# Phân loại và khuyến nghị bài báo khoa học dựa trên phân tích ngữ cảnh của các trích dẫn

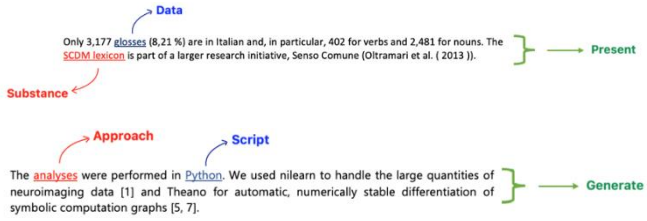
Nguyễn Minh Tiến[1], Hồ Anh Khôi[2]
{[1]20522010, [2]20521477}@gm.uit.edu.vn
Trường Đại học Công nghệ Thông tin – Đại học Quốc gia Thành phố Hồ Chí Minh

***Tóm tắt :*** Bài báo này giới thiệu một nghiên cứu mới trong việc xây dựng hệ thống khuyến nghị bài báo khoa học trực tuyến. Mục tiêu của chúng tôi là cải thiện hiểu biết và quản lý tài liệu khoa học, đồng thời cung cấp các đề xuất tài liệu liên quan và cùng thể loại một cách hiệu quả hơn. Chúng tôi đã thu thập và xây dựng một bộ dữ liệu gồm 2.498 trích dẫn, được sử dụng trong quá trình nghiên cứu của chúng tôi. Để phân loại tài liệu, chúng tôi đã sử dụng các mô hình học máy phổ biến như K-Nearest Neighbor, Random Forest và Support Vector Machine,... Cùng với các mô hình học sâu như BERT-base, Electra và DistillBERT,... Kết quả thực nghiệm đã chứng minh hiệu quả của mô hình của chúng tôi trong cả nhiệm vụ phân loại và đề xuất tài liệu. Bộ dữ liệu mà chúng tôi đã xây dựng sẽ tiếp tục mở rộng và chia sẻ với cộng đồng nghiên cứu để thúc đẩy phát triển và ứng dụng trong lĩnh vực này.

***Keywords: Hệ thống khuyến nghị, tài liệu khoa học, đề xuất, phân loại, trích dẫn mô hình học máy, học sâu.***

## I. GIỚI THIỆU

Trong bài báo này, chúng tôi trình bày một nghiên cứu mới liên quan đến nhiệm vụ phân loại và đề xuất tài liệu trực tuyến dựa trên ngữ cảnh trong văn bản khoa học. Với sự tăng trưởng đáng kể của số lượng bài báo khoa học và tài liệu trực tuyến, việc theo dõi và phân loại các tài liệu này trở nên quan trọng để hỗ trợ nhà nghiên cứu tìm kiếm thông tin và đề xuất tài liệu khoa học, cũng như xây dựng đồ thị tri thức về tài nguyên khoa học. Trước đây, đã có nhiều thử nghiệm để phân loại và đề xuất trích dẫn trong văn bản, nhưng chủ yếu tập trung vào trích dẫn bài báo và chưa nghiên cứu kỹ về trích dẫn tài liệu trực tuyến. Do đó, chúng tôi đề xuất một phương pháp dựa trên ngữ cảnh để đánh giá nhiệm vụ và mục đích của trích dẫn.




Hình 1. Ví dụ về hai loại trích dẫn trong bài báo khoa học. Chúng tôi đã phân tích các từ chỉ nhiệm vụ và mục đích của các trích dẫn đó.

Trong dự án này, chúng tôi định nghĩa trích dẫn tài liệu là liên kết trong văn bản khoa học, kết nối đến tài nguyên trực tuyến cụ thể. Đối với các tài nguyên trong ngữ cảnh, chúng tôi đã tiến hành xác định nhiệm vụ và mục đích của chúng. Để đạt được điều này, chúng tôi xây dựng tập dữ liệu mới với hơn 2498 ngữ cảnh tài nguyên được chú thích thủ công. Chúng tôi sử dụng mô hình phân loại (BERT-base) và các biến thể của nó để so sánh.ở tác vụ phân loại. Còn đối với tác vụ đề xuất, chúng tôi sử dụng mô hình đề xuất (BERT-base+RF) để phân loại và đề xuất nhiệm vụ và mục đích tài nguyên dựa trên ngữ cảnh, bên cạnh đó chúng tôi vẫn có áp dụng một số phương pháp phổ biến khác thường được dùng trong các hệ thống khuyến nghị. Kết quả thực nghiệm chứng minh mô hình của chúng tôi vượt trội so với các phương pháp khác. Hơn nữa, chúng tôi đã phát triển mô hình đề xuất (BERT-base+RF) để dự đoán và đề xuất các bài báo cùng thể loại và liên quan. Mô hình của chúng tôi đạt hiệu suất cao nhờ thông tin về mục đích và nhiệm vụ của tài nguyên. Nghiên cứu của chúng tôi đóng góp quan trọng cho việc hiểu và ứng dụng nhiệm vụ và mục đích của trích dẫn tài liệu trực tuyến trong văn bản khoa học, và khuyến khích phát triển các ứng dụng hỗ trợ phân tích và tìm kiếm các bài báo khoa học trong tương lai.

## II. CÔNG TRÌNH LIÊN QUAN

### A. *Phân tích trích dẫn dựa trên ngữ cảnh*

Phân tích trích dẫn dựa trên ngữ cảnh tập trung vào giá trị của mỗi trích dẫn, xem xét cả cú pháp và ngữ nghĩa. Trong đó, chức năng của trích dẫn đóng vai trò quan trọng, đại diện cho lý do tác giả trích dẫn một bài báo cụ thể. Có nhiều phương pháp đã được phát triển để phân loại chức năng của trích dẫn. Ví dụ, Jurgens et al. (2018) đã tiến hành nghiên cứu về sự phát triển trong lĩnh vực khoa học bằng cách xác định chức năng của trích dẫn. Trong thời gian gần đây, đã có nhiều phương pháp và hướng tiếp cận để nhận dạng các lập luận trong các lĩnh vực khác nhau. Dựa trên những nghiên cứu trước đây, chúng tôi đã phát triển một phương pháp chú thích đặc biệt để mô hình hóa nhiệm vụ và mục đích của nguồn tài liệu khoa học.

### B. *Khám phá nguồn tài liệu cho văn bản khoa học*

Hiện nay, có nhiều nghiên cứu tập trung vào việc phát hiện nguồn tài liệu trong văn bản y sinh học. Một số phương pháp đã được sử dụng, như sử dụng biểu thức chính quy và quy tắc heuristic để trích xuất URL từ trích dẫn (Yamamoto và Takagi, 2007), sử dụng mạng chuyển tiếp và biểu thức chính quy để đặt tên, phát hiện chức năng và sử dụng hệ thống rút trích thực thể để phát hiện tên cơ sở dữ liệu và phần mềm (Duck et al., 2012). Tuy nhiên, hiện vẫn chưa có giải pháp nào cụ thể để mô hình hóa nhiệm vụ và mục đích của nguồn tài liệu ở mức độ chi tiết trong văn bản khoa học đa ngành.

## III. BỘ DỮ LIỆU

Các bộ dữ liệu văn bản khoa học hiện có không cung cấp thông tin về vị trí và ngữ cảnh của trích dẫn tài nguyên trực tuyến. Điều này gây khó khăn trong việc thu thập đủ dữ liệu cho nhiệm vụ này. Chúng tôi đã thu thập và xây dựng một bộ dữ liệu lớn với 4167 mẫu và chú thích thủ công. Bộ dữ liệu này giải thích nhiệm vụ và mục đích của trích dẫn tài nguyên từ góc độ phân cấp. Nó cung cấp cả vai trò tổng quát và chi tiết. Hy vọng bộ dữ liệu này sẽ hỗ trợ các nghiên cứu tương

lai về phân tích nguồn tài nguyên dựa trên ngữ cảnh trong văn bản khoa học.

| Source | Collection | Annotation |
|---|---|---|
| ACL | 1531 | 875 |
| arXiv | 530 | 342 |
| NeurIPS | 469 | 293 |
| BioMed Central | 343 | 227 |
| Plos Pathogens | 197 | 150 |
| ResearchGate | 223 | 126 |
| IEEEXplore | 173 | 110 |
| Other | 701 | 375 |
| **Total** | 4167 | 2498 |

Bảng 1. Mô tả số lượng trích dẫn được thu thập và số lượng trích dẫn sau khi lọc để gán nhãn.

### *A. Xử lý dữ liệu*

Chúng tôi đã thu thập các tài liệu trích dẫn từ nhiều nguồn khác nhau, bao gồm Association for Computational Linguistics (ACL), NeurIPS Proceedings (NeurIPS) PubMed, BioMed,..Tổng cộng, chúng tôi đã thu thập 4,167 trích dẫn từ các nguồn này. Nhóm của chúng tôi đã trích xuất những thông tin cần thiết từ phần văn bản chính và các chú thích của mỗi bài báo. Sau đó, hai sinh viên chúng tôi đã chia nhau ra đọc và tìm hiểu nhiệm vụ và mục đích của các trích dẫn, và gán nhãn cho toàn bộ dữ liệu thu thập.

### *B. Gán nhãn*

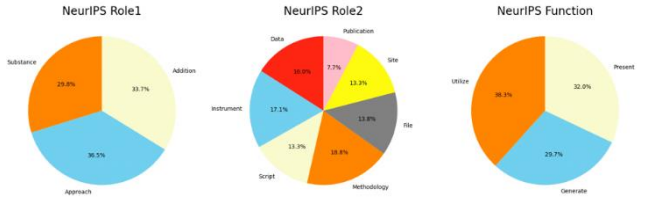
Nhiều mô hình chú thích cho các chức năng trích dẫn đã được tạo ra trong những năm qua. Dựa trên những công trình trước đây và một số mô hình chú thích gần đây, chúng tôi sử dụng mô hình chú thích sau.

- 3 loại nhiệm vụ chung (Task1): Substance, Approach, Addition.id
- 7 loại nhiệm vụ chi tiết (Task2): Data, Instrument, Script, Methodology, File, Site, Publication
- 3 loại mục đích (Purpose): Utilize, Generate, Present.

Quá trình chú thích nhãn được thực hiện bởi một nhóm 2 sinh viên, với chuyên môn như nhau. Tổng cộng, chúng tôi đã chọn ngẫu nhiên 400 mẫu từ bộ dữ liệu cho mỗi sinh viên. Để đảm bảo độ tin cậy, chúng tôi đã loại bỏ các mẫu dữ liệu quá ngắn. Theo kế hoạch thì mỗi trích dẫn sẽ được gán ít nhất một nhãn cho nhiệm vụ chung, một nhãn cho nhiệm vụ chi tiết và một nhãn cho mục đích của trích dẫn đó. Sự đồng thuận giữa hai người gán nhãn đã được đánh giá độ đồng thuận Kohen Kappa cho thấy mức độ đồng thuận tương đối cao, trung bình khoảng 0,71 cho 3 lớp. Cuối cùng, chúng tôi đã thu thập một bộ dữ liệu được chú thích thủ công với 2,498 mẫu dữ liệu. Các loại nhãn mục đích và nhiệm vụ của trích dẫn cùng với số liệu thống kê được hiển thị trong Bảng 2.




Hình 2. Biểu đồ tròn thể hiện số lượng nhãn ở mỗi lớp của hội nghị ACL.




Hình 3. Biểu đồ tròn thể hiện số lượng nhãn ở mỗi lớp của hội nghị NeurIPS.

Phân bố của các loại nhãn nhiệm vụ chi tiết không đồng đều, có một số loại có số lượng rất thấp. Điều này yêu cầu chúng tôi cân nhắc khi xác định và sử dụng lại mô hình chú thích trong tương lai.

## IV. PHƯƠNG PHÁP

Trong lĩnh vực nghiên cứu khoa học, việc xác định nhiệm vụ và mục đích của các trích dẫn tài nguyên trong ngữ cảnh tài nguyên khoa học là một nhiệm vụ quan trọng. Để giải quyết bài toán này, chúng tôi đề xuất hai mô hình cho hai nhiệm vụ tương ứng: Mô hình phân loại (DistilBERT) và mô hình đề xuất (BERT-base+RF).

### *A. Mô hình phân loại*

Chúng tôi sử dụng các mô hình máy học phổ biến cho nhiệm vụ phân loại như SVM, KNN, RF,… Và các mô hình học sâu có khả năng trích xuất ngữ cảnh điển hình như mô hình BERT-base, Electra, DistilBERT, RoBERTa,… Để phân loại các nhiệm vụ 1, nhiệm vụ 2 và mục đích của các trích dẫn tài nguyên. Mô hình này chia sẻ biểu diễn ngữ cảnh tài nguyên để xác định các loại mục đích và nhiệm vụ. Tập dữ liệu của chúng tôi vẫn còn nhiều mặt hạn chế, do các câu thường có độ dài ngắn và thông tin ngữ nghĩa tập trung trong một số danh từ và động từ gần trích dẫn. Vị trí từ trong chuỗi đóng vai trò rất quan trọng. Do những hạn chế đó về bộ dữ liệu nên đa số các mô hình neural không được huấn luyện tốt trên các tham số. Vì vậy, chúng tôi cần phát triển phương pháp hiệu quả để giải quyết những thách thức đặc biệt của nhiệm vụ phân loại tài nguyên khoa học này.

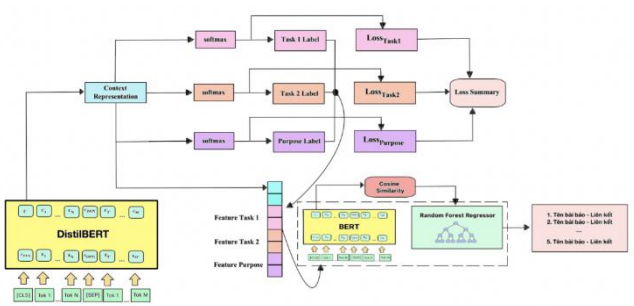
Dựa trên các công trình trước đó, chúng tôi quyết định sẽ tận dụng kiến trúc của mô hình BERT đã được tiền huấn luyện để học các biểu diễn ngữ cảnh cho các tài nguyên trích dẫn, như được hiển thị trong Hình 3. Vì vậy, với mỗi chuỗi từ đầu vào của một ngữ cảnh tài nguyên, hàm mất mát sẽ được tính bằng tổng mất mát của ba nhiệm vụ phân loại và được biểu diễn bởi công thức sau:

$$\text{Loss\_Classification}_{\text{BERT}} = \alpha L_{\text{1st task}} + \beta L_{\text{2nd task}} + \gamma L_{\text{purpose}}$$

### *B. Mô hình đề xuất*

Chúng tôi xây dựng một hệ khuyến nghị nhằm đề xuất các bài báo cùng thể loại và có liên quan dựa trên ngữ cảnh của

câu trích dẫn đang xét, ngoài ra chúng tôi còn cung cấp thêm cho người đọc một đường dẫn URL đến bài báo đó. Mô hình ở nhiệm vụ đề xuất sẽ sử dụng thông tin mô hình ở nhiệm vụ phân loại để hỗ trợ cho việc dự đoán. Đầu tiên, chúng tôi mã hóa chuỗi ngữ cảnh bằng mô hình BERT-base để tạo ra ma trận biểu diễn ngữ cảnh. Tiếp đến, chúng tôi sẽ lọc ra những bài báo nào có cùng loại nhãn với những nhãn vừa được phân loại. Cuối cùng, chúng tôi sẽ kết hợp các biểu diễn ngữ cảnh, tính độ tương đồng giữa các câu trích dẫn để dự đoán các bài báo cùng thể loại và có thể liên quan đến các trích dẫn.




Hình 4. Mô tả sơ đồ kiến trúc mô hình cho bài toán phân loại và đề xuất.

## V. THỰC NGHIỆM

### A. *Đánh giá trên dữ liệu*

Chúng tôi đã tiến hành thử nghiệm trên tập dữ liệu hoàn chỉnh, dùng 80% cho huấn luyện và 20% cho kiểm thử. Để xử lý mất cân đối giữa các lớp, chúng tôi sử dụng phương pháp up-sampling bằng cách nhân bản các mẫu trong các lớp thiểu số trong quá trình huấn luyện. Kết quả là chúng tôi có 2,498 mẫu được dùng để phân loại nhãn cho nhiệm vụ 1, nhiệm vụ 2 và nhãn cho mục đích của trích dẫn. Mặc dù dữ liệu của chúng tôi còn nhỏ, tuy nhiên mô hình của chúng tôi vẫn đạt hiệu suất tốt với dữ liệu được gán nhãn còn bị hạn chế. Chúng tôi báo cáo điểm F1 theo phương pháp macro và độ chính xác (accuracy) để đánh giá các nhiệm vụ và mục đích của trích dẫn.

| Methods | F1-macro for each class | | | F1 average |
|---|---|---|---|---|
| | Task 1 | Task 2 | Purpose | |
| K-Nearest Neighbor | 0,489 | 0,406 | 0,489 | 0,461 |
| Naive Bayes | 0,518 | 0,390 | 0,518 | 0,475 |
| Decision Tree | 0,549 | 0,468 | 0,549 | 0,522 |
| Random Forest | 0,572 | 0,508 | 0,572 | 0,550 |
| Support Vector Machine | 0,590 | 0,516 | 0,590 | 0,565 |
| Electra | 0,191 | 0,546 | 0,620 | 0,452 |
| Bert-base | 0,740 | 0,540 | **0,627** | 0,635 |
| Roberta | 0,743 | **0,631** | 0,577 | 0,650 |
| DistilBERT | **0,764** | 0,601 | 0,600 | **0,655** |

Bảng 3. So sánh kết quả F1-score của các mô hình trong nhiệm vụ phân loại ba lớp.

| Methods | Accuracy | | | Accuracy average |
|---|---|---|---|---|
| | Task 1 | Task 2 | Purpose | |
| Bert-base | 0,752 | 0.601 | 0.603 | 0,652 |
| Roberta | 0.740 | **0.641** | 0.592 | 0,658 |
| DistilBERT | **0.764** | 0.609 | **0.601** | **0,658** |

Bảng 4. So sánh độ chính xác của ba mô hình tốt nhất trong nhiệm vụ phân loại.

Đối với nhiệm vụ đề xuất, chúng tôi sẽ thực hiện thử nghiệm trên hang loạt các hệ thống đề xuất trên tập dữ liệu đã thu thập như: RF (BoW+TFIDF), RF (N-grams+TFIDF), Context-based, Matrix Factorization+SVD và đặt biệt hơn là Bert-based+RF.

| Method | |
|---|---|
| Matrix Factorization + SVD | **Recommend Task** |
| Context-based | |
| RF (BoW+TFIDF) | |
| RF (N-grams+TFIDF) | |
| Bert-based + RF | |

Bảng 5. Thống kê các loại phương pháp được áp dụng để xây dựng hệ khuyến nghị.

### B. *Các mô hình cơ sở và thiết lập*

Trong nhiệm vụ phân loại, chúng tôi đã tiến hành so sánh mô hình DistilBERT với các phương pháp khác trên dữ liệu của chúng tôi. Các phương pháp đó bao gồm các mô hình phân loại trong học máy như SVM, RF, Decision Tree,.. Và các loại mô hình học sâu được tiền huấn luyện như Electra, RoBERTa,... Đối với nhiệm vụ đề xuất, chúng tôi đã tiến hành thử nghiệm trên nhiều phương pháp đề xuất thông thường được dùng trong hệ khuyến nghị và đặc biệt hơn hết là mô hình kết hợp BERT+RF để dự đoán các bài báo cùng thể loại và có độ tương đồng cao.

| Task 1 | F1 | Task 2 | F1 | Purpose | F1 |
|---|---|---|---|---|---|
| Substance | 0,759 | Data | 0,400 | Utilize | 0,610 |
| Approach | 0,759 | Instrument | 0,690 | Generate | 0,580 |
| Addition | 0,773 | Script | 0,680 | Present | 0,610 |
| | | Methodology | 0,500 | | |
| | | File | 0,490 | | |
| | | Site | 0,670 | | |
| | | Publication | 0,780 | | |

Bảng 6. Kết quả (F1-score) trên từng nhãn được dự đoán bởi mô hình tốt nhất cho ba nhiệm vụ.

Chúng tôi đã thử nghiệm mô hình hóa BERT dựa trên phiên bản tham khảo từ Google. Chúng tôi đã tinh chỉnh mô hình DistilBERT trên tập dữ liệu huấn luyện của chúng tôi cho ba nhiệm vụ phân loại. Mô hình đề xuất của chúng tôi sử dụng embedding dimension cho các nhãn nhiệm vụ, nhãn mục đích, và sử dụng embedding đã được tiền huấn luyện trên dữ liệu lớn từ các trích dẫn của chúng tôi.

### C. *Kết quả*

#### 1) *Kết quả phân loại*

Nhìn chung, các mô hình pretrained dường như đạt được kết quả tốt nhất trên cả ba nhiệm vụ phân loại, như được thể hiện trong Bảng 3. F1-score được dùng để tính cho mỗi nhiệm vụ hoặc mục đích của trích dẫn được hiển thị trong Bảng 5. Mặc dù kết quả vẫn còn xa so với các nhiệm vụ phân loại ngữ cảnh tổng quát khác, điều này đặt ra tiềm năng cho sự phát triển của các nghiên cứu trong tương lai. Phân tích chi tiết hiệu suất cho thấy hầu hết các mô hình học máy truyền thống phổ biến được dùng để phân loại có hiệu suất thấp hơn, có thể do tập dữ liệu gán nhãn chưa đủ lớn hoặc còn nhiều hạn chế. Các mô hình pretrained tận dụng được lợi thế của bộ mã hóa

BERT, được tiền huấn luyện trên lượng lớn văn bản và tinh chỉnh trên tập huấn luyện để đạt được hiệu suất tốt hơn.

*2) Kết quả đề xuất*

Chúng tôi đã tiến hành so sánh với các phương pháp phổ biến được sử dụng trong hệ thống đề xuất như RF (BoW+TFIDF), RF(N-grams+TFIDF),Context-based,Matrix Factorization+SVD và đặt biệt hơn là Bert-based+RF. Kết quả cho thấy dường như các mô hình đều đề xuất đều hoạt động tốt trên bộ dữ liệu của chúng tôi, điều này cho thấy được tiềm năng phát triển các hệ thống đề xuất để cải thiện khả năng tìm kiếm và quản lý nguồn tài nguyên khoa học trong tương lai.

## VI. HƯỚNG PHÁT TRIỂN VÀ KẾT LUẬN

Chúng tôi đã giới thiệu một tác vụ mới trong việc phân loại nhiệm vụ và mục đích của tài liệu tham khảo khoa học trực tuyến. Để làm điều này, chúng tôi đã xây dựng hệ thống gán nhãn và thu thập một tập dữ liệu mới thông qua việc gán nhãn thủ công. Đồng thời, chúng tôi đề xuất một số mô hình được tiền huấn luyện kết hợp việc phân loại các nhiệm vụ và mục đích dựa trên ngữ cảnh của tài liệu và đề xuất các bài báo tương tự dựa trên ngữ cảnh. Các mô hình đề xuất của chúng tôi đã cải thiện hiệu suất đáng kể trên tất cả các nhiệm vụ nhờ tập dữ liệu phù hợp. Hơn nữa, chúng tôi cũng đưa ra một nhiệm vụ đề xuất nguồn tài nguyên và phát triển mô hình kết hợp (BERT-base+RF) dựa trên việc sử dụng các nhãn đã phân loại và độ tương đồng của các trích dẫn dựa trên ngữ cảnh. Các model được dùng ở hai nhiệm vụ của chúng tôi có khả năng đóng góp vào việc xây dựng hệ thống tìm kiếm và đề xuất mạnh mẽ hơn cho nguồn tài nguyên khoa học trực tuyến. Trong tương lai, chúng tôi sẽ tiếp tục nghiên cứu và thử nghiệm việc áp dụng BERT vào các lĩnh vực khác và khám phá khả năng kết hợp mô hình của chúng tôi để hỗ trợ các nhiệm vụ như đánh giá, dự đoán và xây dựng đồ thị tri thức cho nguồn tài nguyên khoa học.

## VII. REFERENCES


[1] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: Bert for document classification. CoRR, abs/1904.08398.

[2] Geraint Duck, Goran Nenadic, Andy Brass, David L. Robertson, and Robert Stevens. 2012. bionerds: exploring bioinformatics database and software use through literature mining. In BMC Bioinformatics.

[3] Qi He, Daniel Kifer, Jian Pei, Prasenjit Mitra, and C. Lee Giles. 2011. Citation recommendation without author supervision. In WSDM

[4] Wenyi Huang, Saurabh Kataria, Cornelia Caragea, Prasenjit Mitra, C. Lee Giles, and Lior Rokach. 2012. Recommending citations: translating papers into references. In CIKM. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

[5] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In EACL.Bioinformatics. Geraint Duck, Goran Nenadic, Andy Brass, David L.

[6] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In EMNLP.

[7] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In AAAI.