PHÁT HIỆN ĐÁNH GIÁ GIẢ DỰA TRÊN MẠNG NGỮ NGHĨA TÍCH HỢP THÔNG TIN KHÍA CẢNH

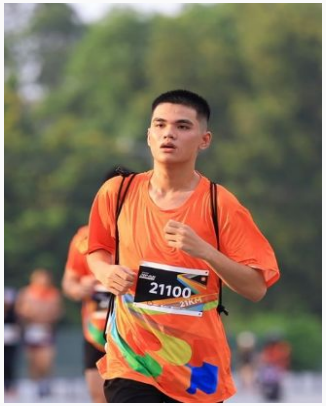
Nguyễn Hữu Đặng Nguyên - 23521045

Đặng Quốc Cường - 23520192

Tóm tắt

- Lớp: CS519.Q11.KHTN
- Link Github: https://github.com/TwinHter/CS519.Q11.KHTN
- Link YouTube video: https://youtu.be/RfCV91t4ecc
- Ảnh + Họ và Tên của các thành viên

Nguyễn Hữu Đặng Nguyên - 23521045



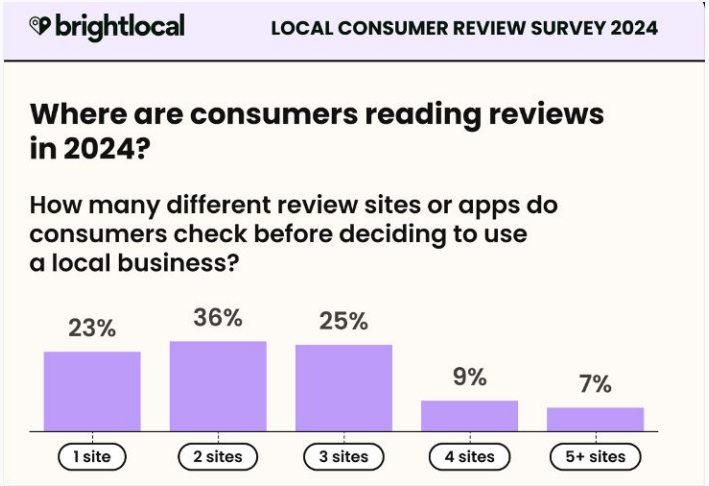
Đặng Quốc Cường - 23520192

Giới thiệu bài toán

Online reviews đóng vai trò quan trọng trong thương mại điện tử. Kéo theo đó sự gia tăng của **các đánh giá giả (fake reviews)** gây nhiều ảnh hưởng tiêu cực.

Fake reviews được viết tinh vi hoặc sinh bằng AI càng phát triển → phát hiện khó khăn.

⇒ Phát triển mô hình **phát hiện đánh giá giả** chính xác và hiệu quả trở nên cấp thiết, có ý nghĩa quan trọng.



Nguồn ảnh: BrightLocal Survey

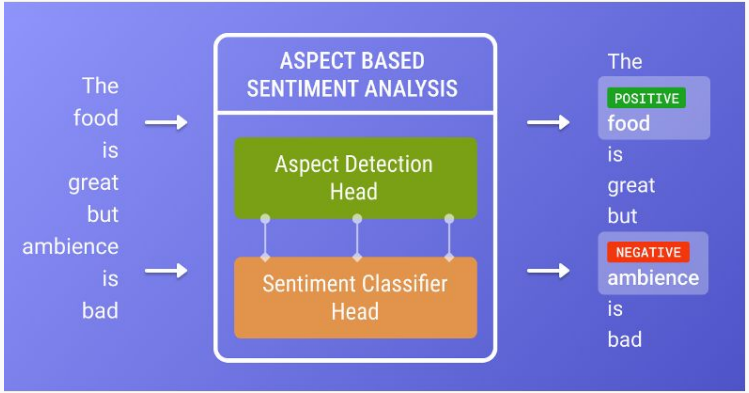
Giới thiệu bài toán

Hướng tiếp cận ban đầu dựa trên đặc trưng thủ công (độ dài câu, tần suất từ khóa,...). Chúng phụ thuộc mạnh vào miền, khó áp dụng cho tổng quát.

Sự ra đời của **Transformer** giúp học biểu diễn ngữ nghĩa sâu. Transformer thường xử lý nội dung như một khối ngữ nghĩa tổng thể ⟶ khó nắm bắt **mối quan hệ giữa các khía cạnh cụ thể trong cùng một review**.

Graph Neural Networks (GNN) trở thành xu hướng:

- Reviewer, review, item có quan hệ phức tạp dạng đồ thị
- Có thể phát hiện pattern vượt ngoài phạm vi ngôn ngữ.



Phần lớn các mô hình GNN hiện nay không tận dụng nhiều thông tin về khía cạnh cụ thể, yếu tố quan trọng trong phân tích mức độ tự nhiên của review.

Nguồn ảnh: https://www.gautamnaik.com/

Các phương pháp liên quan

Spam Review Detection with Graph Convolutional Networks (CIKM 2019)

- Xây dựng đồ thị User-Review-Item.
- Học embedding cho User và Review.
- Nếu User kết nối với nhiều Review giả, User đó sẽ có tính chất giả mạo thông qua quá trình lan truyền của GCN.

PC-GNN - *Pick and Choose Graph Neural Network* **(WWW 2021)**

- **Vấn đề:** Trong thực tế, review giả chỉ chiếm số lượng rất nhỏ (<10%), review thật chiếm đa số..
- **Phương pháp:** Chọn các node nghi ngờ là spam để train kỹ hơn và giảm bớt các node bình thường.

ABSA - Aspect based sentiment analysis: Thư viện **PyABSA** cung cấp nhiều pretrain model.

Heterogeneous Graph Transformer (HGT)

- Là đồ thị có từ 2 loại node hoặc từ 2 loại cạnh trở lên.
- Mô hình học cách giải thích, phân biệt, và định trọng số từng loại quan hệ khác nhau trong đồ thị

Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence (Sun et al., NAACL 2019)

- Đưa thông tin khía cạnh một cách tường minh vào mô hình giúp định hướng quá trình học biểu diễn ngữ nghĩa
- Tích hợp aspect không làm loãng thông tin, mà ngược lại làm giàu representation và cải thiện hiệu quả dự đoán

Mục tiêu

Xây dựng được mô hình phát hiện đánh giá giả dựa trên đồ thị dị thể (Heterogeneous Graph) với khía cạnh và cảm xúc được tích hợp vào đồ thị để khai thác: ngữ nghĩa văn bản, người dùng và mối quan hệ giữa các khía cạnh.

Khảo sát mức độ đóng góp của thông tin khía cạnh đối với hiệu quả của mô hình đồ thị qua việc so sánh kết quả khi có và khi không có các thông tin này.

Thực nghiệm và so sánh mô hình đề xuất với một số phương pháp hiện có trên bộ dữ liệu chuẩn, nhằm đánh giá hiệu quả và khả năng áp dụng.

Research Question: Tích hợp thông tin khía cạnh vào mô hình mạng đồ thị ngữ nghĩa có giúp cải thiện hiệu quả phát hiện đánh giá giả so với các phương pháp hiện có hay không?

Dữ liệu và tiền xử lý

- **Tập dữ liệu:**
  - User (người viết review).
  - Bài viết đánh giá
  - Review metadata: rating, timestamp, Item (kinh doanh, sản phẩm …).
  - Label: spam / non-spam

- **Aspect terms trích xuất từ nội dung review bằng:**
  - Rule-based patterns hoặc từ các mô hình trích xuất khía cạnh được huấn luyện sẵn. (pyABSA)
  - Sentiment theo từng aspect (tích cực/ tiêu cực).
  - Đưa aspect về các nhãn tổng quát.
- **Aspect embedding:** embedding từ BERT / GloVe

Phương pháp

- **Bước 1: Trích xuất và chuẩn hóa Aspect**
  - Trích xuất các aspect terms và sentiment.
  - Chuyển đổi aspect về nhóm khái niệm tổng quát (service_pos, food_neg, …)
  - Tạo aspect nodes để gắn với review node.
- **Bước 2: Xây dựng Đồ thị**
  - Các node: User, Item, Review, Aspect.
  - Các quan hệ: User → Review, Review → Item, Review → Aspect, Item → Aspect…
- **Bước 3: Biểu diễn Node**
  - User / Item: đặc trưng hành vi đơn giản + embedding học được.
  - Review: embedding từ mô hình ngôn ngữ như BERT.
  - Aspect: dùng embedding của token (GloVe) hoặc tổng hợp embedding.
- **Bước 4: Heterogeneous Graph Neural Network**
  - Sử dụng mạng đồ thị dị thể để xử lý nhiều loại nút và quan hệ, đồng thời học trọng số cho từng loại quan hệ.
- **Bước 5: Lớp Phân Loại**
  - Embedding của review từ đồ thị → MLP để phân loại.

Kết quả dự kiến

Phương pháp đánh giá:

- Sử dụng các dataset chuẩn cho bài toán Fake Review Detection (VD: YelpChi)
- Đánh giá bằng: Accuracy, Precision, Recall, F1-score.
- So sánh với kết quả của các nghiên cứu liên quan và các baseline.
- Ablation study để xác định đóng góp của các thành phần, đặc biệt là thành phần khía cạnh trong câu.

Kết quả dự kiến: Bảng so sánh chi tiết hiệu năng mô hình dựa trên ablation study, cùng phân tích hướng mở rộng.

Được kỳ vọng có thể ứng dụng vào thực tiễn như phát hiện đánh giá giả theo thời gian thực.

Tài liệu tham khảo

[1] Shebuti Rayana, Leman Akoglu: Collective Opinion Spam Detection: Bridging Review Networks and Metadata. KDD 2015: 985-994

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171-4186

[3] Ao Li, Zhou Qin, Runshi Liu, Yiqun Yang, Dong Li: Spam Review Detection with Graph Convolutional Networks. CIKM 2019: 2703-2711

[4] Chi Sun, Luyao Huang, Xipeng Qiu: Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. NAACL-HLT (1) 2019: 380-385

[5] Ziniu Hu, Yuxiao Dong, Kuansan Wang, Yizhou Sun: Heterogeneous Graph Transformer. WWW 2020: 2704-2710

[6] Rami Mohawesh; Shuxiang Xu; Son N. Tran; Robert Ollington; Matthew Springer; Yaser Jararweh: Fake Reviews Detection: A Survey. IEEE Access, vol. 9, pp. 65771-65802, 2021

[7] Heng Yang, Chen Zhang, Ke Li: PyABSA: A Modularized Framework for Reproducible Aspect-based Sentiment Analysis. CIKM 2023: 5117-5122