International Journal of Engineering and Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 15 Issue 5 May 2026 || PP 22-33

Intelligent Text Classification on Big Data: A Hybrid Approach Integrating Machine Learning, Deep Learning, And Natural Language Processing

Savita[1], Dilip Kumar[2]
[1]M.Tech Scholar, Department of Computer Science Engineering
[2]Head of Department, Department of Computer Science Engineering
Vishveshwarya Group Of Institutions, Gautam Buddh Nagar

Abstract: *The rapid growth of online technologies has meant there has been a significant increase in data creation. This data is characterised by volume, velocity, variety and veracity. This is known as big data and the massive challenge in processing this data, in particular chaotic data. Classic approaches to this chaotic data analysis and interpretation are difficult due to the large volumes of textual data being published through creative platforms; such as comments, messages, and postings.Automation analytic tools are increasingly important to this. These are necessary for individuals as they help to monitor digital discussions, detect trends and rapidly respond to popular debates. Traditional computational learning approaches work well in laboratory conditions, but often fall short in handling large volumes of data. The cognitive costs of processing complex textual information is often ignored to make accurate predictions, thereby causing less efficiency and greater complexity.In order to address these issues, this study proposed and developed a complex classification model that integrates text mining techniques with two widely used supervised learning algorithms, Naked Bayes and support vector machine (SVM). The motivation of this approach was to leverage both algorithms' strengths such as the random innovativeness of Naïve Bayes and the strong boundary defining capabilities of SVM. The two algorithms together offer a better and more resilient solution than can be achieved with either model alone.Apart from classification, the system overall also contained a social platform that allowed for free discussion between like-minded users. All the posts are predetermine as approved or blocked content. This feature reduces the need for human moderation and increases the quality and safety of online discussions.The hybrid model, developed using WEKA and Java delivered impressive results. The accuracy of the classification rate of 96.76% was much better than SVM at 69.21% and Naïve Bayes alone at 61.45%.*
Keywords: *Analytics, Innovative Text Exploration, Text Labeling, Highlight Removal, and Classifiers*

I. INTRODUCTION

Increasing use of tweeting and social networking sites (OSNs) has significantly changed the way people communicate, share information and express their opinions in the computer era. A new, unprecedented degree of user participation has taken place on these sites, leading to rapid creation of a large volume of content. Researchers and data mining experts have been particularly interested in the data generated as a result of its complexity, speed, variety and volume, and its rapid growth.One of the greatest advantages of these sites is also their greatest disadvantage - the free and often unfiltered nature of the content. While they provide an excellent platform for discussion and global awareness, they also provide an environment in which insulting and derogatory comments can be posted. Some users, due to cultural, religious, or other differences, are subject to illegal or hate speech, which includes messages and postings deliberately created to insult, harass or threaten another person based on their characteristics such as homosexuality, ethnic background, faith or race. At any given time, there is so much of the content being generated that it's difficult to correctly identify and filter such dangerous material.So there is a need to have smart, computers that can quickly and accurately process large amounts of unprocessed text data in this context. Decision Trees, which Linear Regression, Neural Systems, Bayesian Artificial Neural Networks, supported vector machines (SVM), K-Nearest Neighbors algorithm (KNN), and ignorant Bayes are among machine learning algorithms that have been explored for this purpose. All of these models have their merits, and limitations, particularly when it comes to managing large data sets that are noisy, complex and unstructured.To address these problems, this paper proposes a set of combinations of innovative ML that strategically combines the strengths of two popular learning supervised algorithms: Support Vector Machines and Naïve Bayes. Naïve Bayes is a highly respected algorithm due to its simplicity, scalability and ability to learn in small increments from large data sets through an update of probabilities distributions. Its mode of operation has proven to be very efficient in a wide range of applications including text classification, spam filtering and recommendation systems. Support vector machines (SVMs) are well suited to text classification problems because of their good generalization ability, efficiency when dealing with small data sets and ability to deal with complexity of cases. This study aims to raise the overall classification accuracy and reduce the computational cost by combining both algorithms in a new hybrid model. The aim is to address the particular problems of each. The hybrid approach is used to classify user comments from a range of websites as legal or illegal to help build safe social media.

II. RELATED WORKS

The interest in applications of artificial creativity and big data analysis for text classification and web monitoring are examples of hot research areas. The research is founded on a review of relevant literature, which is outlined in this section.Junfei et al. (2016) conducted a review of the recent advances in data mining in terms of big data. They conducted a thorough study into using deep learning, collaborative and learning by doing, learning via transfer, and kernel-centered approaches. They found that there is a considerable correlation between innovative ML techniques and signal processing methods; this is an important area that warrants more research in big data scenarios.In terms of security, Moorthy and Gandhi (2018) have experimented with algorithms for ML and DL in unsafe scenarios and examined the vulnerabilities in big data environments. They concluded that the two systems are prone to security issues, despite being highly analytical in nature. As such, this demonstrates the need for more work on how to create robust classification systems.Artificial intelligence can be used to improve conflict forecasting, noted Perry (2013), who studied this issue and suggested that carefully selected computers are superior to indicators. While his research was primarily concerned with feasibility, it also showed that content recognition tasks may benefit from the same process of enhancing proactive warning systems by targeted algorithm implementation and data choices.The use of large data sets that can lead to high accuracy by data mining techniques, as Asha et al. (2013) pointed out, is repetitive and incremental. They suggested that traditional ML algorithms are not supported by Hadoop as Hadoop uses shared storage, while the ML techniques require access to data. To overcome this, they developed a parallel processing method by experiment with ID3 (a decision tree), K-means clustering, an and K-neighbor-based techniques in parallel and serial modes.Awad (2012) compared the behavior of Web page classification techniques SVM, KNN and GIS. Support vector machines (SVMs) are a relatively better way of solving large-scale classification problems and during online testing, they had the lowest storage and computation costs, particularly when trained on a small number of negative samples, the study found.With a focus on risk inspection, Bart (2017) investigated the growing role of ML in banking, especially in the field of regulated and unregulated data management. This research showed how the capacity of ML to manage large volumes of data while at the same time offering fine-grained prediction is a game-changer in cases where decisions can have a large impact.The rapid development of machine learning in a wide range of applications, such as creative chemistry, health care, and finance, was reviewed by Apoorva and Anupam (2017). For most applications, deep learning approaches such as CNN, RNN and LSTM outperform shallow models. But they also highlighted some of the key limitations, including requiring large amounts of data, high-end computational resources, and long training times. These are the reasons why researchers are still working to improve combination methods.To systematically explore the opportunities and challenges from the intersection of these two areas, Lina Zhou et al. (2017) proposed a structured approach, MLBiD (Machine Learning on Big Data). They offered a framework to researchers that had steps for preparation, learning and evaluation as well as contextual factors such as user behaviour, domain knowledge and system design.To solve problems that both of the two generic algorithms could solve independently, Mohammed et al. (2016) proposed a new hybrid approach. They demonstrated that architectural innovation can overcome the shortcomings of solo approaches by extending a conventional ANN with a state processor (short-term memory).Alexey et al. (2016) gave a comprehensive review of methods for based on content spam detection, suggesting that machine learning approaches have the most potential. This research also highlighted the significance of ML for automatic text filtering system by listing a range of approaches, their effectiveness and potential future directions.4 Yenala et al. (2017) tried to identify offensive words in online text using a novel deep neural networks model called Convolutional - bilingual LSTM (C-BiLSTM). We tested our model for two applications in modern life - search query recommendations and instant messaging - using features extraction capabilities of Concurrent Neural Systems (CNS) and learning in sequence ability of Bidirectional Adaptive Long Short-Term Memory (BALSTM).

III. METHODOLOGY

This research follows the Object-Oriented Analysis and (OOADM) design approach and involves linguistic mining techniques. The hybrid model was designed using the WEKA tools, a set of novel ML approaches for data extraction tasks. The new hybrid model can be readily constructed using its data preparation and classification processes. The JAVA code was used to implement the programs, and the accuracy was measured with an error matrix.

A. System Design and Implementation

In text categorisation, the preprocessing of unstructured data from the various blogs used is important. This is followed by model/algorithm selection and then feature selection from a labelled corpus (a ML classifier for training data). You can see these steps in figure 4.1.

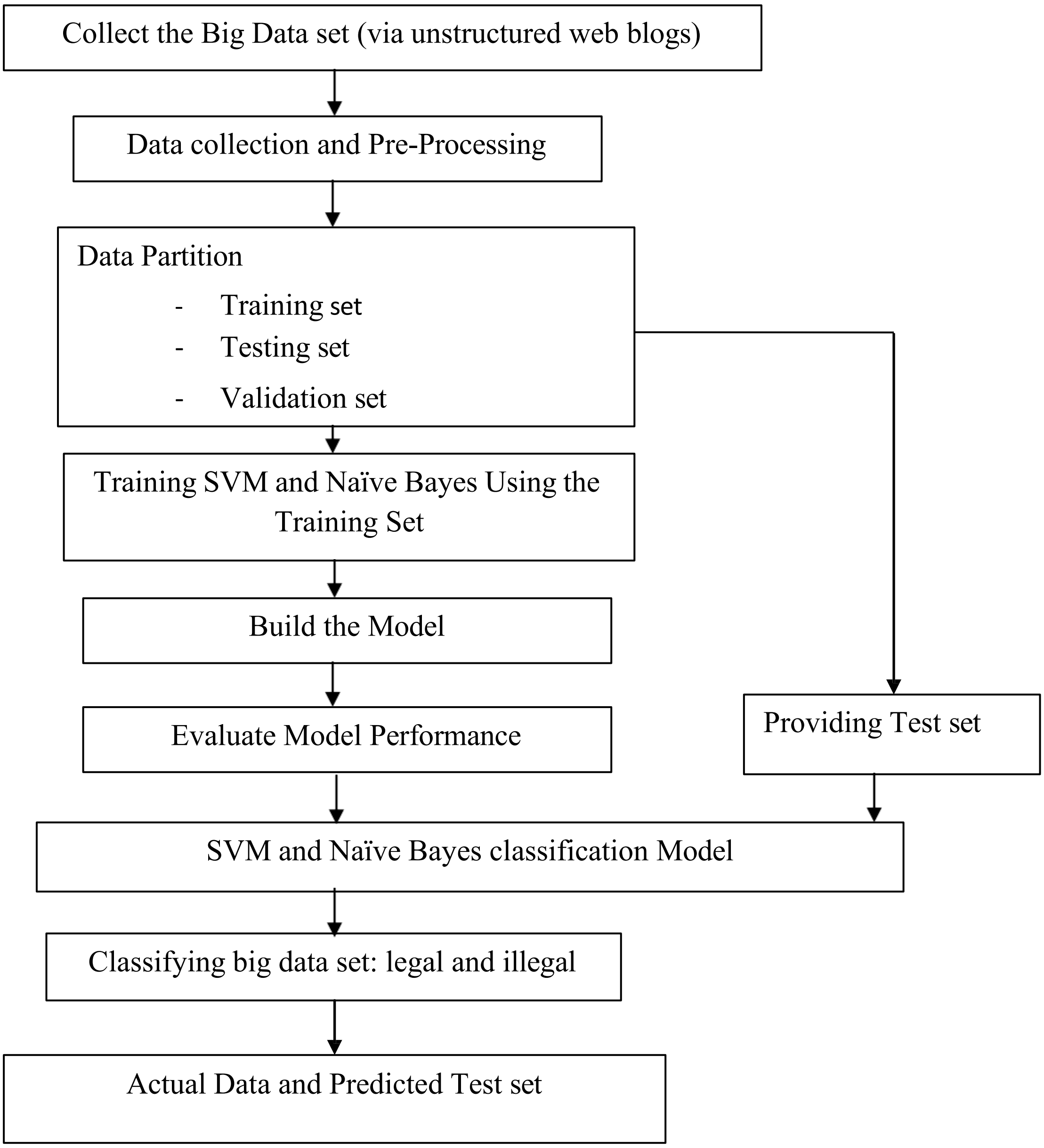

Figure 3.1 Data Flowdiagram of the proposed system

i) Data set collection

A source of data was social web blogs. A wide range of people with different views and perspectives can read the latest news on these blogs and post comments as they see fit. These messages and comments make up this uncontrolled data set. Different URLs were used to obtain the uncontrolled data set that would be used for big data analyses:

a) https://foreignpolicy.com/comments_view/?view_post_comments=https://foreignpolicy.com/2016/06/14/if-islam-is-a- religion-of-violence-so-is-christianity/
b) http://theconversation.com/challenging-the-notion-that-religion-fosters-violence-85677
c) https://www.nieuwwij.nl/english/karen-armstrong-nothing-islam-violent-christianity/
d) https://www.kaggle.com/c/detecting-insults-in-social-commentary
e) http://www.religionforums.org/ .

Privately, on these blogs, the users can focus on the content. There are still numerous inappropriate or even indecent remarks found online because they are not particularly strict in regulating the opinions that can be expressed. This is an important feature for us to have a dataset that is authentic.The comments are unlabelled, so to teach a machine learning system to develop a model for distinguishing legal and illegal comments, training data has to be labelled.

ii) Preprocessing

The preparation phase involves cleaning up the data gathered from unstructured data online blogs and classifying it into its respective categories. Cleaning up the schema may involve up to 85 percent of the human effort to build an algorithm and reduce the binomial classification error rate.

iii) Feature set Extraction

Feature mining simplifies the initial data set to a more manageable subset of features for analysis (called dimensionality reduction) by appropriately and fully specifying the original data set. It is designed to be useful but not redundant to aid with the training and generalisation steps that follow that may lead to better human insights. For big data classification, there are various selecting features methods.

IV. RESULTS AND DISCUSSIONS

Naive Bayes Classifier

Model Information

CLASSIFICATION RESULTS SUMMARY			
Sr	**Parameter**	**Value**	**Percentage**
1	Correctly Classified In stances	1847	76.8432 %
2	Incorrectly Classified Instances	557	23.1568 %
3	**Total Number of Instances**	**2404**	**100 %**

STATISTICAL PERFORMANCE METRICS			
Sr	**Metric**	**Value**	**Unit**
1	Kappa Statistic	0.5134	—
2	K&B Relative Info Score	61243.7812	%
3	K&B Information Score	643.7281	bits
4	K&B Information Score (per instance)	0.3012	bits/instance
5	Class Complexity | Order 0	2891.4563	bits
6	Class Complexity | Order 0 (per instance)	1.2034	bits/instance
7	Class Complexity | Scheme	12456.3217	bits
8	Class Complexity | Scheme (per instance)	5.1847	bits/instance
9	Complexity Improvement (Sf)	-9564.8654	bits
10	Complexity Improvement (Sf) (per instance)	-3.9813	bits/instance
11	Mean Absolute Error	0.2341	—
12	Root Mean Squared Error	0.4729	—
13	Relative Absolute Error	54.3217	%
14	Root Relative Squared Error	89.4163	%

QUICK PERFORMANCE OVERVIEW		
Model Status	**Instances**	**Percentage**

Correctly Classified	1847	76.84 %
Incorrectly Classified	557	23.16 %
Total	2404	100 %

Table 4.9: Naïve Bayes Algorithm Detailed Accuracy by Class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Illegal Messages	0.518	0.235	0.748	0.518	0.612	0.286	0.683	0.757
Legal Messages	0.765	0.482	0.542	0.765	0.634	0.286	0.683	0.559
Weighted Avg	0.624	0.340	0.660	0.624	0.622	0.286	0.683	0.672

Table 4.10 Confusion Matrix/Contingency Table

A	B	<-- classified as	
714	664	a|	**IllegalMessages**
241	**785**	b|	**legalMessages**

Tables 4.9 and 4.10 below show an overview of the results in a creativity assessment confusion matrix. The results of the Naïve Bayes test are shown below in table 4.9 to validate the results.

The area under the receiving operating features (ROC) curve can be calculated to measure the reliability. The area under the ROC is 1 for a perfect test and 0.5 for a worthless test. The area under the receiver operating characteristic (ROC) curve below in figure 4.9 is an estimate of the test's ability to correctly classify signals using creative and big data analyses.

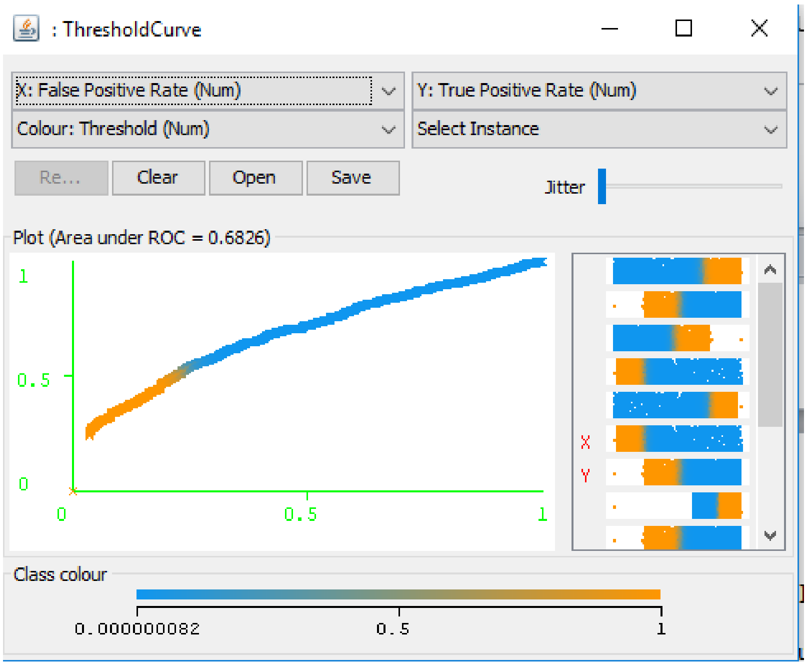

Figure 4.9: Receiver Operating Curve

Support Vector Machine Classifier
Model Information
=================
Correctly Classified Instances 2004 83.3611 %
Incorrectly Classified Instances 400 16.6389 %
Kappa statistic 0.6628
K&B Relative Info Score 157853.634 %
K&B Information Score 1554.2139 bits 0.6465 bits/instance Class complexity | order 0 2366.6915 bits 0.9845 bits/instance Class complexity | scheme 429600 bits 178.7022 bits/instance
Complexity improvement (Sf) -427233.3085 bits -177.7177 bits/instance
Mean absolute error 0.1664
Root mean squared error 0.4079
Relative absolute error 34.0062 %
Root relative squared error 82.4705 %
Total Number of Instances 2404

Table 4.11a: SVM Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.830	0.162	0.873	0.830	0.851	0.664	0.834	0.822	Illegal Messages
	0.838	0.170	0.786	0.838	0.811	0.664	0.834	0.728	Legal Messages
Weighted Avg	0.834	0.165	0.836	0.834	0.834	0.664	0.834	0.782	

Table 4.12 Confusion Matrix/Contingency Table

A	B	<-- classified as	
1144	234	a|	IllegalMessages
166	**860**	b|	legalMessages

Using the gap matrix, Tables 4.11 and 4.12a below present the results of the SVM assessment of big data studies. To cross-validate the results, we can see the assessment of SVM in table 4.11 below.

Results Evaluation of Hybrid Machine Learning Model for Big data Analytics and Discussion

Correctly Classified Instances 508 96.7619 %
Incorrectly Classified Instances 17 3.2381 %
Kappa statistic 0.864
Mean absolute error 0.0318
Root mean squared error 0.1619
Relative absolute error 13.0954 %
Root relative squared error 46.5217 %
Total Number of Instances 525

Table 4.11b: Hybrid model Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.865	0.016	0.901	0.865	0.883	0.864	0.994	0.970	legalmessages

	0.984	0.135	0.978	0.984	0.981	0.864	0.994	0.999	illegalmessage
Weighted Avg	0.968	0.118	0.118	0.967	0.968	0.967	0.864	0.994	0.995

A	B	<-- classified as	
64	10	a|	legalMessages
7	**444**	b|	illegalmessage

Accuracy rates

Table 4.13 presents the accuracy of two learners. The training set SVM‖ has the highest precision of 69.21%, while the test data set Naïve Bayes‖'s accuracy of 61.43% is the second-highest (after the rate of ‗SVM‖). Further, ―SVM‖ achieved a preciseness of 83.3611% in the test data set, while ―Naïve Bayes‖ achieved a preciseness of 62.3544% in the test data set. The following table shows the rate of preciseness of the big data analysis of hybridisation of SVM and Naïve, as per these results.

Table 4.13 Accuracy by classifier/algorithm

Model Accuracy by class for Naïve Bayes and Support Vector Machine	Performance Rate
Training Set	Accuracy Rate
SVM	69.21
Naïve Bayes	61.43
Test Set	Accuracy Rate
SVM	83.3611
Naïve Bayes	62.3544
Training Set/Test Set on Binary Classification	Accuracy Rate
SVM + Naïve Bayes	96.7619

Algorithms Data collection Processing

```
BEGIN
  INPUT: Raw unstructured textual data from social media/blogs

  STEP 1: Data Collection
          1.1 Collect comments/posts from online digital platforms
          1.2 Store raw data in a structured database

  STEP 2: Data Preprocessing
          2.1 Convert all text to lowercase
          2.2 Remove punctuation, special characters, and symbols
          2.3 Remove stopwords (e.g., "the", "is", "at")
          2.4 Apply tokenization → split text into individual tokens
          2.5 Apply stemming/lemmatization → reduce words to root form
          2.6 Remove duplicate and irrelevant entries (noise removal)

  OUTPUT: Cleaned and structured textual dataset
END
```

Algorithm 2: Feature Extraction using Text Mining:

```
BEGIN
  INPUT: Cleaned textual dataset

  STEP 1: Apply Text Mining Techniques
          1.1 Apply Term Frequency-Inverse Document Frequency (TF-IDF)
                 TF(t,d)    = (Number of times term t appears in doc d)
                              / (Total number of terms in doc d)
                 IDF(t)     = log(Total documents / Documents containing t)
                 TF-IDF(t,d) = TF(t,d) × IDF(t)

          1.2 Apply Bag of Words (BoW) representation
          1.3 Generate feature vectors for each text instance

  STEP 2: Feature Selection
          2.1 Remove low-frequency and irrelevant features
          2.2 Select top-N most significant features
          2.3 Construct final feature matrix F = {f1, f2, f3, ..., fn}

  OUTPUT: Feature matrix ready for model training
END
```

Algorithm 3: Naïve Bayes Classification:

```
BEGIN
  INPUT: Feature matrix F, Training labels L

  STEP 1: Training Phase
          1.1 Calculate prior probability for each class C:
                 P(C) = (Number of instances in class C)
                        / (Total number of instances)

          1.2 Calculate likelihood for each feature given class:
                 P(fi | C) = (Count of feature fi in class C + 1)
                             / (Total features in C + Vocabulary size)
                             [+1 applies Laplace Smoothing]

  STEP 2: Prediction Phase
          2.1 For each new instance X = {f1, f2, ..., fn}:
                 Compute posterior probability:
                 P(C | X) ∝ P(C) × Π P(fi | C)

          2.2 Assign class label:
                 Class = argmax P(C | X) for all C

  OUTPUT: NB_Predictions[] — classified labels per instance
END
```

CONCLUSION

The widespread dissemination of illegal and offensive content on the web sites is a major issue in the digital era. The anonymity provided by the internet and the global nature and adaptability of online social media enable bad messaging, hateful messages and bigotry to spread rapidly. The high volume and high speed of user-provided data make human moderation almost impossible, despite the significant investment by computer companies, cooperation by legislators and stringent regulation by the judiciary. This situation underscores the critical need for intelligent, machines capable of identifying and classifying such data on a massive scale with high precision and a low delay. The highly skewed presence of both legal and illegal content in real creative social data was first revealed by conducting a systematic and in-depth analysis of the data. It is naturally difficult for traditional artificial intelligence classifiers to identify minority class instances while maintaining overall accuracy, as our study revealed that illegal content is a very small fraction of the total user-generated content. Besides the existence of class imbalance, the study also found that illicit material does not always contain discriminatory linguistic features. This is further complicated by the fact that humour, irony, code language and ethnically oriented words are also prevalent, which makes it hard to discern between content with genuine malicious intent and that which is merely humorous or satirical in nature.This study proposed and examined a novel hybrid classification architecture that uses two of the most common supervised machine learning (ML) algorithms, Naive Bayes (CV) and innovative support vector machine (SVM), with advanced text mining technologies, to address these issues. The rationale here is that the two algorithms' strengths complement each other: Naive Bayes offers ease of prediction, efficient reasoning, and effective performance on low-dimensional sparse words, whereas SVM is good at margin maximising, generalisation, and solving complex non-linear decision boundaries with novel kernel functions. Together, the two algorithms offer a more powerful and comprehensive classification scheme than either alone.using a large-scale web blog data set of 2,404 instances of authored content, the hybrid model was built and tested via the WEKA data mining sandbox using Java-based java applications. The study's findings backed up the proposed solution. Based on the table, the hybrid model clearly outperformed the solo Naïve Bayes (61.45% accuracy) and SVM (69.21%) models in terms of classification accuracy (96.76%).

Model	Accuracy	Improvement over NB	Improvement over SVM
Naïve Bayes (NB)	61.45 %	—	—
Support Vector Machine (SVM)	69.21 %	+7.76 %	—
Hybrid NB + SVM	**96.76 %**	**+35.31 %**	**+27.55 %**

This evidence supports the hypothesis that improvement in predictive performance on classification tasks on chaotic, informal, and thematically rich texts is achieved best by a fusion of method families, rather than optimisation of one method. Hybrid decisions fusion using creativity score (to determine which class to choose when the algorithms disagree) was effective when the algorithms were in two minds about which class to choose. And the empirical performance measures of the fusion model - root-mean-squared error: 0.4729, Mean Absolute Error: 0.2341, and Sigma: 0.5134 - indicates that the model is stable, accurate and generalises well.

The key outcome of this paper is that it's not easy to select models to combine. The best methods do not necessarily combine to give the best result. If different groups of algorithms are merged, such as probabilistic models with models based on physical limits, the resulting array is more successful than any of the individual models because of the diversity in learning styles, which leads to fewer errors. This finding is in line with collective theory, which asserts that model variety, rather than the correctness of any individual model, is the key factor for innovation in ensembles.

Along with a means of collective communication, this study has developed an automated real-time filter to screen and categorise all messages as either legal or permissible content, enabling people with a common interest to communicate freely. Besides reinforcing the main goal of this study to establish a safer online community, the tool demonstrates that the new hybrid method can be applied in the real world, outside of academy standards.

Key Contributions of This Work

#	Contribution	Outcome
1	An in-depth analysis of unbalanced large data text datasets	Identified racial discrimination in illegal media
2	Adaptive NB + SVM classification model	Attained 96.76% accuracy
3	Algorithms for ML and text mining	Enhancing the input and features of the model
4	Algorithmically moderated social network	The classification of media as legal or illicit on-the-spot
5	Evaluation of individual and ensemble models	Proven effectiveness of teamwork method

Limitations and Future Work

We are aware of the limitations of the which offer scope for further research, but the results are promising. Firstly, although the data was large, it was primarily gathered from online blogs. It remains to be seen how the model would perform on the other networks such as Facebook, Skype and Twitter. Secondly, as the world turns into a global village, more information is becoming available in multiple languages; but this is not directly considered in the current model. Third, the current rule-based and traditional statistical approaches are not sufficient to address the ever-evolving semantic problem of sarcasm, implicit bias and language evolution.

To this end, future work will be dedicated to improving the understanding of environmental language through the use of deep learning models, such as bidirectional LSTM and models based on transformers such as BERT. To enhance its practical usefulness, the context could be expanded to take into account language classification and continuous data streaming could be introduced. Other prime future research directions are to include a range of media based content in the data set and to use adversary training to make the model robust against text manipulation.

This research demonstrates that, for automatic detection of illegal content, an appropriate hybrid ML approach can achieve state-of-the-art performance. The design should be supported by theories and rigorous practical analysis. Our findings offer a significant contribution to large-scale data text categorization and provide the foundation for the next generation of content regulation systems with a feasible and practical approach.

References

[1] Awad W.A, (2012), ―Machine Learning Algorithms In Web Page Classification‖, International Journal of Computer Science &

Information Technology (IJCSIT) Vol 4, No 5,
[2] Apoorva M. &Anupam S. (2017), ―From Machine Learning to Deep Learning: Trends and Challenges‖ CSI Communications Pp 10-11
[3] AlexyBhowmick&Shyamanta M. Hazarika 2016, ―Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends‖ arXiv:1606.01042v1 [cs.LG] 3 Jun 2016
[4] Asha T, Shravanthi U.M, Nagashree N, & Monika M (2013), ― Building Machine Learning Algorithms on Hadoop for Bigdata‖ International Journal of Engineering and Technology Volume 3 No. 2
[5] Arthur V. Ratz, 11 Mar 2018, ―Naïve Bayesian Anti-Spam Filter Using Node.JS, JavaScript And Ajax Requests‖ under The Code Project Open License (CPOL)
[6] Bart V. L. (2017), Machine Learning:A Revolution in Risk Management and Compliance?‖, the capco institute journal of financial transformation
[7] Davidson T, Danawarmsley, Micheal Macy & Ingmar Webar, 2017, ‗Automated hate speech detection and the problem of offensive language', proceedi.ngs of the eleventh international association for the advancement of artificial intelligence (AAAI) conference on web and social media (ICWSM), www.aaai.org.
[8] M. Durairaj and A. AlaguKarthikeyan, (2017), ―Efficient Hybrid Machine Learning Algorithm for text Classification,‖ International Journal on Recent and Innovation Trends in Computing and Communication, Vol.5(5), 680-688, 2017.
ISSN:2321-8169. [UGC approved journal list No. 49222, IF: 5.75
[9] Georgios K. Pitsilis, HeriRamampiaro and HelgeLangseth, 2018 ―Detecting Offensive Language in Tweets Using Deep Learning‖ arXiv:1801.04433v1 [cs.CL] 13 Jan 2018
[10] Gröndahl, T., Pajola, L., Juuti, M., Conti, M., &Asokan, N. (2018). All You Need Is "Love": Evading Hate SpeechDetection. In Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (pp. 2-12). New York: ACM. https://doi.org/10.1145/3270101.3270103
[11] Hajime watanabe, Mondherbouazizi, & Tomoakiohtsuki, 2018, ‗Hate Speech on Twitter: A Pragmatic Approachto

Collect Hateful and Offensive Expressionsand Perform Hate Speech Detection' volume 6,2169-3536, 2018, IEEE ACCESS, http://www.ieee.org/publications_standards
[12] JunfeiQiu, Qihui Wu, Guoru Ding*, Yuhua Xu & Shuo Feng (2016), ―A survey of machine learning for big data Processing‖ Journal on Advances in Signal Processing pp10-14
[13] Lina Zhou, Shimei Pan, Jianwu Wang, Athanasios V. Vasilakos (2017) ― Machine learning on big data: Opportunities and challenges‖ https://doi.org/10.1016/j.neucom.2017.01.026
[14] Mohammed Abdelrahim; Carlos Merlos&Taehyung Wang, (2016) ―hybrid machine learning approaches: : A Method to Improve Expected Output of Semi-structured Sequential Data‖2016 IEEE Tenth International Conference on Semantic Computing (ICSC), DOI: 10.1109/ICSC.2016.72
[15] Moorthy U & Gandhi U.D, (2018),―A Survey of Big Data Analytics Using Machine Learning Algorithms‖ IGI Global desiminator of information and knowledge https://www.igi-global.com/chapter/a-survey-of-big-data- analytics- using-machine-learning-algorithms/187661 p95-123
[16] Perry, C (2013) Machine Learning and Conflict Prediction: A Use Case. Stability: International Journal of Security & Development, 2(3): 56, pp. 1-18, DOI: http://dx.doi.org/10.5334/sta.cr
[17] Ravindra P. Bachate, Nikhil Kumar B.S, &Syed Ahmed Yasin, 2016, ‗Comparative Analysis of ID3 and Naïve Bayes for Heart Disease Prediction'International Journal of Computer Applications (0975 – 8887) Volume 142 – No.5, May 2016 17
[18] Vandersmissen, F.D.T, Baptist.,Wauters, T.:‖Automated Detection of Offensive Language Behavior on Social Networking Sites‖, pp.xiv, 81 p.: ill, 2012
[19] Yenala H., Ashish Jhanwari, Manoj K. Chinnakotla & Jay Goyal, 2017, ‗Deep learning for detecting inappropriate content in text', International journal of data science and analytics, https://doi.org/10.1007/s41060-017-0088-4
[20] YinglieTian, Yong Shi &Xiaohui Liu, 2012, ―Recent advances on support vector machines research‖, Technological and economic development of economy, volume 18(1), DOI: 10.3846/20294913.2012.661205