Executive Summary

Milestone 2 of the TikTok Claims Classification Project

ISSUE / PROBLEM

Tim data TikTok berupaya mengembangkan model machine learning untuk membantu klasifikasi klaim user submissions. Dimulai dengan tim data perlu mengatur raw dataset dan mempersiapkannya untuk penelusuran dimasa mendatang.

RESPONSE

Tim data melakukan investigasi awal terhadap dataset klasifikasi klaim dengan tujuan mempelajari hubungan penting antara variabel.

Mengingat permintaan untuk klasifikasi klaim user, tim data melihat jumlah label klaim dan opini untuk memahami jumlah setiap jenis konten video.

IMPACT

Dampak dari analisis awal ini akan terlihat pada step berikutnya. Untuk memahami dampak setiap video, tim data mengidentifikasi dua variabel penting yang perlu dipertimbangkan: Variabel video_duration (dalam detik) dan video_view_count merupakan faktor yang perlu dipertimbangkan untuk model prediksi mendatang.

UNDERSTANDING THE DATA

Setelah meninjau dataset yang disediakan, variabel claim_status tampak sangat berguna, mengingat proyek yang diajukan klien. Screenshot berikut menunjukkan poin penting analisis yang diperlukan untuk memahami variabel claim_status.

```
data['claim_status'].value_counts()
```

```
claim      9608
opinion    9476
Name: claim_status, dtype: int64
```

Note: Jumlah setiap status klaim cukup seimbang. Terdapat 9608 klaim dan 9476 opini

ENGAGEMENT TRENDS

Tim data mempertimbangkan interaksi penonton dengan setiap video dalam kategori klaim dan opini. Untuk memahami interaksi penonton, tim data mempertimbangkan jumlah penayangan. Rata-rata dan median jumlah penayangan menunjukkan dampak dari setiap kategori video; khususnya, jumlah penayangan rata-rata dan median untuk kedua kategori menunjukkan hubungan antara konten (klaim atau opini) dan penayangan video.

Claims:

```
Avg. view count claim: 501029.4527477102
Median view count claim: 501555.0
```

Opinions:

```
Avg. view count opinions: 4956.43224989447
Median view count opinions: 4953.0
```

KEY INSIGHTS

- Terdapat keseimbangan yang setara antara label opini dan klaim. Dengan pemahaman ini, kami dapat melanjutkan analisis di masa mendatang dengan mengetahui bahwa terdapat jumlah klaim dan opini yang cukup seimbang untuk data video didataset ini.
- Dengan mengidentifikasi key variables dan investigasi awal terhadap dataset klasifikasi klaim, proses analisis untuk eksplorasi data dapat dimulai.

Pie chart visualizes the comparison of the count of claims and opinions