DSC260: Data Science, Ethics, & Society Fall 2024

In-quarter Project #2: Data Deployment, Use, & Revision

Title: **Enhancing Airline Customer Service with Ethical Toxic Speech Detection**

Aakash Agrawal	Ananay Gupta	Sarthak Kala
A69034394	A69033310	A69034307

Executive Summary

Airline customer service agents often face high-stress situations, such as handling toxic calls related to lost baggage, flight cancellations, etc. To address this, our team is developing a toxic speech detection model that not only supports agents by flagging toxic interactions but also enhances their well-being by enabling timely measures such as mental health breaks. Additionally, the model extracts valuable insights to address systemic issues, enhance products and services, and implement targeted customer retention strategies, ultimately driving better customer satisfaction and loyalty.

Our data collection effort involves obtaining informed consent from users before recording customer interactions. To ensure privacy, we take strict measures to remove all Personally Identifiable Information (PII). We define toxicity in collaboration with compliance and legal teams to ensure a clear and consistent understanding. We use state-of-the-art machine learning models for toxicity classification, aiming for both high accuracy and responsible use. It's important to note that the model has its boundaries—it works only in English and should never be the sole decision-maker for high-stakes matters like employee evaluations and taking legal actions.

For this application use, we have defined toxic speech as primarily verbal abuse and harassment. It does not include frustration, irritation, or any other valid negative emotions. The algorithm's outputs include confidence intervals and must be interpreted contextually to ensure fairness. We also leverage interpretability and explainability tools to enhance transparency and trust by highlighting the words that most influence the toxicity classification. Our data retention strategy balances operational needs, privacy, bidirectional educated consent, compliance with anonymization, defined retention periods (e.g., 1-year), access controls and automated deletions, etc. Our data retention practices align with regulations like CCPA[1] and GDPR[2]. We are also aggregating insights before any deletion, to harness the value of the data while simultaneously minimizing compliance risks.

Additionally, our data science effort may face scenarios where revising the AI system becomes inevitable for ethical reasons. These include adapting to seasonal spikes in customer sentiment during festive campaigns and technical glitches, adjusting to evolving definitions of toxicity, discovering new trends of gamification by abusive customers, and addressing biases that affect underrepresented groups. Continuous model monitoring is essential to ensure that the system adapts to these challenges and remains aligned with ethical standards. Finally, we recognize that as our effort scales, it brings several societal-level ethical concerns to the forefront. This includes privileged access while deploying the AI system, data sovereignty, forced participation, regularization of behavior, challenges with navigating data regulation requirements across different regions, etc.

[1] link: https://oag.ca.gov/privacy/ccpa/regs
[2] link: https://gdpr-info.eu/

A. REAL-WORLD CHALLENGES & MODEL USEFULNESS

In the airline industry, customer support agents frequently handle emotionally charged calls related to issues such as lost baggage, missed connections, and flight cancellations. These interactions often involve toxic speech, which can severely impact agents' mental health, leading to stress, burnout, and reduced performance. This not only affects employee well-being but also compromises customer service quality, creating a cycle of dissatisfaction that harms both customers and the company. Additionally, toxic calls provide valuable feedback about customer frustrations and recurring service issues. Without a structured analysis, airlines miss critical insights that could improve their services and policies.

As the data science team at a mid-sized airline company operational only on the West Coast of the US, we are developing a toxic speech detection model to address the impact of high-stress customer calls. Our model aims to support agents by identifying toxic interactions, escalating critical issues, and providing valuable insights to improve service quality, targeted marketing and reduce burnout. This helps enhance both employee well-being and customer experience.

Ways the model enhances justice and equity:

1. **Protecting Employee Well-being**: The model identifies agents handling high volumes of toxic interactions, allowing managers to offer timely support such as mental health breaks or counseling. This ensures that emotional strain is distributed fairly, creating a healthier work environment and enabling agents to perform at their best.

2. **Improving Conflict Resolution**: Toxic calls can be escalated to supervisors trained in conflict resolution. This protects frontline agents from undue stress and ensures customers receive empathetic, effective handling of their grievances, promoting fairness and respect on both sides.

3. **Identifying Systemic Issues**: By analyzing patterns in toxic calls, airlines can uncover recurring service problems—such as baggage mishandling or poor communication—that need addressing. Resolving these issues enhances procedural justice by creating fairer, more transparent processes.

4. **Targeted Customer Support**: The model segments customers based on sentiment, enabling personalized follow-ups or special offers for those who've had negative experiences. This ensures vulnerable customers receive appropriate attention, fostering stronger, more equitable customer relationships.

5. **Balancing Stakeholder Values**: The system minimizes false positives and false negatives to balance the well-being of both agents and customers, ensuring fair and transparent interventions that respect the values of all stakeholders.

B. DATA & MODEL BACKGROUND

First, we will define "toxic" speech in collaboration with legal and compliance teams to establish clear, objective, and culturally sensitive criteria. Toxicity will be characterized by abusive, threatening, or demeaning language, excluding sarcasm and expressions of anger or emotional frustration.

B. 1. Data Collection

The dataset consists of customer support call recordings from an airline's customer service center. Each call is transcribed using speech-to-text models. Educated and bidirectional consent will be obtained from customers, ensuring they understand how their data will be used for model training purposes. To ensure privacy, all Personally Identifiable Information[3] (PII) is removed, and each user is associated with a unique, anonymized identifier. This allows the data science team to analyze customer interactions without compromising user privacy.

B. 2. Data Preprocessing

- **Data Anonymization**: Sensitive information is scrubbed from call transcripts.
- **Exploratory Data Analysis**: We analyze token length distributions to determine optimal padding and truncation strategies for handling sequence limits in transformer[4] models. It also includes examining class imbalances in the dataset.
- **Feature Extraction**: Text embeddings are generated using transformer-based embeddings like BERT.

B. 3. Data labeling for model training

We will outsource the task to a diverse group of annotators who will label the data according to the toxicity definitions established by our compliance team. In this manner, we ensure that personal biases cancel out each other.

B. 4. Model

We are using **BERT**[5] (Bidirectional Encoder Representations from Transformers) for toxic speech classification. BERT is a state-of-the-art language model that captures semantic relationships within text, making it highly effective in identifying toxicity in customer interactions. By leveraging BERT's pre-trained embeddings, the model can accurately detect forms of toxic speech across various customer service scenarios. We further fine-tune the model by leveraging historical call-recording data

C. APPROPRIATE AND INAPPROPRIATE MODEL USE CASES

Understanding the appropriate and inappropriate contexts helps prevent misuse and ensures the model's outputs are interpreted correctly.

Appropriate Use Cases:

1. **Real-time support for agents**: The model can assist customer service agents by flagging potentially toxic speech in real time, enabling them to take appropriate de-escalation measures.
2. **Data-driven insights**: It can be used to analyze trends and patterns in customer interactions, providing insights into recurring issues or negative sentiment.
3. **Targeted Retention Strategies**: The model helps in identifying dissatisfied customers for personalized retention offers and follow-up actions, improving loyalty.

[3] link: https://www.ibm.com/topics/pii
[4] link: https://arxiv.org/abs/1706.03762
[5] link: https://arxiv.org/abs/1810.04805

Inappropriate Use Cases:

1. **Multilingual calls**: The model has been trained exclusively on an English dataset, so it is designed to work only for conversations conducted in English.

2. **Absence of clear toxicity definition**: Without a precise and culturally nuanced definition of toxicity, the model may misclassify speech.

3. **High-stakes decisions**: The model lacks the capability to fully understand cultural nuances, sarcasm, or indirect language. The model will be a supplementary tool for identifying potential toxic interactions, not the sole basis for high-stake decisions such as customer relationship terminations or legal actions.

4. **Evaluating Employees**: There is a risk of toxic speech from customers being wrongly attributed to an employee's responses. Using this system to evaluate employees could lead to unfair judgments and harm morale.

D. CONTEXT-SENSITIVE INTERPRETATION OF ALGORITHM OUTPUTS

To effectively utilize a toxic speech detection algorithm in customer service, outputs are interpreted in ways that provide actionable insights while ensuring fairness and accuracy. A detailed interpretation framework is outlined below:

1. **Confidence Scores and Thresholds**:The algorithm typically assigns a confidence score to its classifications, such as "toxic" (80%) or "non-toxic" (95%). These scores represent the probability that a message belongs to a specific category.

2. **Context Sensitivity**: Contextual understanding is critical for nuanced cases. For instance, sarcasm or cultural idioms may be misclassified as "toxic". Providing the surrounding conversational context helps staff assess whether the algorithm's output aligns with the interaction's tone.

3. **Conversation History**: A single toxic message may not signify a trend. Aggregating scores across a conversation or series of interactions reveals deeper patterns.

4. **Cultural or Linguistic Nuances and Diversity**: Words perceived as harsh in one language may be neutral in another. Algorithms trained on diverse datasets capture these subtleties. Also, the speech-to-text software may struggle to identify and correctly transcribe some accents properly leading to inaccuracies.

5. **Explainability**: Tools like SHAP (Shapley Additive Explanations)[6] and LIME (Local Interpretable Model-agnostic Explanations)[7] are integrated to identify keywords affecting toxic classification for each classification, thus improving transparency.

6. **Real-Time Insights and Trends**: Real-time outputs enable proactive responses, such as defusing escalating situations before they worsen. Aggregated trends from outputs also offer long-term insights for operational teams.

[6] link: https://shap.readthedocs.io/en/latest/
[7] link: https://arxiv.org/abs/1602.04938

7. **Human Oversight and Feedback Loops**: Human Oversight and Feedback Loops: To check for function drift, we take random samples from our real-time toxicity classifier to check its performance with several diverse focus groups periodically. Human-in-the-loop systems are also vital for validating ambiguous cases. Algorithms should assist, not replace, human judgment.

8. **Ethical Considerations**: To avoid harm, we ensure that outputs avoid bias against any group. Regular audits of the algorithm's performance are done to ensure that protected categories are not disproportionately flagged as toxic.

E. DEVELOPING, JUSTIFYING ETHICAL DATA RETENTION PLAN

An ethical data retention plan for a toxic speech detection system must prioritize customer privacy and comply with regulatory and operational needs. The following framework ensures data handling aligns with these goals:

1. **Purpose-Driven Retention**: Data is retained only to meet specific, defined purposes, such as improving the algorithm, analyzing trends, compliance reasons or providing actionable insights. Once the purpose is fulfilled, data is deleted or access restricted.

2. **Legal and Ethical Compliance**: For customer service data of airlines, the US Department of Transportation[8] compliance mandates data storage for 3 years. Our retention policy also aligns with leading data protection laws, such as GDPR and CCPA, to avoid penalties and reputational damage.

3. **Processing and Anonymization**: Wherever possible, we have prioritized processing to derive aggregated insights without retaining raw data. When retention is necessary, we anonymize all personally identifiable information (PII) at ingestion with pseudonyms or irreversible hashes. This approach ensures that even if data is breached, it cannot be traced back to individuals.

4. **Retention Period**: A 1-year retention window suffices for operational and training needs. Data older than 1 year will have restricted access, and any data beyond the compliance period will be automatically deleted.

5. **Access Control and Security**: Retained data is accessible only to authorized personnel like data scientists. Use role-based access controls and encryption to protect sensitive information. Regular audits ensure compliance with security protocols.

6. **Transparency and Consent**: Through a bidirectional educated consent process, we have informed customers about data collection and retention policies through clear, accessible privacy notices.

7. **Bias and Fairness in Retention**: We have avoided retaining toxic interactions only from specific regions or demographics so as not to introduce any biases in training data.

8. **Data colonialism and sovereignty**: To tackle issues with data colonialism and sovereignty, data will be stored in the country of origin of the customer.

[8] link: https://www.ecfr.gov/current/title-14/chapter-II/subchapter-A/part-259

F. MODEL REVISION and MONITORING JUSTIFICATION

There are many plausible scenarios where model revision or re-training is inevitable for ethical reasons.

F. 1. Platform Gamification/Abuse of Toxic Speech to Gain Advantage:

Some users may try to game the system for personal gains. For example: they may intentionally use abusive or toxic speech towards agents to trigger an escalation, hoping to gain advantages such as free services, upgrades, or special treatment. This kind of behavior could be seen in competitive environments, where users exploit the system's mechanisms for their benefit.

It would be unethical to automatically label these instances as toxic without a deeper analysis. It's important to differentiate between genuine toxicity and intentional abuse meant to manipulate the system. If the model fails to detect the game nature of these interactions, it could allow some users to exploit the system while punishing agents and other users unfairly.

Hence, revision of the system is required to discover new trends of gamification on a periodic basis. The model can take into account the number of times a user interacts with the agent, whether the user repeatedly uses toxic words in specific contexts, etc, indicating manipulation rather than genuine toxicity. Here, analyzing and monitoring model predictions can help gain insights into repeated abusage patterns from the same customer across sessions, which can help identify potential patterns of gamification.

F. 2. Context-Driven Sentiment Shifts:

During periods of high traffic, such as festive occasions, holidays, promotional campaigns, etc, or even scenarios like technical glitches, customer sentiment can be dramatically affected. This can lead to a surge in customer-care interactions stemming from frustration due to long wait times, technical glitches during flight booking, confusion over special offers, issues with their redemption, etc. The heightened frustration and impatience among users can naturally lead to the usage of an aggressive tone during customer interactions when they feel they're not getting the fair deal they deserve. This datashift presents many ethical challenges:

1. **Unfair labeling:** In this case, the model might be too sensitive and trigger false positives. It might be unfair to label the heightened tone of emotionally charged users who are simply reacting to temporary disruptions, as toxic speech, without recognizing the context.

2. **Inefficiency of escalations:** If the model escalates these cases without distinguishing between genuine abuse and context-driven frustration, it can lead to an inefficient allocation of resources. Given that many of these interactions may stem from similar issues (e.g., delayed services or confusing promotions), it's unlikely that escalations will resolve the problem faster than addressing it at the initial level of customer service.

Hence, the model should be retrained periodically to take these contexts into account, using data from specific events to help it adapt to these situations. Here, ongoing prediction monitoring is crucial to identify shifts in model performance during high-traffic periods. By tracking the frequency of flagged interactions, we can gain insights if the model is over-estimating toxicity.

F. 3. Evolution of the definition of toxicity:

Toxicity is a slightly subjective argument and its definitions can evolve over time. For example: something considered toxic by Gen-Zs might not be a toxic entity for the millennials, and vice-versa. A term once considered harmless may become offensive due to cultural shifts. This again presents several ethical concerns:

1. **False Negatives:** The model might fail to detect new or adapted toxic language, exposing agents to harmful speech.
2. **False Positives:** The model might inappropriately flag harmless language as toxic due to outdated definitions, leading to wrong escalations.

Hence, retraining of the model is required on a periodic basis to ensure the model adapts to the prevalent notions of toxicity. Here, monitoring can help refresh and annotate a portion of the dataset periodically to capture shifts in language use and toxicity norms.

F. 4. Evidence of Systemic Bias:

Since the concept of toxicity is culturally nuanced, the model may disproportionately flag toxic speech from certain demographic groups due to variations in their linguistic patterns. This could indicate systemic bias in the training data or algorithm. For example, non-native speakers might be flagged more frequently because their speech patterns are misinterpreted. This incurs ethical concerns like:

1. **Exacertabing Inequalities:** Marginalized groups could be unfairly flagged, leading to unequal treatment.
2. **Legal Implications:** Discriminatory behavior could violate anti-discrimination laws and harm the company's reputation.

Hence, revising the model is crucial to remove biases and ensure fairness and equity. Here, monitoring flagged cases can help gain insights into the under-representation of specific groups.

G. SOCIETAL LEVEL ETHICAL AND TECHNICAL ISSUES

G. 1. Privileged Access:

As we aim to deploy our data science effort only across Anglo-sphere countries, there is a risk of creating a polarized distribution of benefits, thereby exacerbating inequalities and technological disparities among non-English-speaking populations. This can also perpetuate systematic biases in AI systems, as the model's success criteria would be tailored to English-speaking users.

G. 2. Bias and Discrimination:

Even though our efforts are deployed in English-speaking countries, the definition of what constitutes toxicity may vary across regions and populations. Since the idea of toxicity is culturally nuanced, the definition of "toxic speech" may also vary significantly across cultures, languages, and contexts. What is deemed offensive or toxic in one culture might be normal in another, leading to misclassification and inadvertent harm. This incurs an ethical concern as the model may unintentionally flag certain demographic groups more frequently (e.g., specific ethnicities, accents, genders), leading to discrimination and unequal treatment.

G. 3. Inevitable False Positives and Negatives:

Issues like false negatives and false positives will always be inevitable, meaning that as our data science effort scales, there will be cases when either the customer or the agent is wrongly treated. These errors can erode trust, as affected individuals, including employees (agents), may feel unfairly judged, and the public may view the company with heightened scrutiny. Such incidents could damage the company's reputation and undermine confidence in the system's fairness.

G. 4. Data Colonialism and Sovereignty:

1. **Consent & Ownership:** As the data science effort gets scaled across countries, issues surrounding data sovereignty may arise. Countries often mandate that sensitive user data be stored and processed on servers located within their national borders to protect their citizens' privacy and maintain control over their digital ecosystems. This could lead to conflicts over how data is accessed, used, and shared across borders. Centralizing or mishandling such data can lead to perceptions of exploitation.

2. **Forced participation:** As our data science initiative expands, the concept of forced participation also becomes a significant ethical concern. Data Colonialism also manifests from the idea that we are forcing people to become data and algorithm subjects during user-agent interaction. While the benefits of improved service and employee well-being may justify minor infringements, we must exercise caution to avoid crossing ethical boundaries.

G. 5. Regularization of Behavior:

As our data science effort scales "**society-wide**", the widespread adoption of toxic speech detection systems could lead to users modifying their natural communication to avoid being flagged. This "uniformity of behavior" might reduce the authenticity of interactions and suppress free expression, making it more calculated. Over time, people may feel constrained in raising valid concerns, fearing scrutiny, thus impacting their ability to communicate grievances effectively. However, this could also foster a more civil and respectful society by encouraging thoughtful and constructive communication.

G. 6. Navigating Data Regulation Challenges:

As our effort expands across regions with varying data protection laws (e.g., GDPR in Europe, CCPA in California, etc.), it's crucial to ensure that customer data collection, processing, and storage practices comply with these local and international standards. Failure to adhere to these regulations could result in significant legal repercussions, loss of customer trust, and reputational damage.

References:

1. David J. Hand, "Aspects of Data Ethics in a Changing World: Where Are We Now?" https://www.liebertpub.com/doi/10.1089/big.2018.0083
2. Butler et al., "Beyond Measure" https://ephemerajournal.org/content/beyond-measure
3. Stanford Encyclopedia of Philosophy, "Privacy." https://plato.stanford.edu/entries/privacy/
4. Acquisti et al., "Privacy and Human Behavior in the Age of Information." https://www.cmu.edu/dietrich/sds/docs/loewenstein/PrivacyHumanBeh.pdf
5. Zuboff et al., "You Are the Object of a Secret Extraction Operation." https://www.nytimes.com/2021/11/12/opinion/facebook-privacy.html
6. Mittelstadt et al., "The Ethics of Algorithms: Mapping the Debate." https://journals.sagepub.com/doi/10.1177/2053951716679679
7. Fazelpour et al., "Algorithmic Bias: Senses, Sources, Solutions." https://compass.onlinelibrary.wiley.com/doi/10.1111/phc3.12760
8. Lipton et al., "The Mythos of Model Interpretability." https://dl.acm.org/doi/pdf/10.1145/3236386.3241340
9. Barocas et al., "Fairness and Machine Learning: Limitations and Opportunities." https://fairmlbook.org/pdf/fairmlbook.pdf
10. Carroll et al., "Indigenous Data Governance: Strategies from United States Native Nations." https://datascience.codata.org/articles/10.5334/dsj-2019-031
11. Schneider et al., "Digital Nudging: Guiding Online User Choices through Interface Design." https://dl.acm.org/doi/pdf/10.1145/3213765
12. Raji et al., "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." https://arxiv.org/pdf/2001.00973
13. Parasidis et al., "A Belmont Report for Health Data." https://uh.edu/ethicsinscience/Media/BelmontReportHealthData.pdf
14. Elish et al., "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." https://estsjournal.org/index.php/ests/article/view/260/177
15. Sloane et al., "Inequality Is the Name of the Game: Thoughts on the Emerging Field of Technology, Ethics and Social Justice." https://www.ssoar.info/ssoar/bitstream/handle/document/62583/ssoar-2019-sloane-Inequality_Is_the_Name_of.pdf