Detectarea optimismului si pesimismului

Diaconescu Alexandra
alexandra.diaconescu1@s.unibuc.ro

Ruști Emilia Noemi
emilia-noemi.rusti@s.unibuc.ro

Farcași George Octavian
george-octavian.farcasi@s.unibuc.ro

1 Abstract

Acest proiect își propune să analizeze sentimentele exprimate în postările de tip tweet, folosind atât o abordare clasică bazată pe algoritmul Naive Bayes, cât și o metodă modernă care utilizează modele de tip transformer. În cadrul abordării moderne este utilizat modelul BERTweet, un model pre-antrenat special conceput pentru limbajul folosit pe Twitter, pe care l-am adaptat pentru a face diferența între sentimente pozitive, negative și neutre. Proiectul compară performanțele celor două metode, rezultatele obținute arătând o îmbunătățire semnificativă a acurateței în cazul metodei bazate pe BERTweet.

2 Introducere

Problema clasificării sentimentelor exprimate în tweet-uri a fost abordată prin încadrarea acestora în trei categorii: pozitiv, negativ și neutru. Analiza acestui tip de conținut este o provocare, având în vedere lungimea redusă a textelor, limbajul colocvial și utilizarea frecventă a simbolurilor, abrevierilor și emoticoanelor.

Am ales această temă deoarece considerăm că are relevanță practică în contexte precum monitorizarea opiniei publice, analiza reacțiilor la produse sau evenimente și în general în înțelegerea comportamentului utilizatorilor din mediul online. În plus, ne-a oferit oportunitatea de a combina metodele tradiționale cu cele moderne, observând creșterea performanței.

În cadrul lucrării, am comparat o abordare clasică, bazată pe algoritmul Naive Bayes, cu una modernă, care utilizează modelul BERTweet, antrenat special pentru limbajul utilizat pe Twitter.

3 Articole asociate

Începuturile s-au bazat pe metode clasice de învățare automată, precum Support Vector Machines sau Naive Bayes, utilizând caracteristici extrase manual, cum ar fi frecvența cuvintelor sau scorurile de polaritate.

În ultimii ani au devenit populare modelele neuronale profunde, în special cele de tip transformer, odată cu introducerea arhitecturii BERT (Vaswani et al., 2017). Aceste modele au fost antrenate pe cantități mari de date și pot fi adaptate cu ușurință la diverse sarcini de clasificare, inclusiv analiza sentimentelor (Devlin et al., 2019).

Un model recent și relevant pentru limbajul din rețelele sociale este **BERTweet** (Nguyen et al., 2020), antrenat pe un corpus mare de tweet-uri în limba engleză. Acesta a demonstrat performanțe superioare față de alte variante de BERT în sarcini precum clasificarea emoțiilor și analiza opiniei în mesaje scurte.

De asemenea, studiile recente (Barbieri et al., 2021) arată că modelele pre-antrenate specializate pe domenii (ex: social media, recenzii, domeniul medical) pot depăși performanțele modelelor generice, în special în cazurile în care datele conțin mult zgomot sau limbaj informal.

Abordarea noastră pornește de la aceste idei și își propune să compare performanța unei metode clasice cu una bazată pe modelul BERTweet, adaptat pentru clasificarea în trei clase de sentimente.

4 Metoda

În această secțiune, vom detalia abordarea clasică utilizată pentru clasificarea sentimentelor, precum și abordarea modernă a acesteia.

Abordarea clasică folosește tehnici bine cunoscute de învățare automată care sunt mai rapid de implementat și mai ușor de înțeles. Totuși acestea nu reușesc să captureze relațiile complexe și subtile dintre cuvinte așa cum fac modelele mai avansate. Pe de altă parte, abordările moderne folosesc modele pre-antrenate, care sunt capabile să înțeleagă mai bine contextul unui text și să îmbunătățească semnificativ performanțele pentru multiple sarcini de procesare.

4.1 Setul de date

Pentru antrenarea și testarea modelelor de clasificare a sentimentelo, am folosit un set de date public disponibil pe Kaggle, intitulat *Sentiment Analysis Dataset* (Shrivastava, 2021). Acesta conține tweet-uri în limba engleză etichetate cu trei tipuri de sentimente: pozitiv, negativ și neutru. Datele au fost colectate automat din tweet-uri publice și etichetarea lor a fost realizată manual, pe baza sentimentului exprimat în fiecare tweet.

Setul de date este structurat în două fișiere: `train.csv`, utilizat pentru antrenarea modelelor, și `test.csv`, folosit pentru evaluarea performanței. Fiecare rând din aceste fișiere reprezintă un tweet împreună cu o serie de metadate relevante. În total, fiecare intrare conține între 9 și 10 coloane, însa relevante pentru noi au fost doar următoarele:

- `text` – tweet-ul propriu-zis, în format text;
- `sentiment` – eticheta sentimentului asociat (pozitiv, negativ sau neutru).

Distribuția etichetelor este relativ echilibrată cu o predominare ușoară a tweet-urilor cu sentiment negativ.

Tweet-urile din acest set de date conțin frecvent limbaj specific platformelor de social media, cum ar fi abrevieri, mențiuni, hashtag-uri și numeroase emoticoane și emoji-uri, ceea ce face problema clasificării sentimentului mai dificilă, dar totodată mai relevantă pentru aplicații reale. Prezența emoticoanelor oferă un indiciu important pentru detecția tonului mesajului, motiv pentru care acest aspect a fost luat în calcul în etapa de preprocesare.

4.2 Preprocesarea

În cadrul abordării clasice, preprocesarea a constat într-un set de tehnici standard, esențiale pentru a reduce zgomotul și pentru a face procesarea mai eficientă, incluzând:

- Eliminarea URL-urilor din tweet-uri.
- Eliminarea caracterelor repetitive și consecutive.
- Modificarea literelor mari în litere mici.
- Transformarea caracterelor speciale în spații (cum ar fi semnele de punctuație).
- Înlocuirea emoticoanelor și emoji-urilor cu descrierea lor textuală.
- Îndepărtarea mențiunilor și a hashtag-urilor.
- Transformarea numerelor în cuvinte.
- Lematizarea cuvintelor pentru a le reduce la forma de bază.

În cadrul metodei moderne au fost utilizate tehnici suplimentare pentru a îmbunătăți preprocesarea datelor, aliniate cu cerințele modelelor de tip Transformer, în special BERTweet. Acestea includ:

- Eliminarea duplicatelor pentru a evita introducerea de bias în modelul de învățare

- Tokenizarea textelor utilizând tokenizer-ul specializat pentru limba folosită în tweet-uri – **vinai/bertweet-base** , care păstrează caracterul colocvial și informal al rețelelor sociale.
- Împărțirea datelor în seturi de antrenare și test, cu stratificare după etichetele de sentiment pentru a asigura un echilibru corect între clase.
- Construirea unui dataset compatibil cu PyTorch, în care fiecare tweet este convertit într-un set de tensori (input_ids, attention_mask și labels), esențial pentru antrenarea rețelelor neuronale de tip Transformer

4.3 Metoda - Classical approach

În cadrul abordării clasice, am implementat un model Naive Bayes pentru clasificarea sentimentelor. Acesta a fost antrenat pe setul de date preprocesat și evaluat utilizând o măsură standard de performanță, cum ar fi acuratețea.

- **Motivul alegerii Naive Bayes**: Am ales această metodă datorită simplității sale și a eficienței sale în contextul în care datele sunt preprocesate corect. Naive Bayes este un model probabilistic care funcționează bine pe seturi de date mici sau medii și este mai ușor de implementat.

- **Cum funcționează**: Modelul se bazează pe calcularea probabilității fiecărei categorii și probabilității unui cuvânt să apară într-o categorie.

Mai întâi, tweet-urile sunt prelucrate și împărțite în trei categorii: pozitive, negative și neutre. Apoi, pentru fiecare clasă de sentiment, se calculează frecvența fiecărui cuvânt apărut în mesajele respective. Pe baza acestor frecvențe, se determină probabilitățile și log-probabilitățile asociate fiecărui cuvânt, în funcție de clasă. Pentru a evita situațiile în care un cuvânt are probabilitate zero (adică nu apare deloc în datele de antrenament pentru o anumită clasă), se aplică tehnica de smoothing Laplace. Astfel, și cuvintele neîntâlnite în setul de antrenament primesc o probabilitate diferită de zero, deși foarte mică.

În final, pentru un tweet nou, se calculează un scor pentru fiecare categorie de sentiment prin însumarea log-probabilităților cuvintelor conținute în mesaj. Categoria cu scorul cel mai mare este aleasă drept eticheta finală a tweet-ului.

- **Acuratețea obținută**: După antrenarea modelului, am obținut o acuratețe de aproximativ 65% pe setul de date de testare.

- **Limitări**: Una dintre principalele limitări ale abordării clasice este faptul că Naive Bayes presupune independența între cuvinte, ceea ce nu este întotdeauna adevărat în cazul limbajului natural, în special în tweet-uri unde dependențele între cuvinte pot fi complexe.

4.4 Metoda - Novel approach

In cadrul abordării novel, am folosit modelul BERTweet-base. Acesta a fost antrenat pe setul de date preprocesat și evaluat utilizând o măsură standard de performanță, cum ar fi acuratețea.

- **Motivul alegerii BERTweet-base**: Acesta este un model pre-antrenat bazat pe arhitectura BERT, dar specializat pe limbajul informal, folosit pe Twitter. Am ales acest model datorită performanței sale in sarcini NLP aplicate pe date de pe social media.

- **Cum l-am folosit?**: Modelul implementat este o rețea neuronală care folosește ca backbone arhitectura vinai/bertweet-base, un model pre-antrenat specializat pe procesarea tweeturilor. Acesta returnează ca ieșire last.hidden.state, adică o reprezentare contextuală a fiecărui token din tweet. Din această ieșire, se extrage doar primul token ([CLS]), accesat prin hidden.state[:, 0], deoarece acesta conține, prin convenție, o agregare globală a întregului text — fiind astfel potrivit pentru sarcini de clasificare la nivel de propoziție.

Această reprezentare trece apoi printr-un strat Dropout, implementat ca dropout rate, care are rolul de a preveni overfitting-ul. Mai exact, acest strat dezactivează aleatoriu o parte din neuroni în timpul antrenării, forțând modelul să învețe reprezentări mai robuste și generalizabile.

În final, urmează un strat Linear, adică nn.Linear(self.backbone.config.hidden.size, 3), care acționează ca un clasificator. Acest strat transformă vectorul ascuns într-un vector de dimensiune 3, corespunzător celor trei clase: negative, neutral, positive. Ieșirea acestui strat reprezintă logits – adică scoruri brute care vor fi utilizate de funcția de pierdere CrossEntropyLoss pentru a calcula cât de departe sunt predicțiile de etichetele reale.

Împreună, aceste componente formează un model adaptat special pentru clasificarea de text scurt (precum tweet-urile), echilibrând între capacitatea de reprezentare a unui model mare ca BERTweet și regularizarea necesară pentru un dataset de dimensiuni moderate.

- **Acuratețea obținută**: Acuratețea obținută în urma antrenării a fost în jur de 80%. Comparând cu acuratețea Naive Bayes, se vede ca Bert are un avantaj în captarea semnificației contextuale și în generalizarea pe texte scurte și zgomotoase.

- **Comparatie SOTA**:Comparativ cu rezultatele din literatura de specialitate, unde modelele SOTA pentru clasificarea sentimentului în tweeturi se situează frecvent în intervalul 82–85%, performanța noastră este competitivă, dar ușor inferioară. Această diferență poate fi explicată prin faptul că nu am aplicat tehnici suplimentare precum fine-tuning avansat, augmentare de date sau ensemble-uri de modele.

- **Limitari**: În primul rând, mărimea relativ redusă a datasetului (28.000 tweeturi de antrenare) nu permite o generalizare robustă la variații mai subtile ale limbajului. În al doilea rând, în ciuda curățării datelor și eliminării duplicate, tweeturile pot conține formulări similare sau redundante semantic, ceea ce poate influența negativ diversitatea semantică din antrenare. De asemenea, distribuția dintre tweeturile pozitive, negative și neutre nu a fost perfect echilibrată.

5 Planuri de viitor

Un aspect pe care nu am reușit încă să îl explorăm este analiza erorilor. Considerăm că ar fi util să identificăm exemplele care sunt cel mai greu de clasificat și să înțelegem cauzele acestor dificultăți. Acest tip de analiză ne-ar putea ajuta să îmbunătățim preprocesarea sau să adaptăm arhitectura modelului pentru a face față mai bine unor situații specifice.

O posibilă direcție de extindere practică ar fi integrarea clasificatorului într-o aplicație care analizează automat mesajele de pe rețelele sociale sau din recenzii, pentru a detecta sentimentele exprimate de utilizatori. Sistemul ar putea fi util în domenii precum analiza de brand, servicii pentru clienți sau monitorizarea opiniei publice. De asemenea, poate constitui o bază pentru alte proiecte de procesare a limbajului natural, cum ar fi generarea de rezumate sau detectarea sentimentului de ură afișat în online.

6 Concluzie

Acest proiect ne-a oferit ocazia să înțelegem mai bine procesul complet al analizei de sentiment, de la preprocesarea textului brut până la antrenarea și evaluarea unor modele de clasificare. Am învățat să comparăm abordări clasice, precum Naive Bayes, cu metode moderne bazate pe rețele de tip transformer, observând avantajele și limitările fiecăreia.

Am apreciat partea practică a proiectului, mai ales lucrul efectiv cu date reale și aplicarea unor modele cunoscute din literatura de specialitate.

Limitări

Una dintre principalele limitări ale metodei moderne este necesarul ridicat de resurse computaționale. Antrenarea unui model bazat pe BERTweet necesită GPU-uri performante. De asemenea, viteza de antrenare și testare poate deveni o problemă atunci când se lucrează cu seturi de date mai mari sau când se dorește optimizarea hiper-parametrilor.

Un alt aspect important este legat de scalabilitate. Deși modelul funcționează bine pe un set de date moderat ca dimensiune, aplicarea sa la scară largă ar presupune o infrastructură mai robustă și costuri mai mari de procesare.

Modelul nostru este antrenat exclusiv pe texte în limba engleză. Astfel, performanța sa pe alte limbi ar fi slabă sau inexistentă, întrucât vocabularul și structura lingvistică diferă semnificativ. Pentru extinderea la alte limbi, ar fi necesare modele pre-antrenate specifice fiecărei limbi sau un model multilingv.

Declarație etică

Modelul nostru de analiză a sentimentelor ar putea fi folosit în moduri neetice, cum ar fi manipularea opiniei publice pe rețelele sociale sau clasificarea automată a utilizatorilor în scopuri comerciale sau politice fără consimțământul lor.

Nu am aplicat măsuri speciale de corectare a bias-ului, însă recomandăm folosirea responsabilă a modelului și testarea lui pe date variate și echilibrate, în special înainte de o eventuală aplicare reală.

References

Francesco Barbieri, Luis Espinosa-Anke, and Jose Camacho-Collados. 2021. Xlm-t: A multilingual language model for twitter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Abhinav Shrivastava. 2021. Sentiment analysis dataset. `https://www.kaggle.com/datasets/abhi8923shriv/sentiment-analysis-dataset`. Accessed: 2025-04-23.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.