Data Mining Regarding Cyberbullying in the Arabic Language on Instagram Using KNIME and Orange Tools

Shumaa Saeed Alzahrani
Computer Science and Engineering Department
College of Computers and Information Systems
Umm Al-Qura University
Makkah, Saudi Arabia
shayma.s.s.alzahrani@gmail.com

Received: 20 July 2022 | Revised: 15 August 2022 | Accepted: 22 August 2022

Abstract-This paper deals with data mining on verbal bullying by Instagram users. It tracks people who repeatedly have abusive behavior and may cause harm to other persons or groups. In this work, a dataset holding verbal bullying in the Arabic language was extracted from Instagram comments, and the entries were classified as regular verbal bullying and suspicious verbal bullying. KINIME and Orange open source data mining tools were utilized to discover comments that involved verbal bullying on Instagram and to delete previous comments while users sent their comments automatically and immediately. Classification algorithms Rule-Based in KNIME and Select Rows in Orange were used.

Keywords-KNIME tool; Orange tool; Instagram; data mining; Instagram comments; cyberbullying; verbal bullying

I. INTRODUCTION

Social media platforms have developed rapidly during the last few years. One of them is the Instagram. It is free, contains online data, and provides an easy form of communication. Users can talk in private and upload, tag, like, comment on, and share posts. Machine learning powers the app. Instagram's feed ranking is constantly adapting and improving based on new data [1]. The Instagram algorithm predicts how much you care about a post. This way, one can find and classify bullies based on trending [1]. Also, the people who use offensive of abusive words can be identified with data mining tools.

A. *How the Instagram Algorithm Works*

Six key factors influence the Instagram algorithm for feed posts: interest, relationship, timeliness, frequency, following, and usage. The Instagram algorithm is constantly changing. The more the Instagram algorithm "likes" a post, the higher it will appear in your feed. This phenomenon is based on the "past behavior on similar content and possibly machine vision analyzing the post's actual content." What you see in your Instagram feed is a mixture of all your behaviors on Instagram [1]: the people you communicate with, the stories you watch, the individuals you are tagged with, and the topics you comment and like. Comments, likes, reshares, and views are the most critical engagements for feed rating, which is beneficial when you prepare content and captions [1].

B. *Comment of All Lengths Count as Engagement*

The Instagram algorithm counts comments that are less than 3 words in length. Instagram comments have become essential sources of knowledge for making fast and informed decisions and understanding how people behave in the real world. National and human rights organizations are now tracking social media [2].

C. *Motivation*

Bullying includes repetitive and violent physical, verbal, or emotional actions. In this paper, the target goal is verbal bulling, which includes calling names, mocking, taunting, threatening, or verbally assaulting. Bullying can make you feel powerless, ashamed, depressed, or even suicidal. Detecting bullying can assist authorities in taking appropriate action by copying data, deleting them from public comments, or imposing fines. Using social media sites as data providers may be an effective mechanism for protecting ourselves.

Due to the evolving technology, bullying is no longer confined to schoolyards or street corners but can happen at home via phone calls, texts, emails, and social media. Cyberbullies stalk, attack, or humiliate victims using digital technology. Cyberbullying, unlike conventional bullying, does not involve face-to-face contact and is not limited to a few people at a time. It also does not necessitate physical strength or many bullies. The embarrassment can be shared by hundreds or thousands of people online with just a few clicks. Cyberbullying may involve sending threatening or degrading messages via text, email, social media posts, or instant messaging, as well as breaking into an email account or stealing someone's online identity. Some cyberbullies may set up a website or a social media account to harass a victim. The approaches used to cyberbully are as diverse and creative as the technologies accessible to bullies. The effects of cyberbullying

Corresponding author: Shumaa Saeed Alzahrani and traditional bullying are similar. They make victims feel angry, hurt, scared, powerless, hopeless, lonely, embarrassed, and guilty. A victim's mental health is likely to deteriorate, and the victim is more likely to experience mental health issues like low self-esteem, depression, PTSD, or anxiety. Because most cyberbullying on Instagram is anonymous, the victims do not know who is targeting them, which can make them feel even more threatened, and it can embolden bullies, who think that because they are anonymous online, they are less likely to be exposed. While cyberbullies cannot see the victim's reaction, they will sometimes go deeper with their harassment or mockery than they would if the victims were personally present.

Arabic speakers post both formal and informal comments in social media. The formal form of Arabic is Modern Standard Arabic (MSA), while the informal form is the regional dialects (DA), the spoken language used for everyday contact in Arab countries. Compared to the most common languages, such as English, dealing with Arabic text poses substantial challenges. Arabic has a wide range of grammatical forms, word synonyms, and meanings, dependent on factors such as word order and diacritics. There is an additional difficulty when dealing with dialect or colloquial language, commonly used in Instagram comments.

D. Research Questions

The main goal of this paper is to demonstrate the ability to detect verbal cyberbullying from social media comments, with Instagram being used as a case study. To identify bullies' comments, suspicious verbal cyberbullying terms, and verbal cyberbullying words, the following are the key research questions and sub-questions I hope to answer:

- RQ1: How can an Arabic dataset of verbal cyberbullying be built?
- RQ1.1: How can an Arabic dataset involving verbal cyberbullying in different dialects be built?
- RQ1.2: How can a verbal cyberbullying dataset be appropriate for this study?
- RQ1.3: How can the reliability of dataset classes be ensured?
- RQ2: How is verbal cyberbullying distinguished from non-related event comments?
- RQ2.1: What is the most effective tool for detecting verbal cyberbullying?
- RQ2.2: What are the best machine learning tools that improve the performance of the approaches?

E. Research Goals

This section addresses the paper's research objectives, followed by a discussion of the targets and their justifications. The following are the main objectives of this paper:

- To create an Arabic Instagram dataset of verbal cyberbullying comments that identifies known and suspicious verbal cyberbullying words in comments in Arabic.
- To determine the best method for detecting verbal cyberbullying by comparing KNIME and Orange tools results to discover the most successful supervised learning strategy.

Most relevant studies have created Instagram datasets for testing verbal cyberbullying detection approaches. Many datasets have been built for commonly used languages such as English, but Arabic has received less attention. To the best of my knowledge, no one has performed or identified verbal cyberbullying detection in Arabic. As a result of the increasing demand for Arabic datasets, I created a dataset specifically to assess my targeted verbal cyberbullying detection system. Because unsupervised methods are typically ineffective at detecting verbal cyberbullying, researchers must monitor their approaches. Most unsupervised approaches use burst detection, which compares constructed words to verbal bullying word frequencies in comments. The burst behavior of specific words may not be verbal cyberbullying. For instance, not all sentences in Arabic that include animal words (حيوان) are verbal cyberbullying. Table I shows examples of verbal cyberbullying-related comments and non-related verbal cyberbullying comments.

TABLE I. INSTAGRAM COMMENT EXAMPLES

Non-related verbal cyberbullying comment	"Mara [a type of rodent] animal" "حيوان المارا"
Related verbal cyberbullying comment	"An animal, may God suffice us of him" "حيوان حسبي الله عليه"

Both comments in Table I use the same word. This word may indicate bursts, but it is not always indicative of verbal cyberbullying. Compared to unsupervised approaches, the detection domain of supervised approaches is small.

II. RELATED WORK

Most published studies concentrate on cyberbullying identification strategies for commonly used languages like English, with Arabic gaining less attention. The first subsection of this chapter is a related work overview of the most influential cyberbullying identification studies in English social media or SNS. The following subsection presents studies on cyberbullying identification in Arabic social media or SNS. In terms of cyberbullying detection, I have divided the reviewed papers into supervised and unsupervised approaches.

Authors in [11] proposed a solution to dispose of verbal cyberbullying. They suggested using a new feature selection technique for the closest neighbor classifier, which involves summarizing the original training materials using a measure of sentence importance. The two measures of sentence similarity used in their method for summarizing a single document were the frequency of the terms in a sentence and the similarity of that sentence to other sentences. After the researchers ranked all sentences, they chose the best-ranking sentences for a summary (within a threshold limitation). The researchers took every document's summary from the corpus and entered it into a new document used for summarization evaluation. In [12], the effort focused on classifying documents, a guided learning technique. Text preprocessing, feature extraction, and classification are the phases that make up the document categorization process. The study evaluated the performance of two classifiers (KNN and Naive Bayes) and specific feature selection strategies with or without combining accuracy, average precision, precision, and recall. The researchers trained each experiment's classifiers using a custom data set. The results showed that the Naive Bayes classifier outperformed the other classifiers in several instances. In [13], the authors employed the comparative study's performance evaluation metrics of accuracy, precision, and F-measure. Three algorithms were used for cyberbullying classification, i.e. Naive Bayes, SVM, and C4.5.

A. *Event Detection in Arabic Social Media*

In [6], the researchers built a text corpus focusing on two common Arabic dialects on Twitter. They proposed a 3-level hierarchical annotation schema for hate and offensive language characterization. For hate speech, their emphasis was on 4 types: religion, ethnicity, nationality, and gender, for offensive speech, they focused on posts containing nonacceptable language or general profanity. Based on machine learning (SVM, Naive Bayes, logistic regression) and deep learning (CNN, LSTM, and GRU), they trained numerous 2-class, 3-class, and 6-class hate speech classifiers using a panoply of feature extraction techniques, including unigram, word, and character n-grams and word embeddings (random, skip-gram, CBOW, and fastText) and contextual word embedding (multilingual BERT). The researchers observed that deep learning was superior to machine learning across the 3 classification tasks. In deep neural networks, the CNN+mBERT model outperformed all the other learned models across the 3 prediction tasks, with 87.05% for the 2-class task, 78.99% for the 3-class task, and 75.51% for the 6-class task. In [7], the researchers presented a scheme to detect cyberbullying messages in Arabic social media streams (Twitter and YouTube). The detection algorithm used a corpus of offensive words most used among Arab youth. The proposed scheme involved the following steps: (i) data cleaning and preprocessing, (ii) extracting bullying keywords and attributing weights, (iii) detecting cyberbullying comments, and (iv) calculating the bullying strength and classifying the comments. This scheme only focused on labeling the comments as bullying or non-bullying and decision making.

B. *Existing Arabic Datasets*

Authors in [3] published one of the first studies on Arabic abusive language identification, which included the development of an Arabic dataset of abusive comments. The dataset contained 1,100 tweets gathered from controversial accounts and hashtags. Three annotators classified the dataset as pornographic, offensive, or clean. The authors used a pattern-based Twitter search to build a seed word list of 228 obscene Arabic words for classification. Then, based on their use of obscene words in the seed list, they separated Twitter users into clean and obscene. They also compiled a longer list of potentially obscene terms by removing only unigram and bigram words. They assessed the utility of both lists as features for categorizing tweets as obscene or clean. Experiments revealed that combining the seed word list with the extended list generated the best F1 score of 60%. Authors in [4] presented a more detailed dataset for religious hate speech in dialectical Arabic. The dataset comprised 6,600 tweets using religious-related keywords. The researchers used crowdsourcing to classify the tweets as hateful or not and if they were religious extremist targets. They looked at various features, including lexicon-based and n-gram features, as well as standard machine learning algorithms. They also used neural networks like GRU and LSTM to evaluate their theories. Their experiments showed that GRU had the highest prediction accuracy of 77%. Authors in [5] created the Levantine Hate Speech and Abusive Behavior (L-HSAB) Twitter dataset, which included 5,812 tweets categorized as average, hateful, or abusive. The researchers divided the learning tasks into 2-class (abusive, normal) and 3-class classification tasks (abusive, hateful, normal) for model validation. They tried various n-gram ranges, such as unigrams, bigrams, and trigrams, as functions. They compared SVM and Naive Bayes classification performance and found that Naive Bayes outperformed SVM with an F1 score of 89.6% for 2-class classification and 74.4% for 3-class classification.

III. ARABIC CYBERBULLYING DATASET

One of the difficulties in detecting Arabic cyberbullying is the small number of Arabic datasets. To detect cyberbullying, I wanted to create a dataset with comments written in MSA, Saudi, and other dialects. In this section, the dataset of Arabic comments from Instagram that I made to detect cyberbullying is presented and the research question and sub-questions RQ1, RQ1.1, RQ1.2, and RQ1.3 will be answered.

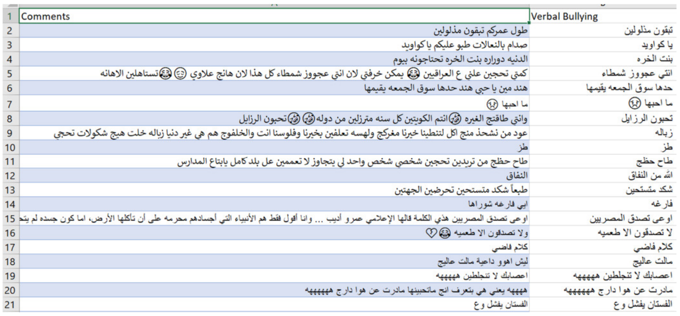
A. *Verbal Cyberbullying Instagram Comment Collection*

Two key measures were followed to establish a dataset to detect cyberbullying. In the first step, comments by keywords were manually collected. In the second step, the collected comments were filtered into two classes: bully or positive. The other way was by verbal keywords. The comments were manually collected from the Instagram website through the author's personal account. The reason for choosing Instagram was because it has a high volume of cyberbullying comments, although it is not open source. Thus, I created the dataset in an Excel file to use in KNIME and Orange tools. Comments were collected for a period of 11 months between January and November of 2021. By using search terms for cyberbullying, a keywords list was prepared, for example:

- { غبي- غبيه- غبية- اغبياء-أغبياء-إغبياء-ياغبي-ياغبيه-ياأغبياء- ي غبي -ي غبيه -ي اغبياء- ي أغبياء}. These words mean stupid in different ways in the Arabic language.
- { ياحمار- يا حمار-ياحماره-يا حماره- ياحمارة-يا حمارة- انت حمار- أنت حمار-إنت حمار -ي حمار -ي حماره – انتي حماره-انتي حمارة- لانك حمار} These words mean donkey in different ways in the Arabic language.

Cyberbullying comments were collected based on a set of keyword lists related to calling names, mocking, taunting, threatening, or verbally assaulting in singular and plural. There were approximately 1,500 comments received for cyberbullying, with about 1,000 comments filtered by 1,857 keywords. Figure 1 shows a part of this dataset. Most comments were collected in 6 months. They were saved in an Excel spreadsheet, which is an excellent choice for storing

Instagram data in tables. In addition, because Excel is a spreadsheet that consists of field and value pairs, storing comments is easy.



1	Comments	Verbal Bullying
2	طول عمركم تبقون مذلولين	تبقون مذلولين
3	صدام بالنعالات طبو عليكم يا كواويد	يا كواويد
4	الدنيه دوراره بنت الخره تحتاجونه بيوم	بنت الخره
5	كمتي تحجين علني ع العراقيين 😂 يمكن خرفتي لان انتي عجووز شمطاء كل هذا لان هانج علاوي 😌 😂تستاهلين الاهانه	انتي عجووز شمطاء
6	هند مين يا حبي هند حدها سوق الجمعه يقيمها	حدها سوق الجمعه يقيمها
7	ما احبها 😢	ما احبها 😢
8	وانتي طاقتج الغيره 🤣انتم الكويتين كل سنه مترزلين من دوله🤣 🤣تحبون الرزايل	تحبون الرزايل
9	عود من نشحذ منج اكل لتنطينا خيرنا مغركج ولهسه تعلفين بخيرنا وفلوسنا انت والخلفوج هم هي غير دنيا زباله خلت هيج شكولات تحجي	زباله
10	طز	طز
11	طاح حظج من تريدين تحجين شخصي شخص واحد لي يتجاوز لا تعممين عل بلد كامل يابتاع المدارس	طاح حظج
12	النفاق	الله من النفاق
13	طبعاً شكد متستحين تحرضين الجهتين	شكد متستحين
14	ايي فارغه شوراها	فارغه
15	اوعى تصدق المصريين هذي الكلمة قالها الإعلامي عمرو أديب ... وانا أقول فقط هم الأنبياء التي أجسادهم محرمه على أن تأكلها الأرض، اما كون جسده لم يتح	اوعى تصدق المصريين
16	ولا تصدقون الا طعميه 😂💔	لا تصدقون الا طعميه
17	كلام فاضي	كلام فاضي
18	ليش اهوو داعية مالت عاليج	مالت عاليج
19	اعصابك لا تنجلطين هههههه	اعصابك لا تنجلطين هههههه
20	هههه يعني هي بتعرف انج ماتحبينها مادرت عن هوا دارج هههههه	مادرت عن هوا دارج هههههه
21	الفستان يفشل وع	الفستان يفشل وع

Fig. 1. A section of table comments and verbal bullying keywords.

IV. PROPOSED APPROACHES

The goal was to examine cyberbullying comments using keywords and categorize them into two types: cyberbullying (known and suspicious) and non-cyberbullying. I used KNIME and Orange tools to evaluate two different methods. Both tools detect cyberbullying comments distinguishable from non-cyberbullying comments by performing workflows using many nodes to classify data. The two tools were evaluated and compared. In this section, the RQ2, RQ2.1, RQ2.2 question and sub-questions will be addressed.

A. *Utilized Methodologies*

In cyberbullying detection, classification is usually used for specific cyberbullying detection, while clustering is generally used for unspecific cyberbullying detection. Two methods were assessed to detect cyberbullying on Instagram. The comments were manually gathered and the dataset was created. The suggested methods aimed to identify a specific form of cyberbullying.

As a result, only supervised learning methods were used. Because social media features such as follower counts, mention counts, and message lengths do not apply to the cyberbullying detection task, I focused the cyberbullying detection task on the textual content of the comments. The first method used the KNIME tool to identify cyberbullying and non-cyberbullying comments. The second method used the Orange tool. The two methods were compared to see if breaking down the issue of cyberbullying identification into two phases improved its effectiveness. Regarding the negative impact of noisy and informal comment text in both proposed methods, writing the keywords in different ways, such as in chatting manners, to detect relevant comments more effectively, is recommended.

B. *Data Mining Tools*

In this paper, cyberbullying is detected through data mining using the open-source tools KNIME and Orange. The reason behind the existence of these tools is the existence of massive amounts of data. As a result, the traditional statistics methods are no longer useful. In the late '80s, many pieces of research appeared to solve these problems, in addition to searching for solutions that combined several disciplines, including statistics, databases, artificial intelligence, distinguishing different models, or analog computing. Then, data mining and knowledge discovery emerged, which proved to be successful solutions for analyzing vast amounts of data by transforming them from accumulated and incomprehensible data into valuable information that could be exploited and used [8]. Data mining is the process of analyzing data from different perspectives, drawing relationships between them, and summarizing them into useful information.

1) *KNIME Tool*

KNIME makes understanding data and developing data science workflows and reusable components accessible by being intuitive, transparent, and constantly incorporating new technologies [9].

2) *Orange Tool*

The Orange tool is open-source machine learning and data visualization software. With a comprehensive and diverse toolbox, it builds data analysis workflows visually [10]. Data visualizations that are parts of the Orange help find hidden data patterns, provide intuition behind data analysis procedures, or support collaboration between data scientists and domain experts. Scatter plots, box plots, and histograms are among the visualization widgets available, as are model-specific visualizations such as dendrograms, silhouette plots, and tree visualizations. Many other visualization tools, such as network visualizations, word clouds, and geographical maps are available as add-ons. Interactive visualizations allow exploratory data analysis. A user can pick interesting data subsets directly from plots, graphs, and data tables and mine them in downstream widgets. For instance, a user can perform cross-validation logistic regression on a data set and map some misclassifications to the two-dimensional projection. It is simple to transform Orange into a tool that allows domain experts to explore their data, even if they have little experience with statistics or machine learning.

C. *Cyberbullying Approaches*

1) *KNIME Tool*

a) *Phase 1*

The workflow in Figure 2 represents the data flow between different nodes, starting with Excel Reader, then moving on to Tika Language Detector (to recognize used languages), Column Filter, Filter Apply Row Splitter, Row Filter, Rule-based Row Filter, and finally Excel Writer.

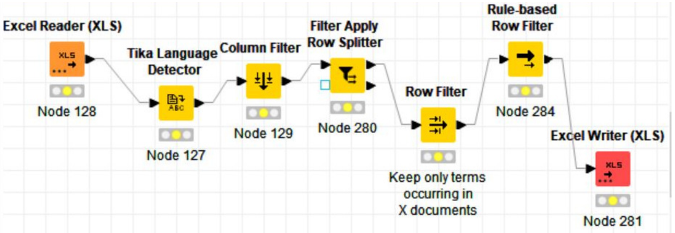
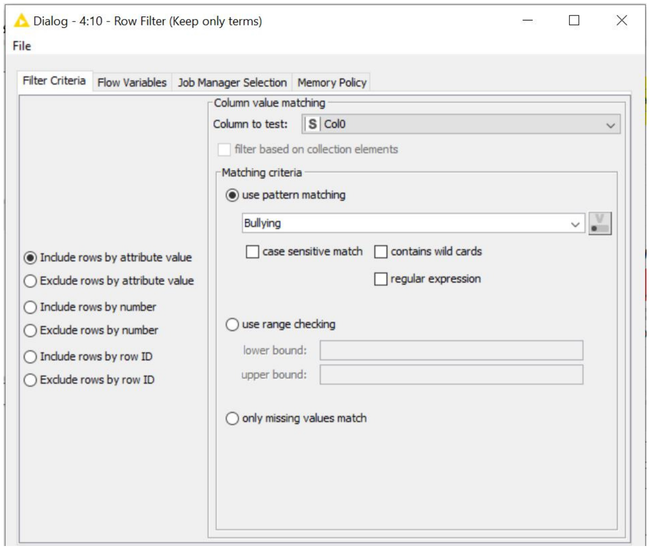


Fig. 2. The workflow of extracting bullying terms by rule.

The working of nodes 128, 127, and 281 has been explained above.

- Column Filter: This node allows columns to be filtered from the input table while the remaining columns are passed to the output table. Within the dialog, columns can be moved between the Include and Exclude list.
- Filter Apply Row Splitter: This node splits the input according to the filter definitions, either given in the input table itself or optional as an additional model input. Filter definitions are only applied if an additional model is provided as input. If the input contains a filter defined on a column not present in the input table, the node will not fail but will display a warning message.
- Row Filter (Figure 3): This node allows for row filtering according to specific criteria. It can include or exclude certain ranges (by row number), rows with a particular row ID, or rows with a specific value in a selectable column (attribute). The node does not change the domain of the data table. In other words, the upper and lower bounds or the possible values in the table spec are not adapted, even if one of the bounds or values is fully filtered out. Figure 3 shows the configuration dialog of the node.



Dialog - 4:10 - Row Filter (Keep only terms)

File

Filter Criteria | Flow Variables | Job Manager Selection | Memory Policy

● Include rows by attribute value
○ Exclude rows by attribute value
○ Include rows by number
○ Exclude rows by number
○ Include rows by row ID
○ Exclude rows by row ID

Column value matching

Column to test: S Col0

□ filter based on collection elements

Matching criteria

● use pattern matching

Bullying

□ case sensitive match □ contains wild cards

□ regular expression

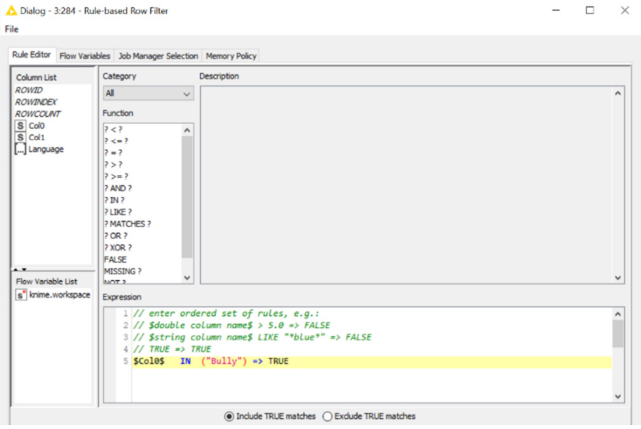
○ use range checking

lower bound:

upper bound:

○ only missing values match

Fig. 3. Row Filter output.



Dialog - 3:284 - Rule-based Row Filter

File

Rule Editor | Flow Variables | Job Manager Selection | Memory Policy

Column List: ROWID, ROWINDEX, ROWCOUNT, S Col0, S Col1, [...] Language

Flow Variable List: s knime.workspace

Category: All

Function: ? < ?, ? <= ?, ? = ?, ? > ?, ? >= ?, ? AND ?, ? IN ?, ? LIKE ?, ? MATCHES ?, ? OR ?, ? XOR ?, FALSE, MISSING ?, NOT ?

Description

Expression

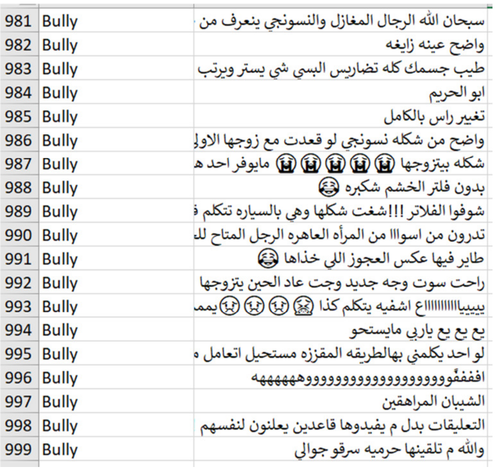
```
// enter ordered set of rules, e.g.:
// $double column name$ > 5.0 => FALSE
// $string column name$ LIKE "*blue*" => FALSE
// TRUE => TRUE
$Col0$   IN  ("Bully") => TRUE
```

● Include TRUE matches ○ Exclude TRUE matches

Fig. 4. Ruled-based Row Filter output.

- Rule-based Row Filter (Figure 4): This node takes a list of user-defined rules and tries to match them to each row in the input table. The row is selected for inclusion if the first matching rule has a TRUE outcome. Otherwise (i.e. if the first matching rule yields FALSE), it will be excluded. If no rule matches, the row will be excluded. Inclusion and exclusion may be inverted (see the options in Figure 4).

In the dialog in Figure 4, I used the simple rule **$Col0$ IN ("Bully") => TRUE** to extract just cyberbullying comments by classifying the comments as bullying (1) or positive (0). Figure 5 shows a part of the cyberbullying comment results using the rule Bully term, which extracted 999 comments out of 1,500.



981	Bully	سبحان الله الرجال المغازل والنسونجي ينعرف من
982	Bully	واضح عينه زايغه
983	Bully	طيب جسمك كله تضاريس البسي شي يستر ويرتب
984	Bully	ابو الحريم
985	Bully	تغيير راس بالكامل
986	Bully	واضح من شكله نسونجي لو قعدت مع زوجها الاول
987	Bully	شكله بيتزوجها 😭😭😭😭😭 مايوفر احد ه
988	Bully	بدون فلتر الخشم شكبره 😂
989	Bully	شوفوا الفلاتر !!!شغت شكلها وهي بالسياره تتكلم ف
990	Bully	تدرون من اسواا من المرأه العاهره الرجل المتاح لل
991	Bully	طاير فيها عكس العجوز اللي خذاها 😂
992	Bully	راحت سوت وجه جديد وجت عاد الحين يتزوجها
993	Bully	يييياااااااااع اشفيه يتكلم كذا 😫😢😢😢يمم
994	Bully	يع يع يع ياربي مايستحو
995	Bully	لو احد يكلمني بهالطريقه المقززه مستحيل اتعامل م
996	Bully	افففففووووووووووووووووووووههههههه
997	Bully	الشيبان المراهقين
998	Bully	التعليقات بدل م يفيدوها قاعدين يعلنون لنفسهم
999	Bully	والله م تلقينها حرميه سرقو جوالي

Fig. 5. Part of cyberbullying comments results by Bully term.

b) Phase 2

The workflow in Figure 6 represents the flow data to extract cyberbullying comments by verbal bullying keywords.

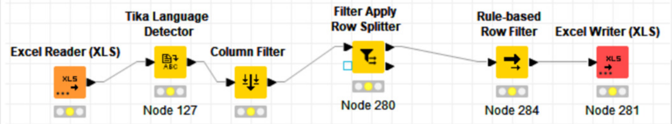
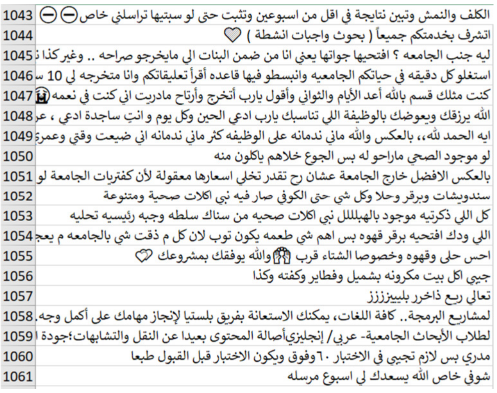


Fig. 6. The workflow used to extract cyberbullying by VB Keywords.



1043	الكلف والنمش وتبين نتايجة في اقل من اسبوعين وتثبت حتى لو سبتيها تراسلني خاص ⊖⊖
1044	اتشرف بخدمتكم جميعاً (بحوث واجبات انشطة) 🤍
1045	ليه جنب الجامعه ؟ افتحيها جواتها يعني انا من ضمن البنات الي مايخرجو صراحه .. وغير كذا ن
1046	استغلو كل دقيقه في حياتكم الجامعيه وانبسطو فيها قاعده أقرأ تعليقاتكم وانا متخرجه لي 10 س
1047	كنت مثلك قسم بالله أعد الأيام والثواني وأقول يارب أتخرج وأرتاح مادريت اني كنت في نعمه 😭
1048	الله يرزقك ويعوضك بالوظيفة اللي تناسبك يارب ادعي الحين وكل يوم و انتِ ساجدة ادعي ، عر
1049	ايه الحمد لله،، بالعكس والله ماني ندمانه على الوظيفه كثر ماني ندمانه اني ضيعت وقتي وعمر
1050	لو موجود الصحي ماراحو له بس الجوع خلاهم ياكلون منه
1051	بالعكس الافضل خارج الجامعة عشان رح تقدر تخلي اسعارها معقولة لأن كفتريات الجامعة لو
1052	سندويشات وبرقر وحلا وكل شي حتى الكوفي صار فيه نبي اكلات صحية ومتنوعة
1053	كل اللي ذكرتيه موجود بالهبلللل نبي اكلات صحيه من سناك سلطه وجبه رئيسيه تحليه
1054	اللي ودك افتحيه برقر قهوه بس اهم شي طعمه يكون توب لان كل م ذقت شي بالجامعه م يعج
1055	احس حلى وقهوه وخصوصا الشتاء قرب 🤲والله يوفقك بمشروعك 🤍
1056	جيبي اكل بيت مكرونه بشميل وفطاير وكفته وكذا
1057	تعالي ريع ذاخرر بلبييززززز
1058	لمشاريع البرمجة.. كافة اللغات، يمكنك الاستعانة بفريق بلستيا لإنجاز مهامك على أكمل وجه.
1059	لطلاب الأبحاث الجامعية- عربي/ إنجليزيأصالة المحتوى بعيدا عن النقل والتشابهات؛جودة ا
1060	مدري بس لازم تجيبي في الاختبار ٦٠وفوق ويكون الاختبار قبل القبول طبعا
1061	شوفي خاص الله يسعدك لي اسبوع مرسله

Fig. 7. Results extract cyberbullying comments by VB Keywords.

The results in Figure 7 represent 1,061 cyberbullying comments, 64 of them being positive. The KNIME tool is inaccurate because it does not deal perfectly with Arabic (as with English). Phase1 in KNIME has 999 right results out of 1500. In phase 2 there are 61 wrong comments.

2) *Orange Tool*

a) Phase 1

The workflow in Figure 8 represents extracting cyberbullying comments by using the Bully term through a categorical type target role in the file node description. The file classifies bully and positive comments. The results are fairly accurate. There were 999 cyberbullying comments out of 1,498, and the workflow ignored the rest.

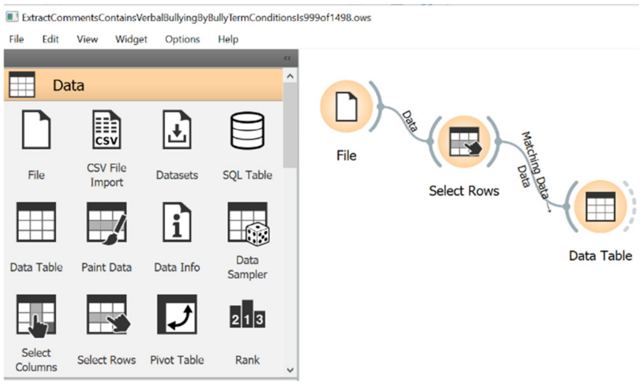
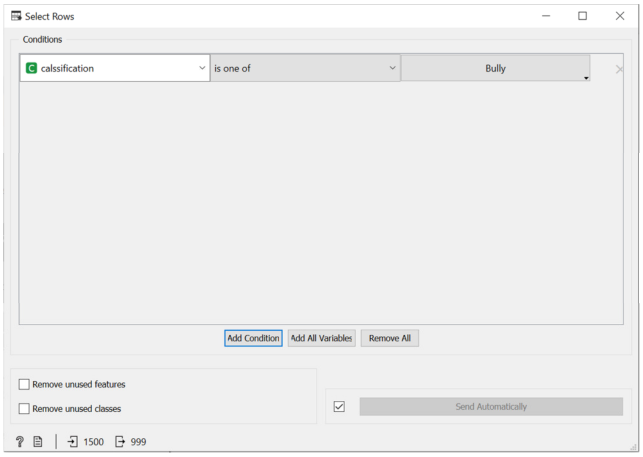


Fig. 8. Workflow for classifying cyberbullying and non-cyberbullying comments by the Bully term.

The Select Rows dialog in Figure 9 represents the condition pattern in which the classification is a column in the original file and the type of condition. Bully is one of the values in the classification column. The other value is Positive for non-cyberbullying comments.



Conditions

calssification | is one of | Bully

□ Remove unused features

□ Remove unused classes

☑ Send Automatically

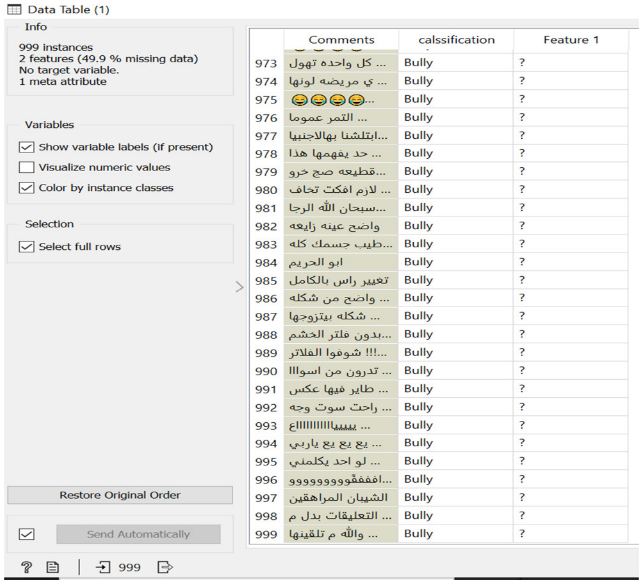
1500 → 999

Fig. 9. Select Rows dialog with condition pattern.

Figure 10 displays the results for extracting 999 cyberbullying comments by Bully term out of 1,500 comments.

b) Phase 2

I changed the Select Rows node description in Figure 8. I used the condition in Select Rows node that the verbal bullying must contain the cyberbullying called (اوعى تسدى المصريين ،ايران ماتتلام فيكم عيل، استوت مثل المومياء، اسلوبها زج، اسلوبج الخايس، اعصابك لاتتجلطين هههههههه،اعطتها اكبر من حجمها، افلس اخلاقيا من بعد مافلس فنيا ، اقلبي وجهك ، اذا تسئل على صراويل العراقين اسئل امك عنها اكيد محتفضه فيهم للذكره) as shown in Figure 11.



Info

999 instances
2 features (49.9 % missing data)
No target variable.
1 meta attribute

Variables

☑ Show variable labels (if present)

□ Visualize numeric values

☑ Color by instance classes

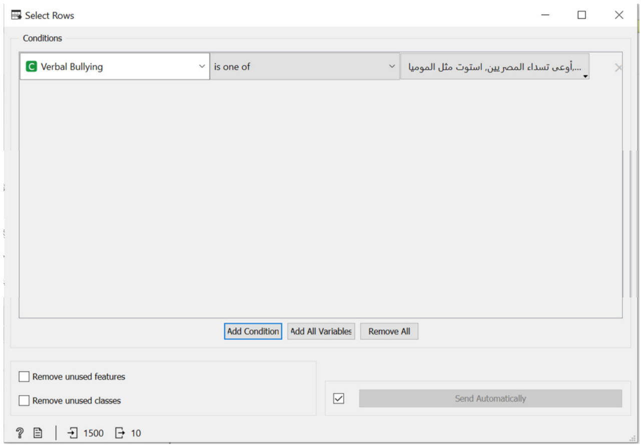
Selection

☑ Select full rows

	Comments	calssification	Feature 1
973	... كل واحده تهول	Bully	?
974	... ي مريضه لونها	Bully	?
975	😂😂😂😂...	Bully	?
976	... التمر عموما	Bully	?
977	...ابتلشنا بهالاجنبيا	Bully	?
978	... حد يفهمها هذا	Bully	?
979	...قطيعه صج خرو	Bully	?
980	... لازم افكت تخاف	Bully	?
981	...سبحان الله الرجا	Bully	?
982	واضح عينه زايغه	Bully	?
983	...طيب جسمك كله	Bully	?
984	ابو الحريم	Bully	?
985	تغيير راس بالكامل	Bully	?
986	... واضح من شكله	Bully	?
987	... شكله بيتزوجها	Bully	?
988	...بدون فلتر الخشم	Bully	?
989	...!!! شوفوا الفلاتر	Bully	?
990	... تدرون من اسواااا	Bully	?
991	... طاير فيها عكس	Bully	?
992	... راحت سوت وجه	Bully	?
993	... يييييااااااااااع	Bully	?
994	... يع يع يع ياربي	Bully	?
995	... لو احد يكلمني	Bully	?
996	...افففقوووووووووو	Bully	?
997	الشيبان المراهقين	Bully	?
998	... التعليقات بدل م	Bully	?
999	... والله م تلقينها	Bully	?

999

Fig. 10. Data Table node results for extracting 999 cyberbullying comments by Bully term.



Conditions

Verbal Bullying | is one of | ...أوعى تسداء المصريين, استوت مثل الموما

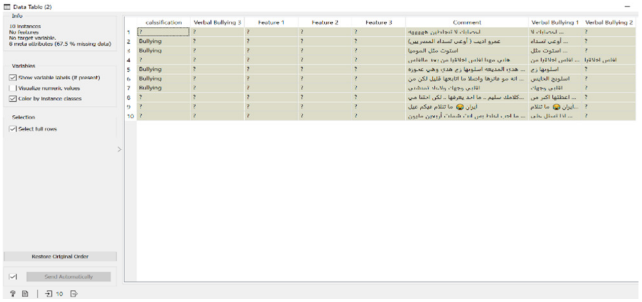
□ Remove unused features

□ Remove unused classes

☑ Send Automatically

1500 → 10

Fig. 11. Extracting cyberbullying comments with more than one VB keyword.



Info

10 instances
No features
No target variable.
8 meta attributes (67.5 % missing data)

Variables

☑ Show variable labels (if present)

□ Visualize numeric values

☑ Color by instance classes

Selection

☑ Select full rows

	calssification	Verbal Bullying 3	Feature 1	Feature 2	Feature 3	Comment	Verbal Bullying 1	Verbal Bullying 2
1	?	?	?	?	?	اعصابك لا تتجلطين هههههه	... اعصابك لا	?
2	Bullying	?	?	?	?	عمرو اديب (أوعى تسداء المصريين)	... أوعى تسداء	?
3	Bullying	?	?	?	?	استوت مثل الموميا	... استوت مثل	?
4	?	?	?	?	?	هادي مودا افلس اخلاقيا من بعد ماافلس	... افلس اخلاقيا من	افلس اخلاقيا
5	Bullying	?	?	?	?	... هدي المديعه اسلوبها زج هدي وهي عجوزه	اسلوبها زج	?
6	Bullying	?	?	?	?	... انه مو فاترها واصلا ما اتابعها قليل لكن من	اسلوبج الخايس	?
7	Bullying	?	?	?	?	اقلبي وجهك ولاحاد تمشي	اقلبي وجهك	?
8	?	?	?	?	?	...كلامك سليم .. ما احد يعرفها .. لكن احتنا مي	... اعطتها اكبر من	?
9	?	?	?	?	?	ايران 😭 ما تتلام فيكم عيل	...ايران 😭 ما تتلام	?
10	?	?	?	?	?	... ما احد اعطاها بس انت شمات أربعين ماتون	... اذا تسئل على	?

10

Fig. 12. Result of extracting cyberbullying comments with more than one VB keywords.

The output in Figure 12 represents 10 chosen rows applied by the condition in the original file of 1,500 comments classified as cyberbullying

c) Phase 3

I changed the Select Rows node description in Figure 8. I used the condition in the Select Rows node that the comment must contain the cyberbullying term (تبقون مذلولين), as shown in Figure 13. The output in Figure 14 represents one row in the original file of 1,500 comments classified as cyberbullying.

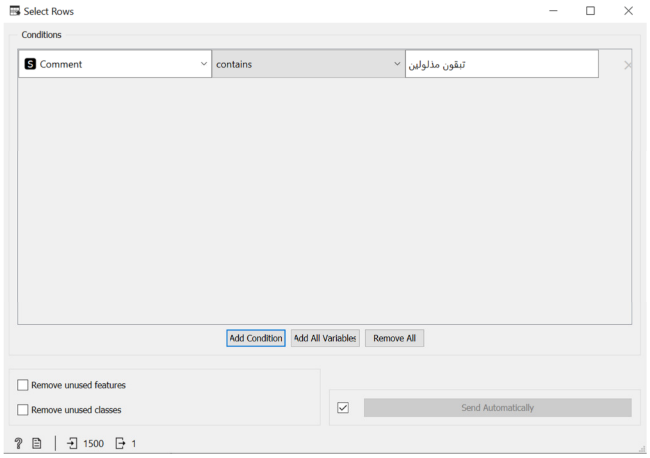
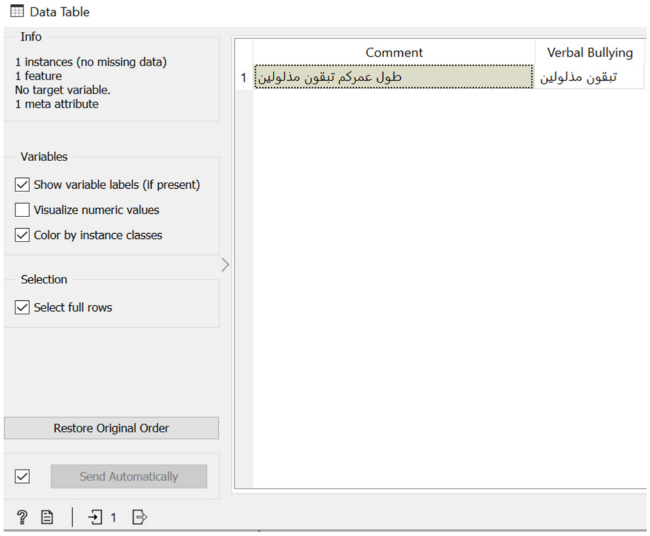

Comment contains تبقون مذلولين

Fig. 13. Select Rows with condition comment contains VB keywords.


Info
1 instances (no missing data)
1 feature
No target variable.
1 meta attribute

	Comment	Verbal Bullying
1	طول عمركم تبقون مذلولين	تبقون مذلولين

Fig. 14. Workflow result.

Phase 1 in Orange tool has a right result. Phase 2 extracts cyberbullying comments by just one cyberbullying term. In Phase 3, only the right answer can be shown when the cyberbulling terms are less than 100.

D. Comparison and Evaluation

In this paper, Orange and KNIME tools were used to classify Instagram comments under the cyberbullying and non-cyberbullying categories. Two data mining methods were applied: the first used two classes, Bully and Positive. The second used VB keywords. Each tool has its advantages and disadvantages. The advantages of the KNIME tool are that it can deal with an extensive dataset, different types of files, a huge variety of components, has easy code-to-write conditions that make the tool more developed, and has excellent performance. The disadvantages are the need to use a special node to define the language, accurate results within the English language but inaccurate results within the Arabic language, the inability to use the Remove Punctuation node to get the correct result in the Arabic language, and unclear descriptions for using the tool. The advantages of the Orange tool are the lack of need to define the language, accurate results when the VB keywords are less than 100, easy dealing with nodes, and easy understanding of the concepts of the tool. The disadvantages are that it cannot deal with a large dataset, every single VB keyword must be chosen every time, and it cannot make a condition with comments containing more than one VB.

TABLE II. COMPARISON BETWEEN KNIME AND ORANGE

Tool	Algorithm	Phase1	Phase2	Phase3	Notes
KNIME	Rule-based	999 of 1500	1061 of 1500		Phase2 has wrong result
Orange	Select Rows	999 of 1500	Cyberbullying term<100 of 1500	1 of 1500	Phase2 allows small data. Phase3 allows one condition every time

E. Conclusion

Orange and KNIME tools were used in this paper to data mine cyberbullying comments and distinguish them from non-cyberbullying comments on Instagram. I extracted cyberbullying in two ways, one using VB keywords and the other classifying comments as Bully or Positive. In KNIME, I got inaccurate data results within the large dataset, while in Orange, I got accurate data results with less than 100 VB keywords. The results in both tools in the second way were accurate.

V. CONCLUSIONS AND FUTURE WORK

The driving question behind this study is "How can you detect cyberbullying in social media?" In this paper, emphasis was given on detecting name calling, mocking, taunting, threatening, or verbal abuse on Instagram. I addressed a complex problem in Arabic social media and carried out the key research goals.

To assess cyberbullying detection methods, I created a cyberbullying dataset that included written comments in MSA, Saudi, and other Arabic dialects. I created the dataset taking cyberbullying into consideration. I evaluated two supervised learning approaches to detect cyberbullying—the KNIME tool and the Orange tool. The keywords were the same in both approaches, as they are on social media. Both methods produced positive evaluation outcomes. Regarding detecting cyberbullying, the Orange tool outperformed the KNIME tool. To address the issue of cyberbullying detection tasks, I suggest using the KNIME tool with raw data.

Regarding future work on cyberbullying detection, the following two directions are suggested for further investigation:

- Dataset expanding. A second version of the cyberbullying keyword dataset can be published by extracting extra samples and performing labeling process (known verbal, suspicious verbal, and non-cyberbullying).
- Cyberbullying detection in audio files using the KNIME tool with the created cyberbullying keywords dataset.

The utilized cyberbullying detection approach will be used in the future on the Twitter platform with its open-source API.

REFERENCES

[1] J. Warren, "This Is How the Instagram Algorithm Works in 2022," *Later.com*, Jun. 21, 2022. https://later.com/blog/how-instagram-algorithm-works/.

[2] F. Chen and D. B. Neill, "Human Rights Event Detection from Heterogeneous Social Media Graphs," *Big Data*, vol. 3, no. 1, pp. 34–40, Mar. 2015, https://doi.org/10.1089/big.2014.0072.

[3] H. Mubarak, K. Darwish, and W. Magdy, "Abusive Language Detection on Arabic Social Media," in *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, BC, Canada, Dec. 2017, pp. 52–56, https://doi.org/10.18653/v1/W17-3008.

[4] K. E Abdelfatah, G. Terejanu, and A. A Alhelbawy, "Unsupervised Detection of Violent Content in Arabic Social Media," in *Computer Science & Information Technology (CS & IT)*, Mar. 2017, pp. 01–07, https://doi.org/10.5121/csit.2017.70401.

[5] L. Kaati, E. Omer, N. Prucha, and A. Shrestha, "Detecting Multipliers of Jihadism on Twitter," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, Aug. 2015, pp. 954–960, https://doi.org/10.1109/ICDMW.2015.9.

[6] S. Alsafari, S. Sadaoui, and M. Mouhoub, "Hate and offensive speech detection on Arabic social media," *Online Social Networks and Media*, vol. 19, Sep. 2020, Art. no. 100096, https://doi.org/10.1016/j.osnem.2020.100096.

[7] D. Mouheb, R. Ismail, S. A. Qaraghuli, Z. A. Aghbari, and I. Kamel, "Detection of Offensive Messages in Arabic Social Media Communications," in *2018 International Conference on Innovations in Information Technology (IIT)*, Aug. 2018, pp. 24–29, https://doi.org/10.1109/INNOVATIONS.2018.8606030.

[8] A. Sayed, "Data mining tool open source: Analytical evaluation study," *Journal of Taibah University Arts and Humanities*, vol. 5, no. 10, pp. 791–865, Jun. 2016, https://doi.org/10.12816/0032954.

[9] "Data Analytics Platform: Open Source Software Tools," *KNIME*. https://www.knime.com/knime-analytics-platform.

[10] "Orange Data Mining." https://orangedatamining.com/.

[11] S. R. Basha, J. K. Rani, and J. J. C. P. Yadav, "A Novel Summarization-based Approach for Feature Reduction Enhancing Text Classification Accuracy," *Engineering, Technology & Applied Science Research*, vol. 9, no. 6, pp. 5001–5005, Dec. 2019, https://doi.org/10.48084/etasr.3173.

[12] S. R. Basha and J. K. Rani, "A Comparative Approach of Dimensionality Reduction Techniques in Text Classification," *Engineering, Technology & Applied Science Research*, vol. 9, no. 6, pp. 4974–4979, Dec. 2019, https://doi.org/10.48084/etasr.3146.

[13] M. Alghobiri, "A Comparative Analysis of Classification Algorithms on Diverse Datasets," *Engineering, Technology & Applied Science Research*, vol. 8, no. 2, pp. 2790–2795, Apr. 2018, https://doi.org/10.48084/etasr.1952.