

# サーベイ論文で引用すべき論文の推薦

柴 克樹 笹野 遼平 武田 浩一  
名古屋大学大学院情報学研究科

shiba.katsuki.c5@s.mail.nagoya-u.ac.jp {sasano,takedasu}@i.nagoya-u.ac.jp

## 概要

サーベイ論文の自動生成は新たな研究分野に取り組もうとする研究者にとって、分野全体の俯瞰を助ける重要な役割を持つ技術である。しかし、既存のサーベイ論文自動生成に関する研究では、サーベイ論文が引用する論文の情報から、サーベイ論文を生成するタスクに焦点が当てられており、引用すべき論文をどのようにして見つけるかについては十分な検討がなされてこなかった。本研究では、サーベイ論文で引用すべき論文の推薦システムの構築に取り組む。具体的には、生成したいサーベイ論文に関するキーワードが与えられたとき、キーワード検索と論文の引用関係に基づき候補論文集合を作成し、そこから引用推薦の枠組みを用いてサーベイ論文が引用すべき論文を推薦するシステムを構築する。

## 1 はじめに

大規模言語モデル (LLM) をはじめとする人工知能技術の急速な発展に伴い、新たな研究分野が生まれている。新たな分野に取り組もうとする研究者にとって、対象の研究分野を俯瞰するサーベイ論文は非常に有益であることから、サーベイ論文の自動生成技術の重要性は高まっていると言える。既存のサーベイ論文自動生成の研究 [1, 2] では、サーベイ論文が引用する論文の情報から、サーベイ論文を生成するタスクに焦点が当てられており、引用論文は実験に使用するサーベイ論文が引用している論文をそのまま利用するか、引用論文をもとにヒューリスティックに収集されたものが使われている。しかし、実際にサーベイ論文を生成することを考えると、引用すべき論文を見つける技術は必要不可欠であることから、本研究ではサーベイ論文が引用すべき論文を推薦するタスクに取り組む。

既存の引用推薦 (citation recommendation) の研究 [3-6] は論文のタイトル、概要、引用関係、著者

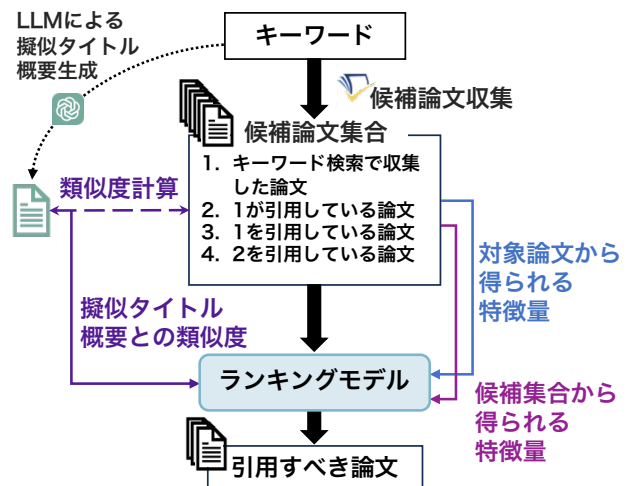


図1 提案する引用論文推薦システム概要

情報などが与えられ、それをもとに推薦を行う設定で行われてきた [7]。近年では SciBERT [8] をベースにタイトル、概要、引用関係から学習を行なった SPECTER [9] が広く利用されている。一方で、20万本程度の中規模な論文プールに対する引用推薦では SPECTER のような TAE (Transformer-based Article Encoders) は高い性能を示すが、200万本程度の大規模なデータセットでは BM25 に性能が劣るという指摘もある [10]。サーベイ論文の自動生成においては、通常の引用推薦と異なり論文のタイトルや概要は存在していないため、対象分野のキーワード等のみから推薦を行う必要があり、また、サーベイ論文と通常論文では引用傾向が異なると考えられるが、サーベイ論文を対象とした引用推薦の研究は行われていない。

そこで、本研究ではキーワードを入力とするサーベイ論文の引用推薦のシステムを構築に取り組む。システムの概要図を図1に示す。提案システムではキーワードをもとに Web 上から候補論文集合を収集し、ランキングモデルにより候補論文集合の順位付けを行うことで引用すべき論文の推薦を行う。

## 2 サーベイ論文における引用の分析

引用推薦システムの構築に先立ち、サーベイ論文が引用する論文の傾向について分析する。

### 2.1 データセット

既存研究で広く利用されてきたデータセットである DBLP-Citation-network (v14) データセット<sup>1)</sup>、および、ACL Anthology Network Corpus<sup>2)</sup>における論文の平均引用本数を調べた結果、いずれも 5.25 本と、引用論文情報の欠落が多く存在していた。このため本研究では、2022 年 8 月までに ACL anthology に掲載された 80,013 本の論文の全文コーパスである ACL OCL Corpus [11] と、Semantic Scholar<sup>3)</sup>を利用して、分析に使用する論文データセットを作成した。

まず、分析対象のサーベイ論文として ACL OCL Corpus から、以下の 3 つの条件を満たす 60 本の論文を抽出した。

1. タイトルに“survey”、“literature review”、“a review”のいずれかのフレーズを含む
2. 引用文献数が 15 以上である
3. ACL、EMNLP、NAACL、AAACL、CL のいずれかで発表されている

続いて、比較対象とする通常論文として、ACL OCL Corpus から無作為に 900 本の論文を抽出した。各論文が引用する論文の一覧は ACL OCL Corpus から獲得し、引用先の論文が出版された年月と被引用数の情報は Semantic Scholar から収集した。

### 2.2 サーベイ論文における引用の特徴

まず、平均引用論文数を調査した。その結果、通常論文の平均引用論文数は約 30 本であるのに対し、サーベイ論文の平均引用論文数は約 85 本と、通常論文の 3 倍近い論文を引用していることが分かった。

次に、通常論文とサーベイ論文、それぞれについて、引用している論文の出版経過日数を調査した。結果を図 2 に示す。出版経過日数を 3 ヶ月ごとにまとめた結果で、縦軸は各グループに属する論文の割合を表す。図 2 から、サーベイ論文は通常論文よりも新しい論文を引用する傾向があり、その傾向は出版経過日数が 1 年以内の論文において顕著であることが分かる。また、通常論文の引用論文の出版経過

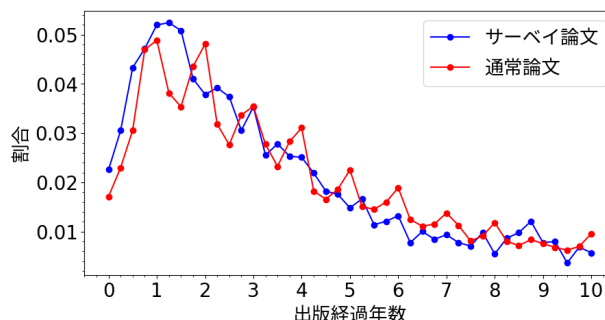


図 2 サーベイ論文と通常論文の引用論文の出版経過日数

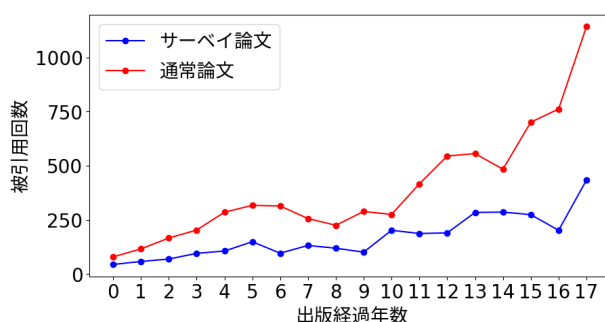


図 3 サーベイ論文と通常論文の引用論文の被引用回数

日数については 1 年単位の周期性があることが確認できた。これは、通常論文は毎年近い時期に開催される同じ学会に投稿された論文を多く引用する傾向があるためだと考えられる。一方、サーベイ論文は多くの学会やジャーナルを横断的に引用するため、このような傾向は確認できなかった。

続いて、通常論文とサーベイ論文、それぞれについて、引用している論文の被引用回数を調査した。被引用回数の中央値をまとめた結果を図 3 に示す。サーベイ論文は引用論文の出版経過年数に依らず、通常論文よりも被引用回数の少ない論文を引用する傾向が確認できた。これは、通常論文は各分野の代表的な論文のみ引用する傾向が強いのに対し、サーベイ論文では対象とする分野の論文を広く引用するためであると考えられる。

## 3 サーベイ論文のための引用推薦

本研究で構築するサーベイ論文に対する引用推薦システムでは、まずキーワードをもとに推薦候補論文集合を収集する。続いて、集まった候補論文をランキングモデルに入力し、得られたランキング結果をもとに引用すべき論文を推薦する。

1) <https://www.aminer.cn/citation>  
2) <https://clair.eecs.umich.edu/aan/index.php>  
3) <https://www.semanticscholar.org/>

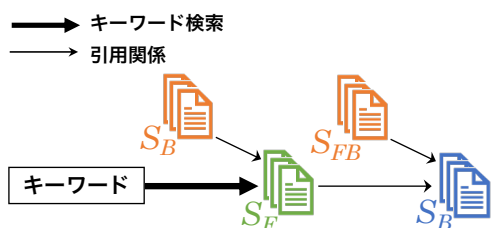


図4 論文集合間の引用関係のイメージ

### 3.1 候補論文収集

候補論文集合として、以下で規定される4つの論文集合  $S_0$  (初期集合)、 $S_F$  (前向き)、 $S_B$  (後向き)、 $S_{FB}$  (前向き+後向き) を収集する。

- $S_0$ : 与えられたキーワードで Semantic Scholar を検索した結果得られた上位  $n$  件の論文
- $S_F$ :  $S_0$  中の論文が (前向き) 引用している論文
- $S_B$ :  $S_0$  中の論文を (後向き) 引用している論文
- $S_{FB}$ :  $S_B$  中の論文を (後向き) 引用している論文

収集された論文集合間の引用関係のイメージを図4に示す。以降では論文集合  $S_0$  と  $S_F$  の和集合を  $S_{0+F}$  などのように記載するものとする。また、1本の論文の引用論文数は多くとも数百程度である一方、一部の論文の被引用数は1万を超えるため、前向き収集と比較して後向き収集では候補論文集合の爆発的な増加が発生しうる。そのため、後向き収集は引用回数が300回未満の論文のみを対象とする。

### 3.2 候補論文のランキング

本研究ではランキングモデルとして LightGBM を利用する。LightGBM に入力する特徴量を表1に示す。特徴量は大きく、対象論文のみから得られる特徴量 ( $F_{\text{対象論文}}$ )、候補論文集合から得られる特徴量 ( $F_{\text{候補集合}}$ )、擬似論文を用いて算出される特徴量 ( $F_{\text{擬似論文}}$ ) の3タイプに分けられる。

$F_{\text{候補集合}}$  のうち“ $S_0$  または  $S_F$  中の論文から引用された回数”は、キーワード検索で上位となった論文 ( $S_0$ )、または、それらの論文が引用している論文 ( $S_F$ ) から多く引用されている論文はサーベイ論文を作成する際に引用すべき論文である可能性が高いという考えに基づく特徴量である。また、 $F_{\text{擬似論文}}$  は引用推薦において広く利用されている SPECTER2 [9] に基づく特徴量である。SPECTER2 では、引用関係にある論文同士のベクトルの類似度が相対的に大きくなるような論文ベクトルを、論文のタイトルと概要から作成する。生成対象のサーベイ論文が存在して

表1 LightGBM に入力する特徴量

#### 対象論文のみから得られる特徴量 ( $F_{\text{対象論文}}$ )

- 出版経過年数・日数
- 会議・論文誌名

#### 候補論文集合から得られる特徴量 ( $F_{\text{候補集合}}$ )

- $S_0$  または  $S_F$  中の論文から引用された回数
- 対象論文が含まれる候補論文集合

#### 擬似論文を用いて算出される特徴量 ( $F_{\text{擬似論文}}$ )

- 擬似サーベイ論文との類似度

いる場合、そのタイトルと概要を用いて SPCETER2 のベクトルを生成できるが、本研究の設定では存在していない。このため、LLM を用いてキーワードから疑似的なタイトルと概要を生成し、生成されたタイトル、概要から SPECTER2 のベクトルを生成し、候補論文集合中の論文の論文ベクトルとの類似度を算出し、素性として利用する。疑似的なタイトルと概要の生成に使用した LLM とプロンプト、生成例は付録Aに示す。

## 4 実験

2.1 節で述べたサーベイ論文60本を対象に、収集された候補論文集合の評価、および、最終的な引用論文推薦結果の評価を行った。

### 4.1 候補論文集合の収集と評価

まず、評価対象とするサーベイ論文60本それぞれについてキーワードを手で付与した。具体的には、タイトルに含まれる単語列から、サーベイのトピックを適切に表現できる2つ、または、3つの単語列を手で抽出した。続いて、付与されたキーワードをもとに候補論文集合を作成した。論文集合  $S_0$  および  $S_F$  は Semantic Scholar を用いて収集し、全60論文に対して収集された結果をまとめた論文プールを構築した。一方、論文集合  $S_F$  および  $S_{FB}$  の収集には、後向きに引用を辿る必要があるが、Semantic Scholar の引用検索機能では1%程度しか収集できなかったため、構築した論文プールを対象に引用情報を利用し、疑似的に論文収集した。ただし、実際にサーベイ論文が引用している論文については、論文プールに含まれていない場合も、 $S_F$  または  $S_{FB}$  の条件を満たすか調査し、条件を満たす場合は候補論文集合に追加した<sup>4)</sup>。

4) 後向きに引用を辿れば追加可能な論文であり、非現実的な設定ではない。

表2 論文収集カバー率のマクロ平均 (括弧中の数値はサーベイ論文1本あたりの収集した論文数)

$n$	$S_0$	$S_{0+B}$	$S_{0+F}$	$S_{0+F+B}$	$S_{0+F+B+FB}$
100	10.31 (100)	26.12 (401)	60.20 (2191)	64.81 (2316)	81.51 (7477)
200	13.44 (188)	30.34 (508)	65.20 (4269)	70.07 (4487)	85.15 (12772)
300	16.23 (268)	35.74 (1044)	68.65 (5780)	74.75 (6356)	88.76 (21330)
500	18.40 (431)	38.52 (1372)	71.06 (9219)	76.27 (9866)	89.15 (28379)

論文収集の結果を表2に示す。 $S_0$  収集の際の閾値  $n$  として 100, 200, 300, 500 の4つの値に対する結果を記載している。 $n = 10000$  まで増やした実験も行ったが、カバー率は 31.92 に留まったことから、引用関係を用いて候補論文集合を収集する有効性が確認できる。

## 4.2 引用論文推薦の評価

引用論文推薦の評価は、実際の引用文献のうち、順位付けされた論文候補集合の上位  $k$  件に含まれる割合を、全 60 論文で平均したスコアにより行った。 $k$  の値は、実際の引用文献数を  $r$  倍した値とした。また、5 分割交差検証を実施し、LightGBM で使用する損失関数は binary log loss、学習率は  $5e-3$ 、num\_leaves は 46 とした。

まず、 $r = 1$  とする設定において、候補論文集合の違いによる性能の違いを検証した。結果を表3に示す。収集対象とする論文集合の種類が少ない場合 ( $S_0, S_{0+B}$ ) は、 $n$  を大きくした場合の方が高い性能となる一方で、論文集合  $S_B$  や  $S_{FB}$  まで収集対象とした場合は、 $n$  が小さい方が高い性能となった。全体でもっとも高い性能となったのは、 $n = 100$  としすべての論文集合を収集対象とした場合であった。

続いて、提案手法の有効性を確認するための比較実験を行った。まず、提案手法における LightGBM の特徴量から、 $F_{対象論文}$ 、 $F_{候補集合}$ 、 $F_{疑似論文}$  を1つずつ除いたモデルとの比較を行った。さらに、疑似的な論文タイトルおよび概要に基づく BM25 ベースのモデル、SPECTER2 ベースのモデルとの比較も行った<sup>5)</sup>。実験は、 $S_0$  収集の際の閾値  $n$  を 100、論文収集対象を  $S_{0+F+B+FB}$  とする設定で実施し、 $r$  の値として 0.5, 1, 2, 3, 5 の5つの値に対して評価を実施した。結果を表4に示す。 $r$  の値に依らず提案手法がもっとも高い性能を達成した。この結果から、

5) BM25、SPECTER2 ベースのモデルの詳細は付録Bに示す。

表3 候補論文集合ごとの引用推薦性能 ( $r = 1$ )

$n$	$S_0$	$S_{0+B}$	$S_{0+F}$	$S_{0+F+B}$	$S_{0+F+B+FB}$
100	9.7	20.0	30.8	31.6	<b>31.8</b>
200	12.0	21.1	30.3	30.7	30.8
300	14.9	22.3	29.5	29.7	29.9
500	14.9	22.3	29.5	29.7	29.3

表4 引用推薦の比較実験結果

手法	$r$	0.5	1	2	3	5
提案システム		<b>20.5</b>	<b>31.8</b>	<b>42.5</b>	<b>48.6</b>	<b>55.6</b>
w/o $F_{論文情報}$		19.3	28.9	39.6	45.8	53.5
w/o $F_{候補集合}$		12.3	19.0	27.7	33.7	41.6
w/o $F_{疑似論文}$		19.1	28.4	37.4	42.4	49.4
BM25		5.1	11.4	18.4	24.3	33.7
SPECTER2		12.2	20.2	30.9	38.6	48.1

特徴量  $F_{対象論文}$ 、 $F_{候補集合}$ 、 $F_{疑似論文}$  はいずれも有効に機能していることが確認できる。特に  $F_{候補集合}$  を除いた場合の性能低下の度合いが大きいことから、対象論文がどの候補論文集合に含まれるかや候補論文集合から何回引用されているかの情報は重要な手掛かりであると言える。また、 $F_{疑似論文}$  が性能の向上に寄与していることから、LLM を用いて疑似的に生成したタイトル、概要が有効に機能していることが確認できる。

大規模なデータセットにおいては、SPECTER2 ベースのモデルは、BM25 ベースのモデルに性能が劣ると指摘 [10] されていたが、本実験では引用関係を利用することでカバー率が高く比較的小さな候補論文集合を収集したため、Semantic Scholar 上の全論文 (約 2 億本) を推薦対象としながらも、SPECTER2 ベースのモデルが BM25 ベースのモデルの性能を大幅に上回った。

## 5 おわりに

本研究ではサーベイ論文に対する引用推薦タスクに取り組み、キーワードを入力とするサーベイ論文の引用推薦システムの構築と評価を行った。また、サーベイ論文における引用の分析を行い、サーベイ論文は新しい論文を引用する傾向があること、引用数の小さな論文も引用する傾向があることを明らかにした。今後の課題としては、分析によって得られたサーベイ論文における引用の特徴を捉えたモデルの構築、引用論文情報からの論文生成モデルの利用によるサーベイ論文自動生成システムの構築が考えられる。

## 参考文献

- [1] Rahul Jha, Reed Coke, and Dragomir Radev. Surveyor: A system for generating coherent survey articles for scientific topics. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 29, 2015.
- [2] Huiyan Xu, Zhongqing Wang, Yifei Zhang, Xiaolan Weng, Zhijian Wang, and Guodong Zhou. Document structure model for survey generation using neural network. **Frontiers of Computer Science**, Vol. 15, pp. 1–10, 2021.
- [3] Zafar Ali, Guilin Qi, Khan Muhammad, Siddhartha Bhat-tacharyya, Irfan Ullah, and Waheed Abro. Citation recommendation employing heterogeneous bibliographic network embedding. **Neural Computing and Applications**, Vol. 34, No. 13, pp. 10229–10242, 2022.
- [4] Xiao Ma and Ranran Wang. Personalized scientific paper recommendation based on heterogeneous graph representation. **IEEE Access**, Vol. 7, pp. 79887–79894, 2019.
- [5] Chanathip Pornprasit, Xin Liu, Pattararat Kiattipadungkul, Natthawut Kertkeidkachorn, Kyoung-Sook Kim, Thanapon Noraset, Saeed-Ul Hassan, and Suppa-wong Tuarob. Enhancing citation recommendation using citation network embedding. **Scientometrics**, pp. 1–32, 2022.
- [6] Zafar Ali, Guilin Qi, Pavlos Kefalas, Shah Khusro, In-ayat Khan, and Khan Muhammad. SPR-SMN: Scientific paper recommendation employing specter with memory network. **Scientometrics**, Vol. 127, No. 11, pp. 6763–6785, 2022.
- [7] Zafar Ali, Pavlos Kefalas, Khan Muhammad, Bahadar Ali, and Muhammad Imran. Deep learning in citation recommendation models survey. **Expert Systems with Applications**, Vol. 162, p. 113790, 2020.
- [8] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pre-trained language model for scientific text. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3615–3620, 2019.
- [9] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level representation learning using citation-informed transformers. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2270–2282. Association for Computational Linguistics, 2020.
- [10] Zoran Medić and Jan Snajder. Large-scale evaluation of transformer-based article encoders on the task of citation recommendation. In **Proceedings of the Third Workshop on Scholarly Document Processing**, pp. 19–31. Association for Computational Linguistics, October 2022.
- [11] Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. The ACL OCL corpus: Advancing open science in computational linguistics. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 10348–10361. Association for Computational Linguistics, December 2023.

## A 擬似タイトル、概要生成の詳細

本研究では LLM として gpt-3.5-turbo-0613 を使用した。また、擬似タイトル、および、概要の生成に利用したプロンプトを表 5 に示す。[keywords] の部分に人手で与えられたキーワードを入力して使用した。

実際にキーワードとして “Bias” と “NLP” を与えた場合に生成された擬似論文タイトルと概要を表 6 に示す。参考のため、本キーワード抽出の際に利用した実際の論文のタイトルと概要を表 7 に示す。

表 5 擬似タイトル、概要生成のためのプロンプト

---

Please generate an appropriate title and abstract for a survey paper focus on the keywords ‘[keywords]’ within 100 words.

---

表 6 擬似的に生成されたタイトルと概要

### 擬似タイトル

A Comprehensive Survey on Bias in Natural Language Processing

### 擬似概要

With the rapid growth of natural language processing (NLP) applications, it has become increasingly crucial to address the issue of bias present in these systems. This survey paper aims to provide an extensive overview of bias in NLP, covering various dimensions such as gender, race, and socio-economic status. We explore the different sources of bias, including data collection, pre-processing, and algorithmic decision-making processes. Additionally, we discuss the impacts of biased NLP systems on society, including perpetuating stereotypes and discrimination. Furthermore, we present recent advancements in bias mitigation techniques and highlight the challenges and future directions in this field. This survey paper serves as a valuable resource for researchers and practitioners interested in understanding and mitigating bias in NLP.

---

## B BM25、SPECTER2 モデルの詳細

**BM25 ベースモデル** サーベイ論文の擬似タイトルと擬似概要をクエリとして、候補論文のタイトルと概要に対する BM25 スコアを算出し、順位づけして推薦する手法。

**SPECTER2 ベースモデル** サーベイ論文の擬似タイトルと概要から SPECTER2 のベクトルを作成し、候補論文の論文ベクトルとの類似度をユークリッド距離により算出し、順位づけして推薦する手法。SPECTER2 は huggingfac 上で公開されているモデル<sup>6)</sup>を利用した。

表 7 実際のタイトル・概要

---

### タイトル

Language (Technology) is Power: A Critical Survey of “Bias” in NLP

### 概要

We survey 146 papers analyzing “bias” in NLP systems, finding that their motivations are often vague, inconsistent, and lacking in normative reasoning, despite the fact that analyzing “bias” is an inherently normative process. We further find that these papers’ proposed quantitative techniques for measuring or mitigating “bias” are poorly matched to their motivations and do not engage with the relevant literature outside of NLP. Based on these findings, we describe the beginnings of a path forward by proposing three recommendations that should guide work analyzing “bias” in NLP systems. These recommendations rest on a greater recognition of the relationships between language and social hierarchies, encouraging researchers and practitioners to articulate their conceptualizations of “bias”—i.e., what kinds of system behaviors are harmful, in what ways, to whom, and why, as well as the normative reasoning underlying these statements—and to center work around the lived experiences of members of communities affected by NLP systems, while interrogating and reimagining the power relations between technologists and such communities.

---

6) <https://huggingface.co/allenai/specter2>