Yahoo News Comments Classifier

NLP-Based Classification of Yahoo Comments for Enhanced Content Moderation

Presented by

Harshit Koncharla (01243405)
Hemant Kumar(01243485)
Sairaj Kuchana(01243581)

Agenda

- Introduction
- Problem Statement
- Proposed Solution & Methodology
- Dataset
  - Pre-Processing of data
- Model Description
  - GPT 3.5
  - Llama 2
  - Google Gemini
  - LSTM
- Evaluation Metrics & Results
- Findings & Conclusions
- Application & Future Work

Introduction

- The surge of social media and online forums has transformed communication dynamics, with comments serving as crucial elements for user engagement.
- However, moderating user-generated content poses challenges due to the decentralized nature of these platforms.
- To address this, comment categorization using natural language processing offers a scalable solution to classify comments and ensure the safety of online communities.

Problem Statement

- Proliferation of harmful or inappropriate comments due to platform anonymity.
- Negative impact of inconsistent and biased categorization on user experience.
- Manual categorization is limited by time constraints, scalability issues, and biases.
- Need for an automated solution to overcome these challenges.

Proposed Solution & Methodology

- We are introducing a comment categorization tool as a promising approach to automate content moderation.
- We aim to presents a methodology tailored for Yahoo comments, aiming to deploy a robust system capable of categorizing comments into six distinct categories (Humour, Spam, Neutral, Consolidating, Ideological, Abusive), thereby contributing to the advancement of automated content moderation.

Dataset

- Utilization of web scraping techniques for dataset collection.
- Classification of comments into six categories: Ideological, Humorous, Consolidating, Abusive, Spam, and Neutral.
- Training dataset consisting of 20x6 comments categorized across multiple topics.

```
// Get the child count
const childCount = document.querySelector("#spotim-specific > div > div").
shadowRoot.querySelector("div > div > div > div > div > div > div > div >
div.ToastWrapper_providerContainer--11-4-15" +
 "> div.spcv_conversation > div.ToastWrapper_providerContainer--11-4-15 >
div:nth-child(2) > ul").children.length;

// Create an array to hold the CSV data
const csvData = [];

// Loop starts from index 1 (second child) to account for the initial loop check
for (i = 1; i <= childCount; i++) {
  const cellData = document.querySelector("#spotim-specific > div > div").
  shadowRoot.querySelector(div > div > div > div > div > div > div > div >
    div.ToastWrapper__providerContainer--11 - 4 - 15 >
     div.spcv_conversation > div.ToastWrapper__providerContainer--11 - 4 - 15 >
     div: nth - child(2) > ul > li: nth - child(${ i }) > article > div > div >
 div: nth - child(1) > div > div > div.components - MessageLayout - index__message
 - view > div >div.components - MessageContent - index__messageEntitiesWrapper >
 div > span > div > p).textContent;
  console.log(cellData);
  csvData.push(cellData);
}
```

Dataset Cntd.

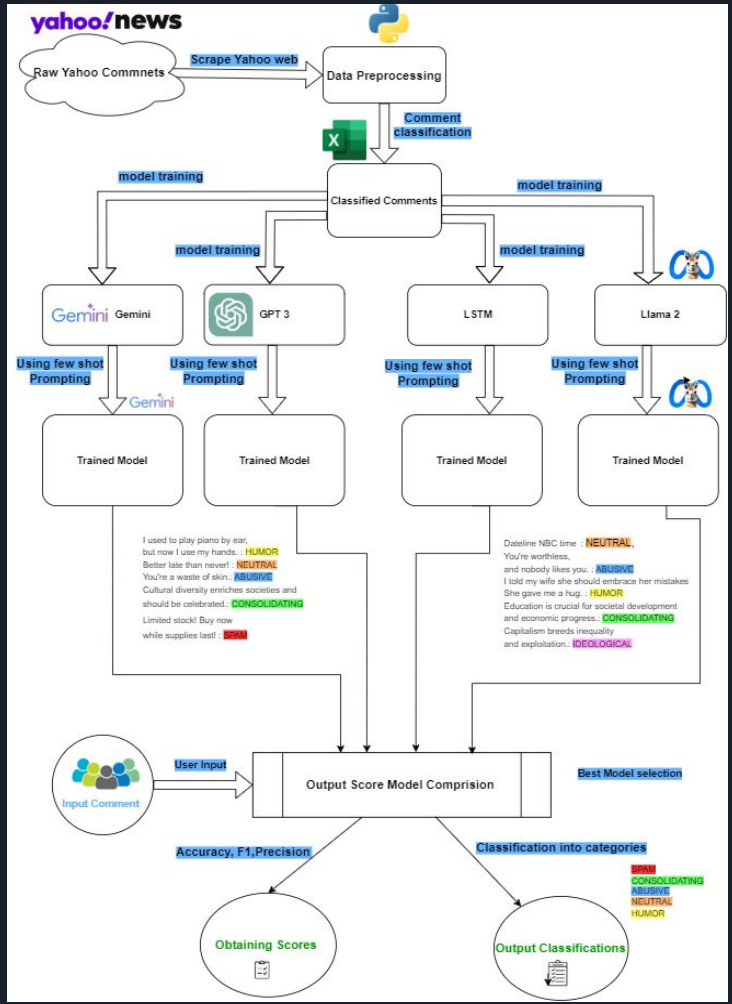
Headline	Comments
Chris Pratt and Katherine Schwarzenegger slamme	They couldn’t just demolish a house without appro
https://www.yahoo.com/entertainment/chris-pra	If it was that big of a deal, they why didn't the city
	Sorry, from the couple pictures I see here the hous
	A 1950's built home would probably suffer from fu
	It is their property. They paid $12 million for it. It
	If the house was so rare and amazing, why was it s
	It is impossible these days to do something withou
	It's THEIR HOUSE... they can do whatever they wa
	I could have sworn the term "McMansion" specific
	If it wasn’t preserved they have a right to do as the
	The thousandaires on the "internet" love to be ap
	They purchased the home and property and were
	I don't always, or even often, approve what peopl
	Hey! It’s NOT even a Craig Ellwood designed prope
	Nobody cared until there was someone they could
	I don't blame them.
	The house wasn’t even designed by Craig Ellwood,
	It's their house. WHO CARES what they did with it
	Properties in the LA area get torn down all the tim
	The Zimmerman House was located in Brentwood
	Like it or not, they paid for that property. It's their
	I would have torn it down, too. They paid for it, th

Scraped_comments | Labeled_comments

Spam	Humor	Abusive	Consolida	Ideologica	Neutral
Get rich q	I told my v	You're wo	Climate ch	Socialism i	Milkmen are a relic of the past fondly r
Lose 20 pc	Why don't	I hope you	Education	Capitalism	I'm happy his family finally got resoluti
Unlock ex	I'm readin	You're so	Access to	Feminism	... but we know there’s another person
Congratul	I used to p	Nobody ca	Income in	Conservat	He survives WW2 only to be murdered
Increase y	Why did tl	You're a w	Racial disc	Libertaria	I understand this case is decades old.b
Make mor	I'm on a w	Go jump c	Freedom (	Anarchism	Neutral
Looking fo	I'm trying	I wish you	Technolog	Environm	Just make sure when you say "no crim
Meet hot	Why don't	You're ugl	Democrac	Nationalis	Was Williams investigated between the
Boost you	I told my c	I hope you	Globalizat	Communi	I have fond memories of when I was a
Limited st	I'm readin	You're not	Mental he	Islamopho	dumpo will just make fun of the Vetera
Get a free	Why did tl	I hope you	Cultural di	Secularism	Better late than never!
Earn $100	I used to b	You're suc	The role o	Liberalism	Price of a first class stamp in 1968 was
Invest in c	Parallel lin	You don't	Art and cu	Populism	...and? It wasn’t solved now. Snitching
Need a loa	I'm writing	I wish you	Corporate	Multicultu	Dateline NBC time.
Upgrade t	Why don't	You're a w	Technolog	Globalism	No crime goes unpunished? Didn't the
Looking fo	I told my v	I hate eve	Investmer	Humanism	Conveniently leave out any photos of c

Model Description

- Comparison of state-of-the-art models: GPT-3.5, Google Gemini, LLAMA2 and LSTM.
- Consideration of model performance metrics including precision, recall, F1 score, and accuracy.
- Evaluation of Model considering the best accuracy from all 4 techniques.



GPT-3.5

- Bidirectional Context Awareness
- Versatility
- Large-Scale Language Understanding

```
def classify_text(text):
    prompt = f"Classify the text: '{text}' into one : {', '.join(categories)}"
    response = openai.ChatCompletion.create(
        model="gpt-3.5-turbo",  # Specify the GPT-3 model as the model
        messages=[{"role": "user", "content": prompt}],
        max_tokens=50  # Adjust as needed
    )
    generated_text = response.choices[0].message.content.strip()
    return generated_text

def extract_score(response):
    # Use regular expression to extract the probability value
    match = re.search(r'Probability: (\d+\.\d+)', response)
    if match:
        return float(match.group(1))
    else:
        return 0.0  # Default score if not found
```

Llama 2

- Transformer Architecture
- Semantic Coherence
- Long-Range Dependencies

```
output = ""
for event in replicate.stream("meta/llama-2-70b-chat", input=
{"system_prompt": "Predict which category does the given comment falls into
Humor or Consolidating or Abusive, Just output the category",
"prompt": "Why don't scientists trust stairs? Because they're always up to something."}):
    output += str(event)

print(output)
```

```
output = replicate.run(
  "stability-ai/sdxl:39ed52f2a78e934b3ba6e2a89f5b1c712de7dfea535525255b1aa35c5565e08b",
  input={
        "system_prompt": "Learn from the given comments. Just learn from them",

        "prompt": """Humor: I told my wife she should embrace her mistakes... She gave me a hug.
                     Humor: Why don't skeletons fight each other? They don't have the guts.
                    Humor: I'm reading a book on anti-gravity. It's impossible to put down!

        Abusive: You're worthless, and nobody likes you.
        Abusive: I hope you fail at everything you do.
        Abusive: You're so stupid, it's embarrassing.

        Consolidating: Climate change is a pressing issue that requires global cooperation to address.
        Consolidating: Education is crucial for societal development and economic progress.
        Consolidating: Access to healthcare should be a fundamental right for all individuals."""
    },
)

print(output)
```

Google Gemini

- Graph Neural Networks
- Comprehensive Understanding
- Semantic Association Extraction

```
def to_markdown(text):
  text = text.replace('•', '  *')
  return Markdown(textwrap.indent(text, '> ', predicate=lambda _: True))

genai.configure(api_key="Enter your API Key")
model = genai.GenerativeModel('gemini-pro')
```

```
def main():
    with Gradient() as gradient:
        base_model = gradient.get_base_model(base_model_slug="nous-hermes2")
        new_model_adapter = base_model.create_model_adapter(
            name="test model 3"
        )
        print(f"Created model adapter with id {new_model_adapter.id}")

        # Define test data with comments and true labels
        test_data = [
            {"inputs": "### Instruction: Looking for a job? Check out our job
            {"inputs": "### Instruction: I told my wife she should embrace he
            {"inputs": "### Instruction: The education system should prioriti
            {"inputs": "### Instruction: Intersectionality recognizes the ove
            {"inputs": "### Instruction: Intersectionality recognizes the ove
            {"inputs": "### Instruction: I hope you die alone and unloved. \n
            # Add more test data as needed
        ]
```

LSTM

- Sequential Data Processing
- Context Understanding
- Long-Term Dependencies

```
def build_and_train_lstm_model(X_train, y_train, vocab_size, max_seq_length, num_classes, embedding_dim=100, lstm_units=128):
    model = Sequential()
    model.add(Embedding(input_dim=vocab_size, output_dim=embedding_dim, input_length=max_seq_length))
    model.add(LSTM(units=lstm_units))
    model.add(Dense(units=num_classes, activation='softmax'))

    model.compile(optimizer='adam', loss='sparse_categorical_crossentropy', metrics=['accuracy'])
    model.fit(X_train, y_train, epochs=10, batch_size=64)  # Removed validation_split parameter

    return model
```

```
sequences = tokenizer.texts_to_sequences(comments)
max_seq_length = max([len(seq) for seq in sequences])
padded_sequences = pad_sequences(sequences, maxlen=max_seq_length, padding='post')

    # Step 3: Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(padded_sequences, labels, test_size=0.2, random_state=42)

    # Convert X_train and y_train to numpy arrays
X_train = np.array(X_train)
y_train = np.array(y_train)
X_test = np.array(X_test)  # Convert X_test to numpy array
y_test = np.array(y_test)  # Convert y_test to numpy array

    # Step 4: Build and train the LSTM model
num_classes = len(categories)
print("Building and training the LSTM model...")
model = build_and_train_lstm_model(X_train, y_train, vocab_size, max_seq_length, num_classes)
```

Evaluation Metrics & Results

GPT 3.5 :

```
Accuracy: 0.16666666666666666
F1 Score: 0.1563087084393097
Precision  0.223160804775673 63
Recall 0.16666666666666666
```

Llama 2:

```
Accuracy: 0.6916666666666667
Precision: 0.6920803782505911
Recall: 0.6916666666666667
F1-score: 0.6618145318066764
```

Gemini:

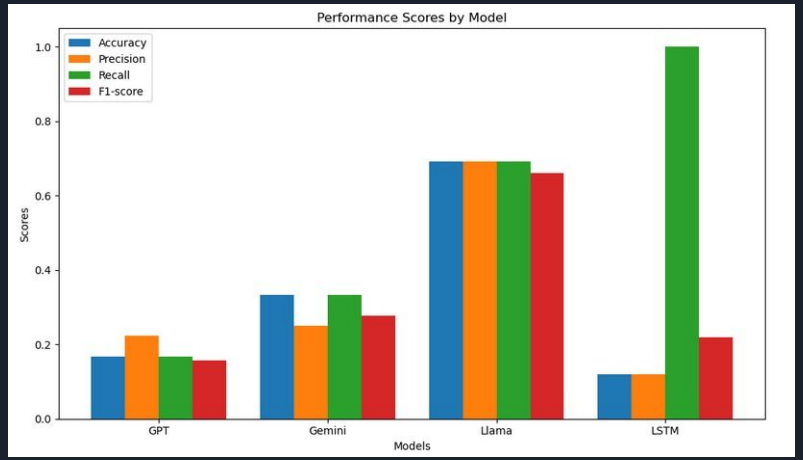
```
Accuracy: 0.3333333333333333
Precision: 0.25
Recall: 0.3333333333333333
F1 Score: 0.27777777777777773
```

LSTM:

	precision	recall	f1-score	support
Ideological	0.00	0.00	0.00	5
Humor	0.00	0.00	0.00	3
Consolidating	0.00	0.00	0.00	5
Abusive	0.00	0.00	0.00	5
Spam	0.12	1.00	0.22	3
Neutral	0.00	0.00	0.00	3
accuracy			0.12	24

Findings & Conclusions

- Llama excels with high accuracy, precision, and F1-score.
- LSTM achieves perfect recall but suffers from low precision.
- Gemini demonstrates decent performance but falls short in precision.
- Overall, Llama emerges as the best-performing model, balancing precision and recall effectively.



Application & Future Work

- Potential applications of the developed comment classification model in automated content moderation.
- Exploration of future research directions, such as enhancing model interpretability and generalizability.
- Consideration of scalability and adaptability of the model to other online platforms beyond Yahoo News.
- Importance of ongoing research and development to address evolving challenges in online content moderation.
- we envision building a Chrome extension based on our implementation, allowing users to seamlessly integrate our comment categorization tool into their browsing experience

References

1. Ona De Gibert, Naiara Perez, Aitor Garc´ıa-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum.arXiv preprint arXiv:1809.04444, 2018.
2. Ahlam Alrehili, Automatic Hate Speech Detection on Social Media: A Brief Survey, https://ieeexplore.ieee.org/document/9035228
3. Anna Wolters, Kilian Müller, Dennis M. Riehle Incremental Machine Learning for Text Classification in Comment Moderation Systems, https://link.springer.com/chapter/10.1007/978-3-031-18253-2_10
4. Ms. Shivani Kadam, Ms. Komal Ghatage, Mr. Aadesh Chaugule, Prof. J. W. Bakal Comment Toxicity Tracker Using NLP with Emphasis on Machine Learning Algorithms, https://www.ijraset.com/best-journal/comment-toxicity-tracker-using-nlp-with-emphasis-on-machine-learning-algorithms
5. D Manikalyan, TOXIC COMMENT CLASSIFICATION USING BI-LSTM, ttps://ijsrem.com/download/toxic-comment-classification-using-bi-lstm/
6. David Allenotor, David Oyemade, A Classification Model Based on Machine Learning for Detecting Racist Comments on Social Media Platforms, https://www.isteams.net/_files/ugd/185b0a_a1fe7284a51e4453a94f1b5f8b5c1179.pdf
7. R. Rooba A.R.Karthekeyan, Youtube Comment Feature Selection And Classification Using Fused Machine Learning, https://www.propulsiontechjournal.com/index.php/journal/article/view/982
8. Julian Risch and Ralf Krestel. Toxic comment detection in online discussions. Deep learning-based approaches for sentiment analysis, pages 85–109, 2020

Thank You