FMI

Sistem de analiză a datelor nestructurate de tip text

Analiza sentimentului

COORDONATOR
Lect.Dr. Monica Sancira

ABSOLVENT
Cristian Marius Pal

18 iunie 2023

CONȚINUT PREZENTARE

- Motivație
- Obiective
- Starea de artă
- Arhitectura sistemul
  - Descrierea modelului propus
  - Descrierea și preprocesarea datelor
  - Descrierea implementării modelului propus
- Soluții obținute
- Concluzii

Motivație

- Analiza sentimentului poate fi utilizată pentru identificarea tendințelor și opiniilor publice.
- Analiza sentimentului din datele nestructurate este crucială, deoarece majoritatea datelor de pe internet sunt nestructurate (aprox. 90%).
- Importantă în cercetare deoarece:
  - Volumul de date se amplifică progresiv cu trecerea zilelor!
  - Există necesitatea de a dezvolta tehnici inovatoare capabile de a gestiona eficient volumul crescător de date.

Obiective

Obiectivul sistemului

Recunoaștere și categorizare eficientă a sentimentelor exprimate în textul utilizatorilor.

Acuratețea sistemului

Comparabilă sau superioară față de standardele actuale în analiza sentimentelor.

Analiza algoritmului PSO

Eficacitate în analiza sentimentului, performanță și rezultate

Contribuție la cercetare

Progrese în domeniul analizei sentimentelor și dezvoltare de metode mai eficiente și precise de recunoaștere și categorizare a emoțiilor și opiniilor umane în text.

Soluții similare I

- Studiile de natură similară:
  - Hybrid Feature Extraction[KKHD18]
  - Attention-emotion-enhanced CLSTM[HLY+22]
  - SVM with Informational Gain[MRI+20]
  - Lexicon with MultiLayer Perceptron[SZX+20]
  - Sentimental analysis on user's reviews using BERT[SL22]

Soluții similare II

- Hybrid Feature Extraction
  - Extragerea caracteristicilor se realizează în etape paralele
    - bazate pe învățare automată
    - bazate pe lexic
  - Metoda de extragere a caracteristicilor bazate pe învățare automată folosește tehnica cunoscută sub numele de "Bag of words".
  - Metoda de extragere a caracteristicilor bazate pe lexic, care extrage 4 caracteristici diferite dintr-o recenzie:
    - Numărul de cuvinte pozitive (PC)
    - Numărul de cuvinte negative (NC)
    - Numărul de conotații pozitive (PCC)
    - Numărul de conotații negative (NCC)
  - Ideea cheie a acestui articol este combinarea caracteristicilor bazate pe învățare automată cu caracteristicile extrase prin utilizarea unui lexic.

Soluții similare III

- Attention-emotion-enhanced CLSTM
  - Emotional Intelligence (EI) - capabil să capteze starea emoțională dintr-un text.
  - Mecanism de captare a atentiei (Attention Layer)
    - Controlează modul în care informația este procesată și propagată prin rețea.
    - Generează o pondere (pondere de atenție) prin intermediul unei funcții de atenție.
    - Ponderile de atenție sunt înmulțite cu valorile corespunzătoare din secvența de intrare.
  - Convolutional LSTM.
  - Combinând mecanismul de captare a atenției cu inteligența emoțională permite modelului (AEC-LSTM) să depășească limitele impuse rețelelor RNN și îmbunătățind capacitatea de învățare.

Soluții similare IV

- SVM with Informational Gain
  - Folosește o combinație între SVM și Informational Gain.
  - Informational gain este folosit ca metodă de selectare a caracteristicilor.
  - SVM este folosit ca metodă de clasificare.
  - Combinarea celor două metode a avut ca rezultat creșterea performanței cu până la 2%.
- Lexicon with MultiLayer Perceptron
  - Mecanismul lexic
    - Calculează orientarea semantică a fiecărei recenzii folosind o bază lexicală (WordNet).
  - Folosește o rețea neuronală multistrat.
  - Combinarea celor două metode de extragere a caracteristicilor a condus la o îmbunătățire a performanței în ceea ce privește rețelele neuronale.
- BERT (Bidirectional Encoder Representations from Transformers)
  - Reprezintă textul de intrare ca o secvență de simboluri și codifică aceste simboluri (tokens) pentru a captura relațiile de context.

Soluții similare V

- Această codificare se face folosind arhitectura Transformers, care permite să proceseze secvențe de intrare într-un mod bidirectional
- BERT creează reprezentări contextuale a textului de intrare, acestea putând fi ulterior clasificate cu ajutorul unui clasificator.

Arhitectura sistemul I

Descrierea modelului propus

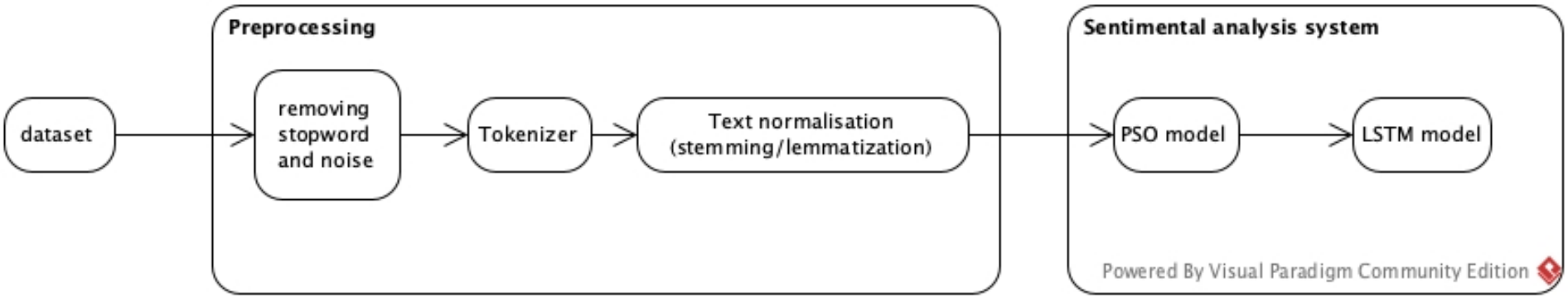

Figura: Arhitectura modelului propus

- Sistem organizat, compus din două părți majore:
  - Preprocessing
  - Sentimental Analysis System/Model

Arhitectura sistemul II

Descrierea modelului propus

- Preprocessing compus din funcții elementare de NLP
  - Stopword removal
  - Tokenizer
  - Text normalisation (Stemming/Lemmatisation)
- Sentimental Analysis System/Model compus din:
  - model LSTM (Long Short Term Memory)
    - Poate determina dependențele între cuvinte în contextul unui text
  - model PSO (Particle Swarm Optimisation)
    - Poate identifica caracteristicile cheie pentru clasificarea sentimentelor.

Arhitectura sistemul I

Descrierea și preprocesarea datelor

- Setul de date utilizat este setul de date IMDB oferit de Stanford AI repository
- Setul de date conține 50.000 de recenzii de filme, împărțite în mod egal
- Setul de date este pregătit pentru analiza prin aplicarea unor tehnici de curățare a textului prin:
  - Eliminarea cuvintelor de oprire
  - Tokenizare
  - Eliminarea punctuației
  - Eliminarea etichetelor HTML
  - Normalizarea textului
    - Stemming
    - Lemmatization

- Tehnica de padding este utilizată pentru a asigura dimensiuni egale ale secvențelor de token-uri și pentru a evita pierderea informațiilor importante.
- Se estimează dimensiunea secvențelor prin $\mu(t) + 2 * \sigma(t)$
  - t reprezintă o listă de dimesiuni pentru fiecare secvență în urma procesului de tokenizare.

Arhitectura sistemul

Descrierea implementării modelului propus

- Modelul LSTM
  - Un strat de încorporare (embedding layer)
  - Două straturi LSTM cu 32 unități respectiv 8 unități
  - Strat Dens cu 2 unități (pentru clasificarea sentimentului)
- Modelul PSO
  - Limitarea vitezei, restricționarea acesteia in domeniul $[0, 1]$ astfel:

$$v(x) = \begin{cases} 1 & \text{dacă } x > 0 \\ 0 & \text{altfel} \end{cases}$$

  - Se aplică restricția ca pozițiile particulelor să rămână în intervalul $[0, 1]$ utilizând funcția *clip* din biblioteca *NumPy*.

- Se propune un set de experimente pentru a determina parametri pentru algoritmul PSO.

Experiment	**Parametri**				
	w	c_1	c_2	it	p
1	0.4	1.2	1.2	5	5
2	0.6	1.4	1.4	5	5
3	0.8	1.8	1.8	5	5
4	0.6	1.4	1.4	10	10
5	0.6	1.4	1.4	10	15
6	0.6	1.2	1.4	10	15

Tabela: Parametri utilizati pentru fiecare experiment

- Rezultatele obținute au demonstrat că:
  - Utilizarea metodei de lematizare pentru setul de date a condus la o acuratețe de 89%, folosind 283 din cele 287 de caracteristici disponibile.
  - Utilizarea metodei de stemming pentru setul de date a condus la o acuratețe de 86%, folosind 214 din cele 216 de caracteristici disponibile.
  - Utilizarea setului de date fără metode de normalizare a condus la o acuratețe de 89%, folosind 281 din cele 282 de caracteristici disponibile.
  - Parametrii optimi identificați pentru algoritmul Particle Swarm Optimization (PSO) sunt:
    - Valoarea w este de 0.6
    - Valoarea c_1 este de 1.2
    - Valoarea c_2 este de 1.4
    - Numărul de iterații este de 10
    - Dimensiunea populației este de 15

Concluzii

- Concluzii
  - Metoda de lematizare s-a dovedit eficientă în experiment și a obținut performanțe mai bune pe setul de validare. Această alegere s-a bazat pe acuratețea echilibrată între seturi.
  - Metoda de stemming poate simplifica prelucrarea textului, dar poate duce la pierderea informațiilor semantice.
  - Selecția de caracteristici (feature selection) joacă un rol important în îmbunătățirea performanțelor modelului.
  - Reducerea dimensiunii setului de caracteristici prin selecția de caracteristici poate elimina redundanța și zgomotul din date și permite concentrarea pe aspectele relevante ale textului.
- Direcții viitoare
  - Optimizarea ponderilor în raport cu caracteristicile selectate poate duce la o mai bună eficiență în utilizarea resurselor și la rezultate îmbunătățite în analiza textului.
  - Complexitatea algoritmului de optimizare și trade-off-urile între timpul de execuție și performanțele obținute trebuie luate în considerare în lucrările viitoare.

Întrebări?

Referințe I

Faliang Huang, Xuelong Li, Changan Yuan, Shichao Zhang, Jilian Zhang, and Shaojie Qiao.
Attention-emotion-enhanced convolutional lstm for sentiment analysis.

IEEE Transactions on Neural Networks and Learning Systems, 33(9):4332–4345, 2022.

H M Keerthi Kumar, B S Harish, and H. Darshan.
Sentiment analysis on imdb movie reviews using hybrid feature extraction method.
International Journal of Interactive Multimedia and Artificial Intelligence, InPress:1, 01 2018.

Referințe II


Reza Maulana, Panny Agustia Rahayuningsih, Windi Irmayani, Dedi Saputra, and Wanty Eka Jayanti.
Improved accuracy of sentiment analysis movie review using support vector machine based information gain.
Journal of Physics: Conference Series, 1641(1):012060, nov 2020.


B. Selvakumar and B. Lakshmanan.
Sentimental analysis on user's reviews using bert.
Materials Today: Proceedings, 62:4931–4935, 2022.
International Conference on Innovative Technology for Sustainable Development.


Zeeshan Shaukat, Abdul Ahad Zulfiqar, Chuangbai Xiao, Muhammad Azeem, and Tariq Mahmood.
Sentiment analysis on imdb using lexicon and neural networks.
SN Applied Sciences, 2, 02 2020.