Pedoman Anotasi untuk Multimodal Hate Speech Detection

Tujuan
Tujuan dari tugas anotasi ini adalah untuk mengklasifikasikan setiap tweet sebagai "Hate Speech" atau "Not Hate Speech." Sebuah tweet dianggap sebagai ujaran kebencian (hate speech) jika berisi teks dan/atau gambar yang menyerang, merendahkan, atau menyakiti individu atau kelompok tertentu.

Definisi Hate Speech
Hate speech mencakup setiap bahasa dan/atau gambar yang:
- Menyerang, merendahkan, atau menyakiti individu atau kelompok berdasarkan atribut seperti ras, agama, etnis, jenis kelamin, orientasi seksual, disabilitas, atau karakteristik lainnya.
- Menggunakan bahasa yang merendahkan, hinaan, stereotip ofensif, atau gambar yang penuh kebencian (termasuk pernyataan agresif yang disampaikan dengan nada sarkastik).
- Menghasut kekerasan atau permusuhan terhadap individu atau kelompok.

Kategori Anotasi
1. **Hate Speech (HS)**
2. **Not Hate Speech (NHS)**

Pedoman Umum
1. **Periksa Kedua Modalitas**: Pertimbangkan baik teks maupun gambar dalam setiap tweet.
2. **Konteks itu Penting**: Pertimbangkan konteks di mana tweet ditulis dan gambar digunakan. Cari makna implisit dan referensi.
3. **Konten Netral atau Ambigu**: Jika kontennya netral atau ambigu, klasifikasikan sebagai Bukan Ujaran Kebencian
4. **Satire dan Sarkasme**: Perhatikan satire dan sarkasme. Meskipun sebuah tweet tampak humoris, bisa tetap diklasifikasikan sebagai ujaran kebencian jika merendahkan atau menyerang seseorang.
5. **Ujaran Kebencian Langsung vs. Tidak Langsung**: Ujaran kebencian langsung mencakup serangan eksplisit, sementara ujaran kebencian tidak langsung bisa mencakup bahasa berkode, permusuhan implisit, atau gambar.

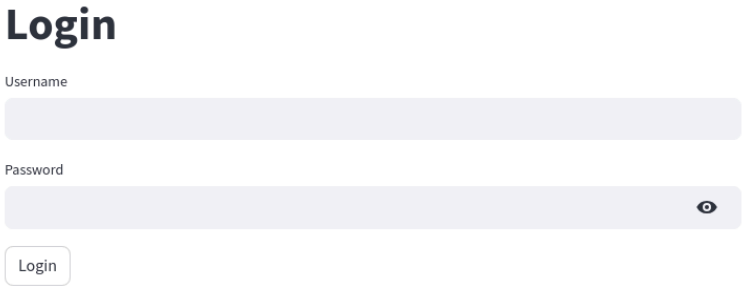
Pedoman Spesifik
1. **Identifikasi Target**
    - Pastikan apakah target yang diberikan sudah sesuai.
2. **Bahasa dan Nada**
    - Kata-kata yang merendahkan bisa menjadi tanda adanya ujaran kebencian, tetapi periksa konteksnya karena ada kemungkinan kata-kata tersebut digunakan dalam konteks yang tidak bermaksud menghina.
3. **Analisis Gambar**
    - Evaluasi gambar untuk simbol, isyarat, atau penggambaran yang penuh kebencian yang merendahkan atau menyerang individu atau kelompok.

- Perhatikan meme, kartun, atau gambar diedit yang mungkin menyampaikan permusuhan atau pesan merendahkan.
- Nilailah konteks di mana gambar digunakan bersama teks.

4. **Informasi Teks**
   - Catat bagian atau kata dari kalimat atau tweet yang mewakili emosi terkait ujaran kebencian. Bisa juga kata-kata makian.
   - Contoh:
     - “Saya **membenci** [kelompok tertentu]”
     - “[kelompok tertentu] memang yang **paling top**, **sampe ngga bisa bedain mana yang bener mana yang salah**”
5. **Contoh Hate Speech**
   - Teks dengan pernyataan eksplisit seperti "Saya benci dengan [kelompok tertentu]" atau "Semua [kelompok tertentu] harus [tindakan kekerasan]."
   - Gambar yang mengandung hinaan, simbol ofensif, atau penggambaran penuh kebencian.
   - Teks dan gambar yang digabungkan yang mempromosikan kekerasan atau permusuhan, misalnya, "Kita harus memusnahkan semua [kelompok tertentu]" dengan gambar tindakan kekerasan.
   - Lelucon atau meme merendahkan yang menargetkan kelompok tertentu.
6. **Contoh Not Hate Speech**
   - Kritik tanpa bahasa merendahkan atau gambar yang bermusuhan, misalnya, "Saya tidak setuju dengan keyakinan [kelompok tertentu]."
   - Pernyataan netral tentang kelompok dengan gambar yang tidak ofensif, misalnya, "Saat ini banyak sekali [kelompok tertentu] yang hidup di daerah ini" dengan foto netral.
   - Diskusi umum tentang topik kontroversial tanpa bahasa atau gambar yang bermusuhan.

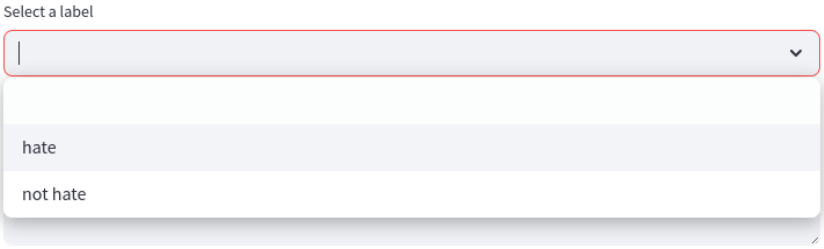
Langkah-langkah Anotasi

1. **Log in dan Akses Alat Anotasi**
   Buka alat anotasi dengan menavigasi ke platform yang ditunjuk tempat tweet di-hosting. Masuk ke platform menggunakan username dan password Anda.

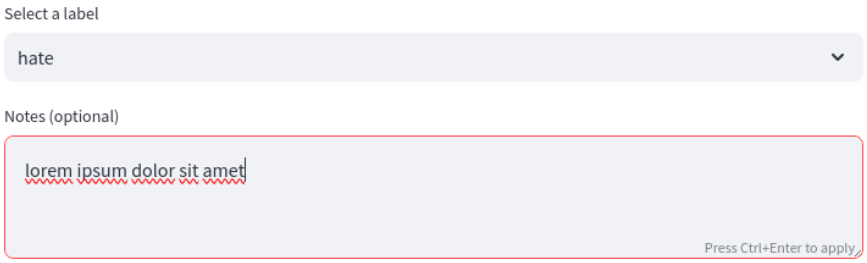


2. **Baca Setiap Tweet dan Periksa Gambar**
   Setelah berhasil log in, Anda akan diarahkan ke dasbor utama. Baca dengan cermat konten setiap tweet yang ditampilkan di layar utama. Periksa gambar yang menyertai dengan klik ikon panah di sudut kanan atas gambar untuk melihatnya dalam ukuran penuh.

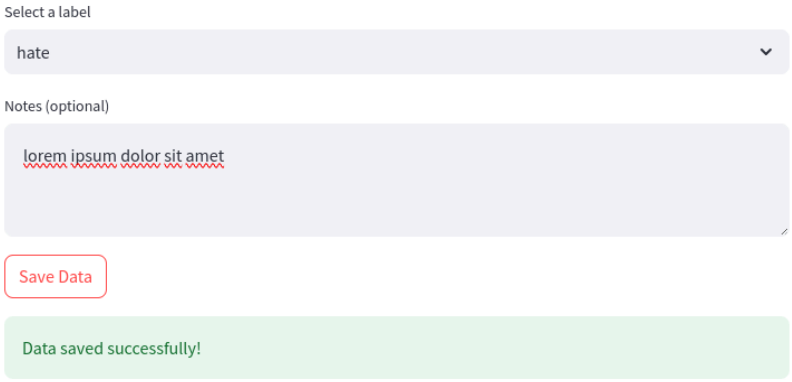
3. **Anotasi Tweet**
   Tentukan apakah tweet termasuk dalam kategori "Hate Speech" atau "Not Hate Speech" berdasarkan pedoman untuk teks dan gambar. Pilih label yang sesuai untuk setiap tweet menggunakan menu dropdown di bawah tweet.
   - Klik menu dropdown yang diberi label "Select a label".
   - Pilih "hate" atau "not hate" berdasarkan penilaian Anda.



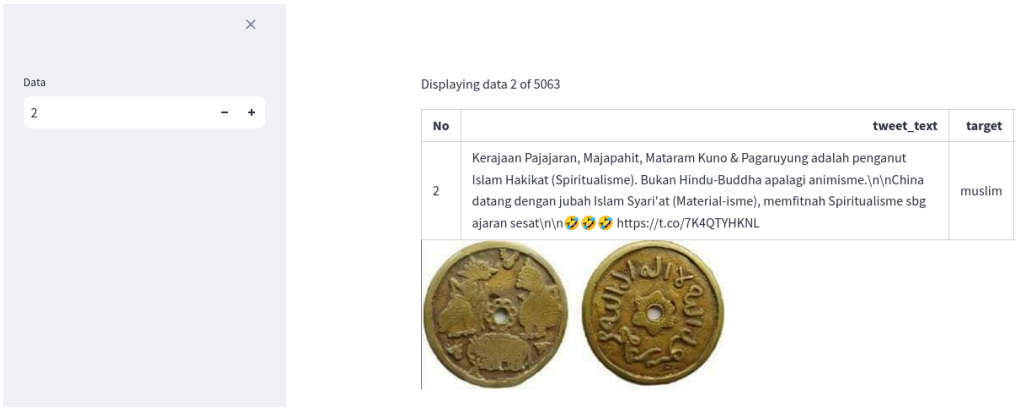
4. **Tambahkan Catatan (Opsional)**
   Jika Anda menemukan tweet yang sangat sulit untuk diklasifikasikan, Anda dapat menambahkan catatan yang menjelaskan keputusan Anda. Gunakan area teks yang diberi label "Catatan (opsional)" untuk memberikan konteks tambahan atau alasan untuk anotasi Anda. Langkah ini opsional tetapi sangat dianjurkan untuk kasus yang ambigu atau berada di batas antara dua kategori.



5. **Simpan Anotasi Anda**
   Setelah anotasi tweet dan menambahkan catatan, klik tombol "Save Data" untuk menyimpan anotasi Anda. Pesan sukses "Data saved successfully!" akan ditampilkan, menandakan bahwa anotasi Anda telah disimpan.



6. **Navigasi ke Tweet Berikutnya**
   Gunakan kontrol navigasi di bilah sisi kiri untuk pindah ke tweet berikutnya.
   - Klik tombol "+" untuk memuat tweet berikutnya.
   - Klik tombol "-" untuk kembali ke tweet sebelumnya.
   - Ulangi proses ini untuk setiap tweet dalam dataset.



Jaminan Kualitas

- **Konsistensi**: Pastikan penerapan pedoman ini secara konsisten di semua tweet.
- **Tinjauan**: Tinjau anotasi Anda secara berkala untuk menjaga kualitas tinggi.
- **Umpan Balik**: Berikan umpan balik jika Anda menemukan tweet yang tidak jelas atau ambigu.

Informasi Kontak

Jika Anda memiliki pertanyaan atau memerlukan klarifikasi lebih lanjut, silakan hubungi Endang Wahyu Pamungkas di ewp123@ums.ac.id