Detectarea Recenziilor False în e-Commerce prin Analiză de Sentiment Bazată pe Aspecte

Autor: Bogdan Szasz
Email: Szasz.Te.Bogdan@student.utcluj.ro
Coordonator: Prof. Viorica Chifu

21 octombrie 2025

Necesitatea detectării recenziilor false în e-Commerce

Detectarea recenziilor false este imperativă pentru a menține integritatea platformelor de e-Commerce. Clientul modern se bazează pe recenzii pentru a lua decizii de cumpărare informate.

- **Impactul recenziilor false:** compromit eficiența recenziilor autentice, afectând atât consumatorii, cât și comercianții.
- **Motivația din spatele recenziilor false:** creșterea vânzărilor, denigrarea concurenței, sau manipularea percepției publice.
- **Provocările detectării:** recenziile false adesea par foarte autentice. Mai nou, LLMs pot genera recenzii false pe bandă rulantă[2].
- **Ce e de făcut?** Predominant, sunt folosite metode cu Natural Language Processing (NLP), Machine Learning (ML) și Deep Learning (DL)[3].

Strategii principale pentru detectare

Detectarea de recenzii false are trei trăsături discriminante de bază: conținutul textual[3], comportamentul autorului și tiparele lingvistice[4].

1. **Trăsături (con)textuale:** modele precum **DeBERTa** sau **BERT**[2], pentru detalii profunde(e.g., limbajul, alegerea cuvintelor și sentimentul general).
2. **Aspect-Based Sentiment Analysis (ABSA):** analizăm sentimentul legat de aspecte specifice ale produsului sau serviciului, dobândind o evaluare mai nuanțată[3].
3. **Reviewer Behavioral Features:** Analizăm date non-textuale, precum frecvența de ratinguri, date ale profilului de utilizator și metrici de activitate (acestea sunt mai greu de replicat de către spammeri și/sau boți)[4][1].

Detaliere a ABSA

Analiza tradițională pe sentiment are dezavantajul că ia în considerare doar sentimente la nivel de document, presupunând că datele sunt consistente și în celelalte aspecte ale recenziei.

- **Limitări alea SA pe document:** O recenzie poate avea mai multe aspecte cu sentimente diferite sau chiar în contradictoriu (e.g., *"I love the watch, but I hate the band"*). SA clasic va percepe aceste sentimente incorect sau cu ponderi exagerate, ducând la erori de clasificare.
- **Avantaje ABSA:** Prin analiza recenziei în toate aspectele și determinând sentimentul pentru fiecare (pozitiv, neutru sau negativ), platformele pot identifica arii bine definite de nivel de satisfacție. Studiille sugerează că ABSA ar fi mai eficientă în detectarea recenziilor false decât SA-urile clasice, pe document (**Ipoteza H1**)[3].

Metoda ABSA clasică

– Studiu de Hajek et al., 2023 –

Metoda ABSA (conform Hajek et al.)

Abordarea propusă de Hajek et al. combină datele output ale unei analize ABSA cu trăsături lingvistice și comportamentale pentru a detecta recenziile false. Acest studiu a folosit un set de date format din recenzii de pe Amazon [3].

- **Extragerea aspectelor:** o rețea nesupervizată de atenție (**ABAE**) a fost antrenată folosind **Word2Vec embedding** folosind $\approx$ 84 mil. recenzii Amazon.
  - Mecanismul de atenție folosit pune accentul pe cuvinte relevante și ia în considerare contextul acestora pentru o mai bună coerență.
- **Clasificarea Sentimentelor Aspectelor:** Modelul **W2VLDA** a fost lansat[3]. Acesta este parțial nesupervizat, folosește W2V embeddings și este ghidat de un model Latent Dirichlet Allocation (LDA) pentru a clasifica polaritatea sentimentelor ale fiecărui aspect.

Rezultate ABSA vs Tradițional SA

- **Performanțe:** Conform experimentelor realizate de Hajek et al. [3], ABSA a depășit abordarea tradițională pe documentla toate cele trei metrici verificate: *Comportamental, Lingvistic, Verified Purchase (VP)* (adunate, ABSA a obținut acuratețe de 82.89%).
- Pragul Wilcoxon signed-rank a confirmat **Ipoteza H1**, ABSA având rezultate semnificativ mai bune decât abordările tradiționale la $p < 0.05$
- Introducerea trăsăturii *Verified Purchase (VP)* s-a dovedit a fi esențială în atingerea performanțelor ridicate.

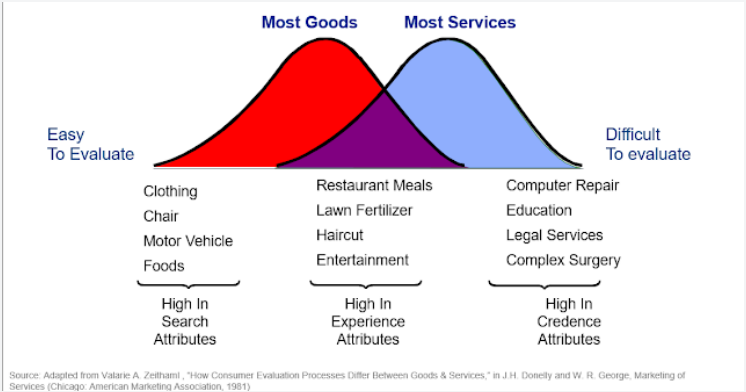
Factorul Verified Purchase (VP)

Studiul a demonstrat importanța atributului VP în îmbunătățirea acurateții.

- **VP** indică dacă acel produs a fost cu adevărat sau nu cumpărat de către autorul recenziei.
- **Impact:** Acuratețea detecțiilor a fost vast îmbunătățită pe toate categoriile de produs.
  - Modelul ABSA a obținut valoarea maximă a acurateții de 82.89% și un F1-Score de 0.829 după ce VP a fost inclus, depășind performanțele modelelor precedente.
- **Interpretare:** Recenziile autentice au la bază experiența consumatorului în folosirea produsului, iar lipsa VP poate duce la neclaritatea asupra existenței acesteti experiențe. Recenziile false sunt, prin definiție, fictive, iar un VP negativ poate evidenția acest fapt în cadrul efortului de detectare.

Impactul Tipului de Produs asupra Atributului VP

Eficiența atributului **VP** prezintă variații semnificative în funcție de tipul produsului vizat, acestea fiind clasificate utilizând frameworkul **Search, Experience, Credence (SEC)**.



The SEC Framework [3]

Impactul Tipului de Produs asupra Atributului VP — cont'd

Impactul diferențial al atributului VP:

- Când VP a fost **exclus**, produsele din clasa "Credence" au avut cea mai mică precizie în detectarea recenziilor false (66.07%)
- Când VP a fost **inclus** în calcul, produsele din clasa "Credence" au avut cea mai drastică îmbunătățire a preciziei ($\approx$ 19.64%), urmate de cele din clasa "Experience".
- **Concluzie:** Modelul de ABSA a dezvăluit ca doi factori sunt fundamentali pentru obținerea unor rezultate precise: **categoria de produs** și **atributul VP**, acestea având cea mai mare contribuție asupra produselor din clasele "Credence" și "Experience".

Modele textuale avansate

– Studiu de Geetha et al., 2025 –

Modele Avansate de Interpretare a Textului: MBO-DeBERTa — Context

Modele curente de ML nu reușesc să interpreteze cu acuratețe textele actuale, fiind copleșite de volumul în creștere de recenzii[2].

Abordarea modelului **MBO-DeBERTa** își propune să rezolve acest impediment prin combinarea unui transformator textual performant cu algoritmi de optimizare.

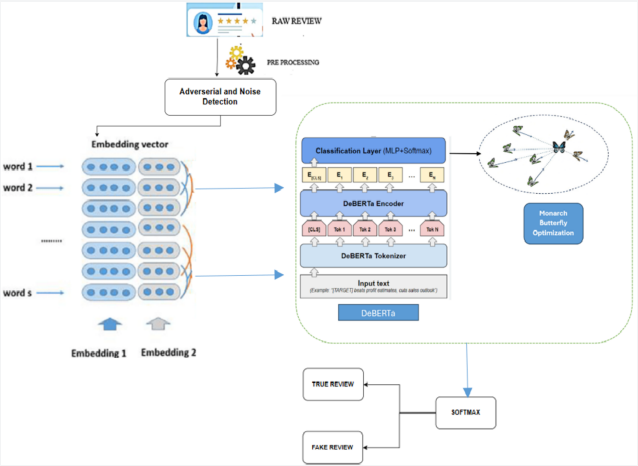


Diagrama functionala a modelului MBO-DeBERTa, propus de Geetha et al. [2]

MBO-DeBERTa — DeBERTa clasic

DeBERTa (Decoding-enhanced BERT with disentangled attention) a fost ales pentru capacitatea avansată de modelare a limbii.

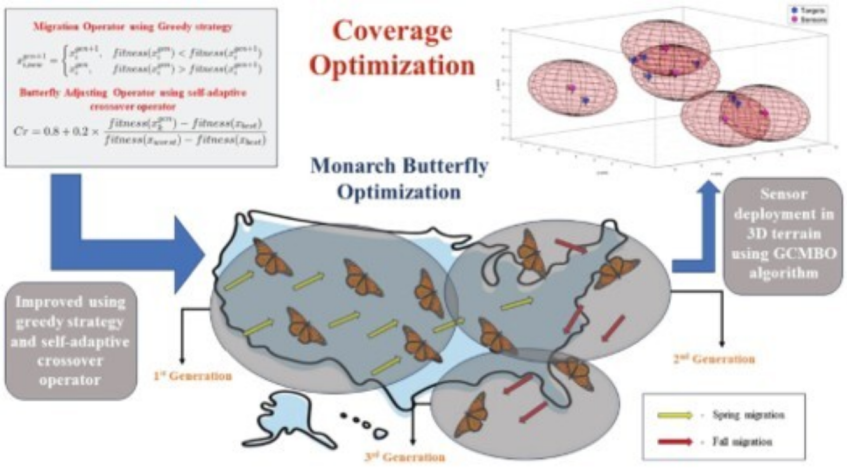
- Folosește un **disentangled attention mechanism** care codifică conținutul ($\rightarrow$ token embeddings) și poziția ($\rightarrow$ relative position embeddings) [2]. Asta permite modelului să interpreteze enunțurile mai precis și să captureze cu succes relații între cuvinte și pozițiile lor, conturând contextul.
- DeBERTa este un model competent de detectare a recenziilor false, observând construcțiile lingvistice ne-naturale și subliniind exagerările nelalocul lor.
- O problemă care persistă este **scalarea** la volum de date și mai mare (deep learning models scalability challenges).

MBO-DeBERTa — Propunerea studiului

Pentru a îmbunătăți performanțele DeBERTa, studiul Geetha et al. a introdus algoritmul de optimizare **Monarch Butterfly Optimization (MBO)**.

- **MBO** este un algoritm meta-euristic inspirat din natură, folosit pentru a optimiza procedeul de ajustare a hiperparametrilor și selecția de trăsături a modelului. Acesta imită tiparele migrării monarhilor, balansând căutările globale (operația de migrație) cu căutările locale (operatorul de ajustare a fluturelui), pentru a găsi soluția cea mai bună[2].
- Modelul propus, **MBO-DeBERTa**, este un model care îmbunătățește capacitatea de a diferenția între trăsăturile care se suprapun între recenziile autentice și cele false. Acest fapt se implementează prin combinarea datelor poziționale din DeBERTa și îmbunătățind fitness-ul rezultatului parțial cu MBO.
- MBO-DeBERTa a obținut o acuratețe a clasificării de 98% în detectarea de recenzii false.

MBO-DeBERTa — MBO ilustrat



Ilustrație a funcționării algoritmului Monarch Butterfly Optimization, utilizat în studiul Geetha et al. [2]

MBO-DeBERTa — Analiză a rezultatelor

MBO-DeBERTa a fost testat pe trei dataseturi publice, vaste și diverse:

1. Amazon (21.000 recenzii)
2. Fake Review (40.000 recenzii)
3. Deceptive Opinion Spam (1.600 recenzii augmentate la 16.000)

Metrici de performanță:

- Acuratețe: 98%
- Precizie: 98%
- Recall (Sensibilitate): 97
- F1-Score: 97%

Acest model a depășit la performanță alte modele de transformare și detectare, precum BERT, RoBERTa (Robustly-optimized BERT approach) și XLNET, iar folosirea algoritmului MBO a obținut rezultate mai bune comparativ cu alți algoritmi (Harris Hawks, Grey Wolf) [2]

MBO-DeBERTa — Analiză a rezultatelor cont'd

- $Accuracy = \dfrac{TP + TN}{TP + FP + FN + TN}$
- $Precision = \dfrac{TP}{TP + FP}$
- $Recall = \dfrac{TP}{TP + FN}$
- $F1Score = \dfrac{2 \times Precision \times Recall}{Precision + Recall}$

Aplicații în lumea reală: Modelul a fost testat cu success pe date nemaiîntâlnite (recenzii de pe Myntra și Amazon) și a dat dovadă de eficiență și precizie în acest scenariu realist.

Analiză bazată pe trăsături comportamentale –

Studiu de Ghulam et al., 2025 –

Analiză bazată pe trăsături comportamentale

Spre deosebire de celelalte studii, cu abordări predominant textuale ale subiectului, studiul realizat de Ghulam et al. explorează recenziile scrise în Urdu cu caractere romanizate.
Urdu este încă săracă din punct de vedere al literaturii de specialitate în ceea ce privesc normele de construcție ale enunțurilor formalizate pentru ML (după care s-ar antrena modelele de detecție).
Astfel, studiul de față pune accentul pe integrarea trăsăturilor comportamentale ale utilizatorului pentru a determina autenticitatea recenziei.

- Rezultatele experimentale inițiale (*Setting-III*)[4] evidențiat faptul că excluderea **Review Textual Features (RTF)** duce la o stabilizare și creștere a valorilor acurateții și F1-Score-ului. Putem concluziona, strict empiric, RTF este non-informativ în determinarea falsității unei recenzii.
- Astfel, trăsăturile pe care s-a pus accentul sunt **Review Lingual Features (RLF)** și **User Behavioral Features (UBF)**

Proiectarea modelului bazat pe trăsături comportamentale

Studiul a descris multiple metode de modelare și inginerie a RLF și UBF, folosite în îmbunătățirea modelului. Aceste metode au fost neglijate în studiile precedente pe Urdu.

RLF: Trăsături derivate din aspectele lingvistice ale textului recenziei.

1. **Sentiment Score:** Maparea sentimentului recenziei (i.e., *Pozitiv, Negativ, Neutru*) la valori numerice;
2. **Is Review Singleton (IRS) Score:** Trăsătură binară care indică dacă utilizatorul a postat un singur review sau mai multe;
3. **Content Similarity: TF–IDF** pentru a determina similaritățile dintre recenzii și a evidenția recenziile făcute „după șablon";

UBF: Trăsături extrase din profilul de utilizator și din interacțiunile sale cu platforma.

1. **Activity Window:** cât timp a fost un user activ;
2. **Rating Deviation:** cât de mult a deviat recenzia investigată față de ratingul mediu al produsului și față de media recenziilor precedente ale autorului;
3. **Extremity Score:** dacă recenzia este extremă (e.g., *1*⋆ sau *5*⋆);

Performanțele modelului comportamental

Studiul compară modele de ML și DL pe trăsăturile RLF și UBF extrase.

- **ML vs. DL**: modelele ML au avut performanțe mai bune decât DL, datorită datasetului mai mare al DL-urilor data-hungry.
- **Strategie de resampling:** Din cauza dezechilibrului mare dintre clase, au fost testate mai multe strategii: **Random Under Sampling (RUS)** a dat rezultate mai bune decât **Random Over Sampling (ROS)** sau **Synthetic Minority Oversampling Technique (SMOTE)**.
- **Performanță optimizată:** modelele de gradient boosting (în special **XGB**) au atins cele mai bune rezultate combinat cu RUS și Recursive Feature Elimination pentru selectarea de trăsături. Acest model a atins un apogeu al preciziei de 81%, F1-Score de 81%, crescând cu 3% față de studiul de bază.

DistilBERT: Eficientizarea transformatoarelor pe text

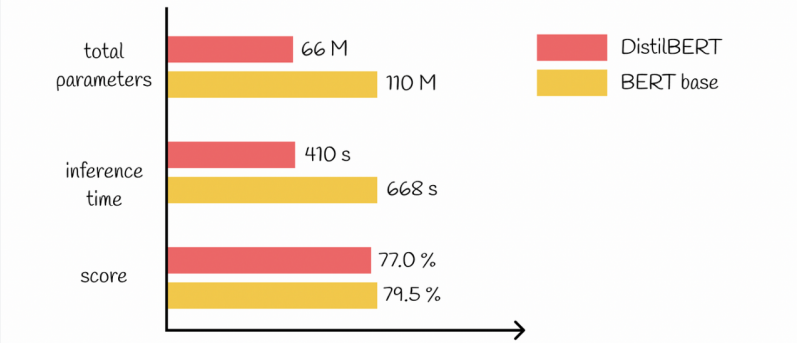
– Studiu de Lovesh et al., 2025 –

DistilBERT: Modelare explicabilă

– Studiu de Shajalal et al., 2024 –

DistilBERT — Context

În timp ce modele complexe de transformare și identificare, precum DeBERTa, obțin acuratețe ridicată, scalabilitatea și reziliența la volume mari de date au rămas problematice. Astfel, abordarea studiului realizat de Lovesh et al. [1] propune modelul **DistilBERT** ca o potențială soluție.



Comparația dintre DistilBERT și modelul BERT de bază [1]

DistilBERT — Proprietăți

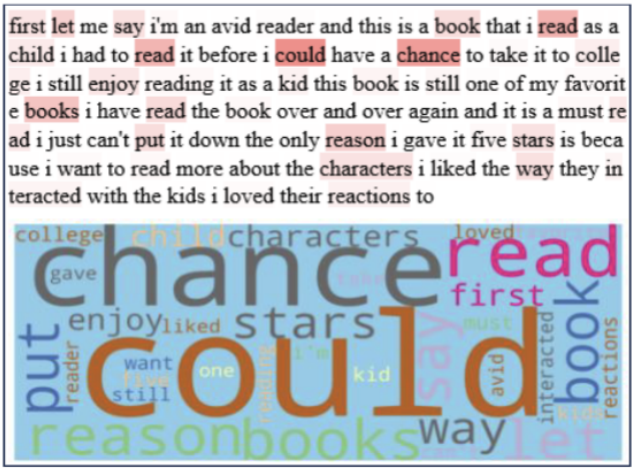
- **DistilBERT** este o variantă "distilată" a modelului BERT, folosind procedeul de *knowledge distillation*. Astfel, devine cu 40% mai *light* și cu 60% mai rapid, menținând în continuare 97% din abilitatea de înțelegere a limbii oferită de BERT.
- Folosește 6 layers în loc de 12, și are 66 de milioane de parametri (mult mai light).
- Într-o comparație directă folosind *Amazon Fake Review Labelled Dataset* (40,000 rec.), DistilBERT a obținut 98% acuratețe, un rezultat mai bun decât modele clasice de ML (SVM $\rightarrow$ 92%, RF $\rightarrow$ 90%, LR optimizat cu Word2Vec embeddings $\rightarrow$ 90%)
- Aceeași performanță, eficientizare a resurselor: DistilBERT a obținut acurateţe de 98% când a fost comparat cu DeBERTa (98%) și XLNet (97%) pe același set de date.

DistilBERT — Deschiderea Black Box-ului

Natura complexă a modelelor de transformare și a celor de DL face ca ele să funcționeze ca **Black Boxes**. Deciziile lor sunt dificil de înțeles și interpretat, mai ales într-un scop ca detectarea recenziilor false[5].

- **XAI:** Tehnicile de **eXplainable AI** sunt necesare pentru a putea interpreta mai uman rezultatele identificărilor.
- **Layer-wise Relevance Propagation (LRP)** a fost adoptat pentru a putea înțelege rațional deciziile modelelor de DL:
  - LRP dezvăluie conținutul Black Box-ului, propagând scorul de relvanță al outputului spre layer-ele precedente de input.
  - Astfel, putem pondera fiecare nivel de procesare intermediară al inputului după relevanță. Explicabilitatea acestui proces este reprezentat prin sublinierea cuvintelor relevante în text[5].

DistilBERT — Sublinierea cuvintelor relevante folosind LRP



Cuvintele relevante sunt subliniate/mărite în word map în vederea vizualizării raționamentului de detectare [5]

Analiză critică empirică — Shajalal et al.

O evaluare empirică a fost realizată pentru a verifica eficiența explicațiilor generate de LRP în vederea ajutării oamenilor în detectarea de recenzii false.

- **Exemplu de output LRP:** Pentru recenziile negative, LRP a subliniat adjectivele specifice exagerărilor (e.g., *awesome, brilliant, disgusting, awful*) sau afirmațiilor redundante (e.g., *recommend, insist, unbelievable*).
- **Feedback-ul utilizatorilor:** Majoritatea participanților la experiment au observat că doar sublinierea cuvintelor relevante nu este suficientă pentru a înțelege contextul deciziei luate de model în privința recenziilor false. De ce?
  - **Contextul contează**
  - **Vocabular similar între recenziile false și negative**
- **Concluzie:** LRP este eficient în alte scopuri de NLP, însă pentru detectarea recenziilor false explicațiile erau reductive, făcând referire doar la cuvinte, eliminând aspecte fundamentale precum autenticitatea sentimentelor sau punerea în context.

COMPARAȚII — Utilitatea proprietăților

Comparativ: abordări și concluzii

Abordare	Focalizare pe trăsături	Componente / Modele	Insight critic
Hajek et al. (ABSA)	Textuale (sentiment pe aspecte) + Comportamentale (VP)	W2VLDA, ABAE, atribut VP	ABSA depășește analiza pe document; VP este critic, mai ales pentru produse de tip *Experience/Credence*.
Geetha et al. (MBO–DeBERTa)	Trăsături textuale contextuale profunde	DeBERTa, optimizare MBO	Acuratețe SOTA obținută prin combinarea encoderului contextual cu optimizarea metaeuristică.
Ghulam et al.	Comportamentale (UBF) + Linguale (RLF)	XGB (Gradient Boosting), eșantionare RUS	Conținutul textual (RTF) a fost neinformativ; trăsături UBF precum *Activity Window* și *Rating Deviation* sunt superioare, mai ales pentru limbi cu resurse reduse.

COMPARAȚII — Raportul eficiență/cost al modelelor

Comparativ: performanțe și complexitate

Model / Studiu	Acuratețe maximă (%)	Bază de trăsături	Observații computaționale
Hajek et al. (ABSA/DNN + VP)	82.89	ABSA + Lingvistice + Comportamentale (VP)	Complexitate moderată; necesită feature engineering extins și modelare pe aspecte.
DistilBERT	98	Textuale (Contextual Embedding)	Arhitectură eficientă și ușoară (≈40% mai mică decât **BERT**); performanță competitivă.
MBO–DeBERTa	98	Trăsături contextuale optimizate (Deep Contextual)	Performanță ridicată; model complex **DeBERTa** optimizat prin **MBO**. Necesită resurse computaționale semnificative.
Ghulam et al. (XGB + RFE)	81	Comportamentale (UBF) + Linguale (RLF)	Abordare **ML** eficientă, preferată pentru viteză și amprentă redusă de memorie comparativ cu **DL**.

Modelele **MBO-DeBERTa** și **DistilBERT** au cea mai mare acuratețe în dataset-urile în engleză, modelele de ML (**XGB**) excelează când aspectele comportamentale sunt prioritare

CONCLUZII

Concluzii

Detectarea automată a recenziilor false a evoluat semnificativ, trecând de la ingineria clasică de trăsături la modele contextuale complexe și puternice. **Constatări cheie:**

- **Context is King:** Modelele avansate de tip **MBO–DeBERTa** și **DistilBERT** obțin acurateţ̦e superioară prin codificarea profundă a informației contextuale.
- **ABSA Matters:** Analiza Aspect-Based Sentiment oferă granularitatea necesară dincolo de analiza la nivel de document, mai ales atunci când este combinată cu atributul **Verified Purchase (VP)**.
- **Behavioral Features are Robust:** Pentru anumite contexte sau limbi cu resurse limitate, combinarea trăsăturilor **User Behavioral (UBF)** și **Review Lingual (RLF)** cu clasificatori **ML** (precum **XGB**) s-a dovedit extrem de eficientă și cu cost computațional redus.
- **The Need for Transparency:** Deși performanțele modelelor **DL** sunt ridicate, metodele actuale de explicare a deciziilor rămân limitate. Este necesară o cercetare suplimentară pentru a comunica eficient *de ce* o recenzie este clasificată drept falsă.

Vă mulțumesc pentru atenție!

Bibliografie I

[1] L. Anand, H.-N. Goh, C.-Y. Ting, and A. Quek. Identifying fraud sellers in e-commerce platform. *JOIV: International Journal on Informatics Visualization*, 9(2):761–769, 2025.

[2] S. Geetha, E. Elakiya, R. S. Kanmani, and M. K. Das. High performance fake review detection using pretrained deberta optimized with monarch butterfly paradigm. *Scientific Reports*, 15(1):7445, 2025.

[3] P. Hajek, L. Hikkerova, and J.-M. Sahut. Fake review detection in e-commerce platforms using aspect-based sentiment analysis. *Journal of Business Research*, 167:114143, 2023.

[4] N. Mughal, G. Mujtaba, M. H. Mughal, A. Manaf, and Z. Kamangar. Fake reviews detection on e-commerce websites using novel user behavioral features: An experimental study. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(9):1–44, 2025.

[5] M. Shajalal, M. Atabuzzaman, A. Boden, G. Stevens, and D. Du. What matters in explanations: Towards explainable fake review detection focusing on transformers. *arXiv preprint arXiv:2407.21056*, 2024.