Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18791-y

CatRevenge: towards effective revenge text detection in online social media with paragraph embedding and CATBoost

Sayani Ghosal[1,2] · Amita Jain[3]

Received: 14 May 2023 / Revised: 19 September 2023 / Accepted: 24 February 2024

Abstract

Huge amount of internet data are produced and consumed by internet users, where most of the data are in natural language and they express their feelings, emotions and thoughts on social media. It is the responsibility of the social media provider to provide healthy communication system among users. It is very challenging job to detect revenge from the social media text due to long sentences where semantic relation dissolves between tokens. Due to that, the social media providers did not provide any attention towards identifying the users spreading revenge. This article propose a novel model named as CatRevenge which identifies both active and passive revenge. This model preprocess with Slangzy internet slang meaning dictionary to detect revenge text more efficiently. CatRevenge assigns impact weight on each of parts of speech in the sentences based on its relevance and TF-IDF score of the words. The novel CatRevenge model also considers the paragraph embedding model for contextual semantic analysis of revenge text. In addition, this research applies gradient boosting CATBoost classifier with categorical features to reduce model overfitting. This feature ranking method can able to reduce the dimensionality of data by ranking the most significant feature. This research considers the revenge posts English language dataset from the Reddit social media where it evaluated with binary and multiclass classification. Results demonstrate achievable performance with a 6—10% increase in binary and a 2.5 -5% increase in multiclass with weighted F1 metric.

Keywords Online Social Network · Paragraph Embedding · Text Classification · CATBoost · Natural Language Processing

✉ Amita Jain
amita.jain@nsut.ac.in

Sayani Ghosal
sayanighosal@gmail.com

[1] NSUT East Campus (Erstwhile A.I.A.C.T.R.), Guru Gobind Singh Indraprastha University, Dwarka, Delhi, India

[2] KIET Group of Institutions, Ghaziabad Delhi-NCR, India

[3] Netaji Subhas University of Technology, New Delhi, India

Published online: 01 April 2024

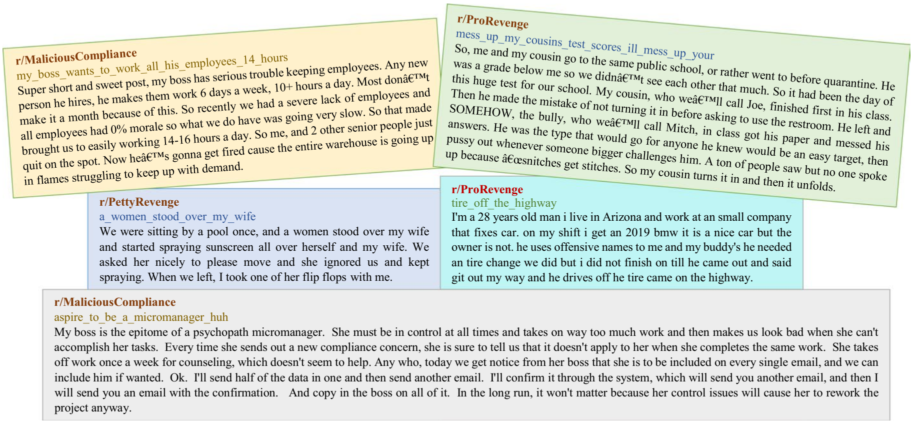
1 Introduction

The rapid growth of information technology and communication network brings various changes in industries, societies and various other sectors. In 2022, 94 zettabytes of internet data were produced [1] and consumed by 5.07 billion internet users and they share their feelings, emotions and thoughts through social media [2]. Nowadays, social networks like Facebook, Twitter, Reddit and MySpace are popular platforms that play significant role for users. Users spend ample time on social media platforms where people interact with others, update their status, share personal experiences, and defame other members [3]. The bright side of social media platforms comes with various negative consequences such as work-life conflicts, anxiety, revenge, depression, aggression, and hate crimes [4]. Various online activities and exposures provide threats for common users. This situation motivates to consider techno-regulation approaches that safeguard society and people. It supports legal norms using technological tools or devices [5]. The two ways to support cyber security norms are nudge and techno-regulation, where nudge push common users to safeguard their future action and techno-regulation forcefully removed the riskier content that harm common users [6]. This techno-regulation comes under various social media research area like privacy, child protection, cybercrime, cyber-security, revenge detection and many others. Revenge is also a root cause of cyberstalking, cybercrimes, and sexting [7].

Revenge is an act of human aggression that harms any person or group of persons in response to a perceived provocation or grievance. Cyber revenge via social media includes various forms of aggression (cyber defamation, cyberbullying, cyber trolling/stalking, revenge porn, and cyber dating abuse) and occurs through email, tweets, social media posts, and text messages [8]. Social media revenge is an almost fresh research direction where some author has implemented revenge porn detection [9, 10]. Interpersonal context-based social media revenge is the new research path and is yet to get due attention. One recent attempt [11] has extracted a dataset from Reddit social media to show the difference between active and passive revenge. Along with that another vengeful or revenge content detection research also considers text from social media, terrorist activity based on religious and school shooters datasets [12]. Revenge posts are increasing because this is easier for users to take revenge through social media. Social media revenge may receive various consequences like job loss, suicide, and lawsuit [8].

Reddit is a popular social media platform where 223 million users are from the United States [13]. It is a controversial social site that contains various communities named Subreddits. Numerous Subreddits from the Reddit website have been banned due to abusive and controversial content [14]. In 2020, most of the data removal requests (Reddit) by the Russian Government and the South Korean Government are 89% and 60% respectively [15]. Various Subreddits represent malicious posts as passive revenge and active revenge. Active revenge consists of petty and pro revenge. Petty revenge is not like criminal revenge, but pro revenge is a more grievous offense than petty revenge.

In Fig. 1, all five posts present revenge stories where two are pro revenge post, two are malicious compliance, and one is petty revenge content. The first malicious compliance post shows an employee's reaction against his boss for daily 14 h of working time. The user wants to defame his boss and share demands through social media. Another malicious post shows an employee's reaction against the manager for her behavior. The malicious post presents passive revenge stories. The petty revenge post shares a personal experience with one women's behaviors, and the user takes active revenge against her. This revenge is not like a criminal offense. Two pro revenge posts indicate bullying



r/MaliciousCompliance
my_boss_wants_to_work_all_his_employees_14_hours
Super short and sweet post, my boss has serious trouble keeping employees. Any new person he hires, he makes them work 6 days a week, 10+ hours a day. Most donâ€™t make it a month because of this. So recently we had a severe lack of employees and all employees had 0% morale so what we do have was going very slow. So that made brought us to easily working 14-16 hours a day. So me, and 2 other senior people just quit on the spot. Now heâ€™s gonna get fired cause the entire warehouse is going up in flames struggling to keep up with demand.

r/ProRevenge
mess_up_my_cousins_test_scores_ill_mess_up_your
So, me and my cousin go to the same public school, or rather went to before quarantine. He was a grade below me so we didnâ€™t see each other that much. So it had been the day of this huge test for our school. My cousin, who weâ€™ll call Joe, finished first in his class. Then he made the mistake of not turning it in before asking to use the restroom. He left and SOMEHOW, the bully, who weâ€™ll call Mitch, in class got his paper and messed his answers. He was the type that would go for anyone he knew would be an easy target, then pussy out whenever someone bigger challenges him. A ton of people saw but no one spoke up because â€œsnitches get stitches. So my cousin turns it in and then it unfolds.

r/PettyRevenge
a_women_stood_over_my_wife
We were sitting by a pool once, and a women stood over my wife and started spraying sunscreen all over herself and my wife. We asked her nicely to please move and she ignored us and kept spraying. When we left, I took one of her flip flops with me.

r/ProRevenge
tire_off_the_highway
I'm a 28 years old man i live in Arizona and work at an small company that fixes car. on my shift i get an 2019 bmw it is a nice car but the owner is not. he uses offensive names to me and my buddy's he needed an tire change we did but i did not finish on till he came out and said git out my way and he drives off he tire came on the highway.

r/MaliciousCompliance
aspire_to_be_a_micromanager_huh
My boss is the epitome of a psychopath micromanager. She must be in control at all times and takes on way too much work and then makes us look bad when she can't accomplish her tasks. Every time she sends out a new compliance concern, she is sure to tell us that it doesn't apply to her when she completes the same work. She takes off work once a week for counseling, which doesn't seem to help. Any who, today we get notice from her boss that she is to be included on every single email, and we can include him if wanted. Ok. I'll send half of the data in one and then send another email. I'll confirm it through the system, which will send you another email, and then I will send you an email with the confirmation. And copy in the boss on all of it. In the long run, it won't matter because her control issues will cause her to rework the project anyway.

Fig. 1 Sample Revenge Posts from Reddit with ProRevenge, PettyRevenge and Malicious Compliance Subreddits (https://www.reddit.com/)

or aggressive behaviors against some person. As per the recent literature, cyberbullying considers an act of active revenge [16]. All posts are long and contain various sentences. Contextual analysis of paragraph-like contents and intensity of aggressive behaviors are significant features of revenge posts. All posts include no offensive terms but show malicious or revenge expressions. Revenge is part of complex human behavior and emotion analysis. Several authors have classified emotions [17], abusive [18], aggression [19], hate [20, 21], cyberbullying [22] content with NLP models, and revenge detection also relates to all these fields. The availability of fewer datasets is the primary concern for this field. It also observes that contextual analysis for implicit revenge content detection requires more NLP research.

With contemplating revenge text research limitations, this study contributes a novel revenge detection model (CatRevenge) for binary and multiclass. The proposed model symbolizes “CatRevenge” where “Revenge” is for feature vectors of revenge post, and “CAT” is for the categorical feature-based CATBoost classifier. This research develop and analyze a novel revenge classification model with the Reddit English social media dataset. It consider paragraph embedding for contextual semantic analysis of paragraph-like contents and POS tag-based impact weight analysis to find the intensity of aggressive behaviors. The contributions of this research are given below in concise form:

- This is the first literature work on revenge posts detection with a combination of syntactic, lexical, and semantic features – POS tag based impact weight, TF-IDF, and paragraph embedding respectively.
- It is the one revenge text detection research where the paragraph embedding model considers for contextual semantic analysis of English revenge stories.
- This research is the first revenge text detection research that considers the impact weight analysis of each POS tagger for each revenge post and that considers the total corpus.
- This is the first revenge text detection research that preprocesses all reviews with Slangzy, internet slang words meaning dictionary.
- It includes an efficient gradient boost classification model CATBoost classifier that considers categorical features of revenge text.

The remainder of the research is prepared as follows: Section 2 illustrates the existing social media research and text-based applications with gradient boosting classifiers. The CatRevenge model architecture and methodology with all features and classifiers are present in Section 3. Section 4 shows the experimentation setup with results and findings. Finally, the last section concludes with future research directions for revenge post-detection works.

2 Related research

This related research work mainly illustrates two sections – Social media text analysis, and text classification with gradient boosting classifiers. With various social media text classification research, revenge text detection is an almost new research direction. So, this study describes various existing social media text research that is related to revenge text analysis. Along with that, this work consider gradient boosting classifiers-based text classification research to portray the existing research path to the readers.

Several terminologies are applied in NLP research to detect negative and harmful text from social media. Active revenge text is the expression of complex human behavior and that related to various NLP research areas like – emotion detection [17], cyber aggression detection [19, 23], abusive language detection [18], hate content detection [21, 24], cyberbullying and stalking detection [22]. Classification of abusive content is also related to malicious content and active revenge. Revenge detection is a fresh NLP research domain, one author has implemented revenge content detection model with POS tagging, word2vec embedding and AdaBoost, KNN classifier [12]. This model has applied three dataset but this model considers short sentence based text and it also not able to detect the importance of each tokens. Both the above models not able to detect slang or offensive words from text. The related work section illustrated various existing research works in the above areas.

Various machine learning, and deep learnings models are applied to the above studies as supervised, unsupervised, semi-supervised, and deep learning models. Most of the studies employed various classification models like – SVM [18], Logistic Regression [21], Random Forest [22], MLP [23], Bert [19, 25], LSTM [17, 26] and CATBoost [24]. One recent emotion detection study [17] has applied the Glove embedding model with the LSTM classifier to detect feelings from the text. Another very recent emotion detection research detects negative emotional text from patients who suffer from mental health. This model also applied the Glove embedding model and bi-directional LSTM classifier along with the CNN model. This study mainly detects negative emotions like stress, anxiety, depression, addiction and shows achievable performance for WebMD and Healthtap datasets [26]. Aggression detection research has employed various Bert models to tackle cyber aggression [19]. One aggression detection study has employed the MLP classifier and deep neural network which achieved 92% accuracy. This study also established that aggression and hate are related to cyber harassment and bullying [23]. Along with emotion and aggression text, one study has detected cyberbullying [22] with the Random Forest classifier which achieved 0.90 F1. Cyberbullying detection has also been explored with sentiment and emotion features that have created a code-switch corpus. The Bert and VecMap based two embedding techniques have outperformed the cyberbullying baseline models [27]. Hate text detection with emotion informed has explored multitask and multi-target approaches. This study has applied sentic computing models, hate speech lexicons, and the BERT model to achieve remarkable performance compared to the baseline models [25].

Hate speech detection research has applied Multinomial logistic regression that achieved 87.68% accuracy [21]. The CATBoost classifier with various features for multiclass hate text classification has also achieved the best performance compared to other machine learning and deep learning classifiers [24]. Abusive language detection is also another NLP research domain that is related to revenge detection employing linear SVM classifier with polarized and generic embedding [18]. Abusive text detection with variations of CNN and LSTM models has also achieved remarkable performance for native and code-switch language [28]. Spam message detection with sentiment analysis has also employed hybrid machine learning classifiers SVM and hybrid KNN algorithm. It has also applied optimization approach to enhance the accuracy of spam message detection and it achieved 99.82% accuracy considering three benchmark datasets [29]. Another sentiment analysis research considered efficient feature selection approach with meta-heuristic genetic algorithm for online customer reviews along with various machine learning classifiers like AdaBoost, XGBoost, Gradient Boost, Random Forest and many more. This approach achieved 77% to 78% accuracy with various benchmark datasets of sentiment classification task [30].

The above social media research works show that revenge text detection is a fresh research path and that needs more findings and analysis. Along with active and passive revenge, pro and petty revenge classification is also a difficult task and needs contextual semantic analysis. This study considers contextual text analysis for improved accurate classification of revenge posts. The proposed work considers paragraph embedding model for revenge text detection that also have applied in various NLP applications. Information retrieval considered PV-DBOW model where it observed that language estimation performance improved for paragraph embedding [31]. Paragraph embedding model also achieved improved performance for social network computation that efficiently approximate closeness centrality measure [32]. Paragraph embedding model also shown effective performance for sentiment analysis domain. Both the models PV-DBOW and PV-DM achieved more than 75% accuracy for sentiment text classification [33]. Along with paragraph embedding model, many NLP application also considered combination of lexical and semantic features that shows effective performance in present state of the art for various NLP applications. In short text classification, top n different words based on each category considered for lexical feature computation and word map with their specific weight considered for semantic feature. The combination of both feature helps to detect right topic for short text classification [34]. Another NLP application for question classification employed combination of semantic, syntactic and lexical features where it considered hypernyms and question category as semantic feature, question pattern and headwords as syntactic feature, and word n-gram and word shape as lexical feature. It was also observed that combination of all approach improved the performance of question classification [35]. Grading system of automatic short answer is another NLP application where combination of lexical and semantic feature was employed and that shows improved performance [36]. Performance of different NLP application with combination of various features motivate this research to consider combination features.

Boosting algorithm is mostly applied to boost weak learners and classifiers to achieve a higher accuracy of the classifier [37]. The most popular boosting model is the Adaptive boosting or Adaboost algorithm. It emphasizes on the misclassified samples where high accuracy based weak classifier have high weight [38]. In very recent research, three ensemble gradient boosting methods [39] show competitive state of the art classification results – XGBoost or extreme gradient boosting [40], Light GBM or light gradient boosting [41] and CATBoost or categorical feature based boosting [42].

This related research section presents the performance of existing gradient boosting based text classification research to show the significance of CATBoost classifier in this research. One recent research has applied Light GBM classifier for large recommendation dataset with 160 M tweets. Author has considered various engagement of tweets with Bert training model. As per the research, Light GBM text classification model is well oriented and fast for large tweet dataset [43]. Another online sexism and harassment detection research has employed XGBoost and CATBoost classifiers with LSTM model. This research shows improved performance with XGB classifier and CAT Boost classifier for social media text classification [44]. Light GBM classifier has also shown achievable performance for sentiment short and natural text classification where it considers domain free data [45]. This research also considers slang words as a feature. Another sentiment based prediction for fluctuation of crypto currency price has considered tree ensemble XGBoost classifier where that also accomplished good performance with tenfold cross validation [46]. Along with above text classification research, another social media aggression detection has employed CATBoost classifier. This research detects aggression and misogyny contents from social media data with TF-IDF and bag of words feature [47].

The CATBoost is a recently developed gradient boosting classifier but it already shows its efficiency in various textual applications as well as other applications. Most of the gradient boosting classifiers show effective performance but CatRevenge considers categorical features based CATboost that reduce overfitting problems for various text datasets.

3 CatRevenge methodology

This proposed CatRevenge model consists of five main tasks where the preprocessing of raw revenge text is an initial task. After preprocessing of text, next part is feature extractions. This model considers three feature extraction approaches to enrich the model efficiency. The three feature extraction approaches are – impact weight analysis with syntactic feature, lexical feature with TF-IDF vector, and last is semantic feature with Paragraph Embedding. POS tagging impact analysis converts data as a sparse row matrix and concatenate with TF-IDF vector. After that, stack paragraph embedding vector in horizontal sequence with above feature matrix. These all feature extraction methods help to analyze the text in depth level where as feature selection methods generally considers to find the subset of existing features. Analysis of text is important part to find actual context. Using various feature selection methods it can removes the important words that contain contextual meaning of text. In existing approach, this research considers three important feature extraction approaches where it considers the relevant importance of word and semantic similarities of text. With feature selection methods the existing long text can remove many important words that effect the model performance. So, final feature matrix considers for classification with CATBoost classifier.

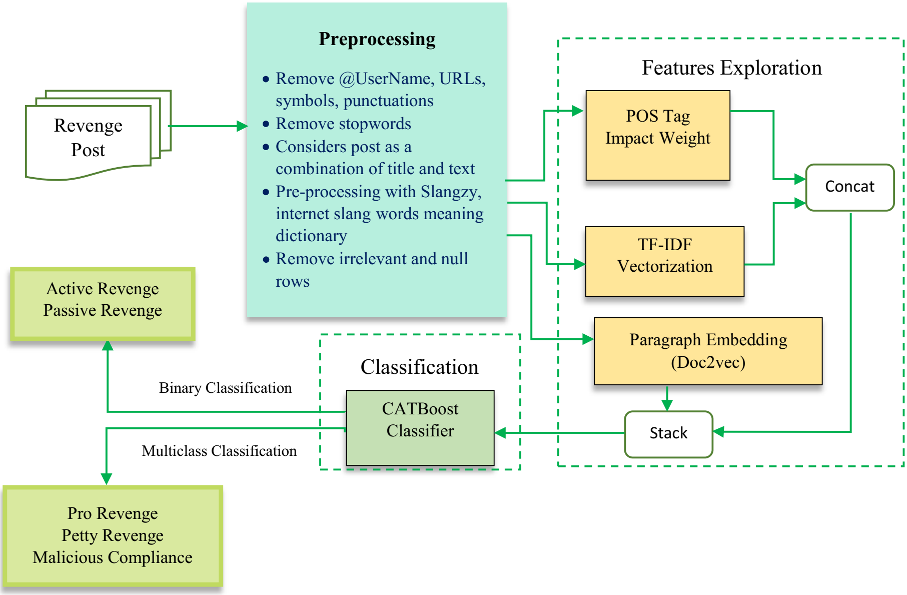

Revenge Post
Preprocessing
• Remove @UserName, URLs, symbols, punctuations
• Remove stopwords
• Considers post as a combination of title and text
• Pre-processing with Slangzy, internet slang words meaning dictionary
• Remove irrelevant and null rows
Features Exploration
POS Tag Impact Weight
Concat
TF-IDF Vectorization
Paragraph Embedding (Doc2vec)
Stack
Classification
CATBoost Classifier
Binary Classification
Multiclass Classification
Active Revenge
Passive Revenge
Pro Revenge
Petty Revenge
Malicious Compliance

Fig. 2 Revenge text detection framework

Both binary and multiclass classification follows same model structure. Figure 2 describes the model architecture for revenge text classification and all the steps to detect revenge text illustrates below. Algorithm 1 presents complete flow of CatRevenge model with all five steps – cleaning and preprocessing, POS tag based impact weight, TF-IDF, Paragraph Embedding and CATBoost classifier.

3.1 Cleaning and preprocessing

This study employs various cleaning and preprocessing steps for social media revenge posts. Analysis of complex human behavior from social media posts initially removes author information columns, symbols, and punctuations. It also removes irrelevant columns and null rows from the dataset. This research removes stopwords using NLTK python library [48]. Elimination of stopwords means removing noise that helps to enhance the speed of processing time. This research also considers slang word meanings from the fuzzy logic based slangzy dictionary to include significant information about tokens for revenge text [49]. Preprocessing steps also combine two columns for more accurate detection of revenge posts. In this analysis, title column and post column considers as a single post column.

```
Input: Revenge Posts $R_1, R_2, R_3, \dots\dots, R_n$
Output: Revenge Posts Binary Classification (Active/Passive), Revenge Posts Multiclass Classification (Pro/Petty/Malicious Compliance)
Value Initialization: vector dimension = 300,
        window size k = 2
Step 1: Preprocessing of Revenge Posts
        Combine title and posts column(R)
        For each post $\{R = R_1, R_2, R_3, \dots\dots, R_n\}$ do
            $R_i$ = cleaning($R_i$)
            $R_i$ = remove stopwords($R_i$)
            $R_i$ = remove irrelevant and null posts ($R_i$)
            $R_i$ = Slangzy ($R_i$)
        Return posts($\{R = R_1, R_2, R_3, \dots\dots, R_n\}$)
Step 2: POS Tag Impact Weight computation
      $t_1, t_2, t_3, \dots\dots, t_m$= Tokenize($R_i$)
      For each term $\{t = t_1, t_2, t_3, \dots\dots, t_m\}$ do
          $p_l$ = POS tag ($t_l$)
      For each post $\{R = R_1, R_2, R_3, \dots\dots, R_n\}$ do
          $RW_i$ = impact weight ($R_i$, p) by equation (1)
      Return impact weight ($RW_1, RW_2, \dots, RW_n$ )
Step 3: Term Frequency and Inverse Document Frequency computation
      For each post $\{R = R_1, R_2, R_3, \dots\dots, R_n\}$ do
          $RT_i$= TF-IDF($R_i$, p) by equation (4)
      Return TF-IDF vector ($RT_1, RT_2, \dots, RT_n$ )
Step 4: Paragraph vector for Revenge Posts
  For each post $\{R = R_1, R_2, R_3, \dots\dots, R_n\}$ do
      $RP_i$= Paragraph Embedding($R_i$, p) by PV-DM
  Return PV-DM vector ($RP_1, RP_2, \dots, RP_n$ )
Step 5: Combine all feature vectors
For all vectors $\{RW, WT, RP\}$ do
   $RW$ = sparse row matrix ($RW$)
   WT = concat vectors(RT, RW)
   PWT = stack (RP, WT, horizontal sequence) Return (PWT)
Step 6: Classification of Revenge Posts
For each post $\{R = R_1, R_2, R_3, \dots\dots, R_n\}$ do
      $RC_i$ = CATBoost Classifier (PWT vectors, $R_i$)
Return($RC_i$)
```

Algorithm 1 Revenge Text classification algorithm

3.2 Impact Weight Analysis with syntactic feature

This research implements syntactical features-based impact weight analysis where it considers POS tagging. Parts of speech tagging classification system reveal the role of a term in a particular context. The English language considers eight POS tags – verb, noun, adjective, adverb, pronoun, preposition, interjection, and conjunction. POS tag is a significant approach for text analysis that helps to show the sentence and word relations. Grammar checking, text to speech, and word sense disambiguation of words are important POS tags applications in text analysis. NLTK python library supervised learning approach to determine POS tag [50]. NLTK POS tagger uses 35 POS tags.

Traditional POS tag does not represent the numerical impact weight analysis of each tag. This study considers the POS tag syntactic feature and computes the impact weight of each tag in a post. Noun, verb, adjective, and adverb POS tags assign for impact weight analysis. It computes the significance of a tag in the corpus. Syntactic feature like POS tags can able to obtain hidden information from textual data. It also extracts the canonical form of a term that helps to analyze syntactical information from a post [50].

$$POS - TAG_{ij} = \frac{|tag_{ij}|}{|tag_j|} \tag{1}$$

where, $|tag_{ij}|$ is a denoted as a total number of frequencies for each POS tag i for each tagged corpus j. $|tag_j|$ denoted as a total number of frequencies for all POS tags and for each tagged corpus j. $POS - TAG_{ij}$ is set of POS tag vectors for each tag in tagged corpus j. The $POS - TAG_{ij}$ vector finally converts into a compressed format of sparse row matrix that considers a syntactic feature.

3.3 Lexical feature with TF-IDF

Lexical feature analysis is an important NLP process that intuits the importance of keywords, meanings, context, and the relation between terms. Keywords and terms represent significant content from documents and sentences that enrich the classification task.

TF-IDF is a statistical-based important keyword or term extraction approach that effectively extracts keywords and terms from documents. It consists of two methods: Term frequency or TF and Inverse Document Frequency or IDF. TF computes the frequency of keyword or term occurrence for a particular document. Variation of document lengths can directly change keywords frequency. So, TF measures the ratio of term frequency and document length.

$$TF_{ij} = \frac{t_{ij}}{\sum_k t_{kj}} \tag{2}$$

where frequency of term i is t_{ij} and term available in document j. IDF computes the relevance of keywords because TF considers all terms with equal importance. So it requires to measures the importance of keywords. IDF considers log value to cut down the weight for less important keywords.

$$\mathrm{IDF}_i = \log \frac{|D|}{1 + |D_i|} \tag{3}$$

where the total documents number is denoted as $|D|$ and the total documents number with term i is $|D_i|$. So, TF-IDF is a combination of term frequency measures and inverse document frequency measures.

$$\mathrm{TF} - \mathrm{IDF} = TF_{ij} * \mathrm{IDF}_i = \frac{t_{ij}}{\sum_k t_{kj}} * \log \frac{|D|}{1 + |D_i|} \tag{4}$$

The revenge post categorization research applies the TF-IDF feature to extract the relevance and occurrence of each word in a context. Numeric representation of a corpus

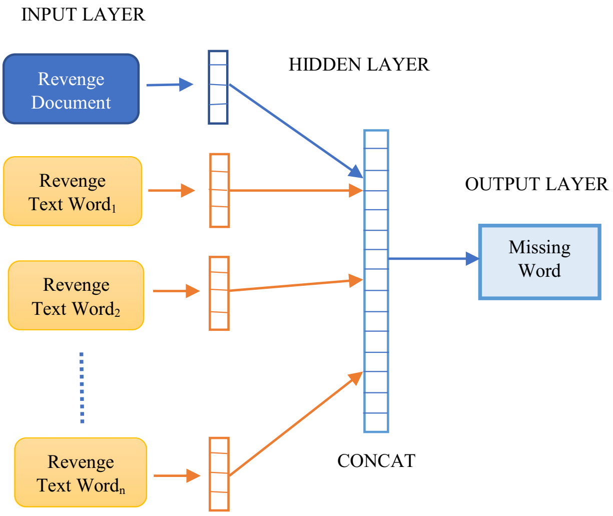


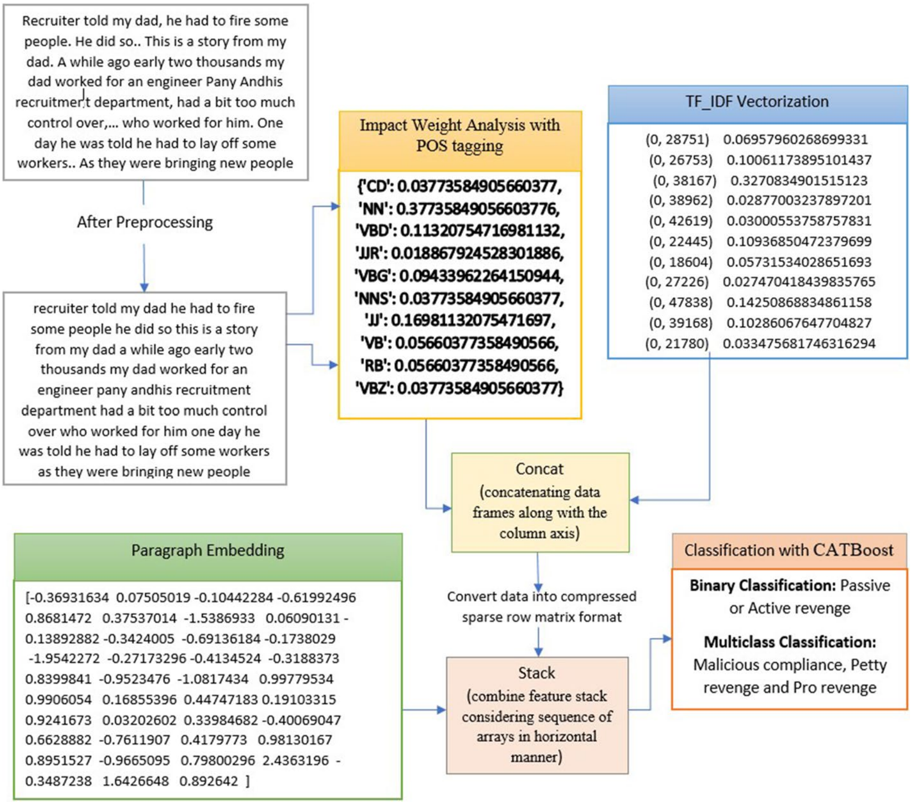
Fig. 3 Paragraph Embedding model with Revenge text

shows characteristics of textual data where TF-IDF represents numeric representation of occurrence. This feature normalizes word occurrence based on document size as well as the contribution of words in the corpus. TF-IDF vectors empower NLP models and have enormous applications in the NLP domain [51].

3.4 Semantic Feature with Paragraph Embedding

Along with syntactic and lexical features, this research requires a suitable semantic feature to extract similarities from revenge text. This research could have applied syntactic and lexical features for classification but to enhance further refinement, it applied semantic features to compute the comparative distance between each term in the revenge context. Paragraph embedding learns vector representations to plot each term in such a manner that it can compute numerical distances between each term based on the context. With lexical and syntactical features, revenge detection model efficiently extracts the semantic meaning of active and passive revenge text to represent the context of each term.

In existing research, various semantic embedding approaches have been applied to several text analysis research. Word2Vec and Paragraph Embedding represent low dimensional vectors for each word or document. Word2Vec model [52] represents high semantic similarities of terms in continuous space whereas the Paragraph Embedding model [53] uses distributed memory model to represent terms and paragraphs as low dimension vectors. This research also implemented the Word2Vec model for semantic embedding features but the Paragraph Embedding model shows better results compared to the Word2Vec model. Along with that BERT is also an efficient embedding model for text analysis. Various NLP research in text analysis including hate speech detection [25], cyberbullying detection [27], short answer grading [36] has utilized BERT embedding model and it has shown improved contextual analysis compared to the other embedding model. This research considers paragraph embedding model due to the long text and each post shared as a paragraph with various connected sentences. In future research, long revenge stories will evaluated with various other embedding models including BERT embedding model to explore the contextual analysis more.


Recruiter told my dad, he had to fire some people. He did so.. This is a story from my dad. A while ago early two thousands my dad worked for an engineer Pany Andhis recruitment department, had a bit too much control over,... who worked for him. One day he was told he had to lay off some workers.. As they were bringing new people
After Preprocessing
recruiter told my dad he had to fire some people he did so this is a story from my dad a while ago early two thousands my dad worked for an engineer pany andhis recruitment department had a bit too much control over who worked for him one day he was told he had to lay off some workers as they were bringing new people
Impact Weight Analysis with POS tagging
TF_IDF Vectorization
Concat (concatenating data frames along with the column axis)
Paragraph Embedding
Convert data into compressed sparse row matrix format
Stack (combine feature stack considering sequence of arrays in horizontal manner)
Classification with CATBoost
Binary Classification: Passive or Active revenge
Multiclass Classification: Malicious compliance, Petty revenge and Pro revenge

```
{'CD': 0.03773584905660377,
'NN': 0.37735849056603776,
'VBD': 0.11320754716981132,
'JJR': 0.018867924528301886,
'VBG': 0.09433962264150944,
'NNS': 0.03773584905660377,
'JJ': 0.16981132075471697,
'VB': 0.05660377358490566,
'RB': 0.05660377358490566,
'VBZ': 0.03773584905660377}
```

```
(0, 28751)   0.06957960268699331
(0, 26753)   0.10061173895101437
(0, 38167)   0.3270834901515123
(0, 38962)   0.02877003237897201
(0, 42619)   0.03000553758757831
(0, 22445)   0.10936850472379699
(0, 18604)   0.05731534028651693
(0, 27226)   0.027470418439835765
(0, 47838)   0.14250868834861158
(0, 39168)   0.10286067647704827
(0, 21780)   0.033475681746316294
```

```
[-0.36931634  0.07505019 -0.10442284 -0.61992496
0.8681472   0.37537014  -1.5386933   0.06090131 -
0.13892882 -0.3424005  -0.69136184 -0.1738029
-1.9542272  -0.27173296 -0.4134524  -0.3188373
0.8399841  -0.9523476  -1.0817434   0.99779534
0.9906054   0.16855396  0.44747183 0.19103315
0.9241673   0.03202602  0.33984682 -0.40069047
0.6628882  -0.7611907   0.4179773   0.98130167
0.8951527  -0.9665095   0.79800296  2.4363196 -
0.3487238   1.6426648   0.892642 ]
```

Fig. 4 CatRevenge Model workflow with sample example

The revenge detection framework employs the Paragraph Embedding vector model that can able to learn semantic relatedness from revenge posts describes in Fig. 3. The semantic embedding Paragraph Embedding model aims to analyze semantic similarities between various terms in the revenge context. The Paragraph Embedding model represents low dimension vectors for text data and that learn token vectors with distributed memory. It considers Distributed Memory Vector or PV-DM to recall missing words in a related context. In comparison with the PV-DBOW or distributed bag of words model, the PV-DM model performs better in revenge text classification context. It maps each post and word to a unique vector.

Along with the above details of paragraph embedding, this research analyzes vector dimensions and window size value, especially for revenge text classification research. Window size value 6 is applied for broader contents whereas 2 is used for smaller and more focused contents. k = 2 considers w-2, w-1, w + 1, w + 2 context words for target term w. The Smaller window size with focused revenge text contains shows better performance. This research experimented with various window sizes k = 2, 3, 5, 6 and vector dimensions 50, 100, 200, 300, 600, and 800. The best performing parameter values are k = 2 and dimension 300.

The Revenge detection model considers three features with various models like Paragraph Embedding, impact weight with POS tagging and lexical feature with TF-IDF. The proposed model considers revenge text from social media. Figure 4 shows the flow of sample revenge text with all feature exploration that considers concatenation between POS tag impact weight and TF-IDF vectorization. It also considers stack with paragraph embedding vectors. “Concat” or concatenation method mainly considers for concatenating data frames of TF-IDF vectors and value of POS tag impact weight along with the column axis. This combine feature stack with paragraph embedding vectors by considering sequence of arrays in horizontal manner to make single array.

3.5 CATBoost classifier

The proposed study considers one gradient boosting machine learning classification algorithm to detect categories of revenge text. Gradient boosting models can efficiently handle learning problems with noise data and heterogeneous features for various NLP research. Generally, gradient boosting considers decision trees for base prediction. Numerical features are convenient for decision trees but various datasets consider categorical features to improve prediction. Categorical features are discrete value sets and not comparable to each other. Generally, categorical feature converts into numeric before training. Gradient boosting effectively accelerates the classification tasks and also reduces the consumption of memory.

The CATBoost classifier [54] is a new non-linear, tree-based gradient boosting algorithm that can effectively handle categorical features [55]. Generally boosting algorithms build new tree to compute the model gradients where CATBoost classifier enhance the existing model by reducing overfitting problems. There are several advantages of the CATBoost algorithm like it supports categorical features, it uses a new schema to reduce model overfitting, and it predicts faster and shows good performance for heterogeneous data. The CATBoost classifier shows achievable performance compare to the Light GBM, XG Boost algorithms for various applications [39].

Categorical boosting or CATBoost aims to reduce the shift in the training phase. This shift arises because gradient boosting applies the same instances for gradient and model estimation to minimize gradients. CATBoost provides the solutions for this shift problem where it estimates gradients by applying sequences of base models and it excludes that particular instance from the training set. CATBoost model considers various hyper parameters for classification tasks – learning rate, depth of the tree, iterations for leaf estimations, and regularization coefficient. This study considers latest optimization tool Optuna for experimentation and improvement of CatBoost model.

Preliminary revenge text research has only implemented some machine learning classifiers, but this research experiments with various deep learning and gradient boosting algorithms. Gradient boosting improves training efficiency and classification accuracy for revenge text. This study considers two classes for binary classification and three classes for multiclass classification. The two classes for binary classification are defined as y ϵ {0, 1} where y = 0 for active revenge and y = 1 for passive revenge. The three classes for multiclass classification defined as y ϵ {0, 1, 2} where y = 0 for Malicious compliance, y = 1 for pro revenge and y = 2 for petty revenge. This research considers this data for training and testing purpose $\{(X_1, y_1), (X_2, y_2) \ldots \ldots .(X_i, y_i), \ldots \ldots (X_n, y_n)\}$, where X_i is the final feature vectors and y_n is the target class.

4 Experiments, results, and findings

This section initially presents the experimentation setup with dataset details, baseline model, and evaluation metrics for this research. Along with that, the next part of this research presents results and findings with various experimentations – impact analysis for various features, performance with paragraph embedding, comparison with other classifiers, and baseline model.

4.1 Experimentation setup

This section initially presents the dataset details and after that it presents the brief of state of the art models and evaluation metrics for comparisons.

4.1.1 Dataset

This study considers one revenge dataset to analyze complex human behavior. The revenge Reddit (social networking site) dataset considers for this research with subreddit (class) of active revenge and passive revenge for binary class and malicious compliance, petty revenge and pro revenge for multiclass. The dataset is extracted from a specified location[1] in csv file format with all details of dataset. It considers three subreddits or topics that pulled the most recent 550 days of data. After preprocessing, the dataset contains 11,189 English posts and task A considers binary classification and task B considers multiclass classification. Both binary and multiclass classification approaches are related to each other. This Reddit dataset is already labeled and to solve any discrepancy, this research appointed two subject matter experts for accurate labelling process.

4.1.2 State-of-the-art model

The revenge text classification is a fresh research direction, so this work can able to consider one State-of-the-art models for this research.

Vengeful Text This study considered Adaboost classifier, POS tag and Word2Vec embedding. KNN classifier with same features also considers as a baseline model [12].

4.1.3 Evaluation Metrics

This revenge detection study applies four important evaluation metrics – Accuracy, Weighted F1 (F1), Precision (P) and Recall (R). Accuracy and weighted F1 metrics consider for best analysis of various classification models and baseline implementation.

$$P = \frac{T_P}{T_P + F_P} \tag{5}$$

[1] https://github.com/ebsiegs/subreddit_nlp/blob/main/data/subreddit_data.csv

$$R = \frac{T_P}{T_P + F_N} \tag{6}$$

$$F1 = 2 * \frac{R * P}{R + P} \tag{7}$$

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \tag{8}$$

where T_P denoted as true positives, T_P denoted as false positives, T_P denoted as true negatives and T_P denoted as false negatives.

This revenge detection model implements with various python libraries. Reddit dataset collected from Github location and that already specified in dataset details section. The experimentation section considers 25% of data for testing in a stratified fashion. The result section is divided into two tasks – the first section shows the performance of the active and passive revenge classification model and the second section analyzes the performance of malicious compliance, petty and pro revenge classification model.

4.2 Results and Findings

This results and findings section analyze the performance of the CatRevenge model. It studies the impact analysis with various features along with CATBoost classifier. In next part it compares various machine learning and deep learning models with a combination of CatRevenge feature set. The final part represents the performance improvement compare to the baseline model.

4.2.1 Impact Analysis for Features

Feature extraction from social media text is an important part for textual analysis. The CatRevenge model consists of various features where this section shows the impact analysis of various features and feature set. This study analyzes the impact of the CatRevenge feature sets with other features for both binary and multiclass classification models. To compare with various other features and feature set this study considers – Bag of Words, Word2Vec, Paragraph Embedding, count vector + TF-IDF, POS tag impact weight + Paragraph Embedding, TF-IDF + Paragraph Embedding, and proposed CatRevenge feature set POS tag impact weight + TF-IDF + Paragraph Embedding. All features and combinations of features consider the CATBoost classification algorithm and (Table 1) presents the comparison with P, R, and F1 evaluation metrics. It also shows the performance of features and feature sets with binary and multiclass classification.

Table 1 shows the performance of the CatRevenge feature set outperformed compared to any feature and feature set. Apart from the CatRevenge feature set, POS tag impact weight + Paragraph embedding feature combination shows a better F1 score compared to other features and feature sets. It was identified that POS tagging based analysis helps to analyze the importance of relevant word in text. It was also observed that semantic analysis with paragraph embedding model achieved better F1 score compare to word2vec embedding model for both cases. In Reddit revenge dataset each post contain long sentences and paragraph like contents, so analysis between sentences and words are more effective with paragraph embedding model. So, the combination

Table 1 Impact analysis of various features and feature set

Sl. No	Features	Revenge Detection (Binary Classification)			Revenge Detection (Multiclass classification)		
		P	R	F1	P	R	F1
1	Bag of Words	0.65	0.58	0.59	0.64	0.67	0.64
2	Word2Vec	0.70	0.67	0.67	0.72	0.72	0.72
3	Paragraph Embedding	0.79	0.77	0.79	0.76	0.77	0.77
4	Count vector + TF-IDF	0.76	0.77	0.77	0.75	0.76	0.76
5	POS tag + Paragraph Embedding	0.84	0.83	0.84	0.75	0.76	0.76
6	TF-IDF + Paragraph Embedding	0.82	0.82	0.82	0.78	0.77	0.78
7	POS tag + TF-IDF + Paragraph Embedding (CatRevenge feature set)	0.87	0.87	0.87	0.80	0.81	0.80

Table 2 Comparison with both Paragraph Embedding Models

Sl. No	Paragraph Embedding	Model	Revenge Detection (Binary Classification)			Revenge Detection (Multiclass classification)		
			P	R	F1	P	R	F1
1	PV-DBOW	CatRevenge	0.81	0.80	0.81	0.73	0.74	0.73
2	PV-DM		0.87	0.87	0.87	0.80	0.81	0.80

between POS tag impact weight analysis and contextual semantic analysis are effective in long paragraphs like posts and that enrich better F1-score. Along with above feature extraction methods, CatRevenge considered TF-IDF feature to improve efficiency of the model more.

4.2.2 Performance with paragraph embedding

This CatRevenge model shows the performance of paragraph embedding vectors with various dimensions in Fig. 4 and Table 2 shows two types of paragraph embedding model performance – PV-DM and PV-DBOW. It was observed that the change of dimensions reflects the CatRevenge model efficiency. It was also observed that the performance of the PV-DM model performs better compared to the PV-DBOW model for the revenge text dataset. As PV-DM model contextually analyzes the contents and is able to find the missing terms. This unsupervised approach can efficiently analyze the text and context of the word. CatRevenge model considers PV-DM that outperformed other classifiers with the same feature set.

With the experimentation of various dimensions, Fig. 5 presents a clear comparison for binary and multiclass classification where 300 vector dimension shows better performance for long text. The CatRevenge model considers model dimension 300 for better results.

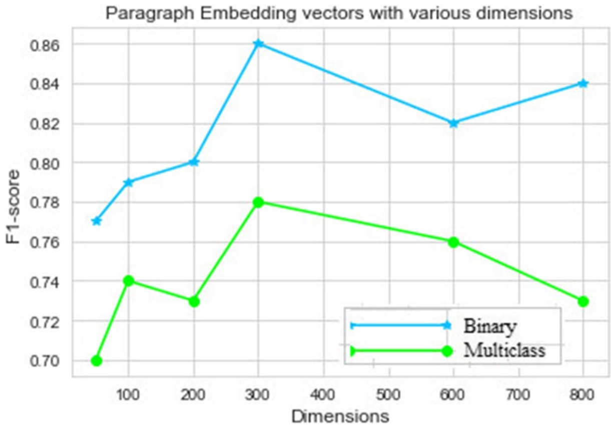


Fig. 5 Paragraph embedding vectors with various dimensions

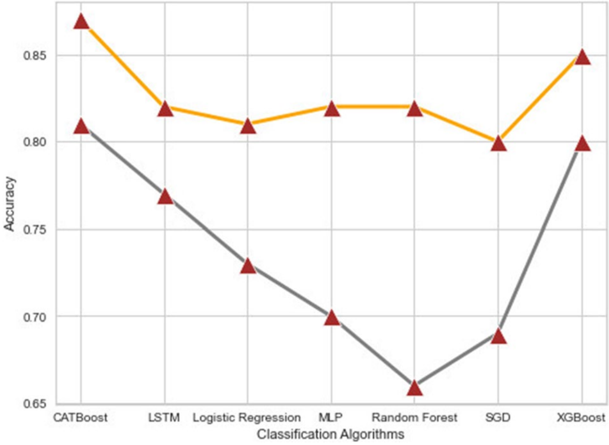


Fig. 6 Accuracy comparison with all other machine learning and deep learning classifiers

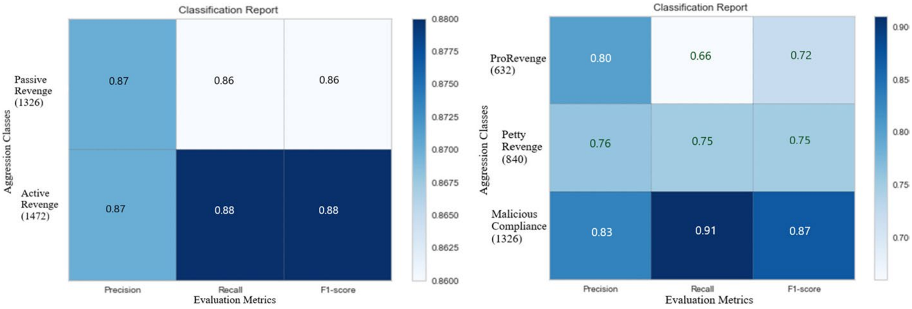
4.2.3 Comparison with Other classification models

This work compare the CatRevenge model with seven machine learning and deep learning models – SVM, Random Forest, MLP, Logistic Regression, XGB Boost, Naïve Bayes, and LSTM. The machine learning and deep learning models are considered the combination of three feature sets for comparing proposed model's performance. Along with that LSTM model considers epoch size 5, batch_size 64, 'crossentropy' loss function and 'adam' optimizer. MLP classifier considers Relu activation function, 'adam' optimizer, and epoch number is 300. XGBBoost classifier considers learning rate 0.05, and max depth of a tree 5. Other all classifiers consider default parameter values. These all machine learning and deep learning models consider to evaluate the performance of proposed CatRevenge model.

Table 3 presents this comparison and considers the CatRevenge feature set with various classifiers. It is observed that Gradient boosting algorithms (XGB Boost and CATBoost) perform better compare to other models. It considers P, R and F1 for evaluation measures and Table 3 shows the comparison for binary and multiclass classification. The CatRevenge model with CATBoost algorithm outperforms all the machine

Table 3 Comparison with various Machine Learning and Deep Learning classifiers

Sl. No	Classification Algorithms	Revenge Detection (Binary Classification)			Revenge Detection (Multiclass classification)		
		P	R	F1	P	R	F1
1	SVM	0.75	0.74	0.74	0.65	0.58	0.59
2	Random Forest	0.82	0.82	0.82	0.70	0.67	0.64
3	Multi-Layer Perceptron (MLP)	0.82	0.82	0.82	0.75	0.75	0.75
4	Logistic Regression	0.81	0.81	0.81	0.73	0.74	0.73
5	XGB Boost	0.86	0.86	0.86	0.79	0.77	0.78
6	Naïve Bayes	0.76	0.77	0.77	0.75	0.76	0.76
7	LSTM	0.82	0.82	0.82	0.78	0.78	0.78
8	CATBoost (CatRevenge)	0.87	0.87	0.87	0.80	0.81	0.80



(a)

(b)

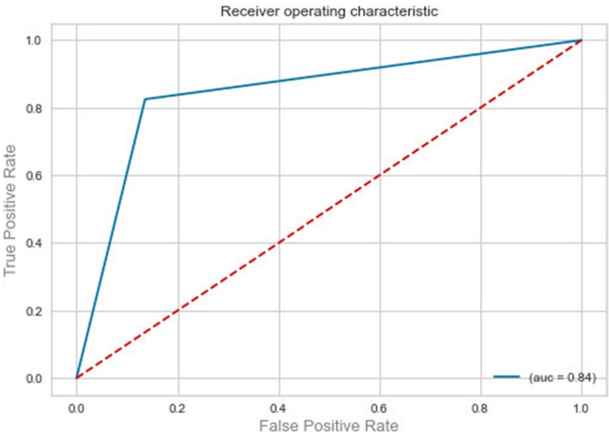
Fig. 7 Classification Reports for CatRevenge Model with (**a**) Binary Classification (**B**) Multiclass Classification

learning and deep learning classifiers. Figure 6 shows the clear evaluation line plot of CatRevenge model with all other machine learning and deep learning classifiers. Here, CATBoost classifier achieved 0.87 F1-score for binary revenge classification. In same manner, CATBoost classifier also achieved 0.80 F1-score for multiclass revenge text classification.

Along with above binary and multiclass revenge post classification results table and bar plots, it also plot two heatmap classification reports. In Fig. 7 (a), it is observed that testing data contains almost equal numbers of active and passive revenge. F1-score is 0.88 for active revenge and precision, recall metrics also show better value for active revenge. In Fig. 7 (b), it is observed that testing data contains more malicious compliance revenge posts compared to the pro and petty revenge posts. F1-score is 0.87 and recall is 0.91 for malicious compliance. Both the multiclass and binary revenge post classification shows achievable performance with CATBoost classifier.

Figure 8 and Fig. 9 present ROC curves for binary and multiclass classifiers. ROC curve shows the comparison between false-positive rates and true positive rates. Figure 9 represents ROC for three class where class 0 denotes pro revenge, class 1 denotes petty revenge and class 2 denotes malfigicious compliance. AUC values for both binary and multiclass classification show achievable performance of CatRevenge model.

Fig. 8 ROC Curve – Binary classification



4.2.4 Performance analysis of CatRevenge models with sample text

The analysis of some example dataset is considered to show the performance of proposed CatRevenge model for binary and multiclass classification. Each of the post from Reddit dataset are very long, so we considered three.

examples in Table 4 and there annotated labels. Along with that Table 4 also shows the predicted labels of this posts. It was observed that some samples are wrongly classified by proposed model. In Table 4, second post is wrongly classified by proposed model. It was also observed that some of the posts are wrongly classified for binary and multiclass classification. Long text contains various contextual information and this may effects in model performance.

4.2.5 Comparison with machine learning models and baseline model

This CatRevenge model compare with other models and state-of-the-art model in Table 4 with considering revenge text dataset. This is the fresh research area, so any standard datasets is not available for evaluation. This research considered other machine learning models with count vector and TF-IDF feature set. It considered two machine learning models Naïve Bayes classifier and Random Forest classifier.

This research also considered one baseline model vengeful text identification [12] with POS tagging, word2vec embedding and KNN, AdaBoost classifiers. This comparison table considered the standard revenge text dataset to evaluate the performance of machine learning models and baseline model Table 5.

This research considers weighted F1 as main metric, it is observed that the CatRevenge model increases 6—10% weighted F1 compared to the baseline model for binary classification. Binary and multiclass classification of revenge post shows achievable performance with Gradient boosting classifiers with all three features. It is observed that the CatRevenge model increases 2.5—10% weighted F1 compared to the baseline model for multiclass classification

In order to evaluate the statistical significance of proposed CatRevenge model, this research considers McNemar's test [56] to analyze the paired observation. This research

Table 4 Performance analysis of CatRevenge model for binary and multiclass classification with sample text

Sl No	Sample Reddit Text	Reddit Text after Processing	Annotated label		Predicted Label		Result
			Binary	Multi-class	Binary	Multi-class	
1	My boss loves to call me at 6. 15 a.m. to ask me, if I would like to fill the shifts of the people, who just called in sick. This is an everyday thing. I was bored and frustrated. So I decided to volunteer at 3.30 a.m. to call this same manager to ask if they needed extra help he got. Super pissed and tried to write me up for it. I showed the gm the time stamps of the calls I has received, I dont get calls anymore	my boss loves to call me at 6 15 a m to ask me if i would like to fill the shifts of the people who just called in sick this is an everyday thing i was bored and frustrated so i decided to volunteer at 3 30 a m to call this same manager to ask if they needed extra help he got super pissed and tried to write me up for it i showed the gm the time stamps of the calls i has recieved i dont get calls anymore	Passive	Malicious Compliance	Passive	Malicious Compliance	It analyse the text correctly
2	VOM bag, this one time I was in a shopping mall with my friend and his little brother his little brother felt sick. So I went into the closest store witch happened to be a high end womans clothing store, I asked the lady behind the counter. May I please have a bag, my friends little brother dosen't feel well. She replied no, I then proceed to bring the little guy into the store. She gave me a bag	vom bag this one time i was in a shopping mall with my friend and his little brother his little brother felt sick so i went into the closest store witch happened to be a high end woman s clothing store i asked the lady behind the counter may i please have a bag my friends little brother dose not feel well she replied no i then proceed to bring the little guy into the store she gave me a bag	Passive	Malicious Compliance	Active	Petty Revenge	Wrongly classified Text

Table 4 (continued)

Sl No	Sample Reddit Text	Reddit Text after Processing	Annotated label		Predicted Label		Result
			Binary	Multi-class	Binary	Multi-class	
3	That’s going to be a good night, my boyfriend is kind of passive aggressive.. When he cannot occupy my main pc to play low game and watch dumb video. I have a laptop.. he could use to but he refuse and he doe this until 3am plus. He scream and shout in the mic like he wa in the middle of Saskatchewan or something like that what’s my revenge. He.. he.. I have to wake up tomorrow at 6am I like to wake up on black metal and or hard bass and I have a very noise coffee machine	that s going to be a good night my boyfriend is kind of passive aggressive when he cannot occupy my main pc to play low game and watch dumb video i have a laptop he could use to but he refuse and he doe this until 3am plus he scream and shout in the mic like he wa in the middle of saskatchewan or something like that what s my revenge he he i have to wake up tomorrow at 6am i like to wake up on black metal and or hard bass and i have a very noise coffee machine	Active	ProRevenge	Active	ProRevenge	It analyse the text correctly

Table 5 Experiment results for Revenge detection. Binary Revenge classification considers active and passive revenge and Multiclass Revenge classification considers Malicious Compliance, Petty Revenge and Pro Revenge

Sl.no	Features	Classification Algorithms	Revenge Detection (Binary Classification)				Revenge Detection (Multiclass classification)			
			Accuracy	P	R	F1	Accuracy	P	R	F1
1	Count vector + TF-IDF	Naïve Bayes	0.77	0.76	0.77	0.77	0.76	0.75	0.76	0.76
2	Count vector + TF-IDF	Random Forest	0.82	0.82	0.82	0.82	0.78	0.78	0.78	0.78
3	POS tag + Word2Vec [12]	AdaBoost	0.72	0.58	0.68	0.72	0.70	0.57	0.67	0.70
4	POS tag + Word2Vec [12]	K-NN	0.79	0.79	0.79	0.79	0.76	0.76	0.76	0.76
5	POS tag + Paragraph Embedding + TF-IDF (CatRevenge)	CATBoost	0.87	0.87	0.87	0.87	0.81	0.80	0.81	0.80

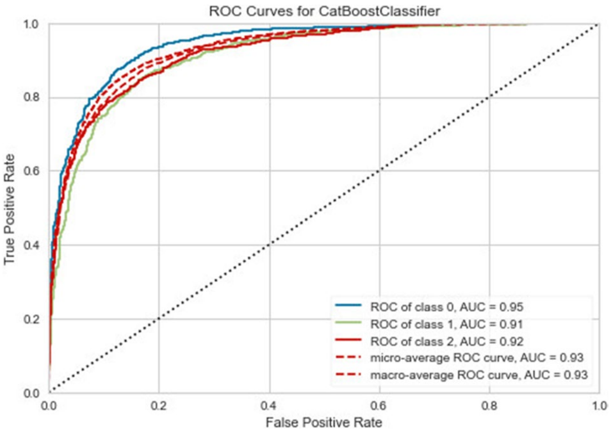
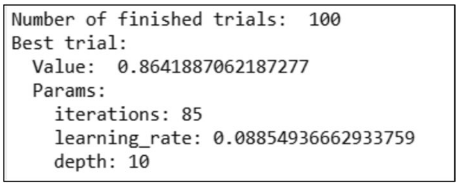


Fig. 9 ROC Curve – Multiclass classification



```
Number of finished trials:  100
Best trial:
  Value:  0.8641887062187277
  Params:
    iterations: 85
    learning_rate: 0.08854936662933759
    depth: 10
```

Fig. 10 Values with hyperparameter optimization

analyzed revenge text classification results for each post before and after considering baseline method in consideration with McNemar's test. This test can check the differences in error rates between two models with statistical analysis. The McNemar's test is mainly based on χ^2 and this is one of the best test for error rate analysis. This research considers McNemar's test because it shown low type 1 error. In experiment 1, proposed model compared with baseline model where it considers AdaBoost classifier [8] and it shown $\chi^2 = 32.33$ and p value < 0.001. In experiment 2, proposed model compared with baseline model where it considers K-NN classifier [12] and it shown $\chi^2 = 30.13$ and p value < 0.001. Both the experiments with McNemar's test show that, statistically significant change is observed in error rates for both methods.

Along with statistical test, the hyperparameter optimization method is considered for proposed CatRevenge model. This research considers latest hyperparameter optimization tool Optuna for experimentation and improvement of CatBoost model [57]. Optuna tool optimize tree based search for hyperparameter and it considers method TPEsampler. Optuna can also able to set superparameter. It can able to find the best hyperparameter. Total number of finish trials is considered as 100, and experimentation also shows the best parameters values for CatBoost classifier.

Above Fig. 10 shows the values with hyperparameter optimization. This research considers this hyperparameter values for k-fold cross validation.

The K-fold cross validation approach is considered for assessing the proposed model performance especially for machine learning classifiers [58]. This research considered 2, 5 and tenfold cross validation for binary and multiclass classification. Table 6 presents the detail results of K-fold cross validation approach. In binary classification, tenfold cross validation outperformed with achieving 0.859 accuracy, compared to the 2 and fivefold cross validation. In same manner, this research also evaluate multiclass classification with 2, 5

Table 6 K-fold cross validation with revenge text dataset

Sl No	Model	K-Fold	Accuracy of Binary Classification	Accuracy of Multiclass Classification
1	CatRevenge	2	0.856	0.801
2		5	0.857	0.806
3		**10**	**0.859**	**0.809**

and tenfold cross validation. It also outperformed for tenfold cross validation with achieving 0.809 accuracy. After considering all validation and testing, the proposed model shows achievable performance compare to other baseline model.

The automatic revenge text detection model is beneficial for society and social structure. Evaluation of proposed CatRevenge framework exhibits the improved performance for revenge text detection compare to the other state of the art strategies. It was observed that efficient classification of active and passive revenge can help to block the strong revenge from social media. The proposed framework also able to detect grievances and revenge stories from paragraph but some text suffers from misclassification. Revenge text consists of long sentences and long paragraphs. Revenge categorization is a challenging task based on long text, as complete paragraph shows mixed categories and ambiguous meaning. It was also observed that some of the revenge text consists of emojis that contain hidden information about revenge. In future research revenge text detection will consider emojis for more accurate classification.

5 Conclusion and Future Scope

The experimentation of the CatRevenge model revealed that the binary and multiclass revenge text detection accomplished satisfactory performance. This research successfully investigated the revenge posts detection approach with contextual semantics and the impact weight of each POS tagger for the English language. This CatRevenge model also considered gradient boosting classifier CATBoost and evaluated the performance with lexical, syntactical, and contextual semantic feature sets.

The result revealed that contextual semantic analysis with paragraph embedding is significant to categorize active revenge for long text. It also observed that Distributed Memory vector critically influenced the performance of the CatRevenge model. The results also indicated that impact analysis of each POS tag enhanced the classification accuracy and weighted F1 for long revenge text. Analysis of revenge posts with NLP, contextual semantic similarity extraction with paragraph vector is significant to produce meaningful classification results. Thus, this work considered a paragraph embedding model that creates vectors for each paragraph to reflect the contexts of revenge terms. Finally this work achieved 6% improvement for binary class and 2.5% improvement for multiclass revenge text classification with the weighted F1 score compared to the baseline model.

Research with revenge datasets is fewer, so the future path is vast. This study is restricted to a single English language revenge dataset, but the CatRevenge model can be used with diverse textual datasets from different sources and languages. Future work may implement negative emotion boundaries to detect implicit active revenge more efficiently. Emotion analysis can able to detect positive and negative emotion along with degree of negative emotion in text. Analysis of emotion level may help to categorize the revenge text in depth level. It was observed that some long revenge stories contain humor contents that misclassified the revenge text. So sarcasm detection or humor detection from long text may improve the misclassification error. In future research, long revenge stories will evaluated with various other embedding models including BERT embedding model to explore the contextual analysis more. Emoji is a symbol that contain hidden information about the complete text and it can also contain hidden expression. Analysis of hidden emotions, sentiments and information from emojis can also enhance the classification performance for revenge text detection. Revenge text in social media may also contain code-mix languages with other low resource languages. So in future, analysis and processing of code-mix revenge text may enhance the performance of revenge text detection model. Minimization of misclassification errors can also enhance the model accuracy. Furthermore, it is acclaimed that the CatRevenge model may improve the efficiency of various aggressive, hate, and cyberbullying detection research.

Funding No funding was received for conducting this study.

Data availability As per request.

Code availability As per request.

Declarations

Competing interests The authors have no financial or proprietary interests in any material discussed in this article.

Conflicts of interest/Competing interests Not Applicable.

References

1. Finances Online [electronic resource] (2022) 53 important statistics about how much data is created every day in 2024 - Financesonline.com. https://financesonline.com/how-much-data-is-created-every-day
2. Statusbrew [electronic resource] (2022) https://statusbrew.com/insights/social-media-statistics/. Accessed 15 July 2022
3. Zhang Z, Gupta BB (2018) Social media security and trustworthiness: overview and new direction. Futur Gener Comput Syst 86:914–925. https://doi.org/10.1016/j.future.2016.10.007
4. Baccarella CV, Wagner TF, Kietzmann JH, McCarthy IP (2018) Social media? It's serious! Understanding the dark side of social media. Eur Manag J 36(4):431–438. https://doi.org/10.1016/j.emj.2018.07.002
5. Zaccagnino R, Capo C, Guarino A, Lettieri N, Malandrino D (2021) Techno-regulation and intelligent safeguards: Analysis of touch gestures for online child protection. Multimed Tools Appl 80:15803–15824. https://doi.org/10.1007/s11042-020-10446-y
6. van Steen T (2022) When choice is (not) an option: nudging and techno-regulation approaches to behavioural cybersecurity. International conference on human-computer interaction. Springer International Publishing, Cham, pp 120–130
7. Clemente M, Padilla-Racero D, Espinosa P (2019) Revenge among parents who have broken up their relationship through family law courts: Its dimensions and measurement proposal. Int J Environ Res Public Health 16(24):4950. https://doi.org/10.3390/ijerph16244950
8. Paulin M, Boon SD (2021) Revenge via social media and relationship contexts: Prevalence and measurement. J Soc Pers Relat 38(12):3692–3712. https://doi.org/10.1177/02654075211045316

9. Zhao J, Shao M, Peng H, Wang H, Li B, Liu X (2021) Porn2Vec: A robust framework for detecting pornographic websites based on contrastive learning. Knowl-Based Syst 228:107296. https://doi.org/10.1016/j.knosys.2021.107296
10. Singh M, Bansal D, Sofat S (2016) Behavioral analysis and classification of spammers distributing pornographic content in social media. Soc Netw Anal Min 6(1):1–18. https://doi.org/10.1007/s13278-016-0350-0
11. Siegel E, Classifying passive vs. active revenge in related subreddits using NLP. https://github.com/ebsiegs/subreddit_nlp. Accessed 31 Mar 2022
12. Neuman Y, Erez ES, Tschantret J, Weiss H (2022) Themes of revenge: automatic identification of vengeful content in textual data. arXiv preprint arXiv:2205.01731
13. Statista [electronic resource] (2020) Ranking of the number of Reddit users by country 2020, https://www.statista.com/forecasts/1174696/reddit-user-by-country. Accessed 25th May 2022
14. Wikipedia [electronic resource] (2020) Controversial Reddit communities. https://en.wikipedia.org/wiki/Controversial_Reddit_communities. Accessed Nov 2022
15. Statista [electronic resource] (2020) Number of content removal requests made to Reddit by governments in 2020, by country, https://www.statista.com/statistics/1255296/government-content-removal-requests-to-reddit-by-country/, Accessed 10 Aug 2022
16. König A, Gollwitzer M, Steffgen G (2010) Cyberbullying as an act of revenge? J Psychol Couns Sch 20(2):210–224. https://doi.org/10.1375/ajgc.20.2.210
17. Alla K R, Kandibanda N, Katta P, Muthavarapu A, Kuchibhotla S (2022). Emotion Detection from Text Using LSTM. In Proceedings of Sixth International Congress on Information and Communication Technology, 545–553. Springer, Singapore. https://doi.org/10.1007/978-981-16-1781-2_49
18. Graumas L, David R, Caselli T (2019) Twitter-based polarised embeddings for abusive language detection. In: 2019 8th international conference on affective computing and intelligent interaction workshops and demos (ACIIW). IEEE, pp 1–7
19. Sharif O, Hoque M M (2021) Tackling Cyber-Aggression: Identification and Fine-Grained Categorization of Aggressive Texts on Social Media using Weighted Ensemble of Transformers. Neurocomputinghttps://doi.org/10.1016/j.neucom.2021.12.022
20. Ghosal S, Jain A (2021) Research journey of hate content detection from cyberspace. In: Natural language processing for global and local business. IGI Global, pp 200–225
21. Ginting PSB, Irawan B, Setianingsih C (2019) Hate speech detection on twitter using multinomial logistic regression classification method. In: 2019 IEEE international conference on internet of things and intelligence system (IoTaIS). IEEE, pp 105–111
22. Novalita N, Herdiani A, Lukmana I, Puspandari D (2019) Cyberbullying identification on twitter using random forest classifier. In Journal of physics: conference series, vol 1192, no 1. IOP Publishing, p 012029
23. Sadiq S, Mehmood A, Ullah S, Ahmad M, Choi GS, On BW (2021) Aggression detection through deep neural model on twitter. Futur Gener Comput Syst 114:120–129. https://doi.org/10.1016/j.future.2020.07.050
24. Qureshi KA, Sabih M (2021) Un-compromised credibility: Social media based multi-class hate speech classification for text. IEEE Access 9:109465–109477. https://doi.org/10.1109/ACCESS.2021.3101977
25. Chiril P, Pamungkas EW, Benamara F, Moriceau V, Patti V (2022) Emotionally informed hate speech detection: a multi-target perspective. Cogn Comput 14(1):322–352. https://doi.org/10.1007/s12559-021-09862-5
26. Dheeraj K, Ramakrishnudu T (2021) Negative emotions detection on online mental-health related patients texts using the deep learning with MHA-BCNN model. Expert Syst Appl 182:115265. https://doi.org/10.1016/j.eswa.2021.115265
27. Maity K, Kumar A, Saha S (2022) A multitask multimodal framework for sentiment and emotion-aided cyberbullying detection. IEEE Internet Comput 26(4):68–78
28. Akhter MP, Jiangbin Z, Naqvi IR, AbdelMajeed M, Zia T (2022) Abusive language detection from social media comments using conventional machine learning and deep learning approaches. Multimed Syst 28(6):1925–1940
29. Srinivasarao U, Sharaff A (2023) Machine intelligence based hybrid classifier for spam detection and sentiment analysis of SMS messages. Multimed Tools Appl 82(20):31069–31099
30. Tripathy G, Sharaff A (2023) AEGA: enhanced feature selection based on ANOVA and extended genetic algorithm for online customer review analysis. J Supercomput, 1–30. https://doi.org/10.1007/s11227-023-05179-2

31. Ai Q, Yang L, Guo J, Croft WB (2016) Analysis of the paragraph vector model for information retrieval. In: Proceedings of the 2016 ACM international conference on the theory of information retrieval, pp 133–142
32. Salehi Rizi F, Granitzer M (2017) Properties of vector embeddings in social networks. Algorithms 10(4):109. https://doi.org/10.3390/a10040109
33. Hidayat THJ, Ruldeviyani Y, Aditama AR, Madya GR, Nugraha AW, Adisaputra MW (2022) Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier. Procedia Comput Sci 197:660–667. https://doi.org/10.1016/j.procs.2021.12.187
34. Yang L, Li C, Ding Q, Li L (2013) Combining lexical and semantic features for short text classification. Procedia Comput Sci 22:78–86. https://doi.org/10.1016/j.procs.2013.09.083
35. Mishra M, Mishra VK, Sharma HR (2013) Question classification using semantic, syntactic and lexical features. Int J Web Semant Technol 4(3):39
36. Del Gobbo E, Guarino A, Cafarelli B, Grilli L (2023) GradeAid: a framework for automatic short answers grading in educational contexts—design, implementation and evaluation. Knowl Inf Syst 65(10):4295–4334
37. Kamarudin MH, Maple C, Watson T, Safa NS (2017) A logitboost-based algorithm for detecting known and unknown web attacks. IEEE Access 5:26190–26200. https://doi.org/10.1109/ACCESS.2017.2766844
38. Li J, Zhang H, Wei Z (2020) The weighted word2vec paragraph vectors for anomaly detection over HTTP traffic. IEEE Access 8:141787–141798. https://doi.org/10.1109/ACCESS.2020.3013849
39. Bentéjac C, Csörgő A, Martínez-Muñoz G (2021) A comparative analysis of gradient boosting algorithms. Artif Intell Rev 54(3):1937–1967. https://doi.org/10.1007/s10462-020-09896-5
40. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 785–794
41. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W et al (2017) Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Proces Syst 30
42. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A (2018) CatBoost: unbiased boosting with categorical features. Adv Neural Inf Proces Syst 31
43. Gilabert P, Seguí S (2020) Gradient boosting and language model ensemble for tweet recommendation. In: Proceedings of the recommender systems challenge, pp 24–28
44. Pereira FS, Andrade T, de Carvalho AC (2020) Gradient boosting machine and LSTM network for online harassment detection and categorization in social media. In: Machine learning and knowledge discovery in databases: international workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, proceedings, part II. Springer International Publishing, pp 314–320
45. Alzamzami F, Hoda M, El Saddik A (2020) Light gradient boosting machine for general sentiment classification on short texts: a comparative evaluation. IEEE Access 8:101840–101858. https://doi.org/10.1109/ACCESS.2020.2997330
46. Li TR, Chamrajnagar AS, Fong XR, Rizik NR, Fu F (2019) Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model. Frontiers in Physics 7:98. https://doi.org/10.3389/fphy.2019.00098
47. Saha P, Mathew B, Goyal P, Mukherjee A (2018) Hateminers: detecting hate speech against women. arXiv preprint arXiv:181206700
48. Loper E, Bird S (2002) Nltk: the natural language toolkit. arXiv preprint cs/0205028
49. Gupta A, Taneja SB, Malik G, Vij S, Tayal DK, Jain A (2019) SLANGZY: A fuzzy logic-based algorithm for English slang meaning Selection. Progress Artif Intell 8(1):111–121
50. Cutting D, Kupiec J, Pedersen J, Sibun P (1992) A practical part-of-speech tagger. In: Third conference on applied natural language processing, pp 133–140
51. Salton G, Yang CS (1973) On the specification of term values in automatic indexing. J Doc 29(4):351–372
52. Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781
53. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013b) Distributed representations of words and phrases and their compositionality. Adv Neural Inf Proces Syst 26
54. Dorogush AV, Ershov V, Gulin A (2018) CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:181011363

55. Everitt BS (1992) The analysis of contingency tables. CRC Press
56. Akiba T, Sano S, Yanase T, Ohta T, Koyama M (2019) Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August; 2623–2631, https://doi.org/10.1145/3292500.3330701
57. kumar, Sahoo , YG (2012) Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA. Int J Inform Technol Comput Sci 4(7):43–49. https://doi.org/10.5815/ijitcs.2012.07.06
58. Nti IK, Nyarko-Boateng O, Aning J (2021) Performance of machine learning algorithms with different K values in K-fold cross-validation. Int J Inf Technol Comput Sci 13(6):61–71

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.