Offensive Language Classification

 # Offensive Language Classification Guidelines
Classify text into three categories based on intent, context, and impact: **"Offensive"**, **"Implicit Offensiveness"**, or **"Non-Offensive"**. Consider cultural context, tone, and intended meaning when making classifications.
Classification Categories
1. Non-Offensive
Definition: Language that maintains respect and social harmony, lacking harmful intent or negative impact.
Key Characteristics:
– Neutral or positive statements – Factual observations – Constructive criticism (when appropriate)
– Professional communication – Personal expressions of emotion or opinion without targeting others
Contextual Considerations:
– News reporting of negative events (when objective) – Professional disagreement (when respectfully expressed)
– Cultural references (when not stereotyping) – Academic or educational discussions
Examples:
אני מאחל לך יום נפלא! I wish you a wonderful day!
האינטרנט שינה את הדרך שבה אנחנו מתקשרים The internet changed how we communicate
אני לא מסכים עם הדעה שלך, אבל אני מכבד אותה I disagree with your opinion, but I respect it
2. Offensive
Definition: Language deliberately intended to harm, discriminate, or cause direct distress.
Key Characteristics:
– Explicit slurs or hate speech – Direct personal attacks – Discriminatory statements
– Dehumanizing language – Explicit threats or violent language – Mockery of protected characteristics
Common Categories:
1. Racial/Ethnic Discrimination 2. Religious Intolerance 3. Gender-based Discrimination
4. Disability-based Discrimination 5. Age-based Discrimination 6. Appearance-based Attacks
7. Socioeconomic Status Attacks
Examples:
– אתה טיפש ומכוער You're stupid and ugly
– כל ה [קבוצה אתנית] הם [סטריאוטיפ שלילי] All [ethnic group] are [negative stereotype]
– נשים לא מתאימות ל [תפקיד] Women aren't suited for [role]
3. Implicit Offensiveness
Definition: Language that appears neutral but carries subtle harmful intent or negative implications.
Key Characteristics:
– Passive-aggressive statements – Backhanded compliments – Coded language or dog whistles – Microaggressions
– Sarcasm with negative intent – Subtle stereotyping – Leading questions with negative implications
Common Patterns:
1. Contextual Triggers: – Ellipses (...) suggesting negative implications – Tone markers that shift meaning
– Cultural references used inappropriately
2. Structural: – False equivalencies – Loaded questions – Subtle stereotyping – "Just asking questions" format
Examples:
למה את תמיד כל כך...מיוחדת? Why are you always so... special?
מעניין שדווקא אתה קובלת את התפקיד... Interesting that you of all people got the position...
את די חכמה... בשביל בחורה You're pretty smart... for a girl
Classification Guidelines
1. Context Analysis: – Consider cultural context – Evaluate speaker intent – Account for power dynamics
– Assess historical context – Consider audience impact
2. Tone Evaluation: – Analyze word choice – Consider delivery method – Look for subtle markers
– Evaluate emotional impact
3. Impact Assessment: – Consider potential harm – Evaluate broader social implications
– Account for group dynamics – Assess perpetuation of stereotypes
Edge Cases and Special Considerations
1. Reclaimed Language: – Consider in-group usage – Evaluate context and speaker identity
– Account for cultural evolution
2. Educational Context: – Academic discussion of offensive terms – Historical documentation
– Anti-discrimination training
3. Artistic Expression: – Creative works – Social commentary – Satirical content
Classification Process
1. Initial Assessment: – Read/hear the complete statement – Note immediate reaction – Identify key terms/phrases
2. Contextual Analysis: – Consider speaker intent – Evaluate situation – Account for cultural factors
3. Final Classification: – Apply category criteria – Consider edge cases – Document reasoning *Remember*:
When in doubt, consider the potential impact on marginalized or vulnerable groups and err on the side of caution.
The text to be classified: `{text}`