☀️ Detecția ironiei cu ajutorul emoticoanelor ☂️

Lăpăduș Daria 333
Radu Raluca 333
Trandafir Alexandru 333

1 Introducere

Am ales această temă deoarece considerăm că în era digitală, în care comunicarea este rapidă, fragmentată și plină de simboluri vizuale, modelele NLP tradiționale nu mai sunt suficiente. În special, modelele actuale ignoră adesea contribuția emoji-urilor, deși acestea sunt adesea folosite pentru a transmite sarcasm, frustrare, ironie sau bucurie exagerată 😀.

În opinia noastră, dezvoltarea unor modele sensibile la prezența și semnificația emoji-urilor este esențială pentru o interpretare mai nuanțată a mesajelor online. Am ales articolul *What a Sunny Day: Toward Emoji-Sensitive Irony Detection* deoarece propune o direcție inovatoare și practică, cu potențial aplicabil în mod direct în social media monitoring, detecția de conținut toxic sau analiza opiniei publice.

Prin acest proiect, ne propunem să înțelegem mai bine cum contribuie emoji-urile la exprimarea ironiei și să implementăm un model capabil să recunoască această relație, îmbunătățind astfel performanța în sarcina de *emoji-sensitive irony detection*.

2 Lucrări Asemănătoare

Detectarea ironiei în limbajul natural a devenit un subdomeniu activ în NLP, întrucât ironia apare frecvent în postări pe rețele sociale, unde exprimarea este informală, prescurtată și adesea acompaniată de emoji-uri sau alte simboluri non-verbale. În trecut, sistemele de detectare a ironiei s-au bazat pe caracteristici lingvistice și reguli sintactice, însă aceste abordări au fost rapid depășite odată cu apariția modelelor de învățare automată și, mai recent, a celor de tip deep learning.

Un pas semnificativ a fost realizat prin organizarea *SemEval-2018 Task 3*, care a oferit un benchmark standard pentru *irony detection in English tweets* (Van Hee et al., 2018). Multe modele au fost testate în acel context, însă majoritatea ignorau complet rolul emoji-urilor, deși acestea apar în peste 20% dintre tweet-urile reale.

Lucrarea de referință pentru proiectul nostru – *"What a Sunny Day: Toward Emoji-Sensitive Irony Detection"* (Chaudhary et al., 2019) – este una dintre primele care propun o analiză sistematică asupra modul în care emoji-urile contribuie la percepția ironiei. Autorii creează un dataset adnotat manual și automat (*Imoji*) și demonstrează că adăugarea sau eliminarea unui emoji poate schimba percepția unui mesaj de la sincer la ironic și invers.

De asemenea, lucrarea *DeepMoji* (Felbo et al., 2017) a demonstrat că emoji-urile pot fi folosite pentru a antrena modele de clasificare a emoțiilor, obținându-se *embedding-uri* de înaltă calitate care reflectă intenția utilizatorului. Această lucrare stă la baza ideii că emoji-urile nu sunt doar decorative, ci conțin *informație afectivă* utilă în analiza tonalității.

Mai recent, modele precum *BERTweet* (Nguyen et al., 2020) și *RoBERTa + emoji*

/ *embeddings* au fost testate pe tweet-uri cu etichete de ironie și sarcasm, arătând îmbunătățiri semnificative când emoji-urile sunt explicit în antrenare.

Prin urmare, stadiul actual al cercetării arată că integrarea emoji-urilor în modelele moderne de NLP nu este doar fezabilă, ci și recomandată pentru sarcini precum detectarea ironiei, sarcasmului și a sentimentelor mixte.

3 Metodologie

A. Descrierea metodei

Modelul de bază implementat este un *biLSTM cu self-attention*, antrenat pe date provenite din SemEval 2018 și extins prin metode de augmentare cu emoji-uri. Pipeline-ul include:

1. **Preprocesarea datelor** – normalizarea emoji-urilor, curățarea textului, tokenizare.
2. **Vectorizarea** – folosirea de *embeddings pre-antrenate* (GloVe pentru text și embeddings speciali pentru emoji-uri).
3. **Modelul** – rețea *biLSTM cu atenție*, urmată de un strat de clasificare softmax.
4. **Antrenarea și evaluarea** – modelul a fost antrenat pe trei subseturi: *original*, *emoji-added* și *emoji-removed*.

B. Compararea cu alte metode

Metodă	Avantaje	Dezavantaje
Reguli lingvistice	Explicabilitate ridicată	Performanță slabă
DeepMoji	Emoții bine reprezentate	Nu e specific pentru ironie
BERTweet	Pre-antrenat pe tweet-uri	Nu e sensibil la emoji
biLSTM + Attention	Sensibil la polaritate emoji	Necesită date adnotate

Tabela 1: Compararea metodelor existente

Modelul nostru se remarcă prin tratarea emoji-urilor ca semnale semantice esențiale. Modelul nostru se diferențiază de abordările generale prin faptul că tratează emoji-ul ca pe o sursă semantică importantă în clasificare. De exemplu, adăugarea unui emoji ironic (ex.🙄) la o propoziție pozitivă poate inversa complet percepția.
Prin testarea pe variante augmentate ale tweet-urilor (cu/ fără emoji), demonstrăm că modelul antrenat pe date „emoji-aware" performează mai bine pe date reale care includ emoji-uri, față de un model antrenat doar pe text.

3.1 Dataset

Pentru acest proiect, am utilizat inițial *datasetul de la SemEval 2018 Task 3*, care conține tweet-uri în limba engleză etichetate ca fiind ironice sau non-ironice. Aceste set de date a fost colectat automat folosind API-ul Twitter, iar etichetarea a fost realizată manual de către voluntari care au analizat conținutul tweet-urilor.

Pentru a evidenția impactul emoji-urilor asupra percepției ironiei, am extins acest dataset conform metodologiilor propuse în articolul studiat, construind trei subseturi:

- **original** – tweet-urile așa cum apar în setul SemEval;
- **emoji-added** – tweet-uri *non-ironice* în care s-au adăugat emoji-uri cu polaritate opusă pentru a sugera ironie;
- **emoji-removed** – tweet-uri ironice din care au fost eliminate emoji-urile pentru a vedea impactul asupra clasificării.

În cazul subseturilor augmentate, datele au fost generate *automat* folosind o combinație de reguli euristice (bazate pe lexicoane de polaritate a emoji-urilor) și apoi validate *manual de adnotatori*, similar procesului folosit pentru construirea *Emoji dataset* (Chaudhary et al., 2019).

Relevanță pentru model:

Ne așteptăm ca prezența unor emoji-uri cu polaritate contrastantă față de text (ex.

😝) să servească drept semnal important pentru modelul nostru. Astfel, includerea acestor exemple în antrenare permite modelului să învețe nu doar din cuvinte, ci și din contextul afectiv transmis vizual prin emoji-uri.

3.2 Preprocesare

Preprocesarea datelor este un pas esențial în orice aplicație NLP, iar în cazul nostru, a fost deosebit de importantă deoarece tweet-urile sunt de multe ori redactate incorect și conțin simboluri speciale și elemente vizuale precum emoji-urile, care pot fi relevante sau, dimpotrivă, pot induce erori.

Am aplicat următoarele metode de preprocesare:

- **Eliminare mențiuni (@), link-uri**: acestea nu contribuie cu informație semantică relevantă pentru detecția ironiei.
- **Eliminare link-uri**: URL-urile prezente în tweet-uri sunt irelevante pentru tonul exprimat.
- **Conversie la litere mici (lowercasing)**: pentru a unifica forma cuvintelor și a reduce dimensiunea vocabularului.
- **Eliminare caractere speciale și a spațiilor multiple**: pentru a curăța datele brute.
- **Tokenizarea textului**: împărțirea tweet-ului în unități semantice de bază (cuvinte, semne).
- **Normalizare emoji-uri**: în loc să le eliminăm, le-am înlocuit cu tokeni speciali precum `[emoji_laugh]`, `[emoji_sad]`, etc., pentru a păstra semnificația lor în rețea.
- **Padding / truncare la lungime fixă**: tweet-urile au fost aduse la o lungime standard pentru a fi procesate de model.

Motivația alegerilor

Am ales să păstrăm și să **codificăm** emoji-urile, spre deosebire de abordările clasice care le elimină, deoarece obiectivul nostru este tocmai evaluarea impactului lor asupra percepției ironiei. Astfel, le-am tratat ca unități semantice proprii în embedding, pe lângă cuvintele din text.

Impactul asupra modelului

Modelul a beneficiat de o acuratețe înaltă în detectarea tweet-urilor ironice atunci când emoji-urile au fost prezente și tratate corect. De exemplu, tweet-uri care nu exprimau clar ironie în cuvinte, dar conțineau un emoji sarcastic (ex.), au fost clasificate corect doar în variantele preprocesate cu emoji păstrate. Astfel, **preprocesarea atentă** a condus la o îmbunătățire a performanței, în special în recall pentru clasa de ironie.

3.3 Modelul

A. Modelul utilizat

Pentru acest proiect, am ales un model de tip *biLSTM* (Bidirectional Long Short-Term Memory) cu mecanism de *self-attention*, inspirat din lucrarea *"What A Sunny Day: Toward Emoji-Sensitive Irony Detection"* (Chaudhary et al., 2019). Alegerea acestui model a fost motivată de:

- **Capacitatea biLSTM-ului** de a captura contextul înainte și după un cuvânt, esențial pentru înțelegerea nuanțelor ironice.
- **Atenția** oferă interpretabilitate și ajută modelul să se concentreze pe tokenii relevanți (emoji sau cuvinte cu polaritate afectivă).
- Este mai ușor de antrenat și interpretat decât modele mai complexe de tip transformer, fiind potrivit pentru experimente rapide și *augmentări controlate*.

B. Parametri aleși:

- Dropout: 0.4
- Optimizator: Adam (lr=0.001)
- Embedding: 200 (text), 50 (emoji)
- Batch size: 64, Epoci: 15

Am utilizat embedding-uri pre-antrenate pentru text și am definit manual vectori speciali pentru emoji-uri, tratați ca tokeni separați.

C. Experimente realizate:

Am comparat performanța modelului pe trei seturi:

- **original**: date brute din SemEval 2018
- **emoji-added**: date augmentate automat
- **emoji-removed**: date cu emoji-uri eliminate

Rezultatele noastre arată că modelul antrenat pe date cu emoji a obținut un F1 mai bun pe tweet-urile care conțin sarcasm indus de emoji, depășind cu 4% scorul modelului de bază. Astfel, se confirmă concluziile din articolul original.

Model	Dataset	Accuracy	F1 (Irony)
biLSTM	Original	72.4%	66.3%
biLSTM + emoji	Emoji-added	75.9%	70.8%
biLSTM + attention	Emoji-removed	73.5%	69.1%

Tabela 2: Performanța modelului pe subseturi

D. Interpretabilitatea caracteristicilor:

Datorită folosirii atenției, am putut analiza importanța fiecărui token în decizia modelului. De exemplu, în tweet-uri precum:

„So happy to be ignored again 🙄”

modelul a alocat atenție crescută pe emoji-ul 🙄 și pe cuvântul „ignored”, ceea ce arată că rețeaua înțelege ironia ca rezultat al contrastului între textul aparent pozitiv și emoji-ul sarcastic.

În schimb, în modelele fără emoji sau fără atenție, acest tweet era adesea etichetat greșit ca fiind sincer.

4 Lucrări Viitoare

Antrenarea unui model de tip transformer, cum ar fi `BERTweet` sau `RoBERTa` adaptat pentru emoji-uri, ar putea aduce îmbunătățiri semnificative. Aceste modele au capacitatea de a capta contexte mai largi și de a înțelege *subtilitățile* limbajului informal din social media.

Integrarea proiectului într-o aplicație practică: de exemplu, un modul pentru detectarea sarcasmului în comentarii online sau un filtru pentru moderarea automată pe platformele sociale. Un sistem capabil să recunoască ironia ar putea preveni interpretările greșite și ar putea contribui la moderarea discursului digital într-un mod mai nuanțat.

Pe termen lung, acest model ar putea fi extins pentru **alte limbi** decât engleza, în special pentru limbi în care emoji-urile joacă un rol cultural semnificativ.

5 Concluzie

Acest proiect ne-a oferit o perspectivă complexă asupra modului în care ironia este exprimată și percepută în mediile sociale, în special prin intermediul emoji-urilor. Deși ironia este subtilă și greu de definit formal, existența unor indicii afective precum emoji-urile poate influența semnificativ interpretarea unui mesaj.

Abordările tradiționale, care ignoră simbolurile vizuale, pot fi ineficiente în fața unei comunicări moderne bazate pe contexte mixte (text + vizual). Astfel, valoarea adăugării emoji-urilor ca tokeni semantici în pro-

cesul de antrenare a rezultat într-o îmbunătățire reală a performanței modelului.

Un impediment a fost lipsa unor dataseturi largi și bine echilibrate care să includă emoji-uri în mod consistent și corect adnotate. Unele tweet-uri pot avea mai multe interpretări, ceea ce face ca și adnotarea manuală să devină subiectivă.

În concluzie, proiectul ne-a ajutat să înțelegem mai bine interacțiunea dintre limbajul natural și emoji-uri.

6 Limitări

Există câteva limitări importante care merită explorate în viitoarele lucrări:

1. **Limitarea lingvistică**
   Modelul antrenat funcționează exclusiv pentru limba engleză, întrucât datele de antrenament sunt în această limbă. Extinderea la alte limbi ar necesita fie traduceri adnotate atent, fie colectarea unor noi seturi de date relevante pentru fiecare limbă, cu particularitățile sale culturale și emoționale legate de utilizarea emoji-urilor.

2. **Generalizarea**
   Datasetul augmentat conține exemple sintetice generate automat, care, deși valide, nu reflectă întotdeauna complexitatea mesajelor reale. Astfel, modelul poate performa bine pe date curate, dar mai slab în situații autentice, cu slang, greșeli gramaticale sau utilizare ambiguă a emoji-urilor.

3. **Resurse computaționale**
   Deși modelul biLSTM folosit este relativ ușor de antrenat, integrarea unor modele mai puternice (precum `BERTweet`) ar necesita resurse GPU considerabile.

4. **Scalabilitate**
   Preprocesarea și augmentarea automată a emoji-urilor sunt încă procese semi-manuale, greu de scalat în lipsa unei infrastructuri dedicate și a unui pipeline complet automatizat de adnotare și validare a ironiei.

5. **Ambiguitatea semantică**
   Un emoji poate avea sensuri multiple în funcție de context. De exemplu, 😝 poate exprima bucurie sinceră sau râs sarcastic. Modelul are dificultăți în a distinge aceste nuanțe fără contexte suplimentare.

Declarație Etică

Având în vedere natura acestui proiect — analiza tweet-urilor și detectarea ironiei în exprimarea online — este important să reflectăm asupra posibilelor implicații etice.

Utilizări potențial neetice

Un model de detecție a ironiei ar putea fi folosit în mod abuziv în aplicații de:

- **supraveghere excesivă a discursului** online, cu scopuri de cenzură automată;
- **manipulare a opiniei publice**, prin clasificarea greșită a mesajelor sarcastice ca fiind negative, ceea ce ar putea distorsiona analiza sentimentelor reale;
- **profilare comportamentală**, în care mesajele ironice sunt analizate fără consimțământul explicit al utilizatorului.

Posibile prejudecăți (bias-uri)

Modelul antrenat pe date în limba engleză, provenite dintr-un context cultural specific (Twitter, 2018), poate încorpora **bias-uri** legate de:

- **cultura limbajului online occidental**;
- **sensul emoji-urilor**, care poate varia în funcție de țară, generație sau context;

- **exprimările informale**, care pot fi interpretate diferit în funcție de grupul social.

Aceste **bias-uri** pot afecta acuratețea modelului și pot conduce la decizii eronate atunci când este aplicat asupra unor populații diverse.

Măsuri luate

Pentru a reduce aceste riscuri:

- Am analizat atent sursa datelor și modul de etichetare;
- Am tratat emoji-urile în mod explicit **ca entități semantice**;
- Am păstrat un echilibru între clasele ironice și **non-ironice** pentru a evita învățarea dezechilibrată.

Recomandări

Recomandăm ca cercetările bazate pe acest model:

- să fie aplicate doar în scopuri educaționale, de cercetare sau pentru îmbunătățirea interacțiunii om-computer;
- să includă o analiză a bias-ului atunci când sunt extinse spre alte limbi sau culturi;
- să fie folosite transparent, fără a înlocui **judecata umană** în interpretarea mesajelor subtile.

Opinie personală

Considerăm că modelele NLP moderne pot aduce beneficii semnificative în înțelegerea limbajului digital, dar trebuie dezvoltate și utilizate cu responsabilitate. În mod special, detectarea ironiei este o sarcină sensibilă, care poate avea consecințe etice serioase dacă este aplicată fără un cadru clar de limitare și interpretare.

Resurse

Articolul principal de referință

- **Chaudhary et al. (2019)**, *"What a Sunny Day ☔: Toward Emoji-Sensitive Irony Detection"* – lucrarea propune un model biLSTM cu self-attention antrenat pe date augmentate cu emoji-uri și introduce datasetul *Imoji*, utilizat pentru a demonstra impactul emoji-urilor asupra percepției ironiei.

Alte lucrări relevante

- **Van Hee et al. (2018)** – organizatorii *SemEval-2018 Task 3*, care a oferit un benchmark standardizat pentru detecția ironiei în tweet-uri.
- **Felbo et al. (2017)**, *DeepMoji* – lucrare care folosește emoji-urile ca sursă de etichetare pentru antrenarea unui model de recunoaștere a emoțiilor, introducând embedding-uri afective.
- **Nguyen et al. (2020)**, *BERTweet* – model de tip transformer pre-antrenat special pentru textul din Twitter, utilizat pentru comparație în literatură și ca sugestie pentru lucrări viitoare.