ZKG INTERNATIONAL

ISSN: 2366-1313

MACHINE LEARNING ALGORITHMS TO CLASSIFY DAMAGING COMMENTS INTO MULTIPLE CATEGORIES

[#1]KOTAWAR SAVITHA, Assistant Professor,

Department of Computer Science and Engineering,

[#2]PARLAPALLI RENUKA, Associate Professor,

Department of Computer Science and Engineering

MOTHER THERESA COLLEGE OF ENGINEERING AND TECHNOLOGY, PEDDAPALLY, TS.

ABSTRACT: When people write harsh, angry, or unfair remarks on internet forums, they are referred to be toxic statements, and many people cease engaging in the discourse as a result. Individuals are less likely to freely express divergent perspectives as a result of cyberbullying and cyber harassment, which limits the spread of ideas. Because the majority of websites are inadequate at enabling meaningful discourse, procedures such as limiting or removing user comments are implemented. The study's goal is to figure out how much internet abuse there is and how it is classified so that machine learning algorithms can evaluate how bad it is.

Keywords: *Accuracy, Multilabel Classification, Machine Learning Algorithms, Toxic Comments.*

1. INTRODUCTION

With the advent of the twenty-first century, the internet and mobile phones enabled individuals all over the world to communicate. This extraordinary achievement is the product of computer science and technology's rapid evolution. Email was largely used for private correspondence in the early days of the Internet. As a result of this style of communication, the number of unwanted or unsolicited communications surged considerably. It was difficult to discern between unwanted and genuine communications during this time period. The rise of social networking sites like Facebook and Reddit has radically impacted how individuals communicate and exchange files online. As a result, categorizing publications as good or harmful is becoming more important as a means of safeguarding society from harm and deterring individuals from engaging in activities that undermine social harmony.

A number of people who propagate damaging and hazardous content online have lately been apprehended by police. The Vadodara police detained popular YouTube user Shubham Mishra last year after a video on his channel showed him making explicit threats against stand-up comic Agrima Joshua in front of a big audience. Donald J. Trump had difficulty using a few social media platforms in January 2021 because he was involved in the instability in Washington. As a result, a scary situation develops, and some information must be validated before being posted online. Internet users are at risk of being damaged as a result of these malicious groups. Can a claim and questions posted on the internet not be verified? Please supply documentation before I can terminate your account, sir. It goes without saying that terms like back up wanker and Bullshit have negative meanings. This statement will go through a specific technique known as preprocessing before we begin analyzing it. The amount of toxicity will then be determined using a categorization scheme.

A variety of classification methods and machine learning algorithms will be used on the supplied data to sort the harmful remarks into the relevant groups. Following that, we'll evaluate and contrast various strategies using metrics like accuracy, log-loss, and hamming loss.

2. RELATED WORK

Recent academic study has concentrated on

ISSN: 2366-1313

categorizing unpleasant comments, particularly those made on the Internet. Researchers successfully classified damaging comments from social media sites using a range of machine learning approaches.

To detect instances of harassment, the authors used supervised learning techniques. The researchers developed a computer framework that uses emotional indicators, local traits, and contextual information to quickly identify potentially dangerous content in online forums and chat rooms.

Ravi used machine learning techniques to identify the severity of negative remarks on social networking sites. Ravi outperformed the WEKA machine learning system, with an 82% success rate.

Researchers used a semi-supervised algorithm to identify objectionable content on Twitter that contained profane language. Researchers used logistic regression to boost the true positive (TP) rate. It climbed from 69.7% to 75% using phrase matching as a baseline. Each had a false-positive rate of 3.77 percent on average.

The researchers were able to extract characteristics from many conceptual layers by combining an automatic flame identification system and a stratified categorization method.

To classify and identify incorrect text and photos on social networking sites, the authors used Naive Bayes and support vector machines (SVM). They cannot, however, always detect inappropriate audio and video on social networking sites.

The goal of this research is to investigate the topic within the context of an academic paper. Researchers in the field of categorization have used both neural network-based and non-neural network-based methodologies independently. They used logistic regression and the Naive Bayes approach as part of their non-neural strategy. Despite greatly improving precision, the approach has a rather low F1 score. The model with a neural network design, specifically a stacked and bidirectional recurrent neural network (RNN), outperformed the one without. This was visible in both the F1 score and the precision assessments.

Instead of the traditional Bag of Words (BoW) method, the authors of the study used Convolutional Neural Networks (CNN) to sort the text. The typical Bag-of-Words (BoW) text classification model performed poorly when CNN and word embedding were employed to classify the text. The Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Latent Dirichlet Allocation (LDA) techniques are used in the BoW model. CNN displayed higher than 90 percent accuracy.

The goal of this query is to gain a thorough understanding of the subject at hand. To safeguard adolescents from online abuse, researchers created a system that combines lexical and parser properties to identify potentially harmful content in YouTube comments sections.

The prevalence of dangerous internet conduct has a substantial impact on people's health. As a result, it is critical to develop a reliable mechanism for spotting potentially harmful comments. We will apply approaches in this work to break the multi-label problem into several single-label problems. We will be able to use existing single-label machine learning algorithms as a result of this.

3. PROPOSED METHODOLOGY

Type of Classification

The dataset for this study is made up of user comments from an online domain. These comments will be separated into six categories. Some of these statements are exceedingly rude, toxic, disrespectful, frightening, and harmful. The information was obtained using Kaggle.

The following step is to decide which of the six labels the given data (note) corresponds to: one, a few, or none. The remark falls into several categories since it has the potential to insult or injure someone. The remark, on the other hand, could have been secure and fit into none of the six categories.

Before you begin, it is critical to understand the difference between multi-class and multi-label labeling. Each class in a problem with many

ISSN: 2366-1313

classes demonstrates reciprocal exclusivity, which means that each input is mapped to only one label. You cannot have both iOS and Android operating systems on the same mobile device.

Multi-label classification is a solution for classification issues in which numerous labels are assigned to each input. This means that more than one name can be connected with the same input at the same time.

Classifying negative comments into categories using the given data can be considered as a task involving several labels.

Exploratory Data Analysis

Exploratory data analysis, or evaluating data in its most basic form, is a critical component of data analysis in general. The basic goal of exploratory data analysis (EDA) is to better understand the presented data and uncover any distinguishing features. Using data visualization tools makes it easier to attain this goal.

.

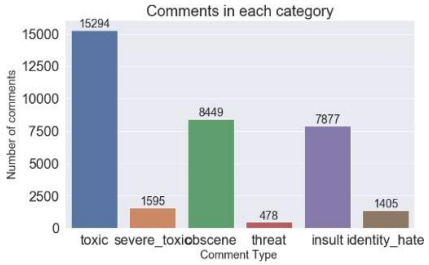


Fig 1. Plot 1

The first graph depicts the distribution of comments across labels. The bulk of comments are negative, according to the data, and the group with the fewest comments is related with threats.

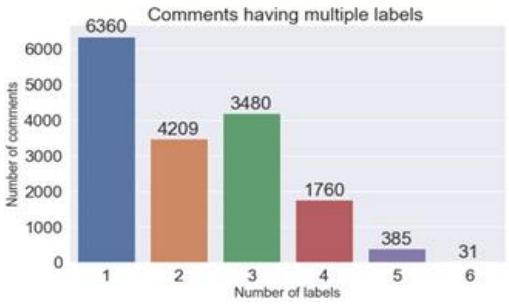


Fig. 2. Plot 2

Figure 2 shows how frequently comments with various labels appear.

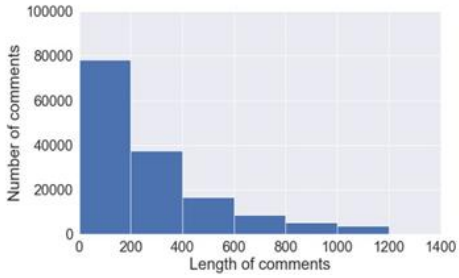


Fig. 3. Plot 3

The frequency distribution of note lengths is depicted in the third graph. Comments might be as short as 100 characters or as lengthy as 1200 characters, depending on what people have seen. However, keep in mind that most comments are less than 200 characters long.

Based on the experimental data analysis results, a pre-processing requirement for picking remarks was established: they had to be less than 400 words long.

Data Pre-Processing

Machine learning and data analysis both rely substantially on data compilation. It entails putting unstructured content into a usable format.

A critical first step is to prepare raw, unstructured data for use in building and training machine learning models. There are two ways to achieve the goals of our collection. Typically, assembling data for machine learning systems is a two-step procedure. The term data cleaning refers to the act of deleting extraneous material from text at the beginning of the process. The second stage is feature engineering, which entails extracting features from data and translating them into formats that machine learning programs can use.

Steps for Data (Text) Cleaning:

- Punctuation and other non-ASCII or special characters are eliminated during the procedure.
- Dissecting the comments into distinct phonemes is part of the procedure.
- The method for removing stop phrases.
- The two phases are cutting and lemmatization.
- The splitting and lemmatization methods.

As a result, the remarks are delivered in the form of groups of cleaned tokens. Each remark must be converted into a vector before it can be used by SciKit Learn's algorithms.

ISSN: 2366-1313

Count Vectorization and TF-IDF Transformation.

	Word 1 Count	Word 2 Count	...	Word N Count
Message 1	0	1	...	0
Message 2	0	0	...	0
...	1	2	...	0
Message N	0	1	...	1

Fig. 4. The Count Vectorizer is used to model a container of words.

Any machine learning model that works with our dataset may now be used to prepare it for the train-test divide.

Finalizing Evaluation Metrics

The Count Vectorizer technique is often used in natural language analysis to generate bundle of words models. This technique considers text data to be a basic list of words, regardless of their order or spelling. Based on the frequency with which a word appears in a text or corpus, the Count Vectorizer provides a numerical value to it. In this way,

Label-based metrics:

The data can now be included into the appropriate machine learning model using the train-test split technique.

Example-based metrics:

These criteria are assessed independently for each label, regardless of their interrelationships. The mean of all identifiers is then computed. Metrics such as average precision and error rate per unit of work can be examined.

These figures are based on actual events. The values we just described are calculated for each instance and then compiled by merging all of the test findings. For example, accuracy, log-loss, and hamming-loss can be used to assess a model's performance.

We are astounded that the vast majority of comments in our dataset are positive and that our data is not distorted. As a result, basing measures entirely on their precision is not a good idea. For example, 92% of the sample comments addressed non-harmful features. This means that for each response, a basic machine learning approach could correctly predict the non-toxic attribute 92% of the time. As a result, it is prudent to choose the metric that represents the quantity of loss. As a result, our machine learning algorithms will measure and rank the performance of various models using accuracy, Hamming-Loss, and Log-Loss.

Applying Multi Label Classification Techniques

Traditional machine learning algorithms are primarily concerned with categorizing occurrences into a single labeled category. As a result, we will employ techniques to break the multi-label problem into several single-label problems. As a result, we can use tried-and-true basic machine learning algorithms.

Binary Relevance Method:

The relationship between categories is neglected in the Binary Relevance Method. As if it were a single label issue, each label is handled separately.

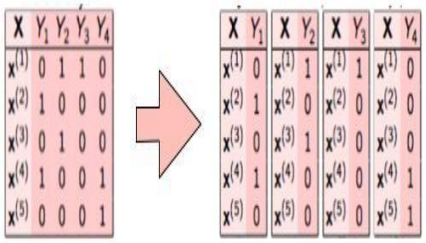


Fig. 5. A Method for Comparing Two Objects

Classifier Chain Method:

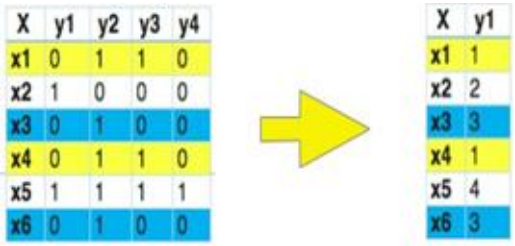
Multiple classifiers are trained using the Classifier Chain Method for a given dataset. The input area is utilized to train not only the next classifier, but also the one preceding it. This technique takes into account the connection between IDs and unprocessed data. Dependence can be expressed in a variety of potentially bad and dangerous ways. Machine learning organizes objects into categories that require several names using the classifier chain approach. It is necessary to train a huge number of binary classifiers, each of which must predict what will happen.

ISSN: 2366-1313

Fig. 6. How to Make Use of a Classification Chain

Label Power Set Method:

The Label Power Set Method takes into account all conceivable label configurations. When we utilize any combination as a label, our multi-label problem becomes a multi-class classification problem.



X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0
x4	0	1	1	0
x5	1	1	1	1
x6	0	1	0	0

X	y1
x1	1
x2	2
x3	3
x4	1
x5	4
x6	3

Fig. 7. The technique for determining the power set.

To achieve the best outcomes, we will apply a set of five machine learning algorithms to each of these strategies.

The following types of machine learning were used in this study:

- Naive Bayes with multiple nodes
- Random Forests' third technique to data organization.
- Different Bernoulli Nave Bayes variants
- Put 5 in the center of the map. Mountain Division Facilities

4. RESULT AND ANALYSIS

Each machine learning method employed three separate methods. Binary Relevance, Classifier Chain, and Label Power Set were the three. Each machine learning algorithm is evaluated using one of three methods: accuracy, log-loss, and hamming-loss.

	Hamming_loss	Accuracy	Log_loss
Binary Relevance	3.861695	88.290936	1.901773
Classifier Chain	3.616800	88.695013	1.374349
Label Powerset	4.167815	88.300119	0.536172

Fig. 8. The method of Multinomial Naive Bayes (NB) is intriguing.

	Hamming_loss	Accuracy	Log_loss
Binary Relevance	3.838736	88.235834	2.174770
Classifier Chain	3.798941	88.281752	2.014295
Label Powerset	4.317813	88.143999	0.538714

Fig. 9. The Bernoulli Naive Bayes (NB) statistical model is well-known in machine learning and natural language processing.

	Hamming_loss	Accuracy	Log_loss
Binary Relevance	2.398445	90.118468	1.982613
Classifier Chain	2.401506	90.146019	1.878685
Label Powerset	2.687728	89.925613	1.241680

Fig. 10. The Random Forest Classifier is a machine learning approach that falls under the umbrella of ensemble learning. It is used for a wide range of classification jobs in a variety of industries, including,

5. CONCLUSION

In this study, accuracy, Hamming-Loss, and Log-Loss were used to compare three approaches to implementing various machine learning methods. A great deal of research has revealed that there is no single optimum way to handle the situation at hand. In every algorithm, however, there is a perfect blueprint for providing the finest possible results. However, when the time required by each method is considered, it is clear that Random Forest is not the best choice for this set of data; other algorithms can get the same results faster.

In future studies, the algorithms could be changed such that multi-label classification is directly achievable. This is accomplished through the use of algorithm adaption approaches. We also aim to investigate how sophisticated deep learning techniques such as recurrent neural networks (RNN), multilayer perceptrons (MLP), and convolutional neural networks (CNN) can be employed in the near future. We are confident that when combined with these cutting-edge deep learning models, our proposed approach would produce better outcomes.

ISSN: 2366-1313

REFERENCES

1. Yin, Dawei, Xue, Zhenzhen, Hong, Liangjie, Davison, Brian, Edwards, April, Edwards, Lynne. (2009), Detection of harassment on Web 2.0
2. Ravi, P. (2012), Detecting Insults in Social Commentary.
3. Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012, Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12). Association for Computing Machinery, New York, NY, USA, 1980–1984. DOI: https://doi.org/10.1145/2396761.2398556.
4. Razavi, A.H., Inkpen, D., Uritsky, S., and Matwin, S. (2010), Offensive Language Detection Using Multi-level Classification. Canadian Conference on AI.
5. Kansara, Krishna B. and N. Shekokar. A Framework for Cyberbullying Detection in Social Network. (2015).
6. Maxime Rivet and Mael Tran, Toxic comments classification, Stanford University journal Year [2016].
7. Spiros V. Georgakopoulos et al. Convolutional Neural Networks for Toxic Comment Classification, Cornell University arXiv:1802.09957 Year 2018.
8. Y. Chen and S. Zhu, Detecting Offensive Language in Social Media to Protect Adolescents, [Online]. Available: http://www.cse.psu.edu/ sxz16/papers/SocialCom2012.pdf
   M. Duggan, Online harassment 2017, Pew Res., pp. 1–85, 2017, doi:202.419.4372