procs.2019.11.155.

[123] R. P. Kusumawardani, S. Priansya, and F. J. Atletiko, "Context-sensitive normalization of social media text in bahasa Indonesia based on neural word embeddings," *Procedia Comput. Sci.*, vol. 144, pp. 105–117, 2018, doi: 10.1016/j.procs.2018.10.510.

[124] S. Xu, S. E, and Y. Xiang, "Enhanced attentive convolutional neural networks for sentence pair modeling," *Expert Syst. Appl.*, vol. 151, p. 113384, 2020, doi: 10.1016/j.eswa.2020.113384.

[125] L. Zhang, H. Wang, and L. Li, "SentPWNet: A Unified Sentence Pair Weighting Network for Task-specific Sentence Embedding," *Arxiv*, May 2020, [Online]. Available: https://arxiv.org/abs/2005.11347.

[126] M. Wang, H. Yang, Y. Qin, S. Sun, and Y. Deng, "Unified Humor Detection Based on Sentence-pair Augmentation and Transfer Learning," *Proc. 22nd Annu. Conf. Eur. Assoc. Mach. Transl.*, pp. 53–59, 2020, [Online]. Available: https://www.aclweb.org/anthology/2020.ea

mt-1.7.

[127] L. H. Ru, T. Andromeda, and M. N. Marsono, "Online data stream learning and classification with limited labels," *Int. Conf. Electr. Eng. Comput. Sci. Informatics*, vol. 1, no. August, pp. 161–164, 2014, doi: 10.11591/eecsi.1.366.

[128] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak, "Ensemble learning for data stream analysis: A survey," *Inf. Fusion*, vol. 37, pp. 132–156, Sep. 2017, doi: 10.1016/j.inffus.2017.02.004.

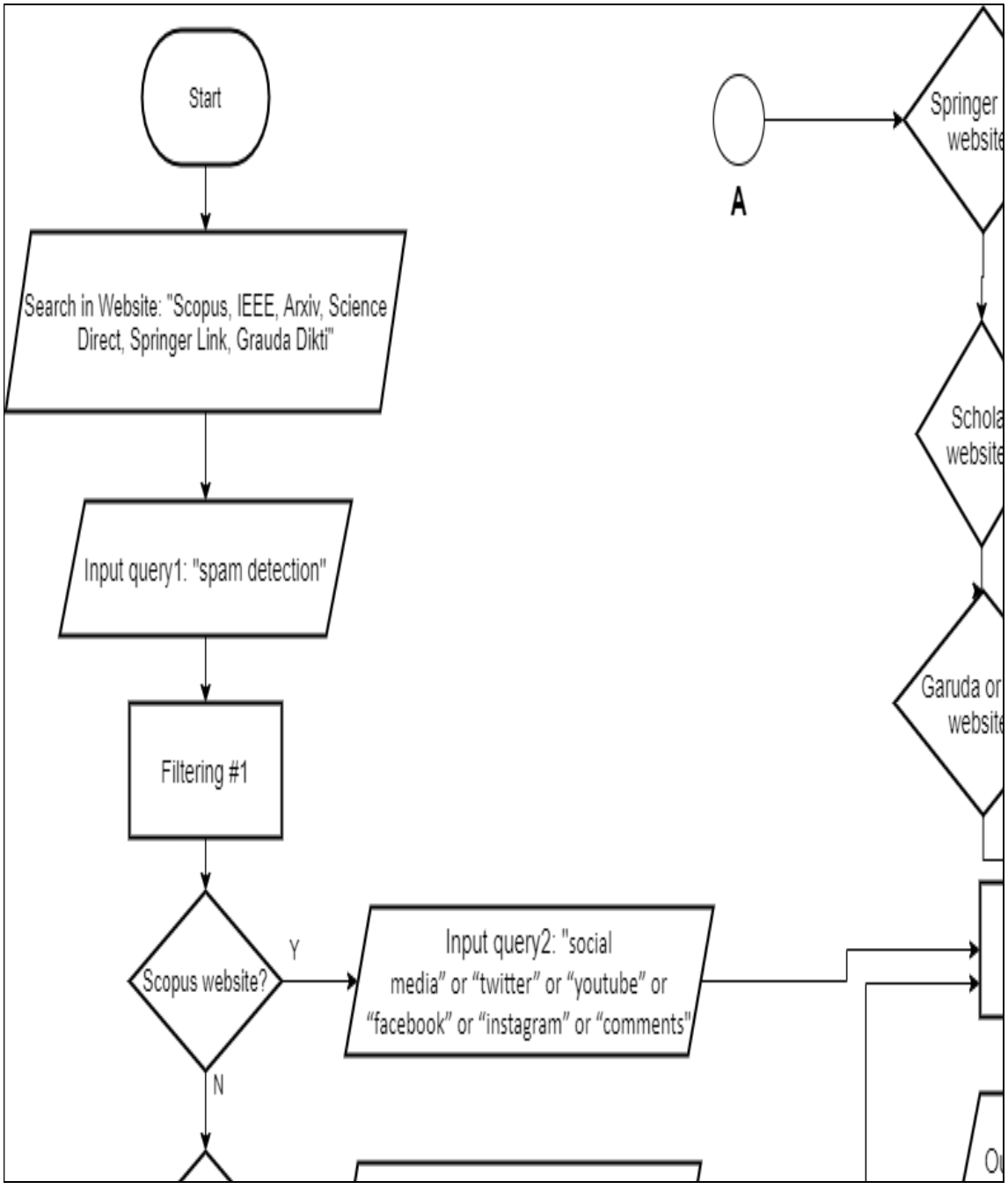

Figure 1: Filtering Method of Literature Search

Journal of Theoretical and Applied Information Technology
30th April 2022. Vol.100. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

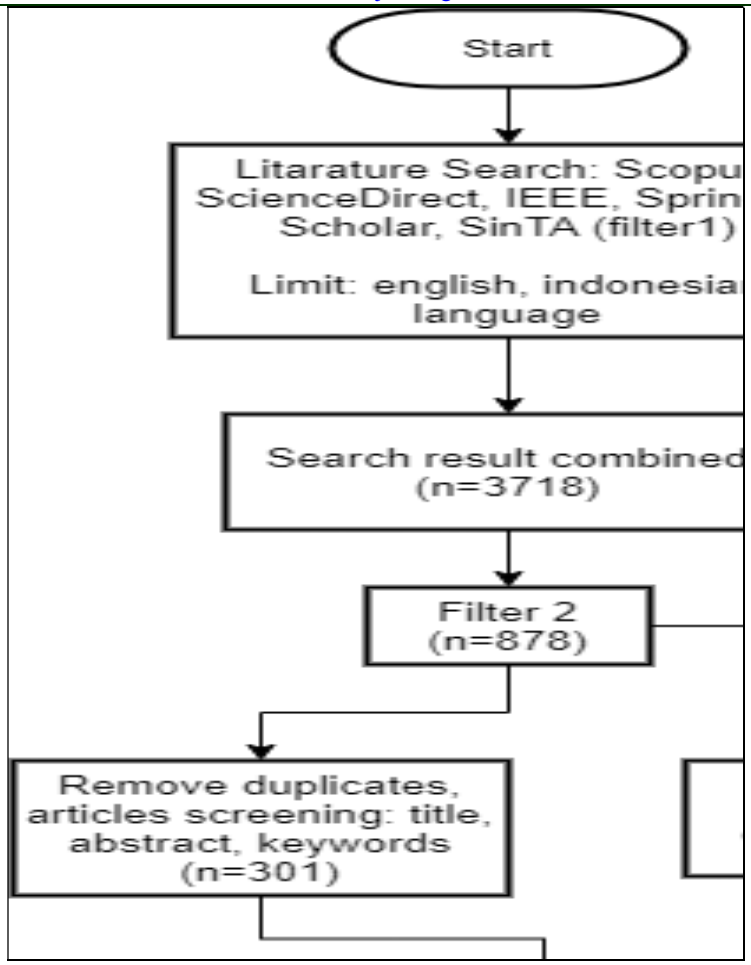

Figure 2: Literature Review Flow

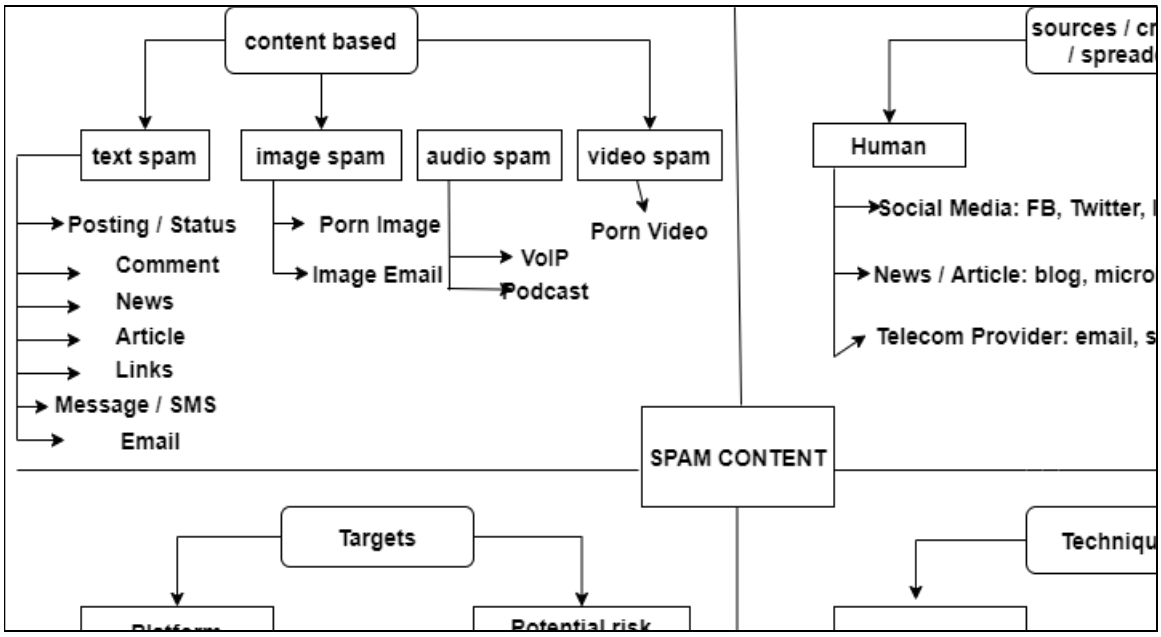

Figure 3: Spam Content Detection Research Categorization

Journal of Theoretical and Applied Information Technology
30th April 2022. Vol.100. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

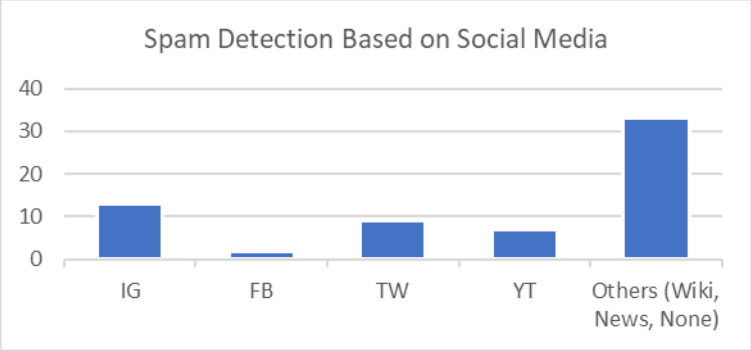

Figure 6: Social Meda Usage in Spam Detection Research

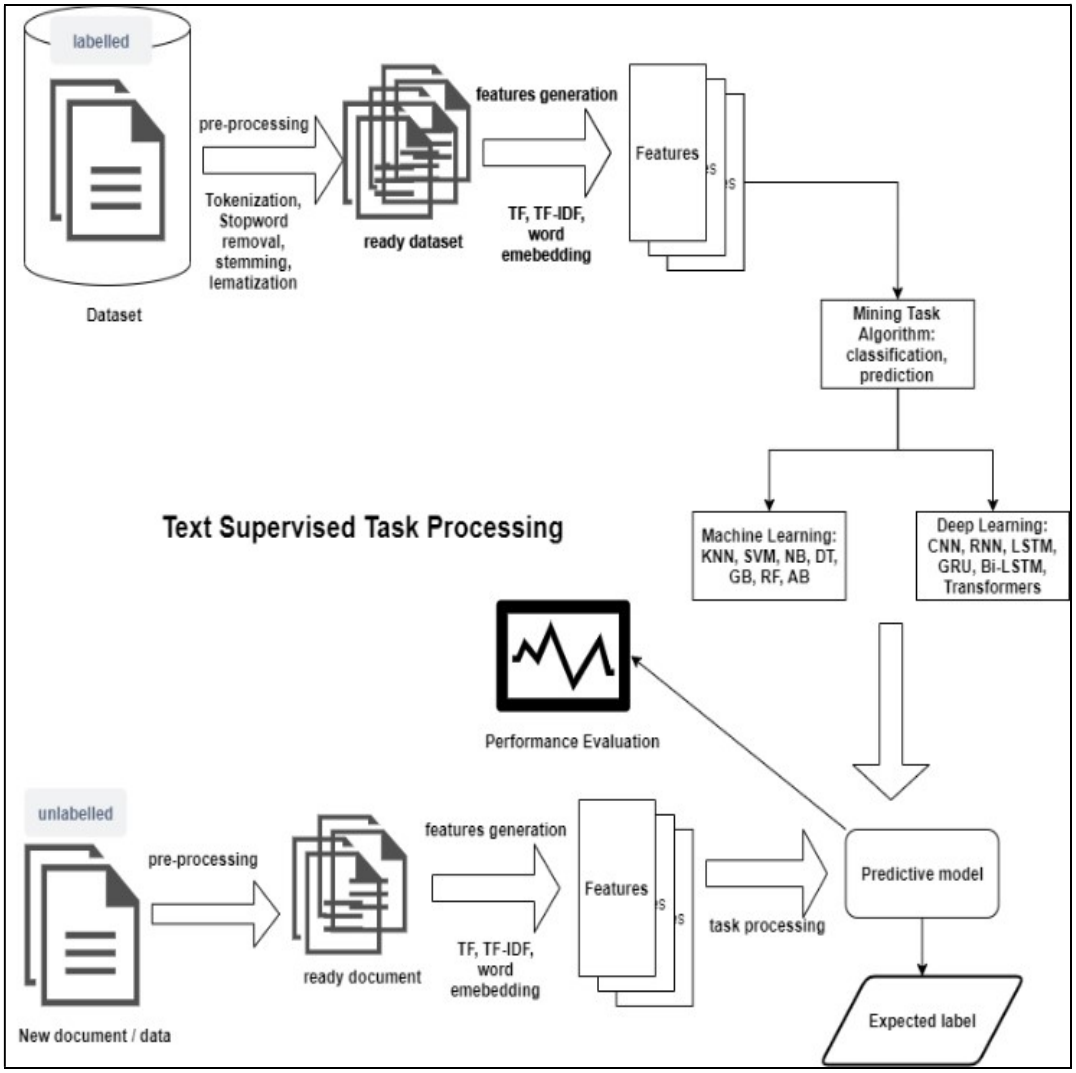

Figure 9: General Concept of Processing Text Data in Supervised Learning

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

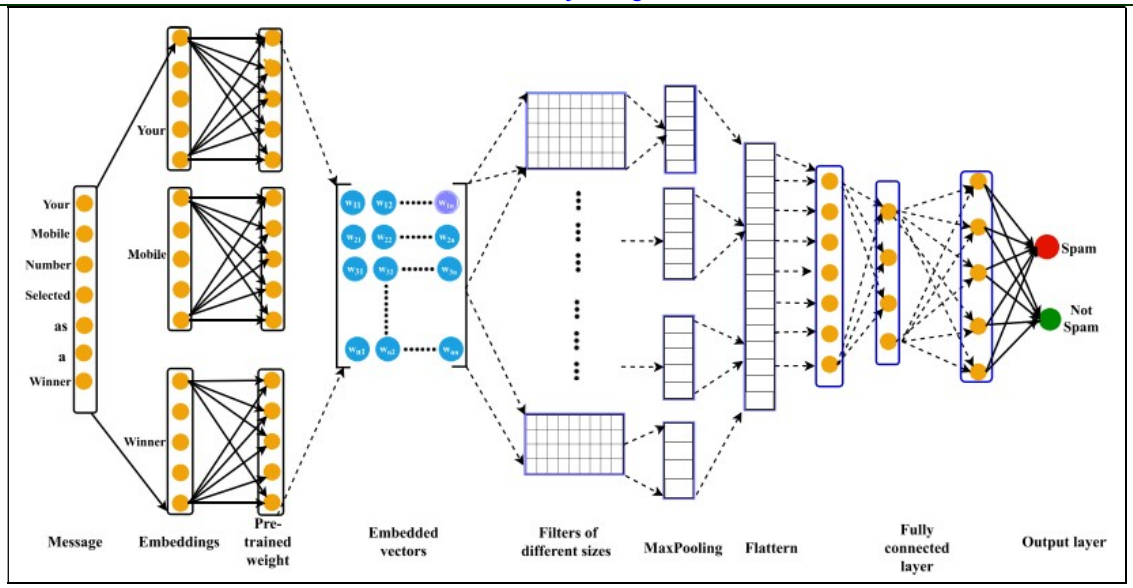

Figure 10: CNN Architecture for Spam Content Detection [10].

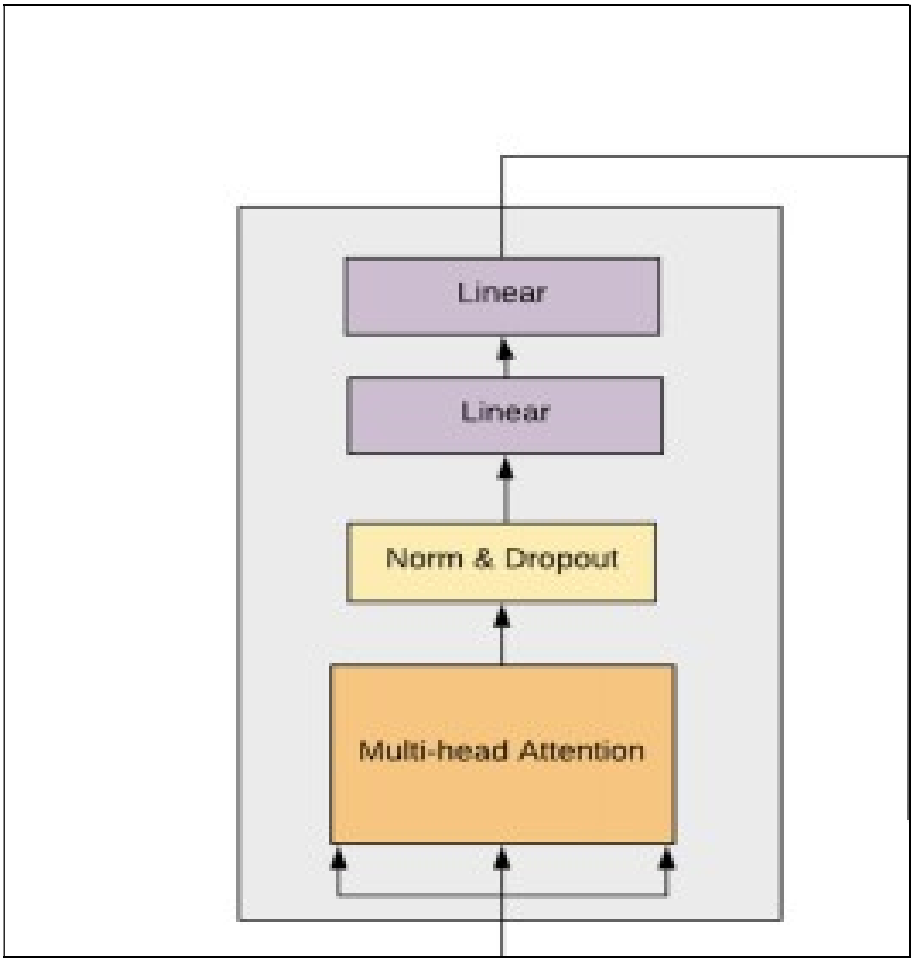

Figure. 11: Transformer Encoder Architecture for Text Classification [26]