Classifying Constructiveness in an Unfiltered Environment

Classification of Constructiveness in Social Media Comments Using Large Language Models and Machine Learning

Member Names: Oscar Wang, Cherish Chen, Shinji Yamashita , Michelle Wu

Team Introduction

Our capstone project team consists of four students—Oscar Wang, Cherish Chen, Shinji Yamashita , Michelle Wu—from the Information School at the University of Washington. We all share a passion for data science and analytics and strive to use this opportunity to deepen our understanding and apply our skills in machine learning, natural language processing, and research methods. Together, we collaborate closely leverage our diverse backgrounds and expertise in order to explore and develop innovative solutions for classifying the constructiveness of social media comments.

Problem Context

In the vast and dynamic landscape of online platforms like Reddit, user interactions vary significantly in quality and constructiveness. While some comments provide valuable insights and foster positive discourse, many others fall short, contributing to unproductive or even toxic exchanges. This variability poses a challenge for maintaining a constructive online environment, essential for meaningful knowledge exchange and community engagement. Previous studies have primarily focused on controlled environments such as news websites, but the unfiltered nature of social media presents a broader context for analyzing constructiveness. Leveraging principles of constructive feedback, foundational knowledge in text analysis, and advanced AI techniques like GPT-3.5, our research aims to develop a robust system for evaluating comment constructiveness across diverse social media platforms.

Problem Statement

Our project seeks to address the challenge of automatically identifying constructive comments on Reddit to enhance user experience and assist moderation efforts. By combining manual annotations based on educational feedback theories with automated annotations using advanced AI, we aim to create a robust model that effectively evaluates comment constructiveness, promoting meaningful engagements and fostering a positive online community.

Key Research Insights

Effectiveness of Hybrid Annotation

Manual and Automated Annotation: Combining manual annotations influenced by educational feedback theories and automated annotations using large language models (LLMs) proved effective. This hybrid approach balances human insight with AI efficiency, leading to more nuanced understanding and classification of constructiveness in social media comments.

Model Performance

GPT-3.5 Turbo and Random Forest Classifier: Initial testing with models like GPT-3.5 Turbo and the Random Forest classifier showed promising results in identifying constructive comments. These tools effectively handle large datasets, providing scalable solutions for real-time moderation and feedback systems.

Importance of Detailed Analysis

Feature Significance: The analysis highlighted the significance of various features, such as comment length, readability, and presence of evidence, in contributing to the constructiveness of comments. These findings can guide the development of more effective content moderation strategies and community guidelines.

Key Research Insights

Feature Correlation Analysis

Structured Comments and Constructiveness: There is a moderate correlation between constructiveness and paragraph count (0.38), suggesting that clear and organized comments are perceived as more constructive.

Credibility through Detail: Comments that include specific references or factual data are seen as more constructive, indicating the importance of detail and evidence in enhancing the perceived quality of comments.

Personal Insights: Personal insights and opinions add perceived value to discussions, suggesting that sharing personal experiences can boost engagement and constructiveness.

Principal Component Analysis (PCA) and Entropy Weight Method (EWM)

Weighting Criteria: Combining PCA and EWM provided a robust and objective evaluation of the criteria's relative importance. This integrated approach ensured a balanced consideration of both dispersion and structural importance, leading to more reliable weighting results.

Significance of Evidence: Both PCA and EWM analyses indicated that 'Evidence' is the most influential criterion, reinforcing its significant role in the dataset.

Key Research Insights

Effectiveness of Hybrid Annotation

Manual and Automated Annotation: Combining manual annotations influenced by educational feedback theories and automated annotations using large language models (LLMs) proved effective. This hybrid approach balances human insight with AI efficiency, leading to more nuanced understanding and classification of constructiveness in social media comments.

Model Performance

GPT-3.5 Turbo and Random Forest Classifier: Initial testing with models like GPT-3.5 Turbo and the Random Forest classifier showed promising results in identifying constructive comments. These tools effectively handle large datasets, providing scalable solutions for real-time moderation and feedback systems.

Importance of Detailed Analysis

Feature Significance: The analysis highlighted the significance of various features, such as comment length, readability, and presence of evidence, in contributing to the constructiveness of comments. These findings can guide the development of more effective content moderation strategies and community guidelines.

Ethical Considerations

1. **User Data Handling and Privacy**
    - To ensure that users' privacy is protected and the data is anonymized, it is crucial that the collection and analysis of Reddit comments—which involves the handling of user-generated content—adheres to the platform's terms of service and considers the ethical implications of using publicly available data without direct consent.
    - Prior to collecting our data using the Reddit API and the Python Reddit API Wrapper (PRAW), we investigated the tools being used and confirmed that they internally follow all of Reddit's API rules and complies with its terms and conditions.
1. **Algorithmic Bias and Representation**
    - Machine learning models can inherit biases present in the training data, so it is important to actively identify and mitigate biases related to certain attributes to ensure the model's fairness and accuracy.
    - To avoid bias in model training and evaluation, we compiled a diverse dataset of different sources (subreddits) and manually annotated 500 comments.
1. **Impact on Users**
    - The automated classification of comments as constructive or non-constructive could influence moderation practices on social media platforms. Hence, it's important to consider the implications of such automated systems on freedom of expression and the potential for unjust censorship.
    - Such novel algorithm-based approaches may cause users to feel surveilled or judged, which could impact their willingness to participate freely in discussions. To mitigate this feeling and maintain user trust, we strive to sustain transparency regarding how our model works and how decisions are made.

Ethical Considerations

4. Model Transparency and Accountability

- Ensuring that the decision-making process of the machine learning and large language models are interpretable means that users and moderators should be able to understand how and why a particular comment was classified as constructive or non-constructive.
- By establishing clear lines of accountability for the outcomes of the automated system and implementing mechanisms to address and correct errors, we prioritize algorithmic accountability and model transparency.

4. Research Integrity

- In order to avoid unintended consequences, models should be rigorously tested and validated in different scenarios and backgrounds. The research conducted must also adhere to emerging standards and guidelines for ethical AI.
- To preserve the integrity of our research, we carefully monitor our model's performance in various scenarios and use different validation techniques such as cross-validation. We also engage with peer review and seek ethical oversight from our mentors to ensure the research adheres to ethical standards.

4. Ethical Use of AI

- It's important to reflect on the broader societal implications and potential consequences of using AI to enhance the quality of online discussions. Both the positive impacts and potential negative consequences must be considered and compliance with ethical AI standards is fundamental.
- Our project aims to develop a responsible, fair, and transparent system that enhances online discussions while mitigating potential risks and negative consequences through the aforementioned ethical considerations. We are constantly researching and evaluating the system's impact, staying informed about emerging ethical standards and guidelines, and engaging with the community to stay vigilant and ensure that our project remains aligned with broader societal expectations.

Next Steps Beyond Capstone

- **Enhancing Model Performance**
  - **Fine-Tuning Models**: Continue refining machine learning models, to improve accuracy in classifying constructiveness. This includes experimenting with different architectures and fine-tuning on diverse datasets to capture nuanced patterns in comments.
  - **Reducing False Positives**: Focus on minimizing false positives to ensure that non-constructive comments are not misclassified as constructive. This involves refining feature extraction processes and enhancing the sensitivity of the model to subtle indicators of constructiveness.

- **Expanding Data Collection**
  - **Increasing Data Volume**: Collect a larger and more diverse dataset to train the models. This will help improve the robustness and generalizability of the models across different types of comments and platforms.
  - **Diversifying Data Sources**: Extend data collection beyond Reddit to include other social media platforms like Twitter, Facebook, and YouTube. This will provide a broader spectrum of interaction styles and content types, enriching the training data.
  - **Incorporating Multilingual Data**: Collect and analyze comments in multiple languages to create a multilingual model capable of assessing constructiveness across different linguistic contexts. This will enhance the model's applicability globally.