Audio-Visual Sentiment Analysis for Learning Emotional Arcs in Movies

I. Pendahuluan

Selain menjadi hiburan, cerita memiliki kekuatan untuk memicu ketertarikan dan menggerakkan kita. Bagaimana sebuah cerita dapat menyentuh penonton mungkin berada pada perjalanan emosional yang membawa penonton tersebut.

Menggunakan *deep convolutional neural net*work untuk menganalisa sentimen video dan audio, mungkin dapat memprediksi keikutsertaan penonton berdasarkan potongan/busur cerita (arc).

Untuk mendorong penelitian ini, beberapa kontribusi diberikan oleh peneliti seperti:

- Datasets. Diberikan dataset dari *Spotify* yang mengandung lebihd ari 600.000 sampel musik, yang dapat digunakan untuk klasifikasi audio, yang juga disertai anotasi dari 1000 klip film berdurasi 30 detik.
- Memodelkan busur/naratif emosional. Peneliti memberikan model pengklasifikasian sentimen yang dilatih, untuk mengkomputasi Arc visual dan audio.
- Analisis keikutsertaan. Peneliti menyiapkan contoh bagaimana Arc naratif film, dapat memprediksi jumlah komentar dengan signifikan secara statistik.

II. Penelitian Terkait

Penelitian serupa juga pernah dilakukan oleh Andrew J. Reagan (2016) bagaimana emosional arc naratif didominasi oleh enam bentukan awal cerita. Selain itu, penulis pada *Dramatica*.com menganalisa secara manual banyak buku dan film yang menggunakan teori cerita *Dramatica* yang setelah itu menjadi basis dari pembuatan perangkat lunak yang memandu penulis. Penelitian sentimen analisis kebanyakan menggunakan data tekstual singkat, atau artikel panjang, dan jarang yang menggunakan gambar.

III. Gambaran

Pemodelan terjadi di dua ukuran, yaitu kecil (mikro) dan besar (makro). Analisis sentimen mikro dilakukan pada potongan video dan audio yang mengukur keselarasan antara prediksi naik-turunnya emosional dengan penilaian “annotator” (manual); dan makro yang dilakukan dengan clustering dan menggunakan busur (arc) naratif visual tersebut untuk analisis keikutsertaan (engagement) dari penonton. Prediksi audio tidak digunakan karena kurangnya kebenaran mendasar mengenai penggabungan prediksi audio-visual.

IV. Dataset

a) Video

Digunakan film *Hollywood* dan film singkat termasuk video dari platform *Vimeo*. Data film atau video digunakan karena mereka dibuat untuk menjelaskan cerita, dimana terdapat teori yang menjelaskan bagaimana alur film dapat memicu respons emosional

b) *Sentibank* dataset

Setengah juta gambar dari dataset *Setibank* digunakan, disertai dengan pelabelan dengan struktur kata sifat-benda, seperti “rumah indah” atau “ikan jelek”. Setiap label ini di kaitkan dengan ukuran sentimen menggunakan kamus *SentiWordnet*

c) *Spotify* dataset

Satu juta lagu digunakan, karena sifat lagu yang memiliki data tempo, kunci minor atau major, yang dapat memicu respons emosional. Dataset dari *Last.FM* juga digunakan, yang berisi lagu yang disertai label seperti “gembira” atau “sedih” dengan jumlah yang terbatas.

V. Metodologi

d) Membuat busur (arc) emosional

Setelah model visual dan audio dilatih untuk prediksi sentimen, model diterapkan pada film untuk membuat busur emosional.

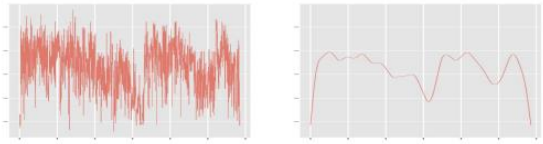


Fig. 2: Effect of smoothing values for arcs: no smoothing versus $w = 0.1 * n$

e) Pemodelan Gambar

Gambar dimodelkan menggunakan *deep convolutional neural network* yang didasarkan oleh arsitektur *AlexNet.*

f) Pemodelan Suara

Suara dimodelkan dengan memvisualisasikan setiap sampel suara menjadi 96-bin spektogram, seperti yang digunakan pada teknologi *tagging* musik, yang menggunakan lima lapisan konvolusi deengan ELU dan *batch normalization.*

g) Mencari keluarga busur (arc) emosional

Menggunakan pendekatan *k-medoids* dan *dunamic time warping* menjadi fungsi jarak untuk meng-kluster-kan.

VI. Evaluasi – Mikro

Model mikro di evaluasi dengan presisinya dalam mengekstraksi momen-momen yang memicu emosi dari busur emosional.

a) Video

Dengan dataset yang dilabeli dengan skor 1 sebagai perasaan negatif dan 7 sebagai perasaan positif, ditentukan titik tengah 4. Prediksi benar ketika klip video yang diekstrak dari bentuk “jurang/lembah” pada busur emosi di labeli sebagai perasaan negatif. Didapatkan tabel seperti berikut:

Set	Precision
All clips	0.642
No ambiguous clips	**0.681**

TABLE IV: Precision of clips: overall

Dengan baris “No. ambiguous clip” yang memiliki momen dimana perasaan tidak dapat ditentukan.

b) Audio

Dihasilkan tabel presisi seperti berikut :

Stddev	Audio-peak precision	Audio-valley precision
[0, 0.02)	4 / 4 = **1.0**	81 / 88 = **0.921**
[0.02, 0.04)	11 / 11 = **1.0**	38 / 56 = **0.679**
[0.04, 0.06)	28 / 40 = **0.7**	21 / 35 = **0.6**
[0.06, 0.08)	39 / 60 = **0.65**	13 / 21 = **0.619**
[0.08, 0.1)	43 / 68 = **0.632**	8 /23 = **0.615**

TABLE V: Precision of clips extracted from audio emotional arc: smaller confidence intervals are more precise

Mengikuti hipotesis peneliti yang menyatakan bahwa *confidence interval* yang lebih kecil akan menciptakan tingkat presisi lebih akurat.

c) Presisi potongan dan Presisi Genre

Cut	Precision
Audio-peaks	0.683
Audio-valleys	0.758
Visual-peaks	**0.508**
Visual-valleys	0.757

(a) Cuts

Genre	Overall	Visual-peak
Action	0.678	**0.264**
Science Fiction	0.699	**0.333**
Thriller	0.678	**0.382**
Adventure	0.726	**0.443**
Drama	0.660	**0.520**
Fantasy	0.769	**0.590**
Comedy	0.705	**0.667**
Animation	0.798	**0.667**
Family Film	0.760	**0.722**
Romance	0.678	**0.757**
Romantic Comedy	0.677	**0.823**

(b) Genres

TABLE VI: Precision of clips: cuts and genre

Tingkat presisi dari *visual-peaks* rendah, karena terikat dengan presisi genre yang lemah juga. Diketahui genre yang memiliki tingkat presisi *visual-peak* yang tinggi berada pada genre yang ringan, seperti romance dan film keluarga. Namun, pada genre aksi, terdapat banyak momen dimana kekerasan, sadis, dan kematian kemungkinan besar tidak ditemukan di dataset *Sentibank.*

d) Gabungan visual dan audio

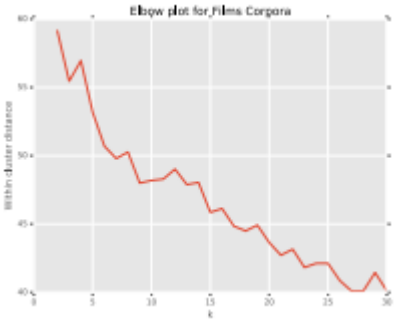
Feature set	Overall	Aud-peak	Aud-valley	Vis-peak	Vis-valley
All features	**0.894**	0.940	0.884	0.872	0.886
No *peakiness*	0.815	0.836	0.828	0.765	0.824
No *movie-embedding*	0.784	0.869	0.786	0.722	0.752

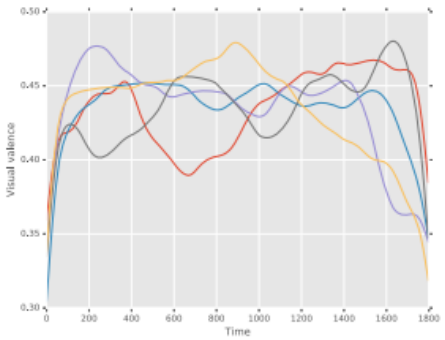
TABLE VII: Precision of combined audio-visual model

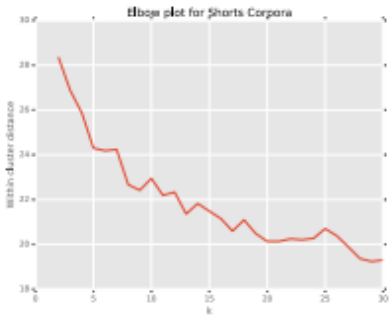
Tingkat presisi gabungan terlihat lebih tinggi, dengan semua fitur digunakan didapat nilai presisi 0.894. "*No movie-embedding*" didasari oleh hasil pada tabel sebelumnya yang melibatkan genre dari film. "No peakiness" yang menghitung dataran tinggi, lembah dan infleksi, dengan rumus. [1]

VII. Evaluasi – Makro

Menggunakan teknik kluster, ditemukan lima busur emosional yang tipikal pada film, yang mengikuti alur cerita, seperti naik tajam di awal dan turun tajam di bawah menandakan awal cerita dan *credits* atau akhir cerita.



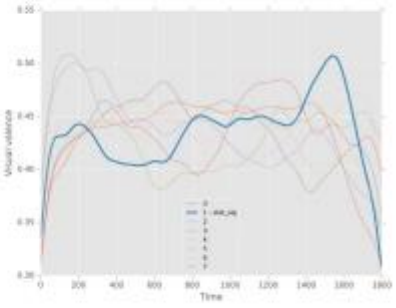




Sedangkan, pada potongan video kecil, ditemukan busur emosional yang tidak terlalu rumit, namun lebih ekstrim

Sedangkan, pada rumusan masalah analisis keikutsertaan, didapatkan tabel seperti berikut: Diketahui bahwa durasi dan tahun menjadi prediktor yang signifikan secara statistik, dan

Feature	Coefficient	p-value
Duration	**-0.1209**	**0.001**
Year	**-0.1720**	**0.000**
Month	*-0.0664*	*0.058*
Day	-0.0079	0.820
Hour	0.0085	807
Author_num_comments	0.0048	0.891
k=8 cluster 0	-0.0594	0.680
k=8 cluster 1	**0.3280**	**0.012**
k=8 cluster 2	0.0852	0.434
k=8 cluster 3	0.0968	0.471
k=8 cluster 4	-0.0831	0.245
k=8 cluster 5	-0.0121	0.951
k=8 cluster 6	-0.0480	0.552
k=8 cluster 7	-0.0224	0.741



beberapa arc juga signifikan secara statistik, yang berhubungan langsung dengan jumlah komentar pada video.

VIII. Kesimpulan

Dengan model prediksi yang telah dibuat, hasil dapat dilihat pada sebagian data yang berbentuk video singkat dari platform *Vimeo*, yang memprediksi jumlah interaksi seperti jumlah komen yang diterima video tersebut. Uji ini dilakukan menggunakan busur visual, dan metrik yang sangat sederhana.

IX. Saran Penelitian Selanjutnya

Seperti yang telah disebutkan sebelumnya, data yang lebih banyak dan berjangkauan lebih luas seperti ekstraksi data visual dan audio dimana keduanya bersifat netral, akan menciptakan hasil akurasi yang lebih akurat, selain itu masih terdapat banyak faktor yang mempengaruhi busur emosional pada sebuah cerita atau film, dengan menganalisa unsur intrinsik seperti dialog, atau alur.

X. Referensi

1. Chu, E., & Roy, D. (2017, November). Audio-visual sentiment analysis for learning emotional arcs in movies. In 2017 IEEE International Conference on Data Mining (ICDM) (pp. 829-834). IEEE.