CSC-496/696: Natural Language Processing and Text as Data

Department of Computer Science, American University

Semester: 2025 Fall
Date: December 9, 2025
Student: Anna Hyunjung Kim
Instructor: Patrick Wu

Propose

Title: Measuring Ethical Risks in AI-Generated News Using NLP with the UNESCO Ethics of AI Framework

Research Question: How many problematic errors occur ethically in news articles generated by AI to some extent. Also, which category of the AI ethics principles proposed by UNESCO do these issues correspond closest to?

Data Set

1. Train Data(data_1):
    a. https://huggingface.co/datasets/hendrycks/ethics (hendrycks, 2021)
    b. 21.8k rows
    c. Use this data to label sentences ethically appropriate/inappropriate, and train a text classification model.
2. Test Data(data_2):
    a. https://huggingface.co/datasets/lvulpecula/ai_watermarked_fake_news-v2 (lvulpecula, n.d.)
    b. 1.5k rows
    c. A trained ethics model is applied to this data to identify ethically problematic texts among AI-generated articles.
3. Ethical category(data_3:
    a. https://huggingface.co/datasets/ktiyab/ethical-framework-UNESCO-Ethics-of-AI (ktiyab, n.d.)
    b. 483 rows

For articles classified as "problematic" as a result of the test, match each principle description in this data with the nearest ethical category meaningfully.

Abstract

This project is the final assignment for the course 'CSC-496/696: Natural Language Processing and Text as Data'. This topic started because the diagnosis and caution regarding the ethical problems included in news automatically generated by Large Language Models (LLMs) are thought to be an important issue for our society. This project process uses a model called DistilBERT to train two different specialized detectors. One is for ‘justice(for ethical)’ issues and the other is for ‘HateXplain (hate/aggression)’ to identify dangerous sentences in the text. Then, the sentences concluded as risky are compared one by one with the 11 UNESCO AI Ethics Principles using Cosine Similarity, and the one with the highest similarity is extracted and used for classification. This becomes the beginning of identifying not just the detection of risk, but also what kind of ethical nature the detected problems possess.

1. Introduction

1.1. Project Objectives

As ChatGPT became widespread, people started creating many sentences using GenAI. This use spread beyond individuals to industry and politics, and even news, which must maintain neutrality, reached a level where AI produces it. This started from the question of whether there might be misinformation or biased information in the news that people read and should judge together.

The massive articles produced by AI brought positive effects like increased productivity, but they can also deviate from human moral and social consensus. Ethical issues that could be problematic for someone and lead them to have wrong prejudices have emerged. A more serious problem is that the persistence of this type of misinformation can expand beyond individuals to cultural discrimination, racism, and social discrimination. People might believe the wrong thing is right and might end up having wrong moral indicators. That is why we must continuously study the fairness of the outputs generated by AI. To visualize these risks, this project will subdivide news articles into sentence units, numerically represent the risk index of each one using NLP, and then compare them with the UNESCO AI Ethics Principles standard.

1.2. Project Structure

The goal of this project is to construct a pipeline that analyzes the ethical risks of AI-generated news through NLP modeling and to demonstrate its effectiveness. To achieve this goal, it will proceed with the following detailed components:

1.2.1. Two Risk Classifiers:

Two deep learning classifiers based on two datasets—Justice and Hate/Aggression (HateXplain)—will be created individually, and their performance will be tuned. Then, the best one will be used. And sentences with a high-risk score will be classified.

1.2.2. UNESCO Framework Mapping:

Detected risky sentences will be semantically mapped to the 11 principles of the UNESCO AI Ethics Principles to interpret the specific 'ethical content' of the risk.

1.2.3. Integrated Pipeline Construction and Visualization:

The constructed classifiers and mapping system will be applied to actual AI-generated news data, and the risk will be aggregated and visualized at the article level.

The entire implementation process and analysis of the project are all in three Jupyter Notebook files to ensure reproducibility. (1of3-Ethics-models.ipynb, 2of3-HateXplain.ipynb, 3of3-Justice+hate_applied_analysis.ipynb)

2. Data Preparation and Preprocessing

2.1. Key Datasets Utilized

2.1.1. ETHICS Dataset (Justice):

In the dataset, there are ‘Common Sense’ and ‘Justice’. Among them, the Justice data deals with social fairness and rights issues and is labeled with 0 and 1. This is used for training the fairness violation detection model. Through the tuning process, Justice was used instead of Common Sense or the entire set.

2.1.2. HateXplain Dataset:

This is the dataset that will be used for modeling to detect hate speech and offensive language. Originally, only Justice was intended to be used, but it was realized that a single feature can absolutely not evaluate moral problems, so this was added. According to the data, each sentence goes through the votes of multiple annotators and has one final label among "hatespeech," "offensive," or "normal," and also includes the rationale for the label. We do not use the label rationale.

2.1.3. UNESCO AI Ethics Principles:

This is data consisting of the explanatory text of the 11 ethical standards that present an international standard for AI ethics, composed in English. This is used to concretize the ethical violation content of the detected risky sentences.

2.1.4. AI-Generated News Article Data:

By utilizing lvulpecula/ai_watermarked_fake_news-v2 from HuggingFace Datasets, this is used as the target data to actually apply the learned models. This data includes metadata such as the title, body, and AI model name.

2.2. Data Processing (Cleaning)

2.2.1. ETHICS Dataset (Justice):

- **Column Name Normalization:** input -> text, scenario -> text (justice and common sense had different column name)
- **Removal of Unnecessary Columns:** Dropping columns other than label/text/source, such as 'is_short', 'edited'.
- **Type Conversion:** Converting the label column to int
- **Source Distinction Column Addition:** Creation of a column distinguishing whether it is 'commonsense' or 'justice'.
- **Dataset Merging:** Merging the train/val/test splits of commonsense and justice respectively, and converting them into a dataset dictionary.

2.2.2. HateXplain Dataset:

- **JSON Download:** Converting the original data(json) files into a pandas dataframe.
- **Label Type Mapping:** Since strings and integer labels are mixed, mapping to consistent integer labels.
- **Removal of Unnecessary Columns:** Dropping fields unnecessary for Generator training, such as 'post_tokens'.

2.2.3. UNESCO AI Ethics Principles:

- Only the principle description text was needed for embedding, so only unnecessary fields were removed.

The three datasets, excluding the news data, were very clean. Only the column names needed to be changed for joining.

2.2.4. AI-Generated News Article Data:

- **Sentence Segmentation by Article:** Splitting the article body (text) into sentence units (e.g., based on period, line break, etc.), and separating the title and body.
- **Adding Field Information per Sentence:** Adding fields such as article_id, sent_idx, is_title, title, source_model, sentence, and sentence length.
- **Handling Unnecessary Characters/Spaces/Special Characters:** Deleting/refining line breaks, the special characters "\n", "\t", etc., in some code.

3. Modeling and Tuning: Construction of Dual Risk Classifiers

To find the hidden multi-dimensional risks in AI-generated news articles, the ETHICS-based ethical classifier and the HateXplain-based hate/aggression classifier were created independently.

3.1. Ethical Classifier (Justice) Modeling: 1of3-Ethics-models.ipynb

3.1.1. Data Basis and Initial Experimentation

The training basis for the ethical classifier is the ETHICS dataset, which is for labeling the moral judgment ability of AI. This dataset is composed of two sub-axes: Everyday Moral Judgment (Commonsense) and Social Fairness/Rights Issues (Justice). In the initial modeling attempt, I tried to train a single classifier model by combining the entire ETHICS data to cover these two moral aspects, but a problem was discovered during the experiment. The ethical situations and text distributions of the two detailed datasets (Commonsense and Justice) were different, and when this data was mixed for training, the model's predictions showed a tendency to become generally unstable. So, I tried together and seperately.

3.1.2. Model Separate Training and Final Model Selection

When comparing the validation (VAL) and test (TEST) performance metrics of the six model versions provided (full_v1, commonsense_v1, justice_v1, full_v2, commonsense_v2, justice_v2), the rationale for finally selecting the Justice_v1 because the project's goal was the detection of Social Fairness/Rights Issues (Justice).

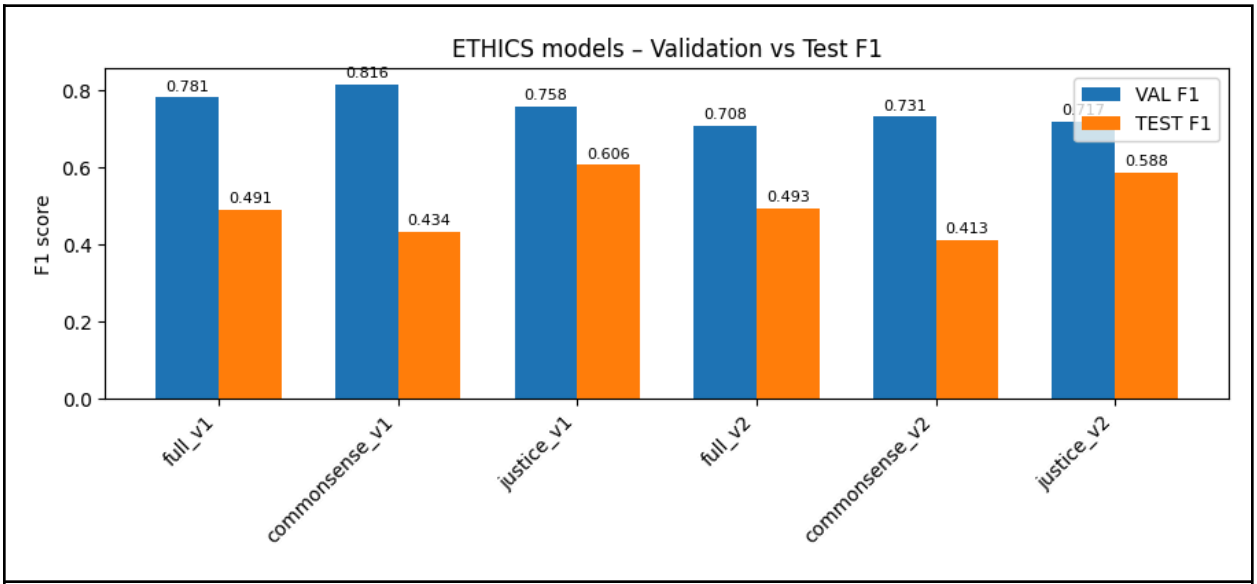


Figure1. ETHICS models – Validation vs Test F1

1) Justice_v1 Model

Justice_v1 (Epoch 3.0) showed the best performance among the models trained using only the Justice dataset. Following the earlier analysis that training on the full data (Full) makes the model unstable, a model specialized for Justice was necessary. Full_v1/v2 models showed generally lower Test Acc, F1, and Recall metrics than the Justice model, reaffirming the problem of mixing datasets. (Full_v1 TEST F1: 0.4914, Recall: 0.4715)

Commonsense_v1/v2 models are specialized in everyday moral judgment, making them unsuitable for analyzing news articles dealing with social fairness, and their test performance was also lower than the Justice model. (Commonsense_v1 TEST F1: 0.4344, Recall: 0.3974)

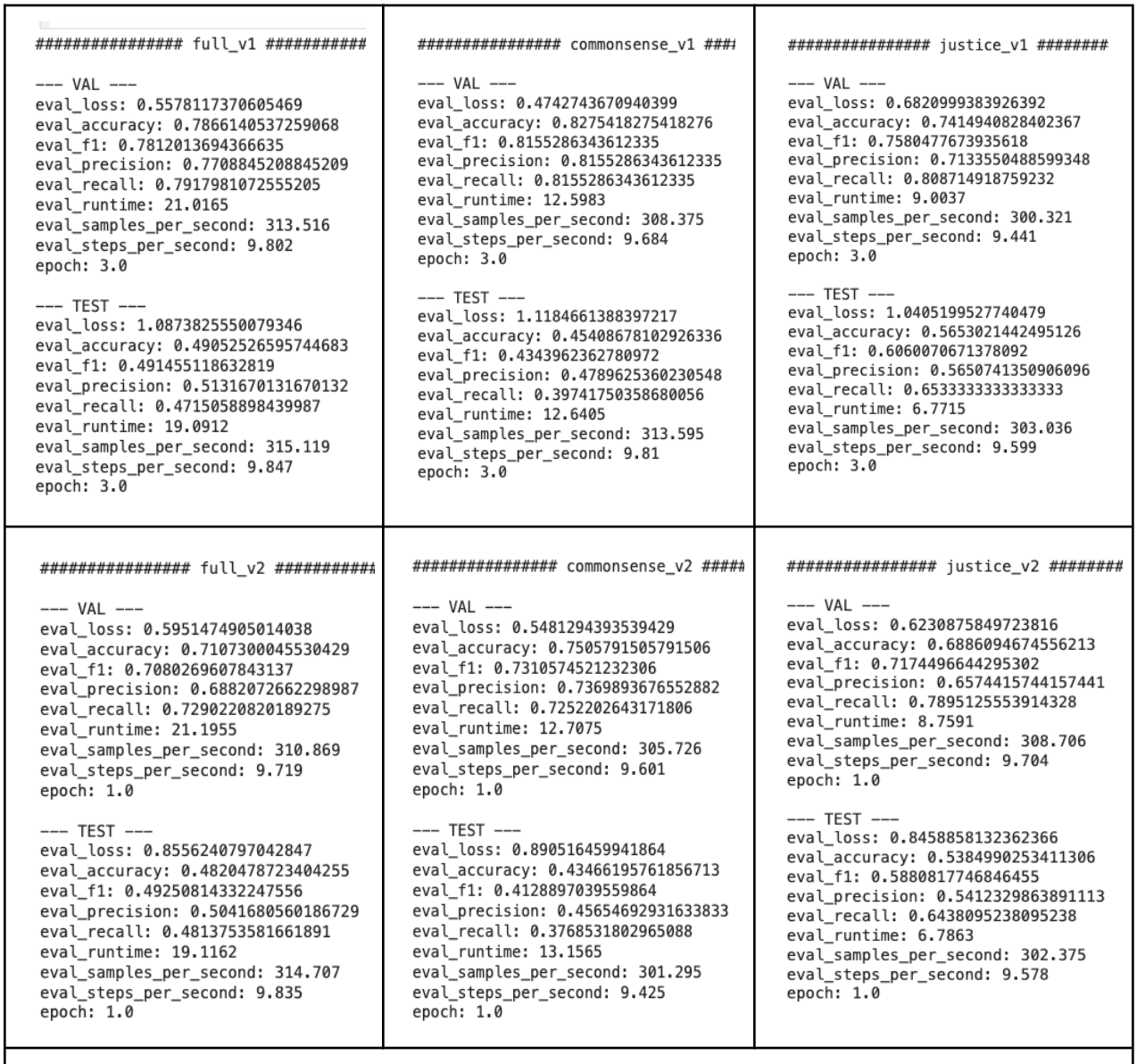


Figure2. Compare models

2) Rationale for Selection: Justice_v1 vs. Other Justice Models

Since the core objective of the project is to detect ethical problems within news articles, the Recall metric is particularly important. Recall indicates how many of the truly risky sentences the model correctly detected as 'risky.' Therefore, it is suitable for an ethical detection model that prioritizes not missing the risk. Justice_v1 was trained for a longer epoch (Epoch 3.0) on the Justice data and recorded a slightly higher Recall (0.6533 vs. 0.6438) and F1-Score (0.6060 vs. 0.5881) on the test data than Justice_v2 (Epoch 1.0), proving the most excellent fairness detection performance.

So, the Justice_v1 model was trained using only the Justice dataset, and by achieving the highest F1-Score (0.6060) and highest Recall (0.6533) on the test set, it was selected as the most suitable model for detecting social fairness violations within AI-generated news articles.

3.1.3. Hyperparameter Tuning Details

For Version 2 (ver2) models, the training loss looked good, but the F1, precision, and recall metrics were uncertain. Also, I thought the given data would be easy to remember the patterns because the labels were attached differently just by changing the words, which indicated overfitting. Therefore, two main changes were applied to reduce overfitting.

First, the dropout rate was increased to 0.2 (dropout=0.2 / attention_dropout=0.2). Increasing the dropout rate reduces overfitting by randomly disabling parts of the model during training, which prevents the model from overly memorizing the training data, helping it generalize better to new sentences.

Second, the number of training epochs was reduced to 1 (num_train_epochs = 1) because the ETHICS dataset is small and noisy. Training for too long leads to overfitting, where the model learns the training data perfectly but performs worse on validation examples. Using only 1 epoch helps prevent this overfitting.

However, upon comparison, Version 1, which did not have (dropout=0.2 / attention_dropout=0.2/ num_train_epochs = 1) applied, proved to be better, and thus features were not used. Consequently, Version 1 was used, which had the following settings applied: weight_decay=0.01, label_smoothing_factor=0.1, num_train_epochs=3

3.2. Hate/Offensiveness Classifier (HateXplain): 2of3-HateXplain.ipynb

3.2.1. Data Preprocessing

The second step is detecting social risks for example, hate speech and offensive language, that cannot be caught by ethics alone.

The HateXplain dataset was used, and each sentence is classified as hatespeech, normal, or offensive. DistilBERT was used identically, and sentences were processed and used with a maximum of 256 tokens. The parameter values were applied exactly as used before.

3.2.2. Model

Validation Set

Metric	Value
Loss	0.8224
Accuracy	0.6946
F1	0.6819
Precision	0.6831
Recall	0.6831
Runtime (sec)	6.55
Samples per Second	293.45
Steps per Second	9.31
Epoch	3

Test Set

Metric	Value
Loss	0.8156
Accuracy	0.6923
F1	0.6762
Precision	0.6758
Recall	0.6792
Runtime (sec)	6.71
Samples per Second	286.79
Steps per Second	9.09
Epoch	3

Figure3. Hate/Offensiveness Classifier model

4. Application to Real Article Data and Pipeline Integration

The final step is to build an integrated pipeline that applies the two constructed models (Justice, HateXplain) to the actual AI news article data to diagnose sentence-level risk and aggregate it into article-wide statistics. At this time, the UNESCO AI Ethics Principles will be compared and matched with each sentence one by one. This process is in 3of3-Justice+hate_applied_analysis

4.1. Model Loading and Dataset Preparation

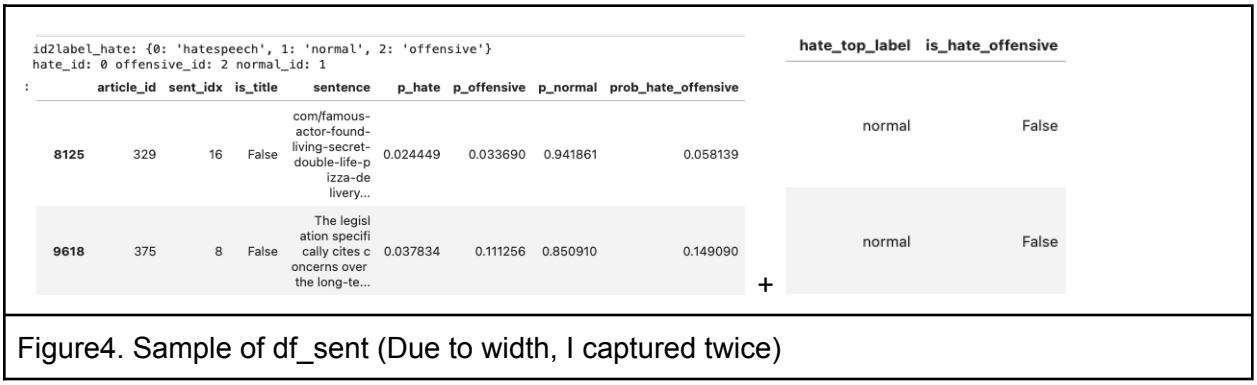
The saved justice v1 and HateXplain models were directly loaded as PyTorch model forms and executed to perform prediction on every sentence.The AI news dataset title and text were decomposed into single sentence units. At this time, it was divided using line breaks, the special characters "\n", "\t".

4.2. Classifier Prediction and Compound Risk Classification

When applying the predictions of both the Justice model and the HateXplain model to each sentence, the following key columns are generated as result values for each sentence. These columns multi-dimensionally contain the sentence's risk level and detailed prediction information.

- **article_id**: The unique number of the article that the sentence belongs to.
- **sent_idx**: The sentence index within the article (which number sentence it is).
- **is_title**: Whether the sentence is the article's title (True/False).
- **sentence**: The original news sentence (the actual text analyzed).
- **p_hate**: The probability of the 'hatespeech' class predicted by the HateXplain model.
- **p_offensive**: The probability of the 'offensive' class predicted by the HateXplain model.
- **p_normal**: The probability of the 'normal' class predicted by the HateXplain model.
- **prob_hate offensive:** The value of p hate plus p offensive (the sum of the prediction probabilities for the two classes: hate slash offensive).
- **hate_ top_label**: The final label selected by the HateXplain model with the highest probability ('hatespeech', 'offensive', 'normal').
- **is_hate_offensive**: Whether the sentence has hate or offensive risk (True/False based on the prediction probability threshold).

By combining these columns with the Justice prediction results, it is possible to analyze the "justice risk (fairness issue)," "hate/offensive risk," and "overlap risk status" for each sentence from various views, and even aggregate statistics at the article level.



id2label_hate: {0: 'hatespeech', 1: 'normal', 2: 'offensive'}
hate_id: 0 offensive_id: 2 normal_id: 1

	article_id	sent_idx	is_title	sentence	p_hate	p_offensive	p_normal	prob_hate_offensive	hate_top_label	is_hate_offensive
8125	329	16	False	com/famous-actor-found-living-secret-double-life-pizza-delivery...	0.024449	0.033690	0.941861	0.058139	normal	False
9618	375	8	False	The legislation specifically cites concerns over the long-te...	0.037834	0.111256	0.850910	0.149090	normal	False

Figure4. Sample of df_sent (Due to width, I captured twice)

4.3. Interpretation via UNESCO Ethical Principle Mapping

To specifically interpret the ethical violation content of risky sentences semantic mapping using the UNESCO AI Ethics Principles is performed

4.3.1. Embedding Based Quantitative Analysis

1) Sentence Embedding Generation: DistilBERT(distilbert-base-uncased) is applied to both the news sentence and the UNESCO principles describing each sentence to obtain embedding vectors.

2) Cosine Similarity Calculation: The embedding vector of each news sentence is compared **one by one** with the embedding vector of each of the 11 UNESCO principles using the Cosine Similarity method. ‘For loop’ was used. Among the similarity scores with each of the 11 principles, the ethical principle with the highest score was finally selected as the principle closest to the sentence's semantic ethical risk.

Using this method, I can do more than just classify a risky sentence. I can also automatically and numerically explain the specific kind of ethical problem it has. For instance, I identify if the risk is about Awareness and Literacy, Safety and Security, or Human Oversight and Determination. I determine this by seeing how closely the sentence matches the international standard principles.

principle	count
Fairness and Non-Discrimination	133
Right to Privacy and Data Protection	72
Proportionality and Do No Harm	55
Transparency and Explainability	39
Safety and Security	37
Sustainability	35
Human Oversight and Determination	31
Responsibility and Accountability	31
Multi-stakeholder and Adaptive Governance & Collaboration	28
Awareness and Literacy	15
Awareness & Literacy	7
Multi-stakeholder and Adaptive Governance and Collaboration	1
Human Dignity and Autonomy	1
Safety and Security, Responsibility and Accountability, and Human Oversight and Determination	1
Safety and Security & Fairness and Non-Discrimination	1

Figure 5. distribution of UNESCO principles dataset

4.4. Article-Level Aggregation and Visualization

Based on the sentence unit analysis results, various aggregation indicators, such as the risky sentence ratio per article, the distribution of major UNESCO principles, and the ratio of each risk type, are calculated, and the results are presented through intuitive visualization, such as histograms and bar charts.

4.4.1. Article-Level Risk Quantification

	article_id	sent_idx	prob_unethical	sentence	unesco_principle	unesco_score
11010	504	5	0.959857	"We are proud to offer a flying car that anyone can pilot, with no special license or training required.	Right to Privacy and Data Protection	0.793193
21630	1442	11	0.955538	ÒBy owning these establishments, theyÕre able to exert influence over peopleÕs daily routines, h elping them feel comfortable while they secretly prepare to expand their reach.	Awareness and Literacy	0.825384
8972	358	2	0.955039	By: Jennifer Solak\nA man in Ireland recently won the lottery, but he's not the kind of guy who likes to brag about it.	Awareness & Literacy	0.792672

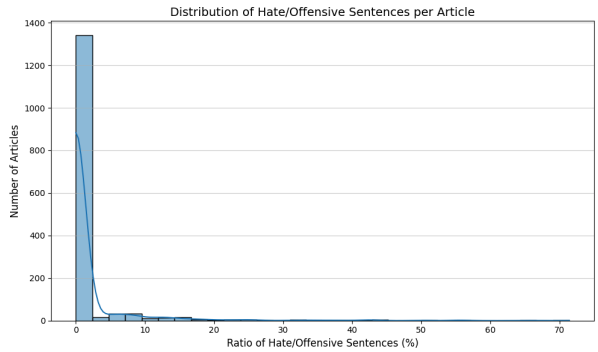
Figure 5. Mapping chart high-risk unethical sentences to UNESCO principles

	article_id	sent_idx	prob_hate_offensive	sentence	unesco_principle	unesco_score
968	47	16	0.968286	"We will be working diligently to bring this Ch icken Nugget Bandit to justice.	Safety and Security	0.793170
972	47	20	0.941983	\nAs the investigation unfolds and the Chicken Nugget Bandit remains at large, l ocals are left contemplating the motivations and cravings that could drive someone to br eak into a fast-food establishment solely for the purpose of devouring frozen nuggets.	Safety and Security	0.819592

Figure 6. Mapping hate/offensive sentences to UNESCO principles

4.4.2. High-Risk Tail Distribution Analysis on hate/offensive risk

The hate/offensive risk inherent in AI generated news articles was confirmed to appear very sparsely. As a result of representing the ratio of hate/offensive sentences in all articles with a distribution, graph histogram the majority of articles about 1300 or more showed an extremely right skewed distribution with a sentence ratio of less than 1 percent or close to 0. This suggests that most articles generated by AI are safe as they do not contain hate or offensive content but the risk can be concentrated in a small number of articles if it occurs. To analyze the characteristics of this high risk group, High Risk Tail in detail the distribution was checked by separating only the articles with a hate/offensive sentence ratio exceeding 5 percent. As a result, in this high risk article group the risky sentence ratio was most concentrated between 5 percent and 15 percent and some articles even observed ratios of over 40 percent up to near 70 percent quantitatively demonstrating that the intensity of the risk can be very high once exposed.



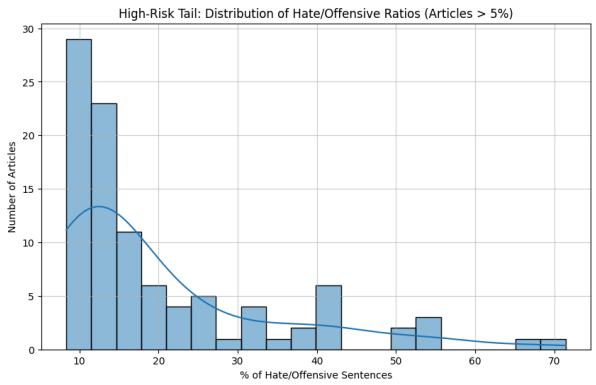


Figure 7 & 8. Distribution of Hate/Offensive Sentences per Article

4.4.3. High-Risk Distribution Analysis on Unethical risk

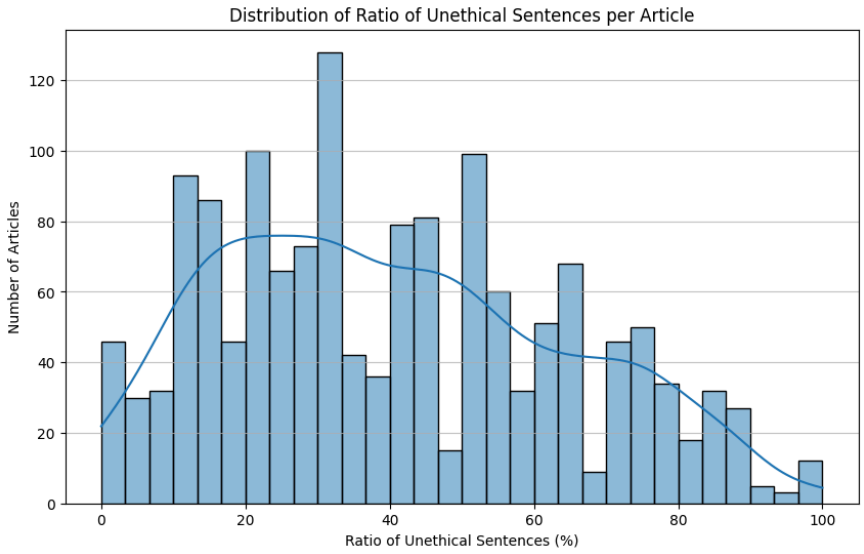


Figure 9. Distribution of Unethical Sentences per Article

This graph shows the distribution of the ratio of unethical sentences within AI-generated news articles, exhibiting clearly different characteristics from the previously analyzed hate/offensive risk distribution. Unlike the hate/offensive risk, which was extremely skewed toward 0 percent, the unethical risk is distributed relatively broadly from 0 percent to 100 percent. The distribution showed a tendency to form major peaks near 30 percent and 50 percent, which means that articles with that sentence ratio are the most numerous. In particular, high-risk article groups with an unethical sentence ratio of 70 percent or more were consistently observed across the entire range. This broad distribution suggests a high probability that if an AI-generated news article contains an ethical problem, the risk is widely spread at a considerable ratio (30 percent to 50 percent) throughout the entire article.

4.4.4.Analysis of Risk Frequency by UNESCO Principle and Mapping Cases

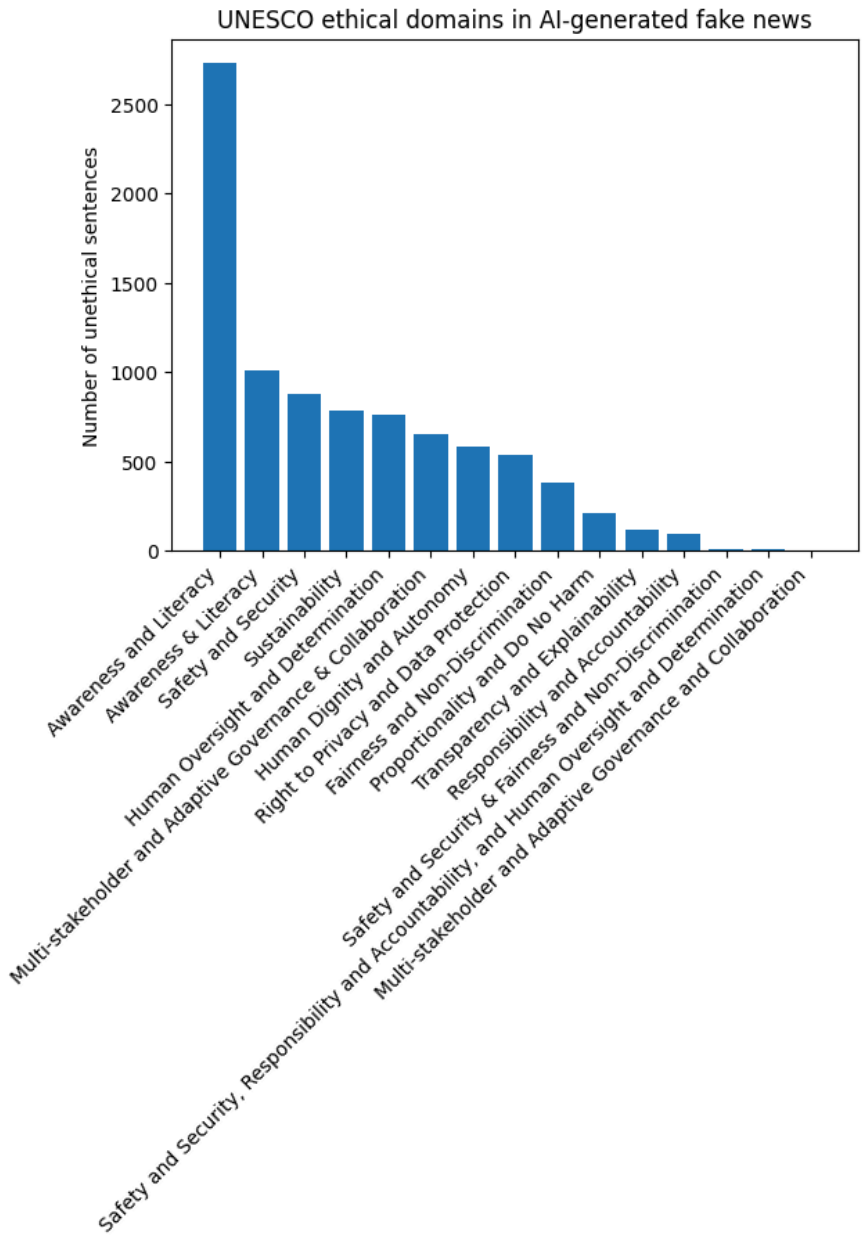


Figure 10. Analysis of Ethical Risk Frequency by UNESCO Principle and Mapping Cases

Overall, the aggregation of detected unethical sentences by UNESCO principle showed that the 'Awareness and Literacy' principle had the highest violation frequency, exceeding approximately 2500 instances. This was followed by 'Safety and Security' with around 1000 instances, and 'Sustainability' was also ranked highly. This suggests that when ethical risks occur in AI-generated news, the highest probability is associated with confusing the reader's level of understanding or perception, or causing problems related to social safety. In contrast, complex

principles such as 'Multi-stakeholder and Adaptive Governance and Collaboration' recorded the lowest frequency.

The results of mapping specific high-risk sentences to UNESCO principles (Figures 5 and 6) support this distribution. Unethical sentences detected by the Justice Classifier (Figure 5) were primarily linked to the 'Right to Privacy and Data Protection' and 'Awareness and Literacy' principles with high scores of 0.79 or higher, capturing ethical issues related to technology misuse (flying cars without licenses) or data influence. Notably, sentences detected by the Hate/Offensive Classifier (Figure 6, related to 'Chicken Nugget Bandit') were consistently mapped to the 'Safety and Security' principle with very high similarity scores of 0.94 or higher.

These results demonstrate that beyond simple hate speech, when AI-generated content targets specific subjects or causes social instability, this risk can be quantitatively interpreted within the ethical context of 'Safety and Security'.

4.4.5. Distribution of UNESCO Principles Mapped to Hate/Offensive Sentences

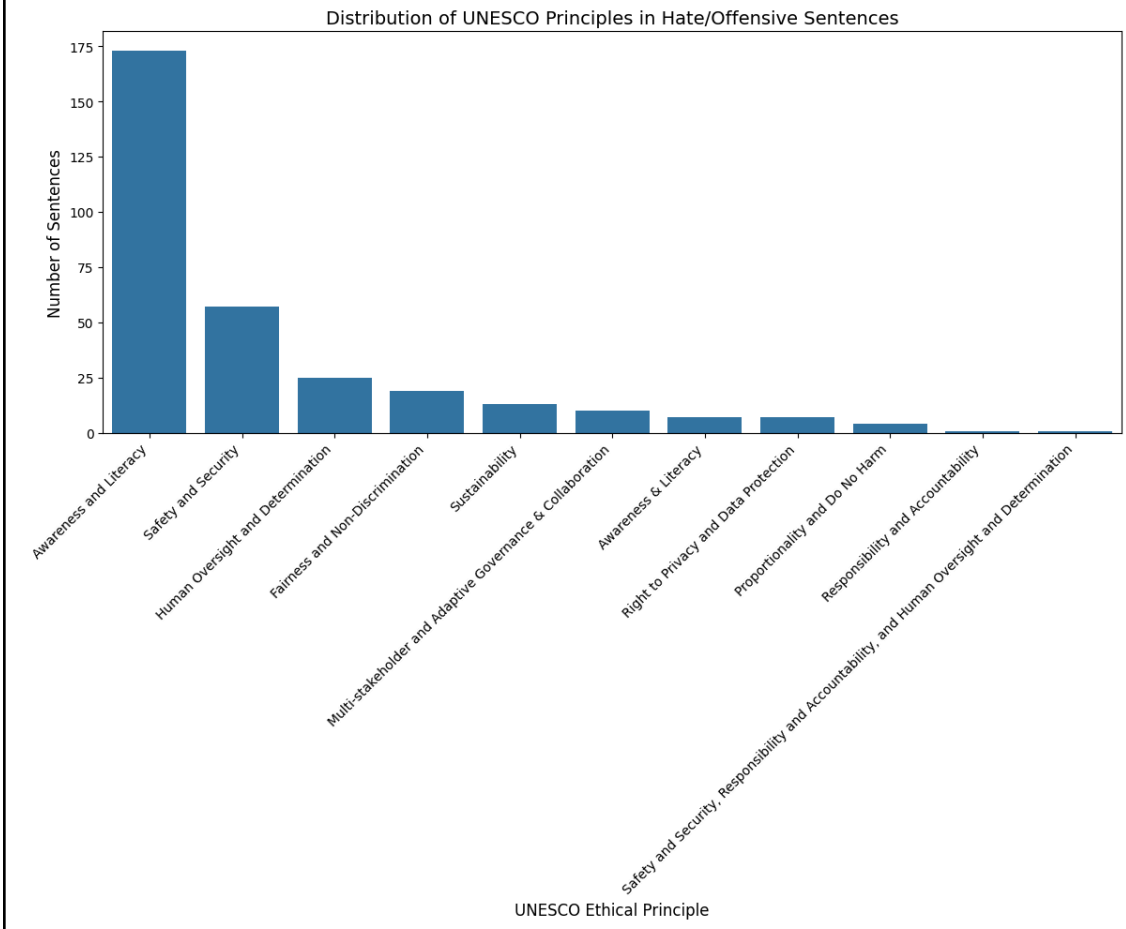


Figure 11. Analysis of Hate/Offensive Risk Frequency by UNESCO Principle and Mapping

The analysis of which UNESCO principles were most frequently mapped to the sentences detected by the Hate/Offensive Classifier (Figure 11) showed a different pattern from the overall distribution of unethical sentences (Figure 10). In the hate/offensive sentences, the 'Awareness and Literacy' principle accounted for the most overwhelming frequency, with approximately 175 instances or more.

This suggests that hate/offensive content is likely associated not just with simple criticism of a target, but also with an intention to manipulate the reader's perception or distort information. The second highest frequency was the 'Safety and Security' principle, recording approximately 60 instances. This supports the finding from the earlier case analysis (Figure 6) that hate/offensive sentences are often interpreted in the context of creating social instability. Other principles followed, including 'Human Oversight and Determination' and 'Fairness and Non-Discrimination'.

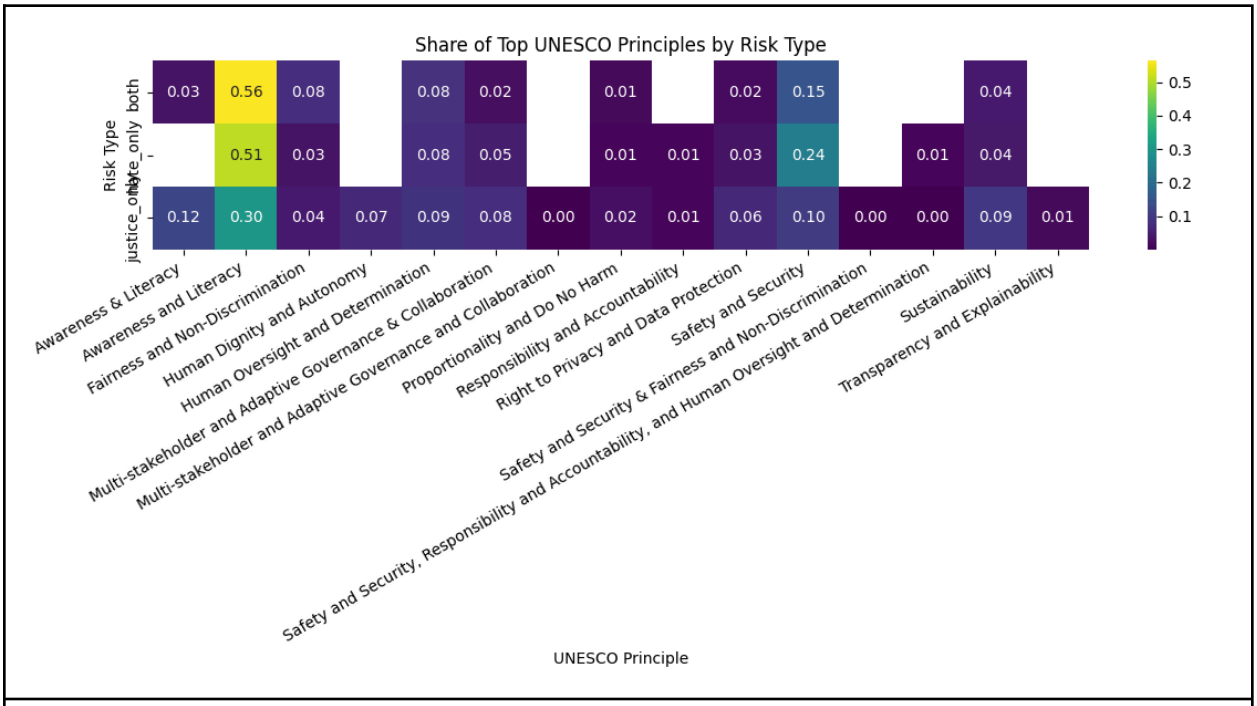
4.4.6. Analysis of UNESCO Principle Share by Risk Type (Hate/Justice)

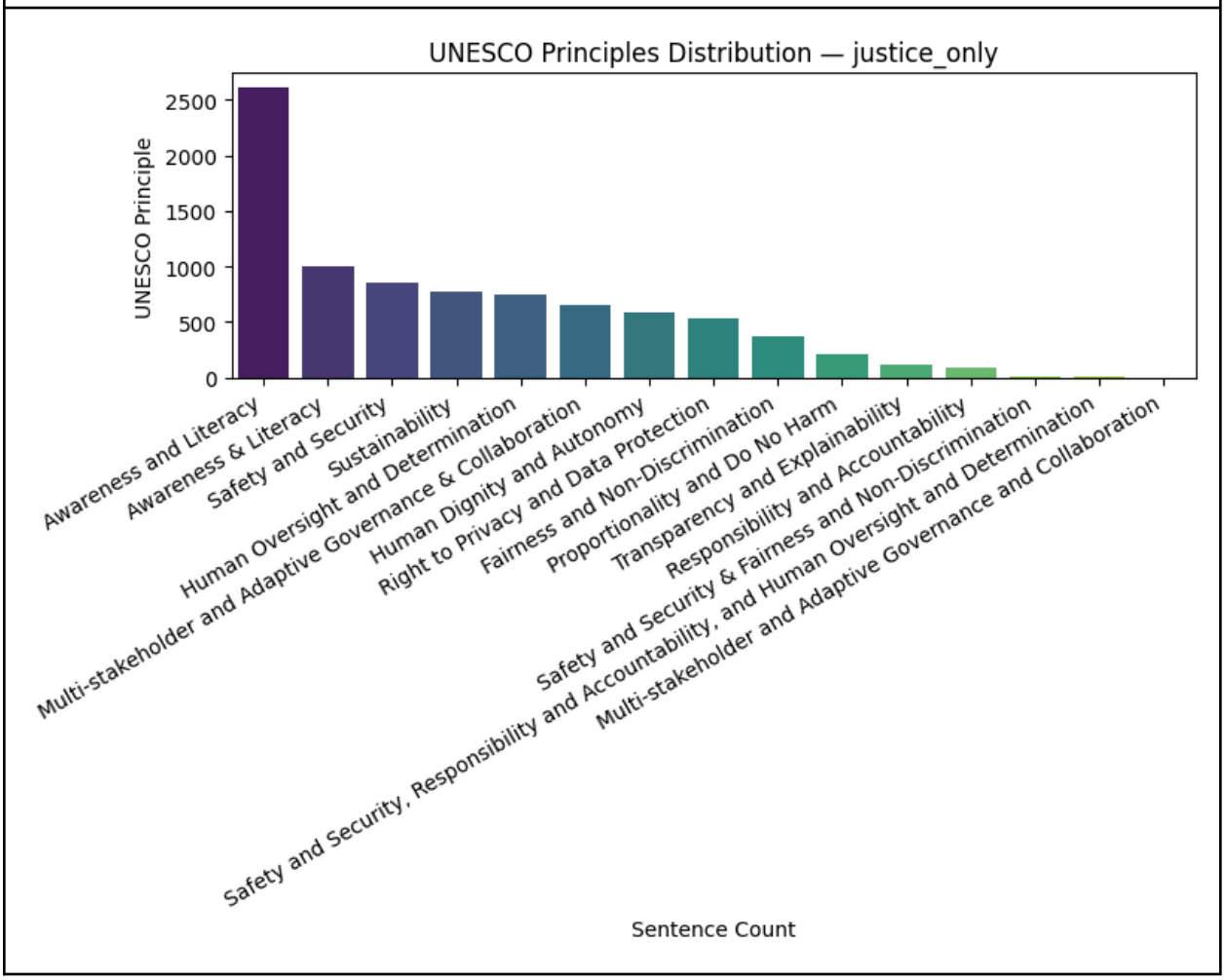
The analysis of the share distribution of detected sentences by risk type (Hate/Justice/Both) across UNESCO principles (Figure 12) clearly separated the ethical characteristics of each risk type.

For the 'Justice Only' sentences, the 'Fairness and Non-Discrimination' principle accounted for the largest share at 0.30, suggesting that the risk detected by the Justice Classifier is most deeply related to social fairness issues.

Conversely, 'Hate Only' sentences also showed an overwhelming number one share of 0.51 for the 'Fairness and Non-Discrimination' principle, followed by the combined 'Safety and Security, Fairness and Non-Discrimination' principle at 0.24, proving that hate sentences directly cause issues of discrimination and unfairness.

Interestingly, in the 'Both' type (sentences flagged by both Hate and Justice), the 'Awareness and Literacy' principle recorded the highest share at 0.56. This analysis suggests that when a sentence contains complex risks, the 'Awareness and Literacy' issue, which misleads or confuses the reader, emerges as the most critical ethical issue.


Share of Top UNESCO Principles by Risk Type
Risk Type
both
justice_only
justice_only
0.03 0.56 0.08 0.08 0.02 0.01 0.02 0.15 0.04
0.51 0.03 0.08 0.05 0.01 0.01 0.03 0.24 0.01 0.04
0.12 0.30 0.04 0.07 0.09 0.08 0.00 0.02 0.01 0.06 0.10 0.00 0.00 0.09 0.01
0.5
0.4
0.3
0.2
0.1
Awareness & Literacy
Awareness and Literacy
Fairness and Non-Discrimination
Human Dignity and Autonomy
Human Oversight and Determination
Multi-stakeholder and Adaptive Governance & Collaboration
Multi-stakeholder and Adaptive Governance and Collaboration
Proportionality and Do No Harm
Responsibility and Accountability
Right to Privacy and Data Protection
Safety and Security
Safety and Security & Fairness and Non-Discrimination
Safety and Security, Responsibility and Accountability, and Human Oversight and Determination
Sustainability
Transparency and Explainability
UNESCO Principle


UNESCO Principles Distribution — justice_only
UNESCO Principle
2500
2000
1500
1000
500
0
Awareness and Literacy
Awareness & Literacy
Safety and Security
Sustainability
Human Oversight and Determination
Multi-stakeholder and Adaptive Governance & Collaboration
Human Dignity and Autonomy
Right to Privacy and Data Protection
Fairness and Non-Discrimination
Proportionality and Do No Harm
Transparency and Explainability
Responsibility and Accountability
Safety and Security & Fairness and Non-Discrimination
Safety and Security, Responsibility and Accountability, and Human Oversight and Determination
Multi-stakeholder and Adaptive Governance and Collaboration
Sentence Count

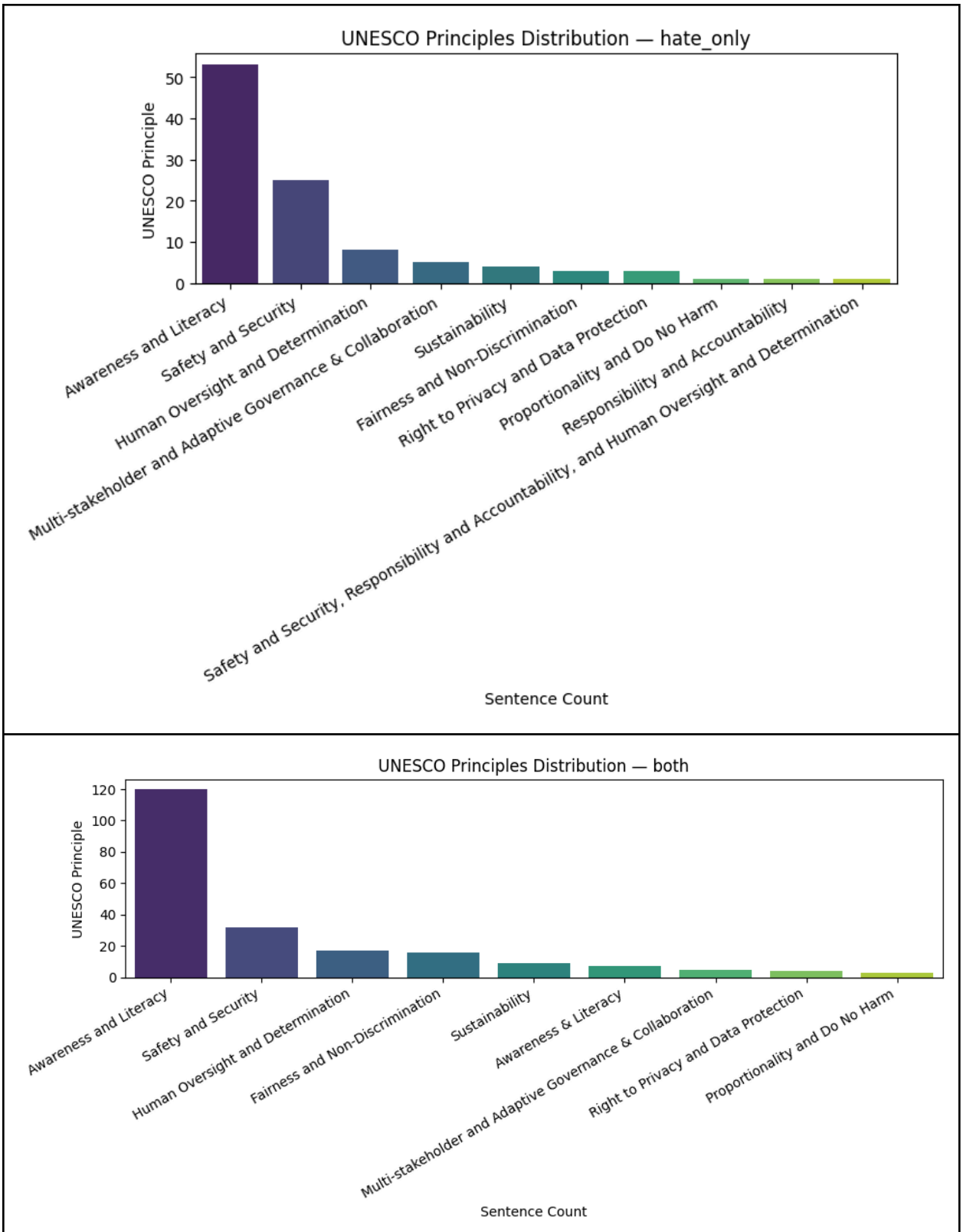


Figure 12-15, UNESCO Principle Share by Risk Type

5. Conclusion and Discussion

5.1. Overall Integrated Flow

This project successfully constructed and demonstrated the effectiveness of a three-stage pipeline designed to diagnose and interpret the complex ethical risks embedded in AI-generated news articles.

The first stage involved building a dual classifier system, independently operating the Justice Classifier and the HateXplain Classifier, aiming for a multi-dimensional approach to risk detection. These DistilBERT-based classifiers simultaneously predict risk types at the sentence level and classify them into compound risk types such as 'both', 'justice_only', and 'hate_only' to ascertain the degree of risk overlap. Significantly, we observed a fundamental difference between the two risk axes, confirming that Unethical risks are broadly distributed, while Hate/Offensive risks exhibit an extreme right-skewed distribution.

The second and last stage is the UNESCO AI Ethics Principle Mapping, which moves beyond simple detection to quantitatively interpret the ethical nature of the risk. This process calculates the Cosine Similarity between the embedding vectors of the risky sentences and the embedding vectors of the UNESCO principles, thereby quantifying the proximity of the sentence to specific international ethical standards. I hope that this process will help society consider which areas should be particularly cautious about and where structural regulations should be applied.

5.2. Limitations of the Experiment and Future Work

The Justice Classifier focuses on fairness/rights issues, while the HateXplain Classifier primarily concentrates on the detection of explicit hate and offensiveness. This constitutes a limitation in that it is difficult to perfectly capture subtle social manipulation, context-dependent incitement, or indirect discriminatory nuances. Therefore, future research must proceed with more dimensions that are better suited to the complex system of language and ethics.

Furthermore, the attempt to map the ethical principles of UNESCO via embedding holds significance, but this method relies on the numerical distance between texts. Consequently, this approach cannot perfectly replace the precision of human intuitive and context-dependent ethical judgment. Therefore, a different, more realistic methodology must be presented in the future.

Lastly, the article data used in this research was collected from Hugging Face and does not possess realism. Therefore, we hope to apply the system to actual news articles published by media outlets in future research.

7. References

Hugging Face. (n.d.). DistilBERT-base-uncased [Pretrained model]. Retrieved from https://huggingface.co/distilbert-base-uncased

Hugging Face. (n.d.). DistilBERT Model Documentation. Retrieved from https://huggingface.co/docs/transformers/en/model_doc/distilbert

Hendrycks, D. (2021). ETHICS Dataset [Data set]. Hugging Face. Retrieved from https://huggingface.co/datasets/hendrycks/ethics

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI With Shared Human Values. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2008.02275

Ktiyab. (n.d.). UNESCO Ethics of AI Principles [Data set]. Hugging Face. Retrieved from https://huggingface.co/datasets/ktiyab/ethical-framework-UNESCO-Ethics-of-AI

Lvulpecula. (n.d.). ai_watermarked_fake_news-v2 [Data set]. Hugging Face. Retrieved from https://huggingface.co/datasets/lvulpecula/ai_watermarked_fake_news-v2

Mathew, B. (2021). HateXplain Dataset [Data set]. GitHub. Retrieved from https://github.com/hate-alert/HateXplain

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 14857-14865. https://arxiv.org/abs/2012.10289

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. https://arxiv.org/abs/1910.01108

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000381137