Human Sentiment and Emotion Classification

Ursu Andrei

Specializarea Matematica-Informatica

ABSTRACT

sentiment = Proces afectiv specific uman, care exprimă atitudinea omului față de realitate

- DexOnline

ABSTRACT

emoție = Reacție afectivă de intensitate mijlocie și de durată relativ scurtă, însoțită adesea de modificări în activitățile organismului, oglindind atitudinea individului față de realitate

- DexOnline

Motivație

- Înțelegerea Comportamentului Uman
- Monitorizarea Sentimentelor pe Social Media
- Sănătatea Mintală
- Susținerea Cercetărilor Psihologice
- Facilitarea Cercetării de Piață
- Îmbunătățirea Asistenților Virtuali

TIMELINE

1980s

Analizarea sentimentelor și a emoțiilor bazate pe reguli lingvistice pentru identificarea și clasificarea sentimentelor și emoțiilor

1990s

Primele încercări de a automatiza detectarea emoțiilor în texte au început să apară odată cu progresul tehnologiilor informatice.

2000s

Au fost dezvoltate tehnici mai avansate de analiză a sentimentelor statistice, utilizând algoritmi de învățare automată și metode statistice

2010s

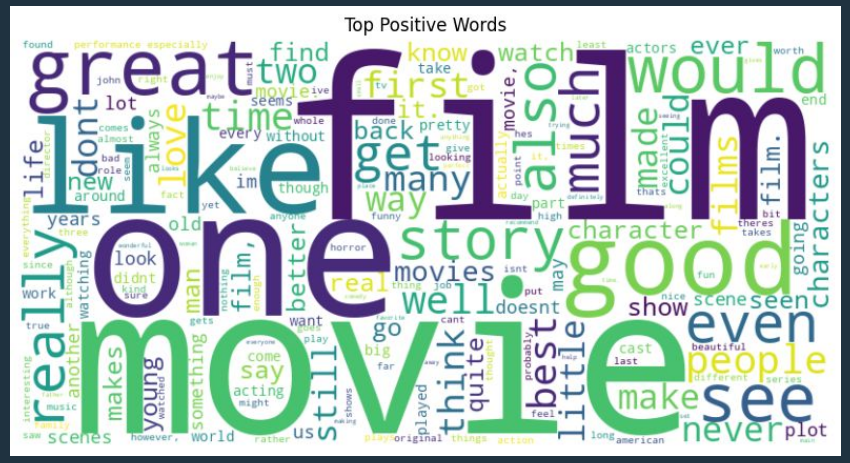
S-au făcut progrese semnificative cu utilizarea rețelelor neuronale și a învățării profunde, ceea ce a condus la o evoluție majoră în detectarea emoțiilor în texte.

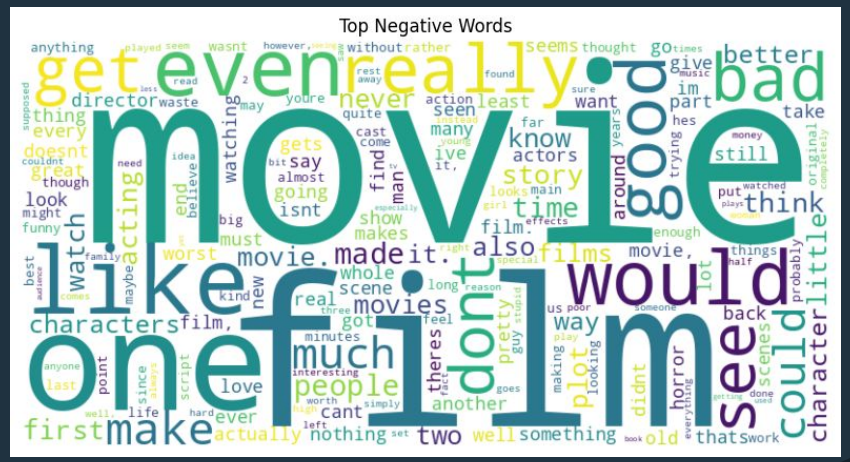
Present

Tehnologiile de procesare a limbajului natural au continuat să avanseze rapid, cu accent pe îmbunătățirea înțelegerii contextului și a nuanțelor semantice din texte.

Dataset

Acest set de date derivă din recenziile de filme colectate de la IMDb, o platformă de recenzii și evaluări a filmelor. Cu scopul de a înțelege și de a categorisi sentimentele exprimate în aceste recenzii, IMDB a furnizat o colecție vastă de date, care a devenit un mediu ideal pentru testarea algoritmilor și a modelelor de NLP. Aceste recenzii au fost adunate și etichetate manual cu o etichetă binară care indică natura sentimentului exprimat - fie pozitiv (1) sau negativ (0).
Alături, putem observa topul celor mai întâlnite cuvinte, atât cele pozitive, cât și cele negative.





Experimente

Word2Vec

FastText

LSTM

ResNet

Custom CNN

Bert Base Uncased

Roberta Base

Twitter Roberta Base Sentiment

Word2Vec

Ce este Word2Vec?

Word2Vec este o tehnică din domeniul NLP, concepută de echipa de cercetare Google sub îndrumarea lui Tomas Mikolov în 2013. Aceasta reprezintă un set de modele care sunt capabile să transforme cuvintele dintr-un text în vectori de caracteristici numerice, proces cunoscut sub numele de 'word embedding'.Cuvintele sunt reprezentate într-un spațiu vectorial continuu unde vectorii apropiați în spațiu sunt semnificativi din punct de vedere semantic.

Cum functioneaza?

Modelul Word2Vec utilizează două arhitecturi principale pentru a produce un embedding: Continuous Bag-of-Words (CBOW) și Skip-Gram. CBOW prezice probabilitatea unui cuvânt dat unui context, în timp ce Skip-Gram funcționează invers, prezicând contextul unui cuvânt.

Word2Vec

Cum functioneaza?

Ambele modele sunt antrenate printr-o metodă numită negative sampling, o tehnică eficientă pentru a optimiza clasificările de probabilitate pentru un set mare de clase. Capacitatea Word2Vec de a identifica sinonimele și relațiile semantice dintre cuvinte îmbunătățește precizia modelelor de analiză de sentiment. Prin utilizarea vectorilor de cuvinte, modelele pot generaliza mai bine peste diferite moduri de exprimare a sentimentelor similare, chiar și când sunt folosite cuvinte diferite. De asemenea, modelele Word2Vec sunt excepțional de bune în a captura aceste nuanțe semantice. De exemplu, cuvintele "bun" și "excelent" vor avea vectori similar plasați, sugerând o polaritate pozitivă.
In final, utilizând Word2Vec, acuratețea obținută pe setul de date IMDB a fost de 76.75% .

FastText

Ce este FastText?

Această tehnică reprezintă o evoluție a conceptului de 'word embeddings', propunând o abordare inovatoare prin incorporarea sub-cuvintelor sau n-gramelor în procesul de vectorizare a cuvintelor. Acest aspect permite modelului să captureze mai eficient morfologia cuvintelor, fiind deosebit de valoros pentru limbile cu o morfologie bogată și complexă.

Cum functioneaza FastText?

Această tehnică reprezintă o evoluție a conceptului de 'word embeddings', propunând o abordare inovatoare prin incorporarea sub-cuvintelor sau n-gramelor în procesul de vectorizare a cuvintelor. Acest aspect permite modelului să captureze mai eficient morfologia cuvintelor, fiind deosebit de valoros pentru limbile cu o morfologie bogată și complexă.

FastText

Cum functioneaza FastText?

De asemenea, FastText excellează în tratarea cuvintelor necunoscute sau insuficient reprezentate în setul de antrenament. Prin utilizarea n-gramelor, modelul poate ghici semnificația și sentimentul asociat cuvintelor noi pe baza segmentelor componente, o proprietate valoroasă când se confruntă cu texte dinamic și în continuă schimbare, cum sunt comentariile utilizatorilor sau, in cazul nostru, recenziile de film.

In final, utilizând FastText, acuratețea obținută pe setul de date IMDB este asemanatoare cu cea obtinuta de Word2Vec, anume de 76.95% .

Long short-term memory (LSTM)

Ce este LSTM?

Long Short-Term Memory, este un tip special de rețea neuronală recurentă (RNN) proiectată pentru a aborda problemele legate de dependențele pe termen lung care apar în secvențele de date. Rețelele RNN tradiționale întâmpină dificultăți în învățarea dependențelor pe distanțe mari din cauza problemelor numite "dispariția gradientului", unde contribuțiile informațiilor mai vechi devin din ce în ce mai diluate pe măsură ce sunt propagate prin timp.

Arhitectură

Această rețea neurală este un model secvențial folosit pentru clasificare binară. Modelul începe cu un strat de **Embedding**, care transformă indicii de intrare în vectori densi de 128 de dimensiuni. Următorul strat este un **LSTM bidirecțional** cu 64 de unități, care procesează secvențele atât în direcția înainte, cât și în direcția înapoi, având o rată de dropout de 0.3 pentru a reduce overfitting-ul. Acest strat este configurat să păstreze secvențele pentru următorul strat.

Long short-term memory (LSTM)

Arhitectură

Apoi, modelul include încă un strat **LSTM** cu 64 de unități, configurat de asemenea să păstreze secvențele, cu o rată de dropout de 0.2. Aceasta este urmat de un alt strat **LSTM bidirecțional** cu 32 de unități și cu dropout de 0.3. Ultimul strat este un strat **Dense** cu o singură unitate și activare sigmoid, folosit pentru a produce o predicție finală binară.

Optimizarea și evaluarea

Optimizarea este realizată cu ajutorul algoritmului Adam, configurat cu o rată de învățare de 0.001. Modelul este compilat cu funcția de pierdere 'binary_crossentropy' și metrica de evaluare 'accuracy', adecvate pentru problemele de clasificare binară.

ResNet

Ce este ResNet?

ResNet, prescurtare de la "Residual Network", este un tip de arhitectură de rețea neuronală care a fost inițial concepută pentru procesarea vizuală, în special pentru clasificarea imaginilor, și introdusă de Kaiming He și colab. în 2015. Principiul fundamental al ResNet este utilizarea de conexiuni reziduale sau "shortcut connections", care permit semnalelor să sară peste unul sau mai multe straturi.

Adaptate pentru NLP

Deși ResNet a fost dezvoltată pentru viziunea computerizată, conceptele sale pot fi adaptate și în procesarea limbajului natural (NLP). În NLP, ResNet poate ajuta în gestionarea rețelelor neuronale foarte profunde, facilitând propagarea eficientă a gradientului prin rețea datorită conexiunilor reziduale.

ResNet

Arhitectură

Stratul de Input - Prima parte este stratul de intrare care primește textul sub formă de secvențe de întregi codificați.

Stratul de Embedding - Transformă indicii întregi (reprezentând cuvinte) în vectori densi de 128 dimensiuni. Acest strat servește la încorporarea semantică a cuvintelor într-un spațiu continuu.

Primul Strat Convolutional - Aplică o convoluție 1D cu 128 de filtre și o fereastră de kernel de dimensiune 5. Activația 'relu' este utilizată pentru a introduce non-linearitatea.

Dropout - Acest strat are o rată de dropout de 0.5, folosit pentru a preveni overfitting-ul prin dezactivarea aleatorie a unor unități din stratul anterior în timpul antrenării.

ResNet

Arhitectură

Blocuri Reziduale - Modelul include mai multe blocuri reziduale, fiecare având un strat de convoluție principal și un strat rezidual. Acestea sunt conectate printr-o operație de adunare cu intrarea blocului, o tehnică inspirată din arhitecturile reziduale (ResNet) care ajută la evitarea degradării gradientului în rețele adânci.

Global Max Pooling - După blocurile reziduale, urmează un strat de Global Max Pooling care extrage caracteristica cea mai importantă din fiecare filtru convoluțional, reducând astfel dimensiunea datelor.

Straturi Dense - Stratul dens final de 128 de unități cu activare 'relu' urmat de un strat de ieșire cu o singură unitate și funcție de activare sigmoidă, folosit pentru a face o predicție binară (de exemplu, pozitiv/negativ).

Custom CNN

Arhitectură

Layer-ul de Embedding: Transformă indicii de cuvinte într-un spațiu vectorial continuu unde cuvintele cu semnificație similară sunt mapate apropiat unul de celălalt. max_features reprezintă dimensiunea vocabularului, iar 128 este dimensiunea spațiului vectorial în care sunt încorporate cuvintele. input_length=maxlen specifică lungimea fixă a secvențelor de intrare.

Layer-ul Conv1D: Acesta este un strat convoluțional pentru date unidimensionale. În contextul textului, acesta "alunecă" peste secvențe de cuvinte încorporate, aplicând 128 de filtre cu fereastra de dimensiune 5, folosind funcția de activare 'relu'. Aceasta ajută la extragerea caracteristicilor locale (de exemplu, grupuri de cuvinte).

Layer-ul MaxPooling1D: Reduce dimensiunea output-ului de la stratul convoluțional, selectând valoarea maximă din fereastra de dimensiuni specificate (în acest caz, 5). Aceasta este o tehnică de subsampling care ajută la reducerea complexității și la prevenirea overfitting-ului.

Custom CNN

Arhitectură

Un alt Layer Conv1D: Un alt strat convoluțional similar cu primul, care continuă să extragă caracteristici mai abstracte din text.

Layer-ul GlobalMaxPooling1D: Redă dimensiunea output-ului la un vector fix, extrăgând cea mai mare valoare de pe fiecare canal al caracteristicilor, independent de dimensiunea lor. Acest lucru este util pentru a obține cele mai importante semnale din fiecare caracteristică în vederea clasificării.

Layer-ul Dense: Un strat dens cu o singură unitate și funcția de activare sigmoid, care este tipic pentru problemele de clasificare binară. Acest strat calculează probabilitatea ca input-ul să aparțină unei anumite clase (de exemplu, pozitiv sau negativ).

Bert Base Uncased

Ce este Bert Base Uncased?

BERT (Bidirectional Encoder Representations from Transformers) este un model de procesare a limbajului natural (NLP) dezvoltat de Google. Arhitectura BERT este bazată pe tehnologia transformer, care folosește mecanisme de atenție pentru a capta contextul din ambele direcții ale unui text (stânga și dreapta).

Base indică versiunea modelului care este mai mică comparativ cu versiunea **BERT Large**. Modelul Base include 12 layere (sau straturi de transformeri), 768 dimensiuni ascunse, și 12 capete de atenție, totalizând aproximativ 110 milioane de parametri.

Uncased semnifică faptul că modelul a fost antrenat pe text convertit la litere mici, adică nu face diferențierea între litere mici și mari. Aceasta este o caracteristică importantă pentru sarcinile în care nu contează distincția între majuscule și minuscule.

Bert Base Uncased

Construcția modelului

Incarcarea Seturilor de Date: Folosind funcția load_dataset pentru a încărca setul de date IMDB, care este împărțit în subansambluri de antrenament, validare și test. Datele sunt amestecate pentru a asigura o distribuție aleatorie.

Preprocesarea Textului: Textul este normalizat prin conversia la litere mici și eliminarea caracterelor nedorite. Această normalizare ajută la reducerea variației în date, ceea ce poate îmbunătățește performanța modelului.

Tokenizarea: Utilizând AutoTokenizer pentru tokenizarea textului. Tokenizer-ul transformă textul într-o formă numerică pe care modelul o poate procesa, incluzând adăugarea de tokeni speciali și ajustarea lungimii secvențelor prin truncare sau padding.

Data Collator: Obiectul DataCollatorWithPadding gestionează padding-ul automat al loturilor de date, asigurând că toate secvențele dintr-un lot au aceeași lungime.

Bert Base Uncased

Construcția modelului

Modelul: Se folosește un model AutoModelForSequenceClassification pre-antrenat (bert-base-uncased). Modelul este configurat pentru clasificarea binară (num_labels=2), adecvat pentru setul de date imdb.

Metrii de Evaluare: Definirea metritelor pentru evaluarea performanței modelului, în acest caz acuratețea și scorul F1, care sunt comune pentru sarcinile de clasificare.

Antrenamentul: Configurarea argumentelor de antrenament cu TrainingArguments, specificând rata de învățare, numărul de epoci, strategia de salvare și dacă rezultatele trebuie împinse pe hub-ul Hugging Face. Apoi, folosim Trainer pentru a gestiona bucla de antrenament, evaluare și aplicarea metritelor.

Roberta Base

Ce este Roberta Base?

RoBERTa (Robustly Optimized BERT Approach) este o îmbunătățire a modelului BERT dezvoltată de Facebook AI, concepută pentru a optimiza performanța în sarcinile de procesare a limbajului natural (NLP). Principalele modificări care contribuie la îmbunătățirea RoBERTa față de BERT includ antrenamentul pe un set de date mai mare și pentru o perioadă mai lungă, eliminarea sarcinii de predicție a următoarei propoziții din antrenament, ajustări ale hiperparametrilor, și utilizarea diferită a strategiilor de tokenizare. RoBERTa Base este versiunea mai compactă a modelului, cu 12 layere, care a arătat îmbunătățiri semnificative în performanță pe diverse benchmark-uri NLP.

Roberta Base

Construcția modelului

Încărcarea și Pregătirea Seturilor de Date: Setul de date IMDB este împărțit în subansambluri de antrenament și test. Datele sunt amestecate și segmentate pentru a crea un subset de validare și pentru a actualiza setul de antrenament.

Preprocesarea Textului: Normalizează textul prin conversia la litere mici și elimină caracterele nedorite, și apoi aplică tokenizarea folosind tokenizer-ul pre-antrenat roberta-base. Aceasta pregătește textul pentru a fi procesat de model.

Tokenizarea: Transformăm textele într-o formă numerică, adăugând tokeni speciali și ajustând lungimea secvențelor prin truncare sau padding.

Data Collator: Ne asigurăm că toate secvențele dintr-un lot sunt de aceeași lungime, gestionând automat aplicarea padding-ului.

Roberta Base

Construcția modelului

Modelul de Clasificare a Secvențelor: Încarcăm RoBERTa cu configurația pentru clasificare binară (num_labels=2), adecvat pentru dataset-ul IMDB.

Metrții de Evaluare: Definim metricile de acuratețe și scorul F1 pentru a evalua performanța modelului.

Configurarea și Antrenamentul Modelului: Stabilim rata de învățare, numărul de epoci, și strategia de salvare. Utilizează Trainer pentru a gestiona antrenamentul, evaluarea și aplicarea metritelor.

Publicarea Modelului: Setările push_to_hub permit publicarea automată a modelului antrenat pe Hugging Face Hub, facilitând partajarea și reutilizarea modelulu.i

Twitter Roberta Base Sentiment

Ce este Twitter Roberta Base Sentiment?

Twitter RoBERTa Base Sentiment este o versiune a modelului RoBERTa optimizată pentru analiza sentimentelor în tweet-uri. Acest model este special adaptat pentru limbajul informal și condensat specific Twitter-ului, fiind antrenat pe date provenite direct de pe această platformă. Scopul său principal este de a clasifica sentimentele exprimate în tweet-uri ca fiind pozitive, negative sau neutre.

Construcția modelului

Preprocesarea Datelor: Datele sunt normalizate prin transformarea textului în litere mici și eliminarea caracterelor nedorite. Apoi, sunt tokenizate folosind tokenizer-ul de la RoBERTa adaptat pentru sentimentele de pe Twitter.

Configurarea Tokenizer-ului și a Modelului: Se folosește un tokenizer pre-antrenat specific pentru tweet-uri, iar modelul de clasificare a secvențelor este, de asemenea, încărcat cu specificația de a ignora dimensiunile nepotrivite..

Twitter Roberta Base Sentiment

Construcția modelului

Antrenamentul Diferențiat: Scriptul împarte parametrii modelului în două grupuri pentru a aplica rate diferite de învățare. Stratul de clasificare și ultimul strat al modelului RoBERTa primesc o rată de învățare mai mare, pe când restul parametrilor sunt actualizați cu o rată de învățare mai mică. Acest lucru este realizat pentru a permite ajustări mai fine la nivelul straturilor superioare, care sunt mai direct implicate în sarcina specifică de clasificare.

Configurarea și Rularea Antrenamentului: Se utilizează clasa Trainer de la Hugging Face pentru a gestiona bucla de antrenament, evaluare și calculul metritelor. Sunt specificate diverse opțiuni, cum ar fi strategia de evaluare, ratele de învățare, dimensiunile loturilor și numărul de epoci.

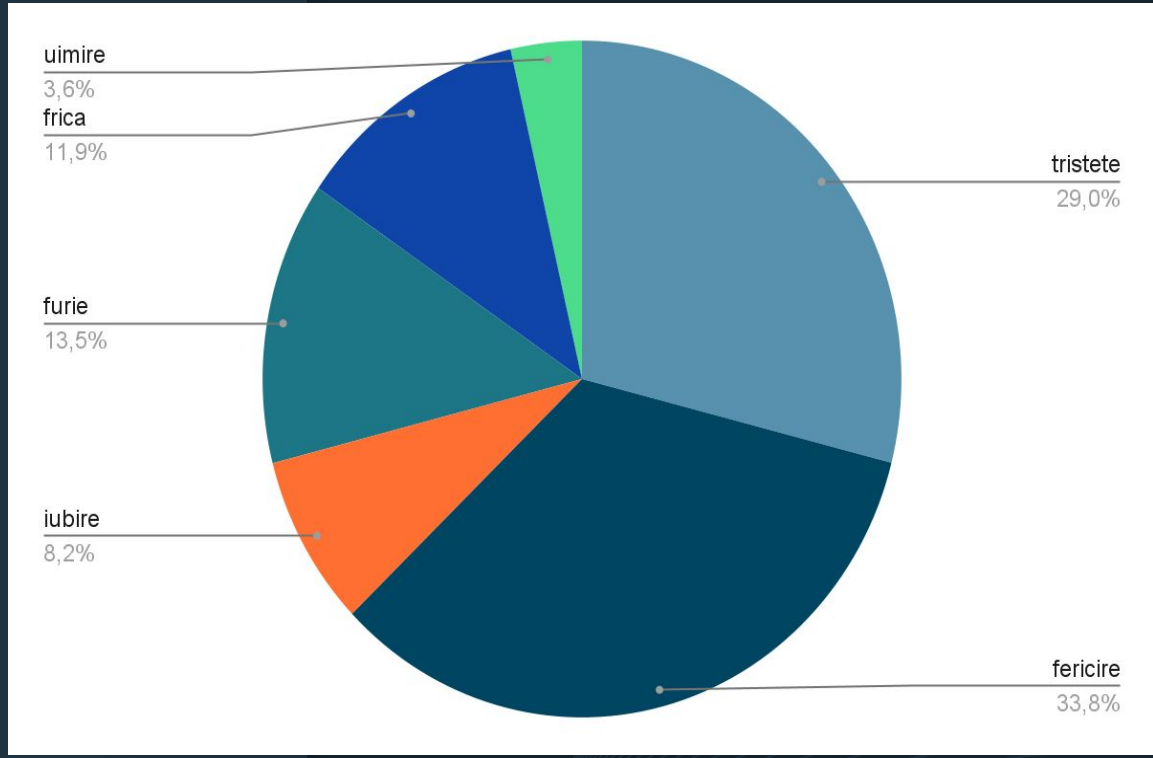
Evaluarea Performanței: Metricele de performanță sunt definite pentru a măsura acuratețea, utilizând un scor F1 și acuratețea pură, care sunt metrice standard pentru evaluarea performanței modelului în sarcini de clasificare binară.

Rezultate

Model	Training Accuracy	Test Accuracy
LSTM	0.9940	0.8099
Custom CNN	**1.000**	0.8492
ResNet	0.9962	0.5899
Bert Base Uncased	0.9531	0.9392
Roberta Base	0.9778	0.9441
Twitter Roberta Base Sentiment	0.9464	**0.9452**

Analiza emoției

Următoarea secțiune implică fine-tuning-ul unui model de clasificare a sentimentelor, bazat pe **Twitter RoBERTa Base Sentiment**, pe un set de date cu șase emoții diferite (tristețe, fericire, iubire, furie, frică și uimire). Scopul este de a evalua dacă modelul poate distinge între cele șase tipuri de emoții, deși inițial modelul a fost antrenat pe un spectru mai îngust de sentimente.



Analiza emoției

Construcția modelului

Pentru acest proiect, modelul a fost adaptat folosind o configurație care permite ajustarea la șase etichete, corespunzătoare celor șase emoții. Antrenamentul a fost realizat utilizând seturi de date încărcate și preprocesate pentru a include descrierile emoțiilor, iar tokenizer-ul RoBERTa a fost folosit pentru a converti textul în formatul necesar pentru model.

Deși modelul tinde să producă destul de mult overfitting atât pe datele de antrenare cât și pe cele de test, clasele prezise de model sunt în general corecte. Cele mai subtile diferențe pe care modelul nu reușește încă să le diferențieze sunt în cazul în care un text se încadrează în categoriile fericire sau iubire.

Future Work

Aplicatie completa

Urmeaza sa imbunatatesc modelul bazat pe TweetRoberta abordand tehnici mai complexe de fine-tunning si sa antrenez modelul pe noi seturi de date. De asemenea, voi crea o aplicatie completa pentru a analiza atat sentimentele/emotiile din texte, cat si din imagini.

Abordare matematica

Intentionez sa studiez mai in profunzime atat setul de date IMDB, cat si setul de date folosit pentru analiza emotiilor dintr-o perspectiva mai teoretica. Voi aplica notiuni din teoria masurii precum spatiile cu masura (in special spatiile de probabilitate) si integrala Lebesgue.

Multumesc!

Link HuggingFace:

https://huggingface.co/AndreiUrsu

INTREBARI