eKNOW 2024

The Sixteenth International Conference on Information, Process, and Knowledge Management

ISBN: 978-1-68558-165-7

May 26 - 30, 2024

Barcelona, Spain

eKNOW 2024 Editors

Susan Gauch, University of Arkansas, USA

Samia Aitouche, Laboratory of Automation and Manufacturing (LAP), University Batna 2, Algeria

eKNOW 2024

Forward

The Sixteenth International Conference on Information, Process, and Knowledge Management (eKNOW 2024), held between May 26th and May 30th, 2024, in Barcelona, Spain, continued a series of events focusing on the complexity of the current systems, the diversity of the data, and the challenges for mental representation and understanding of environmental structure and behavior.

Capturing, representing, and manipulating knowledge was and still is a fascinating and extremely useful challenge from both a theoretical and practical perspective. Using validated knowledge for information and process management and for decision support mechanisms raised a series of questions the eKNOW 2024 conference was aimed at.

eKNOW 2024 provided a forum where researchers were able to present recent research results and new research problems and directions related to them. The topics covered aspects from knowledge fundamentals to more specialized topics such as process analysis and modeling, management systems, semantics processing and ontology.

We take here the opportunity to warmly thank all the members of the eKNOW 2024 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to eKNOW 2024. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the eKNOW 2024 organizing committee for their help in handling the logistics of this event.

We hope that eKNOW 2024 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of information, knowledge, and process management.

eKNOW 2024 Chairs

eKNOW 2024 Steering Committee
Susan Gauch, University of Arkansas, USA
Martijn Zoet, Zuyd University of Applied Science, the Netherlands
Samia Aitouche, Laboratory of Automation and Manufacturing (LAP), University Batna 2, Algeria
Eric J.H.J. Mantelaers, Zuyd University of Applied Sciences, Sittard, the Netherlands

eKNOW 2024 Publicity Chairs
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
Laura Garcia, Universidad Politécnica de Cartagena, Spain

eKNOW 2024
Committee

eKNOW 2024 Steering Committee

Susan Gauch, University of Arkansas, USA
Martijn Zoet, Zuyd University of Applied Science, the Netherlands
Samia Aitouche, Laboratory of Automation and Manufacturing (LAP), University Batna 2, Algeria
Eric J.H.J. Mantelaers, Zuyd University of Applied Sciences, Sittard, the Netherlands

eKNOW 2024 Publicity Chairs

Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
Laura Garcia, Universidad Politécnica de Cartagena, Spain

eKNOW 2024 Technical Program Committee

Rocío Abascal Mena, Universidad Autónoma Metropolitana - Cuajimalpa, Mexico City, Mexico
Malak A. Abdullah, Jordan University of Science and Technology, Jordan
Marie-Hélène Abel, Sorbonne universités - Université de technologie de Compiègne, France
Awais Adnan, Institute of Management Sciences Peshawar, Pakistan
Nitin Agarwal, University of Arkansas at Little Rock, USA
Joyce Aguiar, Center for Psychology at University of Porto (CPUP), Portugal
Abdullah Fathi Ahmed, University Paderborn, Germany
Samia Aitouche, University Batna 2, Algeria
Arnulfo Alanis, Instituto Tecnológico de Tijuana | Tecnológico Nacional de México, Mexico
Mohammed Alqahtani, University of Arkansas, USA
Mohammad T. Alshammari, University of Hail, Saudi Arabia
Bráulio Alturas, Instituto Universitário de Lisboa (ISCTE-IUL) | ISTAR-Iscte (University Institute of Lisbon), Portugal
Gil Ad Ariely, Lauder School of Government, Diplomacy and Strategy - Interdisciplinary Center Herzliya (IDC), Israel
Mohamed Anis Bach Tobji, ESEN – University of Manouba | LARODEC Laboratory – ISG of Tunis, Tunisia
Mário Antunes, Polytechnic of Leiria, Portugal
Jorge Manuel Azevedo Santos, Universidade de Évora, Portugal
Michal Baczynski, University of Silesia in Katowice, Poland
Zbigniew Banaszak, Koszalin University of Technology, Poland
Basel Bani-Ismail, Oman College of Management and Technology, Oman
Dusan Barac, University of Belgrade, Serbia
Peter Bellström, Karlstad University, Sweden
Hajer Ben Othman, National school of computer science - University of Manouba, Tunisia
Asmaa Benghabrit, Moulay Ismaïl University, Meknès, Morocco
José Alberto Benítez Andrades, University of León, Spain
Julita Bermejo-Alonso, Universidad Isabel I, Spain
Shankar Biradar, Indian Institute of Information Technology Dharwad, India
Carlos Bobed, University of Zaragoza, Spain
Karsten Boehm, University of Applied Sciences, Kufstein, Austria

Phivos Mylonas, Ionian University, Greece
Susana Nascimento, NOVA University of Lisboa, Portugal
Samer Nofal, German Jordanian University, Jordan
Issam Nouaouri, LGI2A | Université d'Artois, France
Roy Oberhauser, Aalen University, Germany
Daniel O'Leary, University of Southern California, USA
Eva Oliveira, 2Ai Polytechnic Institute of Cávado and Ave, Barcelos, Portugal
Wiesław Paja, University of Rzeszów, Poland
Jay Patravali, Microsoft Corp., USA
João Paulo Costa, University of Coimbra, Portugal
Jean-Éric Pelet, EMLV and SKEMA Business Schools, France
Rúben Pereira, ISCTE, Portugal
António Miguel Pesqueira, Bavarian Nordic, Denmark
Sylvain Piechowiak, Université Polytechnique Hauts-de-France, France
Salviano Pinto Soares, University of Trás-os-Montes and Alto Douro (UTAD), Portugal
Rodica Potolea, Technical University of Cluj-Napoca, Romania
Adam Przybylek, Gdansk University of Technology, Poland
Paulo Quaresma, University of Évora, Portugal
Lukasz Radlinski, West Pomeranian University of Technology in Szczecin, Poland
Enayat Rajabi, Cape Breton University, Canada
Arsénio Reis, Universidade de Trás-os-Montes e Alto Douro, Portugal
Simona Riurean, University of Petrosani, Romania
Irene Rivera-Trigueros, University of Granada, Spain
Mário Rodrigues, University of Aveiro, Portugal
Alberto Rossi, Huawei, Italy
Polina Rozenshtein, Aalto University, Helsinki, Finland
Inès Saad, Amiens Business School & University Picardie Jules Verne, France
Tanik Saikh, Indian Institute of Technology Patna, India
Virginie Sans, INRISA/IRISA Université of Rennes 1, France
Lalia Saoudi, Msila University, Algeria
Antonio Sarasa Cabezuelo, Universidad Complutense de Madrid, Spain
Hartmut Schweizer, Institute for Applied Computer Science - TU Dresden, Germany
Filippo Sciarrone, ROMA TRE University, Italy
Marcelo Seido Nagano, University of São Paulo, Brazil
Houcine Senoussi, Quartz laboratory - EISTI, Cergy, France
Luciano Serafini, FBK - Fondazione Bruno Kessler, Italy
Nuno Silva, ISEP - IPP (School of Engineering - Polytechnic of Porto), Portugal
Thoudam Doren Singh, National Institute of Technology Silchar, India
Andrzej M.J. Skulimowski, AGH University of Science and Technology, Krakow, Poland
Koen Smit, Hogeschool Utrecht -Institute for ICT, Netherlands
Christophe Soares, Universidade Fernando Pessoa, Portugal
Darielson Souza, Federal University of Ceará (UFC), Brazil
Gautam Srivastava, Brandon University, Canada
Deborah Stacey, University of Guelph, Canada
Efstathios Stamatatos, University of the Aegean, Greece
Abel Suing, Universidad Técnica Particular de Loja, Ecuador
Marta Silvia Tabares, Universidad EAFIT, Medellín, Colombia
Chunxu Tang, Twitter, USA

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

A Bibliometric Analysis of Highly Cited and High Impact Blockchain Publications

Samia Aitouche
Automatic and Production Laboratory (LAP, Laboratoryof Automatio and Manufacturing), Department of Industrial Engineering, Mostapha BenBoulaid University, 05000 Batna 2,
Batna, Algeria
E-mail: s.aitouche@univ-batna2.dz

Yuh-Shan Ho*
Trend Research Centre, Asia University, No. 500 Lioufeng Road, Taichung 41354,
Taiwan
E-mail: ysho@asia.edu.tw
*Corresponding author

***Abstract*— The blockchain technology has been around for two decades already. It is not largely used because it is not enough known by businesses and its cost is still high relatively. It represents one of the emergent technologies of the fourth industrial revolution reinforcing the digitalisation of businesses by smart contracts and cryptocurrencies, handling monetary and non-monetary transactions. To show its advances in researches, a bibliometric study is performed using articles selected by citations from Web of Science database. The sample is the 100 top cited articles. Citations' indexes are calculated (total, by year, total and their average). *Y*-index is used to evaluate publication performance of authors and rank them. The most productive and cited journals, institutions and countries are identified. The most cited articles and their categories are found. There are several practical implications of this study; it offers a guideline to researchers to determine the most impacting authors, institutions, countries and articles in the domain of blockchain technology. It also helps to know the trends overtime of the blockchain.**

Keywords*— *Blockchain; web of science; Scientometrics analysis; bibliometric analysis; High Impact Blockchain Publications; Highly Cited Blockchain Publications.

I. INTRODUCTION

The blockchain technology is a database shaped in a chain of blocks, in a peer to peer network, where all allowed partners can add information concerning their partnership information or transactions without an intermediate. The modification and suppression are not allowed in a blockchain. Miners are the nodes of the network who validate the candidate block representing the added information. These miners are in competition to retrieve the hash of the block to get a reward. All the data in the blocks are encrypted to ensure more security of the data.

This technology could be used practically for monetary or non-monetary purposes. The use of blockchain fosters security against the hack of private data and its modification or suppression. This data is protected by the private and public keys used for decryption of data by the recipient. The electronic signature of data or document allowed by complicated functions of hash is a manner to authenticate the document when it is received.

The absence of intermediates in a blockchain network, helps to gain time in transactions and decrease costs of intermediation, since the network is confident. The blockchain could be applied in all industries; some of them retain the traditional databases and add a blockchain to save the sensible data, to decrease calculations of hashes and the consummation of energy. There is little bibliometric study of blockchain treating it in its generic aspect [1]. This paper will address this lack, using papers extracted from Web of Science database [2] and applying some bibliometric indicators.

The paper is structured as follows: Section II presents the set of used indicators or indexes for papers, authors, institutions and countries in the domain of blockchain. Section III is a comparative study between the found values of indicators. We will finish by a conclusion in Section IV.

II. METHODOLOGY

Document type of articles used in this study were retrieved from the Clarivate Analytics Web of Science Core Collection, the online version of the Science Citation Index Expanded (SCI-EXPANDED) (data updated on 14 May 2023). Quotation marks ("") and Boolean operator "or" were used which ensured the appearance of at least one search keyword in the terms of TOPIC (title, abstract, author keywords, and Keywords Plus). The search was conducted using a targeted keyword, including "blockchain". To ensure the analysis results are as accurate as possible, uncommon terms, such as "blockchains", "block chain", and "block chains" were also included. This approach was taken to ensure that the search is comprehensive and covers a wide range of documents related to the field of blockchain research.

The total citations from Web of Science Core Collection received since publication year till the end of the most recent year of 2022 (TC2022) [3] was used. Articles with TC2022 of 100 or more were selected as highly cited publications [4]. A total of 306 highly cited blockchain articles were found in SCI-EXPANDED from 1991 to 2022. It was pointed out that documents only searched out by Keywords Plus are irrelevant to the search topic [5]. Ho's research group firstly proposed the "front page" as a filter including the article title, abstract, and author keywords [6]. The full record in SCI-EXPANDED and the number of citations in each year for each document were checked and downloaded into Excel Microsoft 365, and additional coding was manually performed [7]. Finally, 296 articles (97% of 306 articles) including search keywords in their "front page" were defined as highly cited blockchain articles. The journal impact factors (IF2022) were taken from the Journal Citation Reports (JCR) published in 2022.

In the SCI-EXPANDED database, the corresponding author is labelled as reprint author, but in this study, we used the term corresponding author [8]. Single authors in articles with unspecified authorship were both the first as well as corresponding authors [9]. Similarly, in a single-country article, the country is classified as the first as well as the corresponding-author country. In multi-corresponding author articles, all the corresponding authors, institutions, and countries were considered. Affiliations in England, Scotland, North Ireland (Northern Ireland), and Wales were reclassified as being from the United Kingdom (UK) [10].

Publications were assessed using following citation indicators:

C_{year}: the number of citations from Web of Science Core Collection in a particular year (e.g. C2022 describes citation count in 2022) [11].

TC_{year}: the total citations from Web of Science Core Collection received since publication year till the end of the most recent year (2022 in this study, TC_{2022}) [6].

CPP_{year}: average number of citations per publication ($CPP_{2022} = TC_{2022}/TP$), *TP*: total number of publications [12].

Y-index was used to evaluate publication performance of authors. The *Y*-index is defined as [11] [9]:

Y-index (*j*, *h*)

where *j* is a constant related to the publication potential, the sum of the first-author articles and the corresponding-author articles; and *h* is a constant related to the publication characteristics, polar angle about the proportion of *RP* to *FP*.

The greater the value of *j*, the more the first- and corresponding-author contributes to the articles.

$h = \pi/2$, indicates an author that has only published corresponding-author articles, *j* is the number of corresponding-author articles ($RP > 0$ and $FP = 0$);

$\pi/2 > h > \pi/4$ indicates that an author has more corresponding-author articles than first-author articles ($FP > 0$);

$h = \pi/4$ indicates that an author has the same number of first- and corresponding-author articles ($FP > 0$ and $RP > 0$);

$\pi/4 > h > 0$ indicates an author with more first-author articles than corresponding-author articles ($RP > 0$);

$h = 0$, indicates that an author has only published first-author articles ($FP > 0$ and $RP = 0$).

III. Results and Discussion

This section will explain the found results by the bibliometric analysis performed on the set of selected articles.

A. Characteristics of publication outputs

Figure 1 shows the distribution of the highly cited articles. The most 100 highly cited articles were published in 2019. In 2016 with seven articles had the greatest CPP_{2022} of 537 which could be attributed to the most frequently cited blockchain article entitled "Blockchains and smart contracts for the internet of things" [13] with a TC_{2022} of 1,782.

In 1991, CPP_{2022} was 402 attributed to the only article entitled "Ordered structure in mixtures of a block copolymer and homopolymers. 1. Solubilization of low-molecular-weight homopolymers" [14] published in 1991.

B. Web of Science Category and Journal

In 2021, Journal Citation Reports (JCR) indexed 9,649 journals with citation references across 178 Web of Science categories in SCI-EXPANDED.

Total 106 journals published highly cited articles related to blockchain in 45 Web of Science categories in SCI-EXPANDED mainly in information systems computer science with 114 articles (39% of 296 articles), immunology with 150 articles (10%), telecommunications with 104 articles (35%), and electrical and electronic engineering with 97 articles (33%).

It should be noticed that journals could be classified in two or more categories in Web of Science Core Collection, for instance *IEEE Access* was classified in information systems computer science, electrical and electronic engineering, and telecommunications, thus the sum of percentages was greater than 100% [4]. A fuzzy classification of journals could be a pertinent solution for the scientific database or in bibliometric study's authors should consider only one class, but the results will be biased.

Six of the 106 journals had 10 highly cited articles or more, including *IEEE Access* ($IF_{2021} = 3.476$) with 36 articles (12% of 296 articles), *IEEE Transactions on Industrial Informatics* ($IF_{2021} = 11.648$) with 22 articles (7.4%), *IEEE Internet of Things Journal* ($IF_{2021} = 10.238$) with 19 articles (6.4%), *IEEE Communications Surveys and Tutorials* ($IF_{2021} = 33.840$) with 14 articles (4.7%), *Future Generation Computer Systems-the International Journal of Escience* ($IF_{2021} = 7.307$) with 12 articles (4.1%), and *International Journal of Production Research* ($IF_{2021} = 9.018$) with 10 articles (3.4%). According to IF_{2021}, the top three journals have an IF_{2021} of more than 30 were the *Nature* ($IF_{2021} = 69.504$) with one article, the *Joule* ($IF_{2021} = 46.048$) with one article, and the *IEEE Communications Surveys and Tutorials* ($IF_{2021} = 33.840$) with 14 articles.

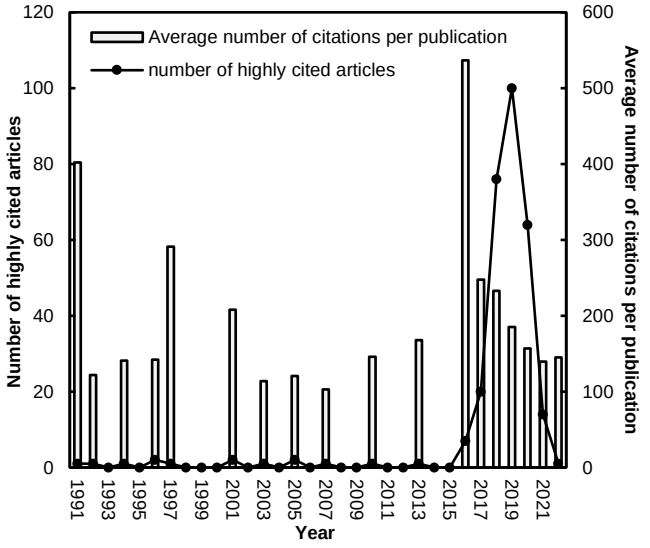

Figure 1. Number of blockchain articles and their average citations per publication by year.

C. Publication performances: countries and institutions

There was one highly cited blockchain articles (0.34% of 296 articles) without affiliations in SCI-EXPANDED. A total of 295 highly cited articles were published by authors affiliated from 53 countries including 125 single-country articles (42% of 1,507 articles) published by authors from 24 countries and 170 internationally collaborative articles (58%) published by authors from 52 countries. Six publication indicators [14] were applied to compare the top 15 productive countries (Table 1).

China dominated in five of the six publication indicators with a *TP* of 136 articles (46% of 295 articles), an *IPC* of 41 articles (33% of 125 single-country articles), an *ICP* of 95 articles (56% of 170 internationally collaborative articles), an *FP* of 114 articles (39% of 295 first-author articles), an *RP* of 99 articles (34% of 293 corresponding-author articles), while the USA ranked top with an *SP* of four articles (40% of 10 single-author articles).

At the institutional level, the determined institution of the corresponding author might be a home base of the study or origin of the paper [11]. Concerning institutions, 65 blockchain articles (22% of 295 articles) originated from single institutions, 60 articles (20%) were national collaborations, and 170 articles (58%) were international collaborations. Seven publication indicators [15] were applied to compare the top 16 productive institutions (Table 2). Out of the top 16 institutions, nine were in China, while the remaining seven were spread across the globe, with two in the USA, and one each in Singapore, Saudi Arabia, Canada, Australia, and Norway.

TABLE I. TOP 15 PRODUCTIVE COUNTRIES.

Country	*TP*	*TP* (%)	IP_C (%)	*ICP* (%)	*FP* (%)	*RP* (%)	*SP* (%)
China	136	1 (46)	1 (33)	1 (56)	1 (39)	1 (34)	2 (10)
USA	86	2 (29)	2 (23)	2 (34)	2 (13)	2 (14)	1 (40)
UK	32	3 (11)	5 (4.0)	3 (16)	4 (4.1)	5 (4.4)	N/A
Australia	31	4 (11)	3 (5.6)	4 (14)	3 (5.1)	3 (7.2)	2 (10)
Singapore	21	5 (7.1)	N/A	5 (12)	7 (3.7)	10 (2.4)	N/A
Canada	21	5 (7.1)	16 (0.80)	6 (12)	24 (0.34)	7 (2.7)	N/A
India	20	7 (6.8)	10 (1.6)	7 (11)	4 (4.1)	10 (2.4)	N/A
South Korea	18	8 (6.1)	3 (5.6)	8 (6.5)	4 (4.1)	4 (5.1)	N/A
Italy	14	9 (4.7)	7 (3.2)	10 (5.9)	7 (3.7)	6 (3.8)	N/A
Japan	13	10 (4.4)	5 (4.0)	14 (4.7)	10 (2.4)	12 (2.0)	N/A
Germany	11	11 (3.7)	16 (0.8)	10 (5.9)	9 (2.7)	7 (2.7)	N/A
Taiwan	11	11 (3.7)	10 (1.6)	12 (5.3)	12 (1.4)	12 (2.0)	2 (10)
Norway	11	11 (3.7)	N/A	8 (6.5)	24 (0.34)	7 (2.7)	N/A
France	10	14 (3.4)	10 (1.6)	14 (4.7)	11 (2.0)	14 (1.7)	N/A
Saudi Arabia	9	15 (3.1)	N/A	12 (5.3)	13 (1.0)	15 (1.4)	N/A

TP: total number of highly cited articles; *TP R* (%): rank of percentage of total number of articles in all articles; IP_C *R* (%): rank and percentage of single-country articles in all single-country articles; *ICP R* (%): rank and percentage of internationally collaborative articles in all internationally collaborative articles; *FP R* (%): rank and the percentage of first-author articles in all first-author articles; *RP R* (%): rank and the percentage of corresponding-author articles in all corresponding-author articles; *SP R* (%): rank and the percentage of first-author articles in all first-author articles; N/A: not available.

The Beijing University of Posts and Telecommunications in China ranked the top with a *TP* of 14 articles (4.7% of 295 articles) and an *FP* of 10 articles (3.4% of 295 first-author articles) while the Beijing Institute of Technology in China ranked the top with an *RP* of nine articles (3.1% of 293 corresponding-author articles). The Worcester Polytechnic Institute in USA ranked the top with an IP_I of three articles (4.6% of 65 single-institution articles). The Kyoto University in Japan had three highly cited articles all of which were single-institution articles. The Hong Kong Polytechnic University in China ranked the top with an *NCP* of five articles (8.3% of 60 nationally collaborative articles) and an *SP* of one article (10% of 10 single-author articles). Only 12 institutions published single-author articles respectively. The Hong Kong Polytechnic University was the only one ranked in the top 16 in total highly cited articles. The Nanyang Technological University in Singapore ranked the top with an *ICP* of 13 articles (7.6% of 170 internationally collaborative articles). Only two of the nine institutions in China the Hong Kong Polytechnic University and the Xidian University had single-institution articles.

TABLE II. TOP 16 PRODUCTIVE INSTITUTIONS.

Institution, Country	*TP*	*TP* *R* (%)	IP_I *R* (%)	*NCP* *R* (%)	*ICP* *R* (%)	*FP* *R* (%)	*RP* *R* (%)	*SP* *R* (%)
Beijing University of Posts and Telecommunications, China	14	1 (4.7)	N/A	6 (3.3)	2 (7.1)	1 (3.4)	2 (2.4)	N/A
Nanyang Technological University, Singapore	13	2 (4.4)	N/A	N/A	1 (7.6)	2 (3.1)	5 (1.7)	N/A
University of Electronic Science and Technology of China, China	12	3 (4.1)	N/A	6 (3.3)	3 (5.9)	4 (2.7)	5 (1.7)	N/A
Hong Kong Polytechnic University, China	12	3 (4.1)	8 (1.5)	1 (8.3)	6 (3.5)	4 (2.7)	13 (1.0)	1 (10)
University of Oslo, Norway	10	5 (3.4)	N/A	N/A	3 (5.9)	N/A	2 (2.4)	N/A
Academy of Sciences, China	10	5 (3.4)	N/A	2 (6.7)	6 (3.5)	7 (1.7)	8 (1.4)	N/A
Guangdong University of Technology, China	9	7 (3.1)	N/A	2 (6.7)	13 (2.9)	6 (2.0)	8 (1.4)	N/A
Beijing Institute of Technology, China	9	7 (3.1)	N/A	6 (3.3)	5 (4.1)	2 (3.1)	1 (3.1)	N/A
University of Texas San Antonio, USA	8	9 (2.7)	8 (1.5)	19 (1.7)	6 (3.5)	16 (0.68)	8 (1.4)	N/A
Shanghai Jiao Tong University, China	8	9 (2.7)	N/A	6 (3.3)	6 (3.5)	8 (1.4)	4 (2.0)	N/A
University of Academy of Sciences, China	7	11 (2.4)	N/A	4 (5.0)	17 (2.4)	N/A	45 (0.34)	N/A
Worcester Polytechnic Institute, USA	7	11 (2.4)	1 (4.6)	6 (3.3)	40 (1.2)	8 (1.4)	8 (1.4)	N/A
University of Technology Sydney, Australia	6	13 (2.0)	N/A	N/A	6 (3.5)	38 (0.34)	20 (0.68)	N/A
Xidian University, China	6	13 (2.0)	3 (3.1)	N/A	17 (2.4)	12 (1.0)	8 (1.4)	N/A
King Saud University, Saudi Arabia	6	13 (2.0)	N/A	N/A	6 (3.5)	38 (0.34)	13 (1.0)	N/A
Carleton University, Canada	6	13 (2.0)	N/A	N/A	6 (3.5)	N/A	20 (0.68)	N/A

TP: total number of highly cited articles; *TP R* (%): rank of percentage of total number of articles in all articles; IP_I *R* (%): rank and percentage of single-institution articles in all single-institution articles; *NCP R* (%): rank and percentage of nationally collaborative articles in all nationally collaborative articles; *ICP R* (%): rank and percentage of internationally collaborative articles in all internationally collaborative articles; *FP R* (%): rank and the percentage of first-author articles in all first-author articles; *RP R* (%): rank and the percentage of corresponding-author articles in all corresponding-author articles; *SP R* (%): rank and the percentage of first-author articles in all first-author articles; N/A: not available.

D. Publication performances: authors

Table 3 lists the top 15 most productive authors with five highly cited blockchain articles or more. Y. Zhang was the most productive author with 16 highly cited articles including two first-author articles, nine corresponding-author articles. Y. Zhang also ranked the top in corresponding-author articles. T.M. Choi and J.W. Kang with six highly cited articles published the most five first-author articles, respectively. T.M. Choi was also the only author had singly-author articles in the top 123 productive authors. Eight of the 15 productive authors including Y. Zhang, L.H. Zhu, J.H. Park, T.M. Choi, Z.H. Xiong, J.W. Kang, P.K. Sharma, and K.K. Gai were found to be the top 15 publication potential authors as evaluated by Y-index.

In the total of 290 highly cited blockchain articles (98% of 296 highly cited articles) had both first and corresponding authors information in SCI-EXPANDED, were extensively investigated based on the *Y*-index. The 290 highly cited blockchain articles were contributed by 1,061 authors in which 664 authors (63% of 290 authors) had no first- and no corresponding-author articles with *Y*-index (0, 0); 144 (14%) authors published only corresponding-author articles with $h = \pi/2$; 12 (1.1%) authors published more corresponding-author articles than first-author articles with $\pi/2 > h > \pi/4$ ($FP > 0$); 98 (9.2%) authors published the same number of first- and corresponding-author articles with $h = \pi/4$ ($FP > 0$ and $RP > 0$); 7 (0.66%) authors published more first-author articles than corresponding-author articles with $\pi/4 > h > 0$

($RP > 0$); and 136 (13%) authors published only first-author articles with $h = 0$.

TABLE III. TOP 15 PRODUCTIVE AUTHORS WITH FIVE HIGHLY CITED ARTICLES OR MORE

Author	rank (*TP*)	rank (*FP*)	rank (*RP*)	rank (*SP*)	*h*	rank (*j*)
Y. Zhang	1 (16)	7 (2)	1 (9)	N/A	1.352	1 (11)
D. Niyato	2 (9)	N/A	N/A	N/A	0	398 (0)
L.H. Zhu	3 (8)	26 (1)	2 (5)	N/A	1.373	2 (6)
J.H. Park	4 (7)	26 (1)	2 (5)	N/A	1.373	2 (6)
J. Sarkis	4 (7)	N/A	7 (2)	N/A	π/2	31 (2)
K.K.R. Choo	4 (7)	N/A	5 (3)	N/A	π/2	19 (3)
Z.H. Xiong	7 (6)	7 (2)	5 (3)	N/A	0.9828	5 (5)
F.R. Yu	7 (6)	N/A	7 (2)	N/A	π/2	31 (2)
S. Maharjan	7 (6)	N/A	N/A	N/A	0	398 (0)
J.W. Kang	7 (6)	1 (5)	N/A	N/A	0	5 (5)
T.M. Choi	7 (6)	1 (5)	33 (1)	1 (1)	0.1974	2 (6)
P.K. Sharma	12 (5)	3 (4)	33 (1)	N/A	0.245	5 (5)
V.C.M. Leung	12 (5)	N/A	33 (1)	N/A	π/2	138 (1)
K.K. Gai	12 (5)	3 (4)	33 (1)	N/A	0.245	5 (5)
M. Guizani	12 (5)	N/A	N/A	N/A	0	398 (0)

TP: total number of highly cited articles; *FP*: first-author articles; *RP*: corresponding-author articles; *SP*: single-author articles; *j*: a *Y*-index constant related to the publication potential; *h*: a *Y*-index constant related to the publication characteristics; N/A: not available.

In the polar coordinates (Figure 2), the distribution of the *Y*-index (*j*, *h*) of the leading 137 potential authors in blockchain research with $j \geq 2$ was demonstrated. Every point has a coordinate *Y*-index (*j*, *h*) that could symbolize a single author or multiple authors, for example, X.H. Huang, Y.H. Zhang, L.D. Xu, S. Saberi, K. Salah, M. Holbl, C. Liu, M.S. Hossain, F.J. Luo, K. Fan, and other 80 authors who published only one highly cited article with *Y*-index (2, π/4). Y. Zhang with *Y*-index (11, 1.352) had the greatest publication potential in highly cited blockchain articles (did not show in the figure), followed distantly by L.H. Zhu (6, 1.373), J.H. Park (6, 1.373), and T.M. Choi (6, 0.1974) respectively. Zhu and Park had the same *Y*-index shows that they have the same publication potential and the publication characteristics. Zhu, Park, and Choi had the same *j* of 6.

These authors are located on the same curve ($j = 6$) in Figure 2, indicating that they had the same publication potential in blockchain research with a *j* of 6 but different publication characteristics (Ho and Hartley, 2016b).

Zhu and Park published more corresponding-author articles than first-author articles with an *h* of 1.373 while Choi published more first-author articles than corresponding-author articles with an *h* of 0.1974. Similarly, K.K.R. Choo with *Y*-index (3, π/2); Z. Su, M. Shen, A. Zhang, Q. Xia, Y. Yuan, D. Ivanov, and S. Ding with the same *Y*-index (3, 1.107); Z.B. Zheng, J. Wang, and B. Cao with the same *Y*-index (3, 0.4636); and M.T. Liu (3, 0) are located on the same curve with *j* of 3.

These authors had the same publication potential with an *j* of 3 but different publication characteristics. Choo published only three corresponding-author articles with an *h* of π/2. Su, Shen, Zhang, Xia, Yuan, Ivanov, and Ding had higher ration of corresponding-author articles to first-author articles with an *h* of 1.107. Zheng, Wang, and Cao had higher ration of first-author articles to corresponding-author articles with an *h* of 0.4636. Finally, Liu published only three first-author articles with an *h* of 0. Similar situation for authors located on *j* of 5, 4, and 2 was also found. W. Viriyasitavat, P. Zhang, M. Ul Hassan, N. Kshetri, A. Dorri, M.A. Ferrag, I. Eyal, and C. Esposito with the same *Y*-index (4, π/4) and X.H. Huang and other 89 authors with the same *Y*-index (2, π/4) are located on the diagonal line ($h = \pi/4$) indicating that they had the same publication characteristics but different publication potential.

Viriyasitavat and other seven authors had the greatest publication potential with a *j* of 4 followed by Huang and other 89 authors with a *j* of 2. K.K.R. Choo with *Y*-index (3, π/2) and Y.L. Teng, J.Y. Wang, J. Weng, K. Wang, X.N. Wang, J. Ren, D.I. Kim, M. Kraft, J. Sarkis, F.R. Yu, and P. Wang with the same *Y*-index (2, π/2) are located on the straight line (y-axis with $h = \pi/2$) had the same publication characteristics. Choo had higher publication potential with a *j* of 3 than others with a *j* of 2. Similarly, J.W. Kang (5, 0), M.T. Liu (3, 0), and S. Wang, Y. Xu, Z. Li, J.W. Leng, W. Liang, and Y.L. Lu with the same *Y*-index (2,0) are located on the straight line (x-axis with $h = 0$) also had the same publication characteristics. Kang had the greatest publication potential with a *j* of 5, followed by Liu with a *j* of 3 and Wang, Xu, Li, Leng, Liang, and Lu with a *j* of 2.

The location on the graph along with one of the curves or along a straight line from the origin represents different families of author publication potential or publication characteristics, respectively. A potential for bias in the analysis of authorship might attributes to different authors having the same name, or the same author using different names over time, especially for Chinese authors [8].

E. The top ten most frequently cited articles in blockchain research

Total citations are updated from time to time on the Web of Science Core Collection. To improve bibliometric study, the total number of citations from the Web of Science Core Collection since publication year until to the end of the most recent year of 2022 (TC_{2022}) was applied to improve the bias using data from the database directly [3]. A total of 245 articles (83% of 296 articles), 279 articles (96% of 291 articles with abstract in SCI-EXPANDED), and 242 articles (93% of 261 articles with author keywords in SCI-EXPANDED) contain search keywords in their title, abstract, and author keywords respectively. Table 4 shows the top 10 most frequently cited articles on blockchain research.

The top ten articles were published from 2016 to 2019. Articles by Xu et al. (2018) and Tschorsch and Scheuermann (2016) only contained search keywords in the author keywords. Article by Mengelkamp et al. (2018) contained search keywords in search keywords in the author keywords and abstract. Seven of the top ten articles contained search keywords in the title, abstract, and author keywords. These articles are directly related to blockchain research. Citations of a highly cited article are not always high [4]. It is necessary to understand citation history of a classic article. The citation histories of the top ten blockchain articles are shown in Figure 3. Articles by [13], Xu et al. (2018), Zheng et al. (2018), and Saberi et al. (2019) had sharper citation increasing after their publication. However, all the top articles had citation decreasing after three years. Blockchain is a nascent research topic, and in its initial stages, subjects are being explored and refined through testing.

The highly cited articles were not only the most frequently cited but also the most impactful in the recent

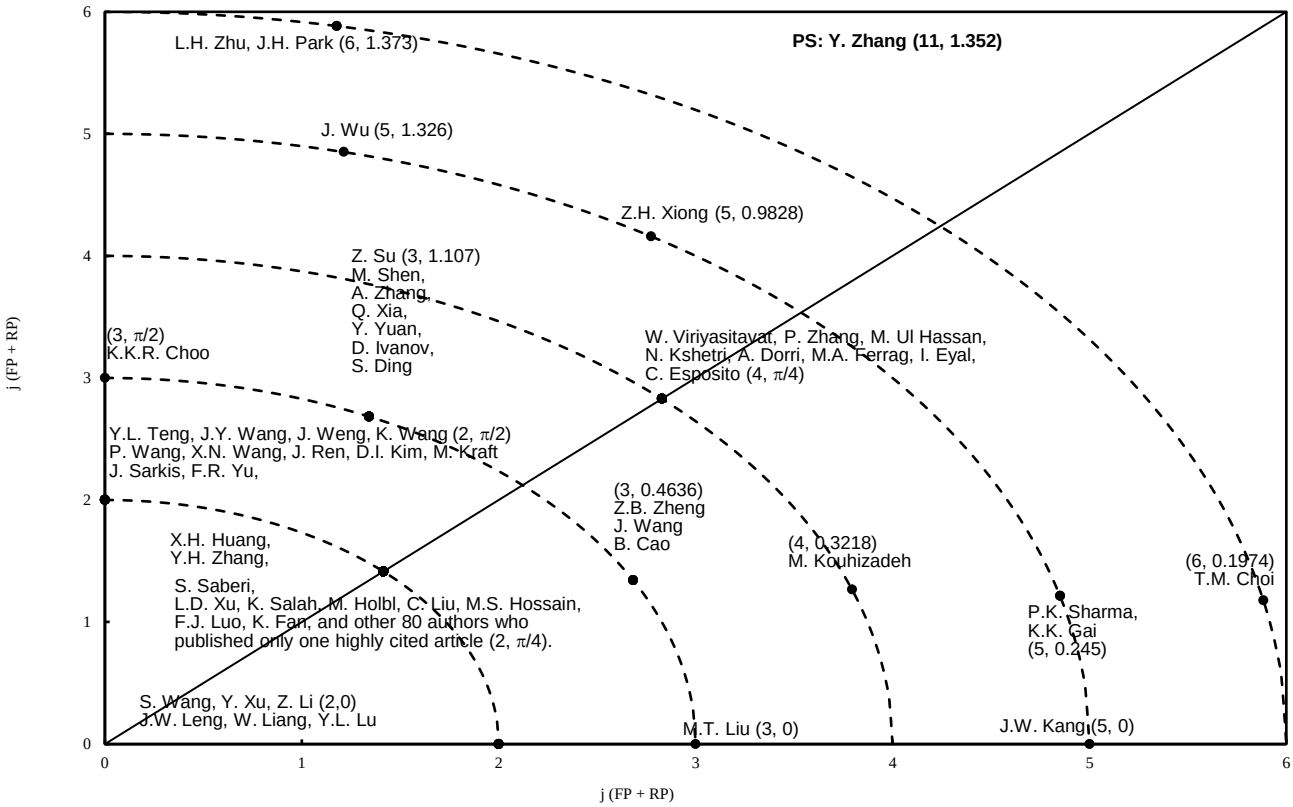

Figure 2. Top 137 authors with Y-index (j ≥ 2).

year 2022 in blockchain research. Six of the top ten most impactful articles were also ranked in the top ten most frequently cited were summarized as:

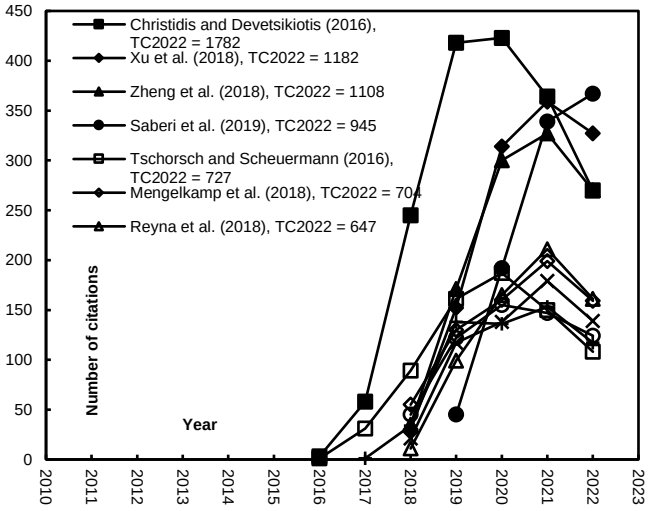

Figure 3. The citation histories of the top ten most frequently cited articles on blockchain research.

- Blockchain technology and its relationships to sustainable supply chain management [16], the articles published by four authors from the Worcester Polytechnic Institute in the USA with a TC_{2022} of 367 (rank 1st in blockchain research) and a TC_{2022} of 945 (rank 4th).
- Industry 4.0: state of the art and future trends [17], the articles published by three authors from the Old Dominion University and the University of Minnesota in the USA with a C_{2022} of 327 (rank 2nd in blockchain research) and a TC_{2022} of 1,182 (rank 2nd).
- Blockchains and Smart Contracts for the Internet of Things [13], the articles published by two authors from the North Carolina State University in the USA with a C_{2022} of 270 (rank 3rd in blockchain research) and a TC_{2022} of 1,782 (rank 1st).
- Blockchain challenges and opportunities: a survey [18], the articles published by five authors from the Sun Yat Sen University, the Macau University of Science and Technology, and the National University of Defense Technology in China with a C_{2022} of 270 (rank 3rd in blockchain research) and a TC_{2022} of 1,108 (rank 3rd).
- On blockchain and its integration with IoT. Challenges and opportunities [20], the articles published by five authors from the University of Malaga in Spain with a C_{2022} of 161 (rank 8th in blockchain research) and a TC_{2022} of 647 (rank 7th).
- Designing microgrid energy markets A case study: The Brooklyn Microgrid [21], the articles published by six authors from the Karlsruhe Institute of Technology in Germany and L03 Energy in the USA with a C_{2022} of 159 (rank 9th in blockchain research) and a TC_{2022} of 704 (rank 6th).

IV. CONCLUSION

The conducted bibliometric study about the blockchain in this paper allowed the calculation of several bibliomtric indicators to rank authors, countries, institutions, articles and their categories according to the database, using essentially the scientific impact on scientific community. It offers a guide for novel scientific researchers on the technology blockchain to know the authors and institutions pioneers in the domain, to establish synergies and collaborations.

TABLE IV. THE TOP TEN MOST FREQUENTLY CITED ARTICLES IN BLOCKCHAIN RESEARCH

Rank (TC_{2022})	Rank (C_{2022})	Title	Country
1 (1,782)	3 (270)	Blockchains and smart contracts for the internet of things [13]	USA
2 (1,182)	2 (327)	Industry 4.0: State of the art and future trends [17]	USA
3 (1,108)	3 (270)	Blockchain challenges and opportunities: A survey [18]	China
4 (945)	1 (367)	Blockchain technology and its relationships to sustainable supply chain management [16]	USA
5 (727)	28 (108)	Bitcoin and beyond: A technical survey on decentralized digital currencies [19]	Germany
6 (704)	9 (159)	Designing microgrid energy markets A case study: The Brooklyn Microgrid [21]	Germany, USA
7 (647)	8 (161)	On blockchain and its integration with loT. Challenges and opportunities [20]	Spain
8 (599)	18 (124)	Security and privacy in decentralized energy trading through multi-signatures, blockchain and anonymous messaging streams [22]	U Arab Emirates
9 (594)	13 (139)	Consortium blockchain for secure energy trading in industrial internet of things [23]	China, Norway
10 (579)	23 (117)	Enabling localized peer-to-peer electricity trading among plug-in hybrid electric vehicles using consortium blockchains [24]	China, Norway, Canada

TC_{2022}: the total number of citations from Web of Science Core Collection since publication year to the end of 2022; C_{2022}: number of citations of an article in 2022 only.

This study helps to know how the researches are in advance about the blockchain to encourage businesses to invest in it and gain security of data, transparency and efficiency of the information system.

REFERENCES

[1] www.scopus.com, Scopus database, March, 2024.

[2] www.webofknowledge.com, Web of Science database, March, 2024.

[3] M. H. Wang and Y. S. Ho, "Research articles and publication trends in environmental sciences from 1998 to 2009," Archives of Environmental Science, vol. 5, pp. 1-10, 2011.

[4] Y. S. Ho, A bibliometric analysis of highly cited articles in materials science. Current Science, vol. 107 (9), pp. 1565-1572, 2014.

[5] H. Z Fu and Y. S. Ho, "Top cited articles in thermodynamic research," Journal of Engineering Thermophysics, vol. 24 (1), pp. 68-85. 2015, DOI: 10.1134/S1810232815010075.

[6] M. H. Wang, H. Z. Fu, and Y. S. Ho, "Comparison of universities' scientific performance using bibliometric indicators," Malaysian Journal of Library & Information Science, vol. 16 (2), pp. 1-19, 2011.

[7] E. A. Al-Moraissi, N. Christidis and Y. S. Ho, "Publication performance and trends in temporomandibular disorders research: A bibliometric analysis," Journal of Stomatology Oral and Maxillofacial Surgery, vol. 124 (1), Article Number: 101273, 2023. DOI: 10.1016/j.jormas.2022.08.016.

[8] W. T. Chiu and Y. S. Ho, "Bibliometric analysis of tsunami research," Scientometrics, vol. 73 (1), pp. 3-1, 2007, DOI: 10.1007/s11192-005-1523-1.

[9] Y. S. Ho, "Classic articles on social work field in Social Science Citation Index: A bibliometric analysis," Scientometrics, vol. 98 (1), pp. 137-155, 2014. DOI: 10.1007/s11192-013-1014-8.

[10] M. Ming Chiu and E. Sui Chu Ho, "Family effects on student achievement in Hong Kong." Asia Pacific Journal of Education, vol. 26 (1), pp. 21-35, 2006.

[11] Y. S. Ho, "Top-cited articles in chemical engineering in Science Citation Index Expanded: A bibliometric analysis," Chinese Journal of Chemical Engineering, vol. 20 (3), pp. 478-488, 2012, DOI: 10.1016/S1004-9541(11)60209-7.

[12] Y. S. Ho, "The top-cited research works in the Science Citation Index Expanded," Scientometrics, vol. 94 (3), pp. 1297-1312. 2013, DOI: 10.1007/s11192-012-0837-z.

[13] K. Christidis and M. Devetsikiotis, "Blockchains and smart contracts for the internet of things," IEEE Access, vol. 4, pp. 2292-2303, 2016, DOI: 10.1109/ACCESS.2016.2566339.

[14] Y. H. E. Hsu, and Y. S. Ho, "Highly cited articles in health care sciences and services field in Science Citation Index Expanded: A bibliometric analysis for 1958-2012,", Methods of Information in Medicine, vol. 53 (6), pp. 446-458, 2014, DOI: 10.3414/ME14-01-0022.

[15] H. Z Fu, X. Long and Y. S. Ho, "China's research in chemical engineering journals in Science Citation Index Expanded: A bibliometric analysis," Scientometrics, vol. 98 (1), pp. 119-136, 2014. DOI: 10.1007/s11192-013-1047-z.

[16] S. Saberi, M. Kouhizadeh, J. Sarkis, L. Shen, "Blockchain technology and its relationships to sustainable supply chain management," International Journal of Production Research, vol. 57 (7), pp. 2117-2135, 2019.

[17] L.D. Xu, E. L. Xu and L. Li, "Industry 4.0: State of the art and future trends,", International Journal of Production Research, vol. 56 (8), pp. 2941-2962, 2018, DOI: 10.1080/00207543.2018.1444806.

[18] Z. B. Zheng, S. A. Xie, H. N. Dai, X. P. Chen and H. M. Wang, "Blockchain challenges and opportunities: A survey," International Journal of Web and Grid Services, vol. 14 (4), pp. 352-375, 2018, DOI: 10.1504/IJWGS.2018.095647.

[19] F. Tschorsch and B. Scheuermann, "Bitcoin and beyond: A technical survey on decentralized digital currencies," IEEE Communications Surveys & Tutorials, vol. 18 (3), pp. 2084-2123, 2016. DOI: 10.1109/COMST.2016.2535718.

[20] A. Reyna, C. Martín, J. Chen, E. Soler and M. Díaz, On blockchain and its integration with IoT. Challenges and opportunities. Future Generation Computer Systems-the International Journal of eScience, 88, pp. 173-190, 2018, DOI: 10.1016/j.future.2018.05.046.

[21] E. Mengelkamp et al., "Designing microgrid energy markets A case study: The Brooklyn Microgrid," Applied Energy, vol. 210, pp. 870-880, 2018.

[22] N. Z. Aitzhan and D. Svetinovic, "Security and privacy in decentralized energy trading through multi-signatures, blockchain and anonymous messaging streams," IEEE Transactions on Dependable and Secure Computing, vol. 15 (5), pp. 840-852, 2018, DOI: 10.1109/TDSC.2016.2616861.

[23] Z. Li et al., "Consortium blockchain for secure energy trading in industrial internet of things," IEEE transactions on industrial informatics, vol. 14 (8), pp. 3690-3700, 2017.

[24] J. W. Kang et al., "Enabling localized peer-to-peer electricity trading among plug-in hybrid electric vehicles using consortium blockchains," IEEE Transactions on Industrial Informatics, vol. 13 (6), pp. 3154-3164, 2017, DOI: 10.1109/TII.2017.2709784.

[25] Y. S. Ho and J. Hartley, "Classic articles published by American scientists (1900-2014): A bibliometric analysis," Current Science, vol. 111 (7), pp. 1156-1165, 2016, DOI: 10.18520/cs/v111/i7/1156-1165.

[26] Z. T. Li, J. W. Kang, R. Yu, D. D. Ye, Q. Y. Deng. and Y. Zhang, "Consortium blockchain for secure energy trading in industrial internet of things," IEEE Transactions on Industrial Informatics, vol. 14 (8), pp. 3690-3700, 2018, DOI: 10.1109/TII.2017.2786307.

[27] S. Saberi, M. Kouhizadeh, J. Sarkis, and L. J. Shen, "Blockchain technology and its relationships to sustainable supply chain management," International Journal of Production Research, vol. 57 (7), pp. 2117-2135, 2019, DOI: 10.1080/00207543.2018.15332.

Intermediate-Task Transfer Learning: Leveraging Sarcasm Detection for Stance Detection

Gibson Nkhata, Susan Gauch
Department of Electrical Engineering & Computer Science
University of Arkansas
Fayetteville, AR 72701, USA
Emails: gnkhata@uark.edu, sgauch@uark.edu

***Abstract*—Stance Detection (SD) in the context of social media has emerged as a prominent area of interest with implications for social, business, and political applications, thereby garnering escalating research attention within the realm of Natural Language Processing (NLP). The inherent subtlety, nuance, and complexity of texts procured from online platforms via crowd-sourcing pose challenges for SD algorithms in accurately discerning the author's stance. Particularly, the inclusion of sarcastic and figurative language drastically impacts the performance of SD models. This paper addresses this challenge by employing sarcasm detection intermediate-task transfer learning tailored for SD. The proposed methodology involves the fine-tuning of BERT and RoBERTa and the sequential concatenation of convolutional, bidirectional LSTM, and dense layers. Rigorous experiments are conducted on publicly available benchmark datasets to evaluate our transfer-learning framework. The performance of the approach is assessed against various State-Of-The-Art (SOTA) baselines for SD, providing empirical evidence of its effectiveness. Notably, our model outperforms the best SOTA models, achieving average F1-score gaps of 0.038 and 0.053 on the SemEval 2016 Task 6A Dataset (SemEval) and Multi-Perspective Consumer Health Query Data (MPCHI), respectively, even prior to sarcasm-detection pre-training. The integration of sarcasm knowledge into the model proves instrumental in mitigating misclassifications of sarcastic textual elements in SD. Our model accurately predicts 85% of texts that were previously misclassified by the model without sarcasm-detection pre-training, thereby amplifying the average F1-score of the model. Furthermore, our experiments revealed that the success of the transfer-learning framework is contingent upon the correlation of lexical attributes between the intermediate task (sarcasm detection) and the target task (SD). This study represents the first exploration of sarcasm detection as an intermediate transfer-learning task in the context of SD and simultaneously exploits the concatenation of BERT or RoBERTa with other deep-learning techniques, establishing the proposed approach as a foundational baseline for future research endeavors in this domain.**

***Keywords*-Stance detection; sarcasm detection; transfer learning; BERT; RoBERTa.**

I. INTRODUCTION

Social media platforms, increasingly popular, enable individuals to freely express opinions and connect globally for real-time updates on diverse topics [1]–[3]. Discourse on emerging subjects yields substantial data valuable for Natural Language Processing (NLP) tasks, notably Stance Detection (SD). SD is the automated identification of an individual's stance on a specific topic based solely on their utterance or authored material [2][4]–[6]. Stance labels categorize expressions into *InFavor*, *Against*, or *None*. This phenomenon, particularly on social media, is a burgeoning focus in social, business, and political applications [3][7].

Previous SD research has been evaluated using the publicly available datasets crowd-sourced from online platforms [2][3][5][8]. However, texts procured from online platforms are often characterized by subtlety, nuance, and complexity, featuring inherent sarcastic and figurative language. This complexity poses challenges for SD algorithms in accurately discerning the author's stance [2]. Additionally, targets are not consistently mentioned in text [4], and stances are not explicitly transparent. Consequently, inferring the author's stance becomes further complicated, often necessitating implicit inference through a combination of interaction, historical context, and social linguistic attributes, such as sarcasm or irony.

Prior work has explored intermediate-task transfer learning, involving fine-tuning a model on a secondary task before its application to the primary task to address the aforementioned challenge [1][9]–[13]. Specifically, [10] and [13] utilized sentiment classification to enhance their models for SD. In a similar vein, [1] incorporated emotion and sentiment classification prior to sarcasm detection. The study by [1] suggested that pre-training a model with sentiment analysis before sarcasm detection enhances overall performance, attributing this improvement to the correlation between sarcasm and an implied negative sentiment. This finding aligns with one of our experimental observations in Section IV, wherein most sarcastic sentences with an "Against" stance were initially misclassified as "InFavor" before the integration of sarcasm pre-training into our model. Nonetheless, sarcasm language in the target tasks has detrimentally affected performance, and previous research has not explored the sarcasm phenomenon for enhancing SD models. In this study, our focus is to experiment with and employ sarcasm detection as an intermediate task tailored to improve SD performance.

Sarcasm detection involves inferring intention or secondary meanings from an utterance, discounting literal meaning [14]. It employs positive words and emotions to convey negative or undesirable figurative attributes, serving as a mechanism to express opinions using seemingly conflicting language [15]–[17]. Sarcasm can alter the stance of a text from *Against* to *InFavor* and vice versa [16][18]. Thus, we propose infusing sarcasm knowledge into the model before SD fine-tuning to enhance performance.

This work employs a model framework consisting of BERT [19] or RoBERTa [20], two convolutional layers (Conv), a Bidirectional LSTM layer (BiLSTM), and a dense layer. Experimental results affirm the efficacy of our approach, demonstrated by improved F1-scores upon the inclusion of sarcasm detection in the model framework. Furthermore, the significance of this approach is emphasized by presenting a sample of sarcastic texts from datasets during a failure analysis of the original SD model results, prior to the incorporation of sarcasm intermediate-task pre-training. Exploring three publicly available sarcasm datasets, we find that different sarcasm detection tasks impact SD performance variably, depending on linguistic and quantitative attributes. Our work makes the following key contributions:

- *Transfer-Learning Framework:* Introducing a novel transfer-learning framework incorporating sarcasm detection as an intermediate task before fine-tuning on SD, utilizing an integrated deep learning model.
- *Performance Superiority:* Demonstrating superior performance against State-Of-The-Art (SOTA) SD baselines, even without sarcasm detection pre-training, indicated by higher F1-scores..
- *Correlation Analysis:* Establishing and illustrating the correlation between sarcasm detection and SD, exemplified through a failure analysis, thereby emphasizing the improvement of SD through sarcasm detection.
- *Impact Assessment:* Measuring the impact of various sarcasm detection models on target tasks based on the correlation between linguistic and quantitative attributes in the datasets of the two tasks.
- *Ablation Study:* Conducting an ablation study to assess the contribution of each module to the overall model framework. The study also reveals a significant drop in performance without sarcasm knowledge, underscoring the importance of our proposed approach.

The remainder of this paper unfolds as follows: Section II reviews related work, Section III outlines our proposed approach, and Section IV delves into comprehensive experiments, covering datasets, results, and subsequent discussions. The conclusion and recommendations for further study are provided in Section V. The final section critically examines the limitations inherent in our study.

II. Related work

This section conducts a literature review on SD and intermediate-task transfer learning.

A. Stance Detection (SD)

The literature on SD has traditionally explored two primary perspectives: Target-Specific SD (TSSD), focusing on individual targets [2][3][21][22], and Multi-Target SD (MTSD), concurrently inferring stances towards multiple related subjects [22]–[25]. Early SD approaches utilized rule-based methods [21][26], followed by classical machine learning techniques [27][28]. Later, the emergence of deep learning models led to neural networks supplanting classical approaches [4][13][29][30]. For instance, a neural ensemble model incorporating BiLSTM, attention mechanism, and multi-kernel convolution was presented in [29], evaluated on both TSSD and MTSD. While our work shares similarities in model framework, it distinctively employs BERT or RoBERTa and introduces an intermediate-task transfer learning technique, deviating from ensemble approaches and multi-kernel usage.

Recent efforts have explored the use of pre-trained language models for SD. While [2] conducted a comparative study, fine-tuning pre-trained BERT against classical SD approaches, [22] employed BERT as an embedding layer to encode textual features in a zero-shot deep learning setting, yielding promising results; however, both studies observed challenges in accurately classifying sarcastic examples. On the other hand, [21] experimented with ChatGPT, prompting the model directly with test cases to discern their stances.

B. Intermediate-Task Transfer Learning

Recent studies have also embraced intermediate-task transfer learning to transfer knowledge from a data-rich auxiliary task to a primary task [12]. This technique has shown significant success in various NLP tasks. For instance, [9] employed supervised pre-training with four-example intermediate training tasks to enhance performance on the primary task evaluated using the GLUE benchmark suite [31]. Furthermore, [13] introduced few-shot learning, utilizing sentiment-based annotation to improve cross-lingual SD performance. Additionally, [1] employed transfer learning by separately fine-tuning pre-trained BERT on emotion and sentiment classification before fine-tuning the model on the primary task of sarcasm detection, leveraging the correlation between sarcasm and negative sentiment polarity.

To our knowledge, prior work has not explored sarcasm detection pre-training for SD, nor has it investigated the concatenation of BERT or RoBERTa with other deep-learning techniques for SD. In this paper, we propose leveraging sarcasm detection for TSSD within a model framework comprising BERT, Conv, BiLSTM, and a dense layer.

III. Methodology

This section outlines our approach, encompassing the intermediate-task transfer learning and the underlying model architecture.

A. Intermediate-Task Transfer Learning

Our model adopts a single intermediate-task training, which consists of two phases: pre-training on an intermediate task and fine-tuning on a target task.

1) Target Task: The focal task in this study is SD, where the objective is to predict the stance expressed in a given text, such as a tweet, towards a specified target, like '*Feminist Movement*'. A tweet, denoted as T, is represented as a word sequence $(w_1, w_2, w_3, ...w_L)$, with L denoting the sequence length. Stance labels are categorized as *InFavor* (supporting the target/topic/claim), *Against* (opposing the topic), or *None* (indicating neutrality towards the target).

2) Intermediate Task: The intermediate task in this study is sarcasm detection. As prior research has not employed sarcasm as an intermediate task, we investigate the following three sarcasm-detection tasks to gain insights into the crucial linguistic attributes for a model to learn from the intermediate task, aiming to enhance SD performance.

Sarcasm V2 Corpus (SaV2C). The SaV2C dataset, introduced by [32], presents a diverse corpus of sarcasm, utilizing syntactical cues and crowd-sourced from the Internet Argument Corpus (IAC 2.0). It comprises 4,692 lines extracted from Quote and Response sentences in political debate dialogues. Our exploration focuses on the General Sarcasm category within the dataset, containing 3,260 instances each of sarcastic or non-sarcastic comments.

The Self-Annotated Reddit Corpus (SARC). The SARC dataset [33] is derived from Reddit. In contrast to the other datasets, sarcasm annotations in SARC are directly provided by the authors, ensuring reliable and trustworthy data. Due to accessibility issues with the original website, we obtained the Main Balanced first version of the dataset directly from the author of [1]. This version comprises 1,010,826 training samples, evenly distributed between sarcastic and non-sarcastic instances.

SARCTwitter (ST). The ST dataset [34] is designed to predict readers' sarcasm understandability using features, including eye movement. In our study, we utilized the dataset variant employed by [35], excluding the eye movement feature. Crowd-sourced from Twitter (X), ST is manually annotated by seven readers and contains 350 sarcastic and 644 non-sarcastic tweets. Our intermediate-task transfer learning pipeline is depicted in Figure 1.

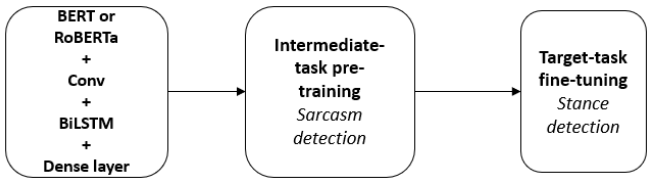

Figure 1. Intermediate-task transfer learning pipeline.

B. Underlying Model Architecture

The entire model framework primarily comprises an input layer, an embedding layer, and deep neural networks.

1) Input Layer: This layer takes a text S encoding the stance information and comprising n words. S is transformed into a vector of words and passed to the embedding layer.

2) The Embedding Layer: We employ BERT [19] and RoBERTa [20] for textual input encoding into hidden state H in our experimentation. Noteworthy achievements of these language models in the literature [1][2][9][12][30][36] motivate their exploration to identify the most suitable model for alignment with our research objectives.

3) Deep Neural Networks: This module utilizes Conv, a BiLSTM layer, and a dense layer, positioned atop the embedding layer. The purpose of incorporating convolution is to discern specific sequential word patterns within a sentence, generating a composite feature map from H. This feature map facilitates the BiLSTM layer in acquiring nuanced higher-level stance representations, which are subsequently mapped into a more differentiable space by the dense layer. Figure 2 depicts the overall model framework.

IV. Experiments

This section delineates the datasets employed, details the data pre-processing procedures, outlines baseline models, presents experimental results, and engages in a subsequent discussion.

A. Datasets

For evaluation purposes, we employed two publicly available SD datasets: 1) the well-established SemEval 2016 Task 6A Dataset (SemEval); and 2) the Multi-Perspective Consumer Health Query Data (MPCHI).

1) SemEval: The SemEval [37] task encompasses tweets manually annotated for stance towards a specified target, a target of opinion, and sentiment. Our experiments exclusively utilize tweets and their corresponding stance annotations. The dataset comprises tweet data associated with five distinct targets: Atheism (AT), Climate Change (CC), Feminist Movement (FM), Hillary Clinton (HC), and Legalization of Abortion (LA).

2) MPCHI: MPCHI [38] serves as a dataset for stance classification to enhance Consumer Health Information (CHI) query search results. Comprising formal texts extracted from top-ranked articles corresponding to queries on a specific web search engine, the dataset includes sentences related to five distinct queries, which are also the targets for stance classification: MMR vaccination can cause autism (MMR), E-cigarettes are safer than normal Cigarettes (EC), women should take HRT post menopause (HRT), Vitamin C prevents common cold (VC), and Sun exposure leads to skin Cancer (SC).

Consistent with [2], the datasets are partitioned into training and test sets following similar proportions. Each text in the datasets is annotated with one of three classes: *InFavor*, *Against*, and *None*. Table I provides statistical details describing the datasets.

B. Data pre-processing

We conducted standard data pre-processing steps, including case-folding, stemming, stop-word removal, and deletion of null entries, across all datasets. Text normalization, following the approach by [39], and hashtag pre-processing, using Wordninja [40], were also performed. However, for neural network models relying on pre-trained embeddings, stemming and stopword removal were omitted, as stemmed versions of terms might not be present in the pre-trained embeddings. The default tokenizer for the corresponding pre-trained language model was employed to tokenize words in tweets before supplying them to the classifier.

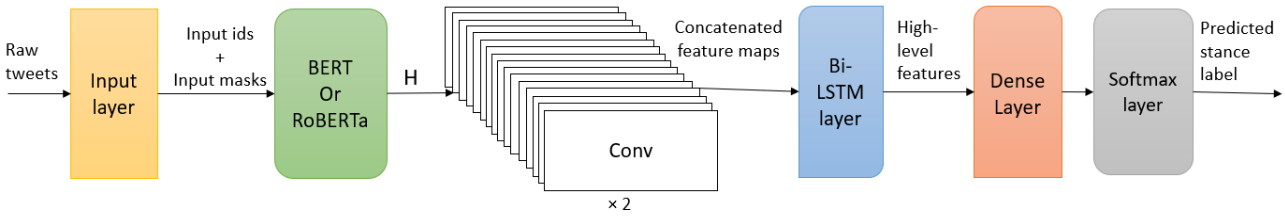

Figure 2. Proposed model framework.

TABLE I
STATISTICS OF THE DATASETS DIVIDED INTO TRAINING AND TEST SETS

Dataset	Target	Training samples			Test samples		
		INFAVOR	AGAINST	NONE	INFAVOR	AGAINST	NONE
SemEval	AT	92	304	117	32	160	28
	CC	212	15	168	123	11	35
	FM	210	328	126	58	183	44
	HC	112	361	166	45	172	78
	LA	105	334	164	46	189	45
MPCHI	MMR	48	61	72	24	33	21
	SC	68	51	117	35	26	42
	EC	60	118	111	33	47	44
	VC	74	52	68	37	16	31
	HRT	33	95	44	9	41	24

C. Baseline models

Our model is evaluated against the top-performing results from the SemEval challenge [41], as reproduced in [2] with minor modifications. Additionally, we compare our model's performance with the most recent SOTA methods in SD.

1) SemEval models: We select the Target-Specific Attention Neural Network (TAN-) proposed by [42], and the 1-D sem-CNN introduced by [43]. Additionally, we adopt Com-BiLSTM and Com-BERT, implementations provided solely by [2].

2) ChatGPT and ZSSD: The work by [21] investigated ChatGPT for SD by directly probing the generative language model for the stance of a given piece of text, focusing on the SemEval task with specific targets: FM, LA, and HC. On the other hand, the Zero-Short SD (ZSSD) technique [22], employing contrastive learning, was similarly implemented on SemEval only.

D. Experimental settings

The inductive approach to transfer learning was applied to the entire model framework, initializing the target task model with parameters learned during sarcasm-detection pre-training. Given the primary focus on enhancing model efficacy for the target task, intermediate tasks were divided into training and validation sets solely for model pre-training on sarcasm detection. In contrast, the target task featured a separate test set for final evaluations and comparisons. As Sav2C and ST are the smallest intermediate-task datasets, five-fold cross-validation was employed on both, while SARC, with its larger size, undergoes an 80/20 train/validation split.

A kernel of size 3, 16 filters, and a ReLu activation function have been employed for the convolutional layer. The BiLSTM layer has been used with a hidden state of 768, matching the hidden state size of the pre-trained language models. The dense layer has employed an output size of 3 and a softmax activation function. All experiments have been conducted on an NVIDIA Quadro RTX 4000 GPU.

Hyperparameter tuning has been performed through multiple experiments, selecting the best intermediate-task training scheme based on holdout development set results. The optimal per-task model has been then evaluated on the test set. Iterating over datasets with a mini-batch of 16 samples, the Adam optimizer [44] has been used for parameter learning, employing cross-entropy loss as the cost function. Training runs span 10 to 50 epochs, with early stopping triggered if validation accuracy on holdout data stagnates for five consecutive epochs. The training schedule involves an initial learning rate of 3e-5, decayed to a final learning rate of 1e-9 for the intermediate-task and 1e-10 for the target task. A dropout of 0.25 is introduced between model layers to address overfitting. Due to imbalanced class distributions, class weights are incorporated during training to enhance model generalization on underrepresented classes. Experimental setups adhere to the original papers for baseline models unless otherwise specified, in which case our experimental configurations are adopted.

E. Evaluation metrics

For consistency with prior works [2][4][41], the evaluation of our model employs the macro-average F1-score for the *InFavor* and *Against* classes.

F. Results

We report averaged results from five experiment runs on each target task. Table II shows experimental outcomes before sarcasm pre-training in our model. Results for ChatGPT and ZSSD are directly transcribed from their original papers, while the results for other baseline models have been replicated in our experimentation. The table illustrates the commendable performance of our BERT-based model across various targets, with notable superiority in all aspects except for HC and CC, where ChatGPT and our RoBERTa-based model excel, respectively. Consequently, we opt to proceed with our BERT-based model in subsequent experimental results.

Table III presents experimental results involving sarcasm detection pre-training with our model only. Model performance improves by **0.050** and **0.003** on SemEval and MPCHI, respectively, when pre-trained with ST, surpassing all baseline models in Table II, but diminishes with Sav2C and SARC.

Table IV presents results of an ablation study using ST only. Different base model components were systematically removed to assess the contribution of each constituent module to the entire model framework. As shown in the table, the model with all components—BERT, Conv, BiLSTM, and sarcasm pre-training—performs the best with average F1-scores of **0.775** and **0.724** on SemEval and MPCHI, respectively.

G. Failure Analysis and Discussion

Subsequent to obtaining results in Table II, a failure analysis was conducted on misclassified test samples. Predominantly, misclassifications on SemEval were associated with texts containing sarcastic content, aligning with prior findings [2]. This observation substantiated the motivation for considering sarcasm-detection pre-training before fine-tuning on SD. On the contrary, misclassifications on MPCHI were associated with samples encompassing colossal and generic health-related facts neutral to the respective target under study. Additional observations stemming from the experiments and results across all tasks are outlined below.

1) Our model outperforms SOTA models even without sarcasm detection: Specifically, it outperforms ChatGPT and Com-BERT, the best models, on SemEval and MPCHI, by **0.038** and **0.053** on average F1-scores, respectively. While Com-BERT employs only BERT and a dense layer as a classifier, our model incorporates Conv and BiLSTM before the dense layer, contributing to the observed performance improvement. Additionally, it was noted that the inclusion of the BiLSTM module in our model yielded better performance than using pooling layers after the Conv module. This suggests the effectiveness of our model architecture and its ability to capture nuanced representations, leading to proper generalization on SD tasks.

2) Sarcasm detection is correlated with SD: Consider an illustrative misclassified example: "*I like girls. They just need to know their place. #SemST*", a sarcastic comment from the FM target in SemEval. The ground truth for this example is *Against*, but it was predicted as *InFavor* before sarcasm-detection pre-training. Notably, most sarcastic samples in the *Against* class were misclassified as *InFavor* due to their explicit positive content. After incorporating sarcasm knowledge into the model through pre-training, 85% of misclassified sarcastic samples were predicted correctly. This observation underscores the relevance of sarcasm-detection pre-training in improving the performance of SD models in our experimentation.

3) Not every sarcasm detection model is a good candidate for intermediate-task transfer learning on SD: The inclusion of SARC and SaV2C knowledge in the model pipeline introduced noise and adversely affected model performance on SD compared to incorporating ST knowledge. An analysis of Sav2C and SARC revealed several discrepancies between the intermediate-task datasets and the target tasks. Firstly, the average sentence length in Sav2C and SARC is longer than in SemEval and MPCHI. Secondly, SARC is sourced from different domains than both SemEval and MPCHI, leading to disparities in topic coverage, vocabulary overlap, and the framing of ideas across datasets. Additionally, SARC, being the largest intermediate task, covers a wide range of topics through various subreddits. In contrast, ST, the best-performing intermediate task, shares a similar average sentence length with the target tasks. Moreover, both ST and SemEval are crowd-sourced from Twitter, which likely contributes to the strong performance observed when using ST as an intermediate task on the SemEval dataset. Consequently, the mismatched attributes render certain intermediate tasks less commensurated and less correlated with target tasks, resulting in a negative impact on model performance. Careful consideration and experimentation are essential when selecting a suitable sarcasm model for transfer learning in the context of SD.

4) Ablation study regarding sarcasm knowledge: The variations in the ablation study results in Table IV help to isolate the effects of each module and determine their individual contributions to the overall improvement in SD performance through sarcasm detection pre-training. Comparing our best average results in Table II and Table IV, the infusion of sarcasm knowledge significantly enhances model performance on the SemEval task compared to the MPCHI task. The SemEval task comprises extensive opinionated and sarcastic texts. Conversely, the majority of examples in the MPCHI dataset encompass extensive health-related facts, unrelated to specific targets, aside from occasional sarcasm-related expressions. Consequently, there is a modest increase in performance on MPCHI even when sarcasm detection is utilized. This observation prompts the consideration of exploring variants of BERT or RoBERTa embeddings pre-trained on health-related data specifically for SD on MPCHI as a potential avenue for future work.

V. Conclusion And Future Work

In this study, we introduced a transfer-learning framework that leverages sarcasm detection for SD. RoBERTa and BERT were individually fine-tuned and sequentially concatenated with other deep neural networks, with BERT delivering

TABLE II
EXPERIMENTAL RESULTS WITHOUT SARCASM DETECTION PRE-TRAINING

Model	SemEval						MPCHI					
	AT	**CC**	**FM**	**HC**	**LA**	**Avg**	**MMR**	**SC**	**EC**	**VC**	**HRT**	**Avg**
Sem-TAN-	0.596	0.420	0.495	0.543	0.603	0.531	0.487	0.505	0.564	0.487	0.467	0.502
Sem-CNN	0.641	0.445	0.552	0.625	0.604	0.573	0.524	0.252	0.539	0.524	0.539	0.476
Com-BiLSTM	0.567	0.423	0.508	0.533	0.546	0.515	0.527	0.522	0.471	0.474	0.469	0.493
ZSSD	0.565	0.389	0.546	0.545	0.509	0.511	-	-	-	-	-	-
Com-BERT	0.704	0.466	0.627	0.620	0.673	0.618	0.701	0.691	0.710	0.617	0.621	0.668
ChatGPT	-	-	0.690	**0.780**	0.593	0.687	-	-	-	-	-	-
Ours-RoBERTa	0.740	**0.775**	0.689	0.683	0.696	0.712	0.692	0.687	0.700	0.701	0.698	0.695
Ours-BERT	**0.767**	0.755	**0.697**	0.704	**0.702**	**0.725**	**0.747**	**0.722**	**0.704**	**0.702**	**0.732**	**0.721**

TABLE III
EXPERIMENTAL RESULTS WITH SARCASM-DETECTION PRE-TRAINING

Task	SemEval						MPCHI					
	AT	**CC**	**FM**	**HC**	**LA**	**Avg**	**MMR**	**SC**	**EC**	**VC**	**HRT**	**Avg**
SaV2C	0.595	0.718	0.596	0.645	0.578	0.626	0.605	0.545	0.545	0.352	0.495	0.508
SARC	0.697	0.612	0.683	0.557	0.641	0.638	0.605	0.545	0.545	0.352	0.495	0.508
ST	**0.769**	**0.800**	**0.774**	**0.795**	**0.741**	**0.775**	**0.749**	**0.727**	**0.704**	**0.703**	**0.739**	**0.724**

TABLE IV
EXPERIMENTAL RESULTS OF AN ABLATION STUDY

Model	SemEval						MPCHI					
	AT	**CC**	**FM**	**HC**	**LA**	**Avg**	**MMR**	**SC**	**EC**	**VC**	**HRT**	**Avg**
BERT	0.674	0.677	0.678	0.609	0.685	0.665	0.568	0.519	0.441	0.482	0.595	0.521
BERT+Conv+BiLSTM	0.767	0.755	0.697	0.704	0.702	0.725	0.747	0.722	**0.704**	0.702	0.732	0.721
ST+BERT	0.712	0.735	0.698	0.687	0.696	0.706	0.687	0.601	0.540	0.466	0.546	0.568
ST+BERT+Conv	**0.770**	0.759	0.689	0.683	0.694	0.719	0.458	0.535	0.479	0.350	0.524	0.469
ST+BERT+BiLSTM	0.747	0.765	0.675	0.657	0.678	0.704	0.640	0.618	0.573	0.528	0.633	0.598
ST+BERT+Conv+BiLSTM	0.769	**0.800**	**0.774**	**0.795**	**0.741**	**0.775**	**0.749**	**0.727**	**0.704**	**0.703**	**0.739**	**0.724**

promising results. The model underwent separate pre-training on three sarcasm-detection tasks before fine-tuning on two target SD tasks. Evaluation against SOTA models demonstrated superior performance, even prior to incorporating sarcasm knowledge. We established the correlation between sarcasm detection and SD, with the infusion of sarcasm knowledge boosting model performance, accurately predicting 85% of misclassified samples in the SemEval task. Failure analysis revealed SemEval's abundance of opinionated sarcastic samples, underscoring the efficacy of sarcasm pre-training, compared to MPCHI, characterized by generic health-related facts unrelated to specific targets. Additionally, we showed that not every sarcasm-detection intermediate task improved SD due to incongruous linguistic attributes. Finally, an ablation study highlighted that optimal model performance is achieved when utilizing all model constituents.

To the best of our knowledge, this is the inaugural exploration of sarcasm-detection pre-training applied to the BERT(RoBERTa)+Conv+BiLSTM architecture before fine-tuning for SD. Serving as a foundational reference, our approach establishes a baseline for future researchers in this domain. Future investigations will assess variant BERT or RoBERTa embeddings tailored to health-related text data for the MPCHI task. The research will also concentrate on cross-target SD for both tasks and a more comprehensive examination of other intermediate tasks, including sentiment and emotion knowledge.

VI. LIMITATIONS

Despite the significant advancements this study brings to NLP applied to social media contexts, several limitations merit consideration. Firstly, the extent of model performance enhancement is contingent upon the attributes of both the intermediary sarcasm detection task and the ultimate SD task. The divergence in linguistic characteristics across datasets utilized for sarcasm detection and SD potentially constrains the broader applicability of the study's outcomes. Secondly, although the integration of BERT or RoBERTa with other deep-learning methodologies represents an innovative approach, the intricate nature of the model architecture may present computational resource challenges and interoperability issues in certain contexts. Lastly, the extensive reliance on fine-tuning techniques and specific datasets raises concerns regarding the model's capacity to generalize effectively across diverse text types or domains not encompassed within the training data corpus.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation under Award number OIA-1946391, Data Analytics that are Robust and Trusted (DART). We express our sincere appreciation to our three anonymous reviewers for their valuable insights and constructive feedback. Additionally, we extend our gratitude to all individuals who contributed to this study in various capacities.

REFERENCES

[1] E. Savini and C. Caragea, "Intermediate-task transfer learning with bert for sarcasm detection," *Mathematics*, vol. 10, no. 5, p. 844, MDPI, 2022.

[2] S. Ghosh, P. Singhania, S. Singh, K. Rudra, and S. Ghosh, "Stance detection in web and social media: a comparative study," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*. Springer, 2019, pp. 75–87.

[3] A. ALDayel and W. Magdy, "Stance detection on social media: State of the art and trends," *Information Processing & Management*, vol. 58, no. 4, p. 102597, Elsevier, 2021.

[4] I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva, "Stance detection with bidirectional conditional encoding," arXiv preprint arXiv:1606.05464, 2016.

[5] D. Küçük and F. Can, "Stance detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 1, pp. 1–37, ACM, NY, USA, 2020.

[6] D. Küçük and F. Can, "A tutorial on stance detection," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, ACM, 2022, pp. 1626–1628.

[7] D. Biber and E. Finegan, "Adverbial stance types in english," *Discourse processes*, vol. 11, no. 1, pp. 1–34, Taylor & Francis, 1988.

[8] D. Küçük and F. Can, "Stance detection: Concepts, approaches, resources, and outstanding issues," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2021, pp. 2673–2676

[9] J. Phang, T. Févry, and S. R. Bowman, "Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks," *arXiv preprint arXiv:1811.01088*, Cornell University, 2018.

[10] Y. Li and C. Caragea, "Multi-task stance detection with sentiment and stance lexicons," 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)*, ACL, 2019, pp. 6299–6305.

[11] M. Sap, H. Rashkin, D. Chen, R. LeBras, and Y. Choi, "Socialiqa: Commonsense reasoning about social interactions," *arXiv preprint arXiv:1904.09728*, Machine Learning, ICML, 2019.

[12] Y. Pruksachatkun et al., "Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work?" *arXiv preprint rXiv:2005.00628,* ACL, 2020.

[13] M. Hardalov, A. Arora, P. Nakov, and I. Augenstein, "Few-shot cross-lingual stance detection with sentiment-based pre-training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, AAAI, 2022, pp. 10 729–10 737.

[14] A. Ghosh and T. Veale, "Fracking sarcasm using neural network," in *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, ACL, 2016, pp. 161–169.

[15] S. M. Sarsam, H. Al-Samarraie, A. I. Alzahrani, and B. Wright, "Sarcasm detection using machine learning algorithms in twitter: A systematic review," *International Journal of Market Research*, vol. 62, no. 5, pp. 578–598, Sage Journals, 2020.

[16] R. Jamil et al., "Detecting sarcasm in multi-domain datasets using convolutional neural networks and long short term memory network model," *PeerJ Computer Science*, vol. 7, p. e645, National Library of Medicine, 2021.

[17] A. Kumar, V. T. Narapareddy, V. Aditya Srikanth, A. Malapati, and L. B. M. Neti, "Sarcasm detection using multi-head attention based bidirectional lstm," *IEEE Access*, vol. 8, pp. 6388–6397, IEEE, 2020.

[18] C. Liebrecht, F. Kunneman, and A. van Den Bosch, "The perfect solution for detecting sarcasm in tweets# not,", in *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, ACL, 2013, pp. 29–37.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, ACL, 2018.

[20] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, ACL, 2019.

[21] B. Zhang, D. Ding, and L. Jing, "How would stance detection techniques evolve after the launch of chatgpt?" *arXiv preprint arXiv:2212.14548*, ArXiv. /abs/2212.14548, 2022.

[22] B. Liang et al., "Zero-shot stance detection via contrastive learning," in *Proceedings of the ACM Web Conference 2022,* ACM, 2022, pp. 2738–2747.

[23] P. Sobhani, D. Inkpen, and X. Zhu, "A dataset for multi-target stance detection," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, ACL, 2017, pp. 551–557.

[24] H. Liu, S. Li, and G. Zhou, "Two-target stance detection with target-related zone modeling," in *Information Retrieval: 24th China Conference, CCIR 2018, Guilin, China, September 27–29, 2018, Proceedings 24,* Springer, 2018, pp. 170–182.

[25] P. Sobhani, D. Inkpen, and X. Zhu, "Exploring deep neural networks for multitarget stance detection," *Computational Intelligence*, vol. 35, no. 1, pp. 82–97, Computational Intelligence, 2019.

[26] M. Walker, P. Anand, R. Abbott, and R. Grant, "Stance classification using dialogic properties of persuasion," in *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, ACL, 2012, pp. 592–596.

[27] D. Küçük and F. Can, "Stance detection on tweets: An svm-based approach," *arXiv preprint arXiv:1803.08910*, ArXiv. /abs/1803.08910, cs. cL, 2018.

[28] I. Segura-Bedmar, "Labda's early steps toward multimodal stance detection." in *IberEval@ SEPLN*, ACL, 2018, pp. 180–186.

[29] U. A. Siddiqua, A. N. Chy, and M. Aono, "Tweet stance detection using an attention based neural ensemble model," in Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (*long and short papers)*, ACL, 2019, pp. 1868–1873.

[30] L. H. X. Ng and K. M. Carley, "Is my stance the same as your stance? a cross validation study of stance detection datasets," *Information Processing & Management*, vol. 59, no. 6, p. 103070, ACM, 2022.

[31] A. Wang et al., "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, ACL, 2018.

[32] S. Oraby et al., "Creating and characterizing a diverse corpus of sarcasm in dialogue," *arXiv preprint arXiv:1709.05404*, ArXiv. /abs/1709.05404 [cs.CL], 2017.

[33] M. Khodak, N. Saunshi, and K. Vodrahalli, "A large self-annotated corpus for sarcasm. *arxiv," arXiv preprint arXiv:1704.05579,* arXiv:2312.04642v1 [cs.CL], 2018.

[34] A. Mishra, D. Kanojia, and P. Bhattacharyya, "Predicting readers' sarcasm understandability by modeling gaze behavior," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, AAAI, 2016, pp. 3747–3753.

[35] N. Majumder et al., "Sentiment and sarcasm classification with multitask learning," *IEEE Intelligent Systems*, vol. 34, no. 3, pp. 38–43, IEEE, 2019.

[36] G. Nkhata, "Movie reviews sentiment analysis using bert," Masters thesis, University of Arkansas, Fayetteville, AR, USA, December 2022.

[37] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*, vol. 17, no. 3, pp. 1–23, ACM, 2017.

[38] A. Sen, M. Sinha, S. Mannarswamy, and S. Roy, "Stance classification of multi-perspective consumer health information," in *Proceedings of the ACM India joint international conference on data science and management of data*, ACM, 2018, pp. 273–281.

[39] B. Han and T. Baldwin, "Lexical normalisation of short text messages: Makn sens a# twitter," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, ACL, 2011, pp. 368–378.

[40] Keredson, "Wordninja," https://github.com/keredson/wordninja, 2017, [Online; accessed 19-April-2024].

[41] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "Semeval-2016 task 6: Detecting stance in tweets," in *Proceedings of*

the 0th international workshop on semantic evaluation (SemEval-2016), ACL, 2016, pp. 31–41.

[42] J. Du, R. Xu, Y. He, and L. Gui, "Stance classification with target-specific neural attention networks." *International Joint Conferences on Artificial Intelligence*, ICAI, 2017, pp. 3988–3994.

[43] Y. Kim, "Convolutional neural networks for sentence classification," *in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ACL, 2014, pp. 1746–1751.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, ArXiv. /abs/1412.6980[cs.LG], 2014.

 ISBN: 978-1-68558-165-7

Sequence Graph Network for Online Debate Analysis

Quan Mai, [1] Susan Gauch, [1] Douglas Adams, [2] Miaoqing Huang [1]
[1]*Department of Electrical Engineering and Computer Science,* [2]*Department of Sociology and Criminology*
University of Arkansas
Fayetteville, Arkansas, USA
{quanmai, sgauch, djadams, mqhuang}@uark.edu

***Abstract*—Online debates involve a dynamic exchange of ideas over time, where participants need to actively consider their opponents' arguments, respond with counterarguments, reinforce their own points, and introduce more compelling arguments as the discussion unfolds. Modeling such a complex process is not a simple task, as it necessitates the incorporation of both sequential characteristics and the capability to capture interactions effectively. To address this challenge, we employ a sequence-graph approach. Building the conversation as a graph allows us to effectively model interactions between participants through directed edges. Simultaneously, the propagation of information along these edges in a sequential manner enables us to capture a more comprehensive representation of context. We also introduce a Sequence Graph Attention layer to illustrate the proposed information update scheme. The experimental results show that sequence graph networks achieve superior results to existing methods in online debates.**

Keywords-Graph neural networks; dialog modeling; sequence graph network; online debates.

I. INTRODUCTION

Online debate has become an integral part of our digital age, transforming the way we engage in discourse and exchange ideas. In social media platforms (e.g., Facebook, Twitter (currently X), etc.), individuals from diverse backgrounds and geographical locations converge to discuss and deliberate on a wide array of topics, ranging from politics and ethics to music and science. Debating with a wide range of debaters requires participants to research and present well-informed arguments, encourages critical thinking, and challenges preconceived notions.

Like other forms of debate, online discussions are contingent on the flow of time (temporal dependency); each subsequent comment relies on the content of the previous comment it responds to. Participants interactively promote their point while countering the opponent's [4]. Within a turn, debaters employ a variety of strategies, each of which plays a crucial role in determining the outcome of the debate. These strategies involve either directly addressing the opponent's argument, presenting their own viewpoint, or skillfully combining both tactics. The latter approach often appears to be the most effective, allowing the debater to simultaneously achieve both objectives during their turn. However, one cannot always adopt that strategy as it depends on their position in the debate. For instance, if a debater is the first speaker in a debate, their primary task is to present their own ideas coherently and logically, as they do not have the opportunity to directly counter their opponent's arguments at this stage. In such a scenario, the debater's effectiveness lies in the clarity and persuasiveness of their presentation, making it challenging for the opposing side to refute their position. These strategies are also discussed in [4], which examined the dynamics of information flow within online debates.

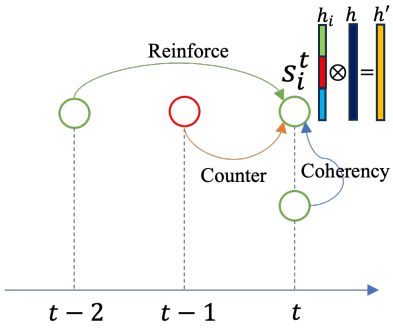

Figure 1. A "what-should-we-mention" information flow scheme that mimics the interaction process of a debater. At each time step t, the node features are updated by considering their peer nodes from the same turn and the connected nodes from previous turns, using Directed Graph Attention Network layers. Nodes associated with different debaters are colored differently. Each type of edge (colored arrows) contributes a corresponding representation, collectively forming $\mathbf{h}_i$. The node's utterance embedding $\mathbf{h}$ and the interaction representation $\mathbf{h}_i$ are used to update the node feature $\mathbf{h}'$.

As the argument process is temporally dependent, Recurrent Neural Networks (RNNs), such as Long Short Term Memory (LSTM) [9] and Gated Recurrent Unit (GRU) [13], have been one of the most widely used techniques in argument-winning research as well as dialog extraction. Several studies employ RNNs as the encoder for utterances [5] [7] [10], leveraging their capacity to capture sequential dependencies and relationships within textual data. In addition to encoding individual utterances, sequence networks are employed to encode entire conversations by sequentially processing the arguments [11].

In a debate, however, participants engage in interactive turn-by-turn rebuttals to counter their opponents' arguments, and sequencing the entire conversation fails to capture this dynamic interaction. In order to model the process of dialogical argumentation, [10] use a co-attention network to capture the interaction between the participants and achieve a promising performance on the prediction task. The focus of [7] is placed on identifying connections between the sentences of debaters. This approach is instrumental in capturing critical argumentative components, making it a pivotal factor for predicting the winner. The aforementioned studies compute

"attention scores" for each pair of sentences belonging to two participants in order to assess the *relevance* of one sentence to another.

An alternative method for capturing these interaction dynamics is through the use of graphs. Graphs are an effective way to represent relationships and dependencies among entities, making them suitable for a wide range of applications, including social networks and recommendation systems [17]–[19]. The connection between two components of an argument can be effectively represented by a link (or edge) within the graph. Graphs can also serve as input to Graph Neural Networks (GNNs) for capturing the contextual information within the conversation. In their work, [12] employ a heterogeneous graph to represent the relationships among entities discussed in multi-party dialogues. In order to model the relationships between argument pairs, [5] incorporate intra-passage and cross-passage links to interconnect sentence nodes. Subsequently, they employ a Graph Convolutional Network (GCN) [15] for efficient information propagation.

Traditional GNNs, including GCNs and Graph Attention Networks (GAT) [3]), may not effectively capture the temporal dynamics within a conversation, particularly in a debate scenario in which participants engage in interactive exchanges to counter arguments or defend their own viewpoints. To tackle this challenge, we integrate the strengths of both RNNs and GNNs within a unified framework. In this framework, we conceptualize the debate as a graph, where argument components are depicted as nodes, and their features undergo sequential updates, according to the turn to which they correspond. We introduce the Sequence Graph Attention (SGA) cell, which resembles the traditional RNN-cell, to capture long-range dependencies in the debate (which is treated as a sequence of subgraphs). The experimental results demonstrate that our approach can capture the interaction between debaters and outperforms state-of-the-art models in accurately predicting the winner in several online debate datasets. The code and models are available at [39].

The structure of the remainder of this paper is organized as follows: Section II describes the process of constructing a graph from a debate. In Section III, we introduce our proposed framework. The effectiveness of this method is evaluated in Section IV. Section V reviews some relevant literature. Finally, Section VI provides a summary of our findings and discusses potential avenues for future work.

II. Preliminary

Before describing the details of the proposed method, we first give a brief introduction to how we construct a graph for an online debate.

A. Debate Format

Our primary focus lies in online debates wherein the victor emerges through the collective votes of an audience or a panel of judges. These debates adhere to the Oxford-style format, featuring two participants representing opposing viewpoints —one in favor of the claim (Pros) and the other in opposition (Cons) — who alternate in presenting their arguments on a given topic. After the debate, a winner is declared, unless a tie occurs. In this study, we define a *turn* as each instance when a debater presents their argument, and a *round* represents the stage in which opposing sides provide their arguments. Consequently, round 0 consists of turn 0 and turn 1, round 1 consists of turn 2 and turn 3, and so forth.

B. Debate-to-Graph construction

Given a debate that contains a total of N sentences, a directed, unweighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{H})$ is constructed based on sentences and their relationships (Figure 2). Sentences in the debate are represented by a set of nodes $\mathcal{V}$ ($|\mathcal{V}| = N$), and a node attribute matrix $\mathcal{H} \in R^{N \times D}$, defined by D-dimensional embedding vectors for each of the sentences. Sentences in the debate may be interconnected and these interconnections are represented by $\mathcal{E}$, the set of edges in the graph.

Edge types: We define three different types of edges to elucidate the participants' strategies throughout the debate. Each type is categorized based on the turn it corresponds to and the strategic role it plays. In Section III, we will delve into how each type contributes to node feature aggregation.

1) Logical and Coherent Edges: These edges emphasize the participants' ability to construct logical and coherent arguments within their turn.
2) Reinforcement Edges: These edges serve to strengthen the points previously made by the debater in their previous rounds. We will interchangeably use the terms *reinforcement edges* and *supporting edges*.
3) Counterargument Edges: These edges highlight the participants' skill in countering their opponents' arguments effectively.

Intra-argument Links These edges connect sentences of the same turn. During a turn, edges are constructed based on the relative position among sentences. These *Logical and Coherent* edges capture coherency in an argument turn. Given two sentences, denoted as s_i^t and s_j^t, both belonging to turn t, we establish an edge e_{ij}^{inter} from s_j^t to s_i^t if the positional difference $\mathcal{D}$ between them is within a specified distance threshold d.

$$e_{ij}^{inter} = \begin{cases} 1 & \text{if } \mathcal{D}(s_i^t, s_j^t) \leq d \\ 0 & \text{otherwise} \end{cases}$$

Cross-argument Links These edges interconnect sentences that belong to different turns and are categorized into two types: *Reinforcement* and *Counterargument* edges. The former connects nodes belonging to the same debater whereas the latter connects nodes belonging to different debaters. For example, nodes in the 3rd turn are connected to nodes from the 1st turn through *Reinforcement* edges and are also linked with their opponent's nodes from the 2nd turn. Unlike intra-argument edges that rely on the relative positions of sentences, cross-argument edges are established using semantic textual similarity between sentences. In this work, we use cosine similarity S_c to capture the semantic relationship of texts.

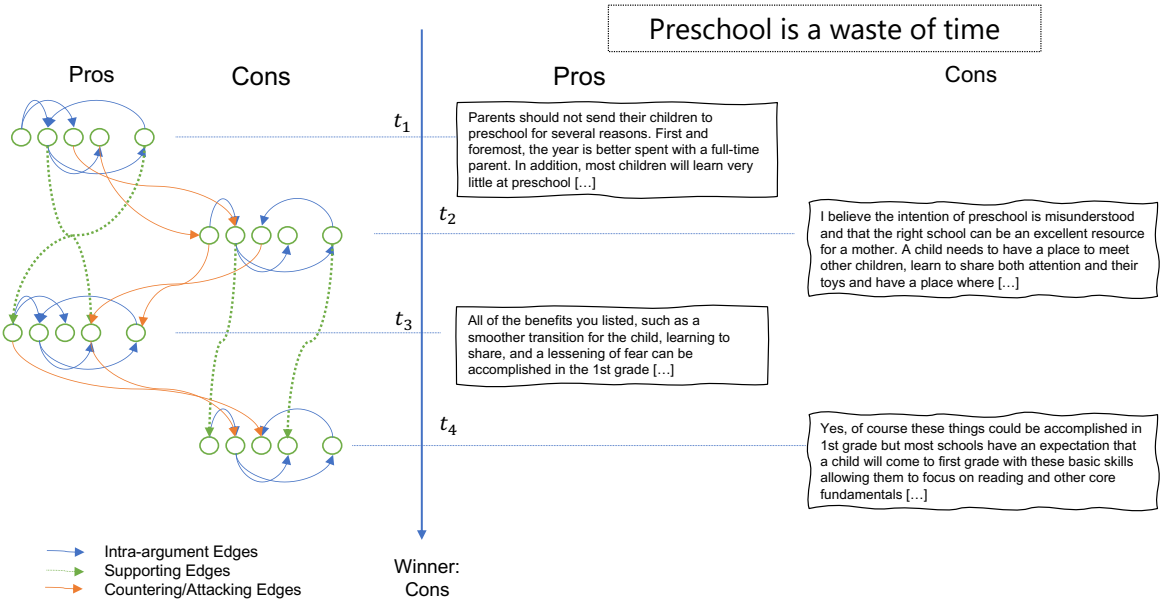

Figure 2. Graph Construction from Debate: Nodes establish connections through three distinct edge types, indicated by colored arrows. Intra-argument edges (blue) link nodes within the same turn, reinforcement edges (green) connect nodes from the same debater across different turns, while countering edges (orange) connect nodes from a debater to their opponent's, illustrating counter-argumentation. The sample debate is taken from data collected by [1].

An edge e_{ij} links 2 nodes v_i and v_j if their similarity score $S_c(\mathbf{h}_i, \mathbf{h}_j)$ meets a threshold value S_{th}

$$e_{v_i,v_j} = \begin{cases} 1 & \text{if } S_c(\mathbf{h}_i, \mathbf{h}_j) \geq S_{th} \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{h}_i$ and $\mathbf{h}_j$ are i^{th} and j^{th} rows in $\mathcal{H}$, representing embedding vectors of sentences v_i and v_j, respectively. S_{th} serves as a crucial hyper-parameter for evaluating the influence of participant interactions on the debate's outcome. An alternative approach is to employ the top k similarities, allowing each node to establish connections with up to k cross-argument nodes that possess the highest similarity scores. We will evaluate the effectiveness of each approach on the predictive performance in Section IV. It is important to note that cross-argument edges consistently flow from nodes in previous turns to nodes in subsequent turns; there is no reverse direction.

III. Proposed Method

A. Utterance Encoder

We encode each sentence using pre-trained sentence embedding (Sentence Transformer (SBERT)) [2]. In preliminary work, we found that this approach works better than using GloVe [6] word embeddings and a bidirectional LSTM to encode semantic vectors for sentences. This step gives us the sentence embedding matrix $\mathcal{H}$, in which each row $\mathbf{h}_i$ is an embedding vector for sentence s_i.

Turn Embeddings: Participants employ distinct strategies during different debate turns. For instance, in the initial round consisting of two turns, the first participant presents their perspective on the topic while the second participant challenges their opponent's arguments and introduces their own viewpoint. We incorporate the temporal turn information into the node features by concatenating it with the sentence embedding $\mathbf{h}_i$. We opt for a 30-dimensional embedding vector $\mathbf{h}_{it} \in R^{30}$ to represent the turn information for each node.

$$\mathbf{h}_i = \mathbf{h}_i \| \mathbf{h}_{it} \tag{1}$$

Let B denote the number of dimensions of the embedding vector of a sentence from SBERT, then $D = B + 30$.

B. Information flow

Graph Attention Layer: We employ a Graph Attention Network (GAT) [3] layer to update the node representation. The attention mechanism allows GAT to focus on and weigh the importance of different neighbors when aggregating information for each node, called the "attention score". We are motivated to use GAT in our model because, intuitively, not all sentences in the debate carry equal importance. One can detect the opponent's argumentative "vulnerable region" [7] and effectively counter it to win the debate. This layer takes as input a set of A ($A \leq N$) node features $\mathbf{h} \in \mathbb{R}^{A \times D}$ and produces a new set of node features $\mathbf{h}' \in \mathbb{R}^{A \times D'}$ ($D' < D$). The attention score of sentence j to sentence i is computed as:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i || \mathbf{W}\mathbf{h}_j]))}{\sum_{k \in \mathcal{N}} \exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i || \mathbf{W}\mathbf{h}_k])}$$

where $\mathbf{W} \in \mathbb{R}^{D \times D'}$ and $\mathbf{a} \in \mathbb{R}^{2D'}$ are trainable weight matrix and vector of the layer. The output features of node i is the weighted sum of the features of its neighboring node set $\mathcal{N}_i$:

$$\mathbf{h}'_i = \sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}\mathbf{h}_\mathbf{j}$$

In this work, we employ three distinct GAT layers, each responsible for aggregating information from a specific type of edge. We refer to these layers as GATI (intra-argument edge), GATC (counterargument edge), and GATS (supporting edge). At each turn, the GAT layer processes a specific set of input node features and produces a new set of features, called *interaction* representation of each sentence:

$$\mathbf{h}_I^t = \text{GATI}(\mathbf{h}_{\mathcal{I}_t}; \mathbf{a}^I, \mathbf{W}^I) \quad (2)$$

$$\mathbf{h}_C^t = \text{GATC}(\mathbf{h}_{\mathcal{J}_t}; \mathbf{a}^C, \mathbf{W}^C) \quad (3)$$

$$\mathbf{h}_S^t = \text{GATS}(\mathbf{h}_{\mathcal{K}_t}; \mathbf{a}^S, \mathbf{W}^S) \quad (4)$$

where $\mathbf{a}^*$ and $\mathbf{W}^*$ are vectors and matrices associated with each layer. Here, we have three sets of node features: $\mathbf{h}_{\mathcal{I}_t}$, $\mathbf{h}_{\mathcal{J}_t}$, and $\mathbf{h}_{\mathcal{K}_t}$, each corresponding to distinct node sets:

- $\mathcal{I}_t$ represents the set of nodes that pertain to the same time step, encompassing nodes within the current turn. $\mathbf{h}_{\mathcal{I}_t} = \{\mathbf{h}_1^t, \mathbf{h}_2^t, \mathbf{h}_3^t, ...\}$ denotes features matrix of a set of nodes at time t.
- $\mathcal{K}_t$ comprises nodes from time steps $t-2$ and t, all originating from the same debater and exhibiting a supportive relationship. This set characterizes argumentative enhancement or promotion. Note that the set of node features at time $t-2$ are **updated** in turn $t-2$. Therefore, $\mathbf{h}_{\mathcal{J}_t} = \{\mathbf{h}_1'^{t-2}, \mathbf{h}_2'^{t-2}, ..., \mathbf{h}_1^t, \mathbf{h}_2^t, ...\}$ denotes the updated features matrix of a set of nodes at times $t-1$ and utterance matrix of nodes at t.
- In contrast, $\mathcal{J}_t$ encompasses nodes from time steps $t-1$ and t and signifies an adversarial relation, capturing how a debater challenges an opponent's position by considering nodes from the opponent's previous turn $(t-1)$. Because nodes feature at time $t-1$ are updated, $\mathbf{h}_{\mathcal{K}_t} = \{\mathbf{h}_1'^{t-2}, \mathbf{h}_1'^{t-2}, ..., \mathbf{h}_1^t, \mathbf{h}_2^t, ...\}$.

a) Sequential Update: The node features are updated sequentially using a temporal attention mechanism. Information propagation occurs along *directed* edges, and the features of nodes at time t are updated based on their neighboring nodes from the same turn (via intra-argument edges) as well as nodes from previous turns (via cross-argument edges) (Figure 1). This information flow scheme illustrates the cognitive process of a debater during their turn, as they must consider the opponent's previous arguments, formulate counterarguments, reinforce their own points, and even introduce new ideas. The node features updated at time t serve as the input when updating node features at times $t+\tau$ ($\tau \in \{1, 2\}$). This process shares similarities with traditional RNNs like LSTM and GRU. However, it is important to note that our work focuses on handling a specific subset of nodes at each timestep. This distinction sets us apart from Gated Graph Sequence Neural Networks [8] that process the entire graph as input at each timestep. Similar to an RNN-Cell, that operates on a single input element at each time step and generates output that serves as a hidden feature for subsequent times, we introduce the SGA layer to manage the processing of a specific subset of nodes at time t. The entire debate graph is processed sequentially subgraph-by-subgraph.

Given a debate $\mathcal{S}$ that has T turns: $\mathbb{S} = \{S_t; t \in [0, T-1]\}$, $S_t = \{s_j^t; j \in [0, M_t - 1]\}$ denotes a debate turn consisting of M_t sentences s_j^t. It is noticeable that $N = \sum_{t=0}^{T-1} M_t$. Let $\mathbf{h}_j^t$ the utterance embedding of the sentence s_j (from 1), the new node feature h_j' is calculated using the SGA layer which executes the following operations (we discard the superscript t for readability):

$$h_j' = SGA(\mathbf{h}_j, \mathbf{h}_{\mathcal{I}}, \mathbf{h}_{\mathcal{J}}, \mathbf{h}_{\mathcal{K}}) = \mathbf{h}_j \otimes \mathbf{h}_j^{\text{X}} \quad (5)$$

where $\otimes$ is the update operator using GRU operations [13]. The $\mathbf{h}_j^{\text{X}}$ denotes the interaction representation feature at time t, encompassing intra-argument coherency, counterarguments against the opponent's points, and reinfordcement of the debater's previous statements. It is calculated by concatenating the node features produced by three component GAT layers (equations 2, 3, 4):

$$\mathbf{h}_j^{\text{X}} = \mathbf{h}_j^{\text{GATI}} \,||\, \mathbf{h}_j^{\text{GATC}} \,||\, \mathbf{h}_j^{\text{GATS}} \quad (6)$$

It is important to observe that during the initial turn, denoted as $t = 0$, there are no counterarguments in the debater's thoughts. As a result, we initialize $\mathbf{h}_j^{\text{GATC }0}$ to be equal to 0. Additionally, a debater does not introduce a reinforcing argument until their second round (or when $t \geq 2$). Consequently, both $\mathbf{h}_j^{\text{GATS }0}$ and $\mathbf{h}_j^{\text{GATS }1}$ are set to 0 during this period. The updated node features h_j' are then employed to update the attributes of nodes in subsequent turns.

C. Readout Layer

Once all the node features have been updated, we employ a readout layer to "summarize" the ideas presented by each participant during the debate. For each debater, we select a set of top r (e.g., $r = 3$) *representatives*, which are used as input for the prediction classifier. The process of selecting these representative nodes is determined by the highest "attention scores" generated by each GATI, GATC, and GATS layers, denoted as α_I, α_C, and α_S, respectively. During the feature update step, each node receives an attention score from its neighboring nodes. These scores emphasize the significance of a node in relation to others. The more significant a node is, the greater its contribution to a debater's overall idea. The total attention received by each node is obtained by summing up its individual attention scores. Consider a node s_l, its attention scores are:

$$\alpha_{s_l}^I = \sum_{i \in \mathcal{I}} \alpha_i, \;\; \alpha_{s_l}^C = \sum_{j \in \mathcal{J}} \alpha_j, \;\; \alpha_{s_l}^S = \sum_{k \in \mathcal{K}} \alpha_k \quad (7)$$

We opt to select the top r nodes with the highest scores for each type of attention. We then concatenate the feature vectors corresponding to these selected nodes to create a $3 \times r \times D'$-dimensional vector, where D' is the dimension of the node feature produced by SGA. The readout layer subsequently generates two "summary" vectors, each serving as a deep representation of each debater's performance during the debate.

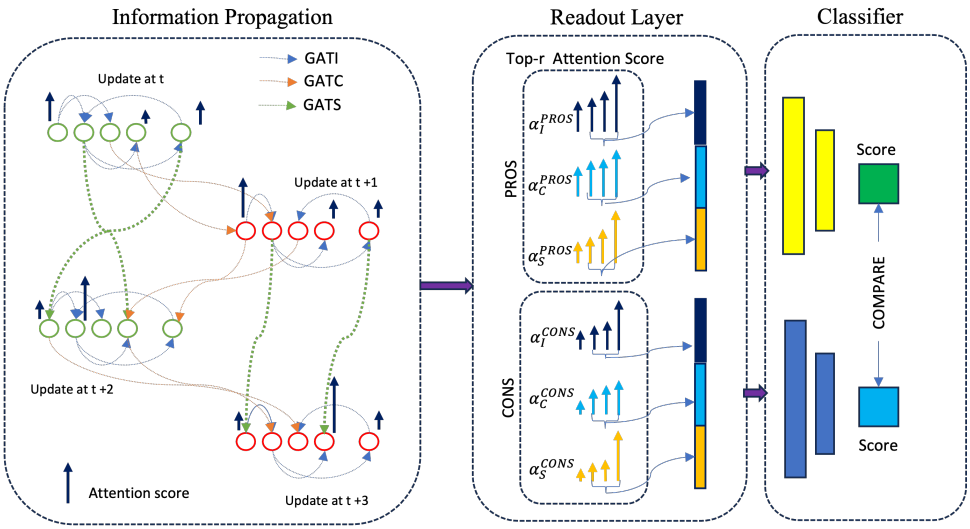

Figure 3. The proposed architecture consists of three key modules: (1) Information propagation is driven by the SGA layers, updating node features sequentially using a graph attention mechanism. (2) The readout layer identifies representative vectors associated with each debater, which are subsequently supplied as input to (3) an MLP classifier for predicting the debate winner.

D. Classification

The two vectors, $\mathbf{Q}^{PROS}$ and $\mathbf{Q}^{CONS}$ achieved by the readout layer are fed to the classifier to perform the prediction task. Each vector is mapped to a score value $c \in \mathbb{R}^1$ by linear transformation using a Fully Connected (FC) layer followed by an activation function (e.g., ReLU), Layer Norm (LN) [14] and dropout layer [24]. Let us denote a series of FC + ReLU + LN + Dropout an MLP, then

$$c^{PROS} = MLP1(\mathbf{Q}^{PROS})$$
$$c^{CONS} = MLP2(\mathbf{Q}^{CONS})$$

If the Pros side wins, we expect that $c^{PROS} > c^{CONS}$, and conversely when the Cons side wins. Here, we denote C^+ and C^- as the scores of the winner and loser, respectively. Our objective is to maximize the difference between C^+ and C^- as much as possible. To achieve this, we employ Pairwise Cross-Entropy (PCE) loss, that minimizes:

$$\mathcal{L} = \text{PCE}(C^+, C^-) = \log(1 + exp(C^- - C^+)) \quad (8)$$

The network architecture is illustrated in Figure 3.

IV. Evaluation

A. Dataset

Our study is conducted on the *debate.org* dataset collected by [1]. The dataset contains 78,376 debates on controversial topics, including *abortion*, *death penalty*, *gay marriage*, and *affirmative action*. Each debate consists of multiple rounds in which two participants from two opposing sides take turns expressing their opinions. Further details can be found in [1].

a) Winning criterion: The winner is determined by the criterion of "Made more convincing arguments". We exclude debates with fewer than 5 voters and tie debates. Additionally, debates in which the winner has just one more vote than the loser are also classified as ties.

TABLE I
THE NUMBER OF SENTENCES, NUMBER OF COUNTERARGUMENT EDGES, AND NUMBER OF SUPPORTING EDGES MADE BY WINNER AND LOSER IN AN ARGUMENT TURN. CROSS-ARGUMENT EDGES ARE CONSTRUCTED USING A SIMILARITY THRESHOLD OF 0.85.

	#Sentences	#Countering	#Supporting
Winner	38.6	6.96	5.93
Loser	36.1	6.78	6.64

b) Preprocessing: To study the interaction among debates, we only keep debates that have at least 3 rounds (equivalent to 6 turns). Short arguments are also eliminated, i.e., we remove debates that have fewer than 5 sentences in each round (each graph thereby has at least 30 vertices). The first 3 rounds of longer debates are used for analysis. The dataset exhibits an imbalance, with the Cons side accounting for 65% of the winners whereas the Pros side wins only 35%. To create a balanced dataset, we also use the final 3 rounds of the debates where the Pros side wins and the debate comprises more than three rounds. This data augmentation step also increases the size of the dataset.

c) Statistics: After the experimental dataset selection step, there are a total of 2,445 debates available for model training and testing. Among these debates, the Pros side wins in 1,130 debates, while the Cons side secures victory in 1,325 debates. Additional statistical information is shown in table I. Observing the table, it becomes evident that the winning side tends to produce more sentences and more counterarguments compared to the losing side. Conversely, the losing side appears to prioritize reinforcing their own ideas rather than generating a higher number of counterarguments.

B. Experimental setup

a) Data Preprocessing: We randomly split the dataset with 60% for training, 20% for validation and 20% for

testing. For text normalization, we employ the following steps: (1) replacing URLs with "website", (2) replacing all the numbers with "number", and (3) lowercasing text. Next, we employed spaCy [23] for sentence tokenization. Sentences are then encoded by SBERT's "all-MiniLM-L6-v2" model that transforms a sentence into a 384-dimensional vector.

b) Parameter setting: We use a similarity threshold of 0.85 for cross-argument edge construction, other approaches regarding edge construction will be further discussed in the ablation study. The intra-argument distance threshold is $d = 3$. Each node within a turn links to nodes that share a relative positive correlation within a 3-node proximity. Node features updated by each GAT layer have $D' = 32$ dimension. For the readout layer, we choose $r = 3$. We use a stack of three MLPs to transform the readout layer's output into a score for each debater. The first layer reduces the vector from $3 \times r \times D$ to half its size. The second layer further reduces the output of the first layer by half, and the final layer maps the second output vector to a real value. We apply the tanh function to ensure the value falls within the range [-1; 1]. For hyper-parameters, we apply the dropout rate of 0.2 for all GAT layers and the classifier. Optimization is performed using Adam [16]. The batch size is 32. We run the model for 50 epochs with early stopping. The learning rate is 0.0001.

c) Other settings: Deep learning frameworks are Pytorch [21] and Pytorch Lightning [22]. We use DGL package [20] as the graph deep learning framework. The networks are trained and tested on an NVIDIA Quadro RTX 8000 GPU with 50GB of memory.

C. Comparison baselines

Given that the Cons side accounts for 52.5% of wins in the test set, it serves as the **majority baseline**, representing the best prediction one can make regardless of the input features. We compare our model's performance to SOTAs in debate winning prediction which adopt sequence approach in their work.

a) Sequence approach: In the study by [11], they aggregate the entire discussion into a single sequence and model it using LSTM with an attention mechanism applied to the sentences, referred to as the **all-LSTM** approach. They also incorporate implicit discourse relations using the Penn Discourse Tree Bank [25] discourse structure. While their research primarily centers on the Reddit dataset [35], we apply the same methodology to our debate dataset. Additionally, we find relevance in the work of [26], denoted as **ASODP**, which shares our focus on Oxford-style debates and employs a sequential approach for debate analysis. Furthermore, [27] introduces the **DTDMN** method, designed to process pairs of conversations and predict their persuasiveness. Similarly, we present the Pros and Cons sides as inputs to facilitate comparative analysis.

b) Graph approach: To highlight the significance of processing the debate on a turn-by-turn basis, we introduce two baseline models for graph analysis. The first baseline employs a 2-layer **GAT** network, while the second baseline utilizes a **GGNN**. These GNNs serve as information aggregators and feature extractors for the debate graph, simultaneously processing all nodes in the graph (and repeating this process 6 times, corresponding to 6 turns in the case of GGNN). In the case of GAT, the initial layer transforms the input into 64-dimensional vectors, and the subsequent layer maps the output from the first layer to 32-dimensional features. In GGNN, we also utilize a 32-dimensional output feature size to align with the output feature size of our SGA layer. To summarize the node features for each debater, we introduce a mean readout operation.

c) Temporal graph approach: Since no other sequential graph approach exists for debate winning prediction, we adopt the information flow method proposed in [33] (**Graphflow**), initially designed for machine comprehension. We utilize the output of the RGNN layer from the final turn, feeding it into the MLP layer for the prediction task.

TABLE II
DEBATE-WINNING PREDICTION RESULTS. THE BEST RESULTS ARE IN BOLD. (**: USING THE TOP 3 HIGHEST SIMILARITY SCORES TO CONSTRUCT CROSS-ARGUMENT EDGES, *: USING A THRESHOLD VALUE OF 0.85 TO CONSTRUCT CROSS-ARGUMENT EDGES).

Models	Acc.	F1
Majority Baseline	0.525	
Sequence Baseline		
all-LSTM	0.635	0.563
ASODP	0.656	0.623
DTDMN	0.660	0.625
Graph Baseline		
GAT	0.541	0.472
GGNN	0.565	0.522
Sequence Graph Baseline		
Graphflow	0.645	0.620
SGA		
w/o GATI	0.621	0.523
w/o GATC	0.562	0.495
w/o GATS	0.629	0.534
FULL MODEL		
*S = 0.85	0.654	**0.667**
**k = 3	**0.675**	0.625

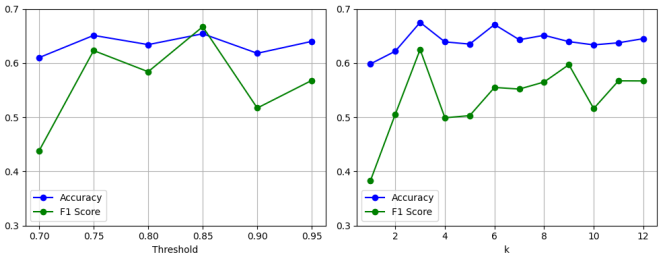

Figure 4. Impact of cross-argument construction values on network performance. Left: Edge construction using a threshold value. Right: Edge construction using top-k highest values.

D. Experimental results

The evaluation results are presented in Table II. The sequence baselines (all-LSTM, ASODP, DTDMN) all perform similarly, with DTDMN producing the best accuracy of this group at 66.0%. The Graph baselines perform more poorly,

with the highest accuracy, 56.5% produced by SSGN. Graphflow, boasting an accuracy of 64.5%, outperforms traditional graph approaches. However, it still trails behind the robust benchmarks set by sequential approaches such as ASODP and DTDMN. Our full model, SGA with k=3, outperforms all baselines with an accuracy of 67.5%, a 1.5% absolute (2.3% relative) improvement over DTDMN. The F1-score, achieved by constructing cross-argument edges with a threshold of 0.85, significantly outperforms the baselines. It reaches 66.7%, representing a 4.2% absolute (or 6.7% relative) improvement over DTDMN. We thus demonstrate that we outperform state of the art models for this dataset.

The performances of all-LSTM and DTDMN are diminished when applied to the *debate.org* dataset. This can be attributed to a fundamental distinction between the two domains. In the context of *debate.org*, the ultimate determination of the winner is not based on subjective criteria but rather relies on the judgments of a panel of judges or the voters. The voters place substantial emphasis on the debaters' ability to rigorously address and counter their opponents' reasoning. Furthermore, they favor debaters who engage in high-quality and dynamic interactions throughout the debates.

a) Sequence matters: The results show a significant superiority of sequence-based baselines over graph-based ones when applied to the debate dataset. This highlights the critical significance of adopting a sequential approach, where the debate is processed turn-by-turn, rather than relying solely on graph-based methodologies.

b) Counter-argument is crucial: We extended our analysis by performing an ablation study to assess the individual impact of each GAT layer on our proposed SGA model. We observe that when we omit the counter-argument edges, the reduction in network performance was more significant compared to scenarios where we exclude either GATI or GATS layers. Specifically, accuracy drops by 11.3%, in contrast to 5.4% and 4.6%, respectively. This outcome can be elucidated by considering that if a debater disregards the opponent's remarks from the preceding turn, their persuasive ability may diminish in the eyes of the voters or judges. In essence, acknowledging and responding to counter-arguments plays a pivotal role in constructing compelling arguments in a debate context.

E. Impact of graph parameters

We conduct a detailed analysis of the impact of graph construction parameters, such as S_{th} and k, on the network's performance (Figure 4). In the context of employing a similarity threshold, it is noteworthy that a threshold value of 0.85 yields the highest performance in terms of accuracy and F1-score.

Regarding the top-k approach, it is worth highlighting that while $k = 3$ achieves the highest accuracy, as well as highest F1-score. These insights into parameter effects contribute to a deeper understanding of how to optimize network performance for specific objectives and trade-offs.

V. Related work

Graph Neural Networks (GNNs) have proven to be powerful tools for harnessing insights into, and making predictions on, data structured as graphs, particularly in the realm of Natural Language Processing (NLP). Within NLP, GNNs have been applied to a wide spectrum of tasks including, but not limited to, dependency parsing [29], sentiment analysis [30] [31], and semantic understanding [15]. In recent developments, researchers in NLP have extended GNNs by integrating them with RNNs to enable sequential processing of graph-structured data. Notably, [32] introduced a graph-to-sequence methodology for the AMR-to-text generation task, wherein they construct an Abstract Meaning Representation (AMR) graph and progressively update the entire graph during sequential generation. Furthermore, [33] made significant strides in the domain of machine comprehension by incorporating conversation history into their model. They adopt a graph-based approach, constructing a graph that evolves with each conversational turn. While our work shares a commonality in the sequential update of subgraphs, it is important to emphasize that the implementation details diverge significantly. Researchers have explored temporal graph approaches for tasks like traffic flow forecasting [36] [37] and skeleton-based action recognition [38]. However, the utilization of sequence graph approaches in conversation analysis, particularly within online debate and argumentative analysis contexts, remains relatively unexplored.

VI. Conclusion and future work

In conclusion, the task of modeling online debates, characterized by the dynamic exchange of ideas, is a challenging endeavor. To tackle this complexity, we introduced a novel approach using sequence-graph modeling. By representing conversations as graphs, we effectively captured the interactions among participants through directed edges, while the sequential propagation of information along these edges enriched our understanding of context. Our incorporation of the SGA layer demonstrated the efficacy of our information update scheme. Our experimental results demonstrate the success of sequence graph networks in outperforming existing methods when applied to Oxford-style online debate dataset.

The proposed method not only advances the ability to model dynamic discussions but also highlights the potential of sequence-graph approaches for a wide range of tasks involving sequential interactions and context-rich data. As online debates continue to evolve, the techniques presented in this paper offer valuable insights into improving our understanding of complex conversational dynamics.

While the proposed method has demonstrated promising results in predicting debate outcomes, it does exhibit certain limitations. Firstly, the construction of cross-argument edges relies solely on similarity scores. While this approach may suffice for reinforcing connections, it may not consistently identify valid counterarguments. High similarity scores between two sentences do not guarantee a counterrelation. Secondly, the method overlooks the utilization of argument structures. The

intra-argument links primarily capture temporal relationships by connecting adjacent sentences. However, this approach fails to account for potential relationships between sentences that are distant within an argument turn. There is room for improvement by incorporating pre-trained models that account for argumentative structures. For instance, [26] enhanced predictability on debate datasets by integrating argument structure introduced by [34].

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Award # OIA-1946391, "Data Analytics that are Robust and Trusted (DART)".

REFERENCES

[1] E. Durmus and C. Cardie, "A corpus for modeling user and language effects in argumentation on online debating," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Italy, pp. 602–607, 2019.

[2] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, pp. 3982–3992, 2019.

[3] P. Veličković et al., "Graph attention networks," Proceedings of the 6th International Conference on Learning Representations, Canada, 2018.

[4] J. Zhang, R. Kumar, S. Ravi, and C. Danescu-Niculescu-Mizil, "Conversational flow in Oxford-style debates," Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, California, pp. 136–141, 2016.

[5] J. Bao et al., "Argument pair extraction with mutual guidance and inter-sentence relation graph," Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3923–3934, 2021.

[6] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Qatar, pp. 1532–1543, 2014.

[7] Y. Jo et al., "Attentive interaction model: Modeling changes in view in argumentation," in Proc. of the 2018 Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies, Vol. 1, pp. 103-116, 2018.

[8] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," Proceedings of the 4th International Conference on Learning Representations, Puerto Rico, 2016.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[10] L. Ji et al., "Incorporating Argument-Level Interactions for Persuasion Comments Evaluation using Co-attention Model," Proceedings of the 27th International Conference on Computational Linguistics, USA, pp. 3703–3714, 2018.

[11] C. Hidey and K. McKeown, "Persuasive influence detection: The role of argument sequencing," Proceedings of the AAAI Conference on Artificial Intelligence, pp. 5173–5180, 2018.

[12] H. Chen, P. Hong, W. Han, N. Majumder, and S. Poria, "Dialogue relation extraction with document-level heterogeneous graph attention networks," Cognitive Computation, vol. 15, no. 2, pp. 793-802, 2023.

[13] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Qatar, pp. 1724–1734, 2014.

[14] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," arXiv preprint arXiv:1607.06450, 2016.

[15] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," Proceedings of the 5th International Conference on Learning Representations, France, 2017.

[16] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in 3rd International Conference on Learning Representations, USA, 2015.

[17] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui, "Graph neural networks in recommender systems: a survey," ACM Computing Surveys, vol. 55, pp. 1-37, 2020.

[18] W. Fan et al., "Graph neural networks for social recommendation." In The world wide web conference, pp. 417-426. 2019.

[19] Q. Cao, H. Shen, J. Gao, B. Wei, and X. Cheng, "Popularity prediction on social platforms with coupled graph neural networks," Proceedings of the 13th international conference on web search and data mining, pp. 70-78. 2020.

[20] M. Wang et al., "Deep graph library: A graph-centric, highly-performant package for graph neural networks," arXiv preprint arXiv:1909.01315, 2019.

[21] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," In Advances in neural information processing systems, pp. 8024-8035. 2019.

[22] W. Falcon et al., "Pytorch lightning," GitHub 3, 2019.

[23] M. Honnibal and I. Montani, "spaCy: Industrial-Strength Natural Language Processing in Python," available at https://spacy.io, 2021.

[24] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," The journal of machine learning research 15, no. 1, pp. 1929-1958, 2014.

[25] R. Prasad et al., "The penn discourse treebank 2.0 annotation manual," December 17, 2007.

[26] J. Li, E. Durmus, and C. Cardie, "Exploring the role of argument structure in online debate persuasion," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 8905–8912, 2020.

[27] J. Zeng et al., "What changed your mind: The roles of dynamic topics and discourse in argumentation process," Proceedings of The Web Conference, pp. 1502-1513, 2020.

[28] L. Wu et al., "Graph neural networks for natural language processing: A survey," Foundations and Trends® in Machine Learning 16.2, pp. 119-328, 2023.

[29] T. Ji, W. Yuanbin, and L. Man, "Graph-based dependency parsing with graph neural networks," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Italy, pp. 2475-2485, 2019.

[30] B. Liang, S. Hang, G. Lin, C. Erik, and X. Ruifeng, "Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks," Knowledge-Based Systems, vol. 235, p. 10764, 2021.

[31] R. Li et al., "Dual graph convolutional networks for aspect-based sentiment analysis," Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 6319–6329, 2021.

[32] L. Song, Y. Zhang, Z. Wang, and D. Gildea, "A graph-to-sequence model for AMR-to-text generation," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Long Papers, Australia, pp. 1616-1626, 2018.

[33] Y. Chen, L. Wu, and M. J. Zaki, "Graphflow: Exploiting conversation flow with graph neural networks for conversational machine comprehension," Proceedings of the 29th International Joint Conference on Artificial Intelligence, Japan, pp. 1230-1236, 2020.

[34] V. Niculae, J. Park, and C. Cardie, "Argument mining with structured SVMs and RNNs," Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, pp. 985-995, 2017.

[35] C. Tan, V. Niculae, C. Danescu-Niculescu-Mizil, and L. Lee, "Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions," Proceedings of the 25th International Conference on World Wide Web, pp. 613-624, 2016.

[36] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 3634-3640, 2018.

[37] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, pp. 922-929, 2019.

[38] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12026-12035, 2019.

[39] https://github.com/quanmai/SGA.

Business Process Completeness

Foundation of Business Knowledge Management

Shuichiro Yamamoto
Information Engineering, IPUT in Nagoya
Nagoya, Japan
e-mail: yamamoto.shu@n.iput.ac.jp

***Abstract*—A clear definition of business processes is required to realize business. A business process is only complete when the problem is addressed. Additionally, it is difficult to address a problem if it is not identified. In this paper, we propose a comprehensive business process completeness concept from the aspects of business process, consisting of business process acceptance/resource/judgement conditions and exceptions that propagate between processes. In addition, a self-process completeness diagram is proposed to analyze the comprehensive process completeness. Furthermore, we confirm the effectiveness of the proposed method using examples.**

Keywords-business process management; knowledge transfer; Self-Process Completeness Diagram.

I. INTRODUCTION

In Japan, several inspection test frauds of manufacturing industry have recently been discovered and have become social problems [1]. The top management of a Japanese automobile company apologized for the inspection test fraud, saying, “We may have misjudged the workload.” It is clear that, if a company accepts orders that exceed its production capacity, it will not be able to produce the required amounts of products or services, or even if it is able to produce them, the quality of the products or services will degrade.

If an organization does not know its production capacity, it cannot know when the number of orders exceeds its production capacity. The production capacity at the time of planning often falls below the organization's planned production capacity at the time of execution due to excessive orders or changes in materials required for production. Not everything goes according to plan. Therefore, it is necessary to design business processes that can detect deviations from the plan as exceptions and respond to them. Conversely, if the upper limit of production capacity is known, it is possible to limit further orders by detecting an excessive number of orders as an exception. In order to correctly execute a business process, it is necessary to know the execution capability of the business process. Therefore, it is important to correctly define and confirm not only business process but also process execution conditions.

In this paper, we propose the Self-Process Complete Diagram (SPCD) as a model for designing the production process in industry and clarify that it can be applied to manage process completeness. Below, Section II describes related research. Next, Section III proposes SPCD as a means to manage comprehensive completeness among whole production processes. Section IV describes an application example of SPCD. In Section V, we discuss our considerations, and in Section VI, we present the conclusion.

II. RELATED WORK

Related studies on Ji-Koutei-Kanketsu (JKK), Knowledge transfer, Business Process Modeling (BPM), Self-Organized Process, and Functional Resonance Analysis Method (FRAM) are explained below.

A. Ji Koutei Kanketsu

In the production process, there is a misconception that local optimization is necessary, as long as one's own process is fine and that unnecessary problems shall not be introduced to one's own department. If a problem is discovered at the final stage of development, the design cannot be modified or the basic structure of the product cannot be changed. Therefore, comprehensive product design and manufacturing is required throughout the entire production process. Ji-Koutei-Kanketsu (JKK) is a method that optimizes the entire production process, not just a specific process. The Japanese words Ji, Koutei, and Kanketsu [2] are self, process, and completion, respectively.

To introduce JKK, it is necessary to define not only business procedures that define the flow of work, but also requirements organization sheets that define business requirements. The requirements organization sheet consists of fields of the necessary items/information, business inputs, and business outputs for each business process. The necessary item and information field clarifies the input, tools, methods, capabilities/authority, and reasons as conditions for the quality of product. The input field describes the receiving criteria, such as when, where, and what. The output field describes where to sink, by when, and what to produce. The criteria field describes criteria for determining that "the output of the process is good."

JKK's production processes can also be seen as business processes. JKK clarifies the completeness conditions for each business process element. The requirement organization sheet is an essential feature of JKK.

B. Knowledge transfer

In order to transfer a company's experiential knowledge, it is necessary to clarify business processes. For this reason,

 ISBN: 978-1-68558-165-7

methods for clarifying business processes have been proposed for knowledge transfer.

From a knowledge perspective, processes need to be defined to provide appropriate knowledge for tasks in an organization's operational business processes. In addition, knowledge must be extracted for the long-term growth, development, and competitiveness of companies. However, unless valuable knowledge within an organization is externalized or formalized, it cannot be used by other employees and disappears from the company. Therefore, Knowledge management shall be established using Business Process Modeling (BPM). Salvadorinhoa and Teixeira [3] pointed that BPM can not only help organizations improve their Industry 4.0 environment, but also facilitate knowledge acquisition and distribution.

C. Business Process Modeling

Ore et al. [4] proposed a Self-managed organization based on Business Process Management. They showed a need for the business process management approach, which would manage the need for keeping critical business processes continuity and self-managed way of working of autonomous teams.

As long as the digitalization of business is promoted, business process documentation becomes vital for business process continuity. The digitalization re-constructs the traditional business processes into a new digitalized business processes [5]. For example, Digital Balanced Scorecard (DBSC) [6] consists of digital business processes.

There are many Business Model notations including Business Process Models. Yamamoto [7] compared the representation capability of Business Model notations by defining fifteen key features of these notations with five interrogatives.

Leonard and Swap [8] defined deep smart as the expertise that allows experts to instantly grasp complex situations and make quick and wise decisions in order to deal with real problems. That is, deep smart is “strong expertise formed by beliefs and social influences that can generate insights based on tacit knowledge grounded in direct experience.” For example, in production process design, the problem is how to transfer defect investigation knowledge from experienced workers to beginners. An example of deep smart is the failure investigation knowledge that experienced engineers have. Leonard and Swap pointed out the importance of acquiring empirical knowledge through experimental learning. However, no concrete experimental learning method has been clarified. In addition, they have not clarified the knowledge representation of deep smarts. If deep smart cannot be expressed, it remains tacit knowledge, and deep smart knowledge transfer from experts to beginners is individual and difficult to spread horizontally.

As a technique for improving production processes in the manufacturing industry, Mono-Koto-Bunseki (MKB) (in Japanese) has been proposed [9]. Mono, Koto, and Bunseki mean Entity, Process, and Analysis, respectively. By treating objects such as materials and products as “entities” and the series of activities that make products from materials as "process," MKB can analyze the production process, discover waste, and optimize it.

Yamamoto and Fujimoto [10] proposed the Production Knowledge Chart (PKC) that expresses the production process to acquire the empirical knowledge necessary for investigating defects in manufacturing processes.

Object Process Methodology (OPM) proposed by Dori includes Object and Process [11][12]. For example, the aircraft design OPM has a Stakeholder Needs Set, Assumptions and Constraints Sets, and Requirements as Objects. There are three types of Processes: Defining, Realizing, and Implementing. In addition, physical Objects include Aircraft, System, Item, and Item component.

D. Self-Organized process

Bussmann and Schild [13] developed a strictly decentralized approach to manufacturing control by using workpiece and machine agents. Machine agents manage a virtual buffer. Workpiece agents manage the state of workpieces. They showed a capacity bottleneck is automatically propagated in the opposite direction of the material flow.

Graessler et al. [14] clarified the process changes and opportunities for the development process by the vision of Self-Organizing Production Systems (SOPS). Main features of SOPS are as follows. SOPS consists of segmented autonomous modules instead of one connected system. Distributed control procedures of SOPS manage to react to unexpected changes of the production system. Connecting to related services and devices allows them to exchange information regarding the execution of their own production processes.

E. Functional Resonance Analysis Method

Functional Resonance Analysis Method (FRAM) [15] has been used to analyze complex functional resonances of socio-technical systems through functional networks. The FRAM function is defined by hexagonal nodes with six sides. These sides correspond to six aspects which are Input, Output, Time, Control, Resource, and Precondition. The output side of a function can be connected to the other five sides of other functions. FRAM provides useful means for safety analysis. Possible aspect relationships are <O, I>, <O, T>, <O, C>, <O, R>, and <O, P>. Here, <X, Y> is where X and Y are functional aspects.

The following three types of FRAM matrix representations have been proposed.

Lundberg and Woltjer [16] proposed a Resilience Analysis Matrix (RAM) to visualize functional dependencies between complex systems. RAM is a square matrix that shows the propagation relationship between functions. The size of RAM is the number of functions in FRAM. Element (i, j) of RAM indicates that some aspect of function i is propagated from the output of function j. The diagonal element (i, i) of RAM is the output of function x.

Patriarca et al. [17] proposed another square matrix composed of aspect combinations of FRAM functions. If

there are n couplings in FRAM, RAM is defined as an n × n square matrix. The value of RAM (i, j) is 1 or 0.

Functional Aspect Resonance Matrix (FARM) is a non-square matrix that shows the propagation relationship between the output of a function and other aspects [18]. The number of rows in FARM is the number of output sides of the function that are propagated to other functions in FRAM. The column size of FARM is the number of sides of a function that are connected from the output sides of other functions. Element (i, j) of FARM indicates that some functional surface j is propagated from the output of function i. In general, the number of rows and columns in FARM are not equal, so there are no diagonal elements. The equivalence of the above three matrices has been shown by Yamamoto [18].

III. SELF COMPLETE BUSINESS PROCESS

A. *Self-Process Complete Diagram*

Self-Process Complete Diagram (SPCD) is defined by hexagonal nodes with six sides. These sides correspond to six aspects which are Input, Output, Acceptance condition, Resource condition, Exception condition, and Judgement condition. The acceptance, input, resource, and judgement aspects represent outside-in flows from external elements. The output and exception aspects represent inside-out flows to external elements.

Figure 1 shows an example of SPCD.

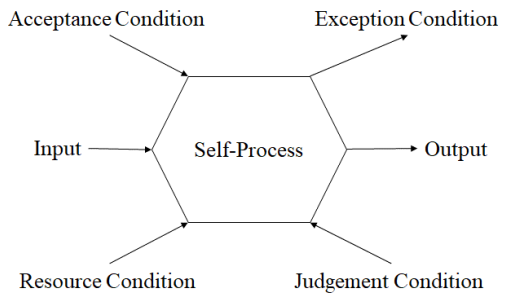

Figure 1. Example of Self-Process Complete Diagram.

The metamodel of SPCD is shown in Figure 2. There are two relationships, i.e., connection and propagation relationship.

The connection relationship defines the binary relationship that flows from the output aspect of a process into the input aspect of other processes. The connection relationship is used to define business process flows.

The propagation relationship defines 1) the exception condition of a process flows into acceptance condition of other process, and 2) the exception condition of a process flows into the exception condition of other processes.

The propagation relationship is used to propagate exceptions of a process into forwardly and backwardly other processes.

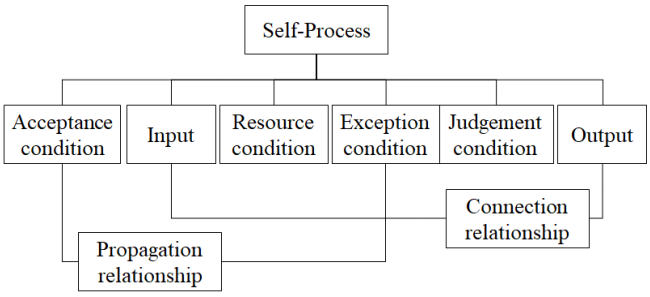

Figure 2. Metamodel of Self-process Complete Diagram

Figure 3 shows an example of propagation relationship. In Figure 3, there are two processes of a production plan and a production for delivery. The production plan process accepts a purpose of plan and generate the production order. If the production for delivery process accepts the production order, then it generates the product. In case of production capacity is not sufficient, the production delay occurs as the exception in the process. The exception is propagated to the acceptance aspect of former process. Then the former process is noticed that the purpose of the plan is no more realized.

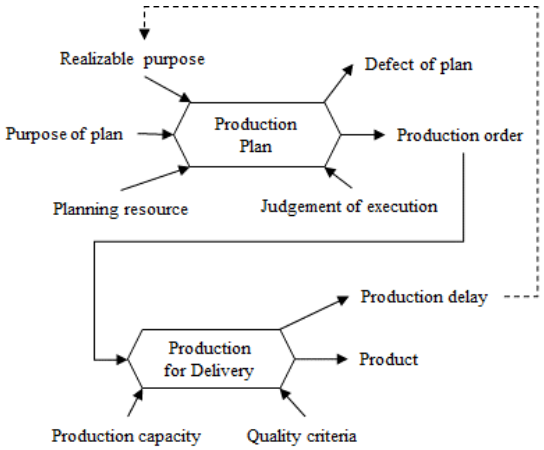

Figure 3. Example of Self-processes for production

B. *Conditions of the Complete Self-Process*

The following are conditions used to check if the process itself is complete.

If the acceptance conditions are not met, the process will not start.

Unless the resource conditions are met, the process will not start.

If the result of the own process does not satisfy the judgment conditions, it will not be output.

Generates an exception condition when the own process cannot start or when the output does not satisfy the judgment conditions.

When the resource conditions are satisfied for the input that satisfies the acceptance condition, generate an output that satisfies the judgment condition of the own process.

C. Business Process Analysis with SPCD

Business process analysis using SPCD is as follows.

[step1] Describe business processes and flows among processes with input and output arrows.

[Step2] Clarify resource, acceptance, and judgement conditions for each process.

[Step3] Analyze the possibility of deviations under above three conditions of each process.

[Step4] Identify exceptions of each process based on deviations analyzed.

[Step5] Analyze propagations of exceptions among processes.

There are two directions of propagation: upward and downward propagation. The upward propagation feeds back exceptions from a downstream business process to its upstream business processes. The downward propagation feeds exceptions from an upstream business process to its downstream business processes.

The exception propagation analysis is used to discover candidates of business process improvement. If an exception is notified to a process, the process should change acceptance, resource, and judgement conditions to handle the exception. This presents an opportunity for process evolution to improve sustainability in response to environmental changes.

IV. CASE STUDY

In this section, two case studies are explained to show the applicability of SPCD.

A. Alcoholic beverage delivery

When I went on a trip to a north region of Japan, I decided to buy two bottles of sake from that region at a local liquor store and send them home. By paying for the local sake and filled out the delivery slip, I asked a courier to send the sake packaged by the liquor store to the address on the slip. A courier delivered the package to a distribution center near my home. At the distribution center, they noticed that local sake was leaking. There was a problem with the sake during delivery, so the distribution center requested the sending liquor store to repack it and redeliver it. Due to sufficient packaging, the two bottles of sake were re-delivered safely. Figure 4 shows the flow of these processes along with backpropagation of exceptions.

The dotted lines show propagations of exceptions. For example, the “insufficient packaging” exception in “Packed for local delivery” process propagates to the acceptance condition of “Receive alcoholic beverages” process. Then the “Alcohol is leaking” exception occurred in “Receive alcoholic beverages” process. The exception again propagates to the acceptance condition of “Packed for local delivery” process.

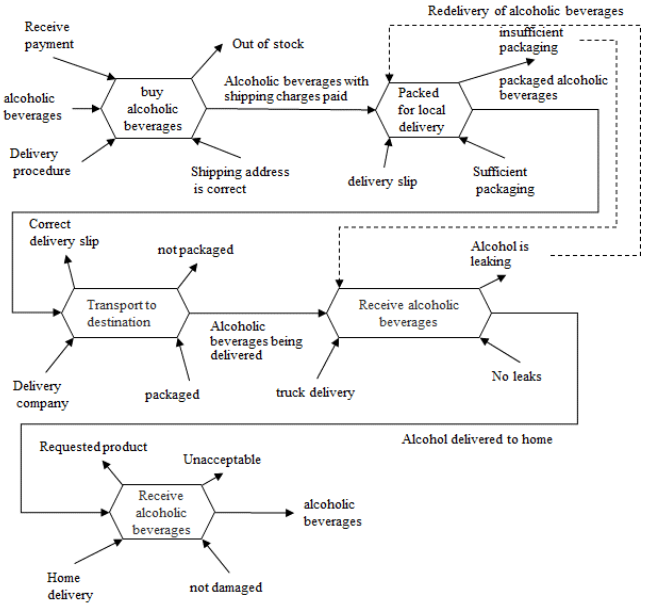

Figure 4. Example of Alcohol beverage delivery.

B. Strawberry cake shipping

There was an online sale in which strawberry cakes for Christmas were delivered on Christmas Eve. This strawberry cake was supervised by a famous pastry chef and became popular, with many orders placed. However, due to the intense summer heat, the strawberry crop failed, and they were unable to procure the strawberries they needed right away. Production of the cake was delayed due to a delay in the procurement of strawberries. Furthermore, during the shipping process, the manufactured cake had to be frozen for a certain period of time to maintain quality. As a result, delays in the procurement of strawberries caused delays in production and insufficient freezing time. As a result, some cakes collapsed when delivered to consumers.

Figure 5 shows the result of describing this process flow from order reception to manufacturing and delivery using SPCD. The SPCD shows cause and result of the accident by the propagation of exceptions. The “procurement delay” exception causes “unmanufactured orders” in “Accept order” process and “delay in production” in “Manufacture cake” process. The “delay in production” exception propagates to “insufficient refrigeration period” exception in “Refrigerate cake” process. Finally, “crumbled cake” exception has occurred in “deliver cake” process because of insufficiently refrigerated cake.

To prevent this event, it is needed to know the “procurement delay” exception in the course of “accept order” process and suspend or stop orders that will cause unexpected troubles. In this way, SPCD will help analyze exception propagation and prevent unexpected matters in business processes.

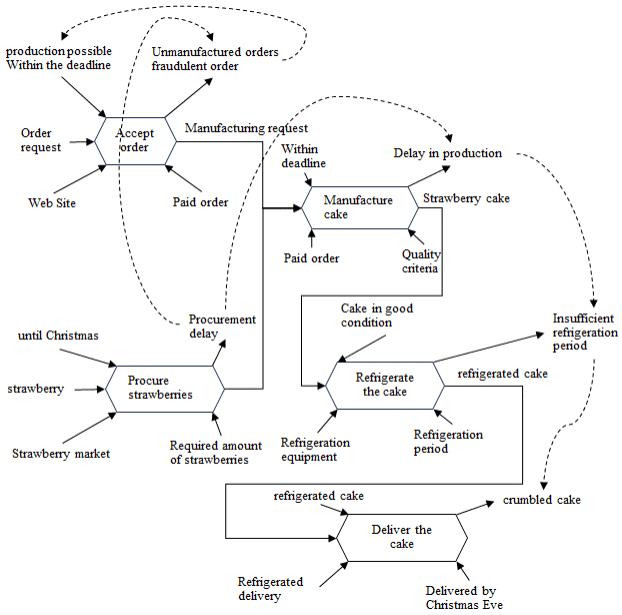

Figure 5. Example of Strawberry cake shipping

V. DISCUSSION

A. Novelty

The SPCD is designed to clarify comprehensive business process completeness by using six aspects. They are input, output and four conditions (acceptance, resource, judgement, and exception). So far, the aspect combination proposed in the paper has never been known. Moreover, exception propagation relationship has been proposed to countermeasure the failure risk of business processes. Acceptance conditions can block further failures by recognizing that an exception has occurred during the course of subsequent processing.

The completeness of business processes has also been defined by using SPCD aspects. Until now, the completeness of business processes has not been clear.

JKK needs to describe not only business process diagrams but also requirements organization sheets for processes. SPCD compactly describe comprehensive business process conditions than JKK in one diagram.

B. Effectiveness

In this paper, we proposed SPCD as a method of analyzing the completeness of business processes. In addition, we clarified the effectiveness of the proposed method by applying it to the simple service delivery and manufacturing examples. It was also revealed that the completeness of business processes can be confirmed by propagating exceptions.

C. Equivalence of SCPD

SCPD is defined by a set of processes P, aspects A, I, R, J, O, E, and relationships set R between elements of P. let P= {(x, a, i, r, j, o, e): x is a process, a, i, r, j, o, and e are aspects of x}. Then R can be the union of the following three set.

Output to Input {(x. o, y. i): x and y are processes of P}

Exception to Exception {(x. e, y. e): x and y are processes of P}

Exception to Acceptance {(x. e, y. a): x and y are processes of P}

Now, let <P1, R1> and <P2, R2> be two SPCDs.

<P1, R1> and <P2, R2> are equivalent if the following condition holds.

P1 = P2 and R1 =R2

D. Comparison FRAM and SPCD

FRAM and SPCD have common aspect as input, output, and resource. FRAM has time, precondition, and control aspects which are not in SPCD. SPCD also has acceptance, exception and judgement condition aspects which are not in FRAM. The output of FRAM is restricted to output aspects. Therefore, the meaning of output in FRAM may be unclear as it is difficult to discriminate exceptional output from normal output by aspects.

Although there are differences between FRAM and SPCD, it is unclear whether they have the same expressive power. As FRAM can be applied to analyze the resonance relationship between processes, the completeness of business processes may also be possible to analyze by FRAM. Sujan, and Felici [19] combines Failure Mode and FRAM. This implies a new method possibility that integrates analysis method using SPCD with Failure mode analysis.

The formal comparison between FRAM and SPCD is an interesting future research theme.

E. Comparison with IDEF0

The comparison of SPCD and Integrated DEFinition 0 (IDEF0) [20] is as follows. IDEF0 describes connectivity of functions with four arrows of input, output, control condition and mechanism conditions. Only output arrow of IDEF0 flows into outside functions from the source function.

SPCD describes six arrows of input, output, acceptance condition, judgement condition, resources condition and exception condition. Output and exception condition arrows of SPCD flow outside from processes.

In IDEF0, it may complicate to distinguish exception flows from output flows. Moreover, acceptance and judgement conditions are difficult to distinguish in control conditions of IDEF0.

F. Digital Transformation

The data driven management is a vision of Digital Transformation. Digital business processes will ease to collect business data in real time. The six aspects of proposed SPCD are business data candidate shall be collected for the digital twin of organizations. For example, a major business process failure incident will not be managed if management is unaware that an exception has occurred. This will be the case which mentioned episode in the beginning of this paper. The incident response should rapidly be executed. Digitalization of incident management is inevitable, because human communication is time consuming task. Moreover, human employees tend to hide incidents where they are cause or responsible. Digital

algorithms do not hide incidents and, if implemented, report them quickly. This shows the importance of identifying aspects of SPCD. If the aspects are not identified, business process data cannot be collected and utilized.

G. Limitations

In this paper, we proposed a method to describe complete business processes by SPCD. We also clarified that SPCD can express the exception handling knowledge in comprehensive business processes. These cases are only based on small cases happened in Japan.

Future work on evaluating the proposed method can be designed an experiment to compare SPCD with JKK, BPM, and IDEF0. For the given same business process, it is needed to compare productivity and quality of these approaches. Moreover, qualitative capability assessment study of these approaches should be conducted.

Although the necessity of digital twin of business organizations was mentioned in the former section using SPCD, the digital twin architecture has not been clarified. The digital twin of SPCD will provide exception events monitoring and activation of appropriate handling processes. It also stores all management data issued across business processes required for data-driven business management.

VI. CONCLUSION

In this paper, we proposed a notion of business process completeness, as well as Self-Process Complete Diagram (SPCD) for describing business processes in industry. As a result, we clarified the following.

(1) SPCD can represent the business process using six aspects

(2) SPCD can represent the defect propagation process

(3) It was also pointed that SPCD has the potential to integrate business process design and data driven management of industry.

ACKNOWLEDGMENT

The author would like to thank Shinichi Sasaki, former vice president of Toyota Motor Corporation, who conceived the JKK concept through Toyota Motor Corporation's manufacturing process. The author also thanks Ms. Hosomi who introduced the JKK to the author. Without her suggestions I would never know JKK.

REFERENCES

[1] The Japan News, Irregularities result from closed corporate culture. Available from: https://japannews.yomiuri.co.jp/editorial/yomiuri-editorial/20220804-49330/

[2] S. Sasaki, “Self-process completion - Quality is built in the process,” JSQC selection, Japan Society for Quality Control, 2014 (in Japanese).

[3] J. Salvadorinhoa and L. Teixeira, “Organizational knowledge in the I4.0 using BPMN: a case study,” CENTERIS 2021, Procedia Computer Science 181, 2021, pp. 981–988.

[4] A. Ore, O. Kuznecova, and A. Jegorova, “Self-managed Organization: A Role of Business Process Management,” 2021 62nd International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS), 2021, doi: 10.1109/ITMS52826.2021.9615260

[5] P. Faraboschi, E. Frachetenberg, P. Laplante, D. Milojicic, and R. Saracco, “Digital Transformation: Lights and Shadows,” IEEE Computer, 2023, pp. 123-130.

[6] S. Yamamoto, “A Strategic Map for Digital Transformation,” KES 2020, Procedia Computer Science, vol. 176, 2020, pp. 1374-1381.

[7] S. Yamamoto, “A Comparative Analysis of Business Model Notations,” Journal of Business Theory and Practice, ISSN 2372-9759 (Print) ISSN 2329-2644 (Online), vol. 7, no. 3, 2019.

[8] D. Leonard and W. Swap, “Deep Smarts: How to Cultivate and Transfer Enduring Business Wisdom,” Harvard Business Review Press, 2005.

[9] Z. Nakamura, “How to conceive a simple job—Succeed in Entity Process Analysis,” Nikkan Kogyo Shimbun, 2003 (in Japanese).

[10] S. Yamamoto and H. Fujimoto, “PKC-- Production Knowledge Chart for Knowledge Transfer,” KES2023, Procedia Computer Science 225, 2023, pp. 414-423.

[11] D. Dori, “Model-Based Systems Engineering with OPM and SysML,” Springer, 2016.

[12] B. Cameron, E. Crawly, and D. Selva, “System Architecture,” PEARSON, 2015.

[13] S. Bussmann and K. Schild, “Self-Organizing Manufacturing Control: An Industrial Application of Agent Technology,” 4th Int. Conf. on Multi-Agent Systems, 2000, pp. 1-8.

[14] I. Graessler, J. Hentze, and A. Poehler, “Self-organizing production systems: Implications for product design,” Procedia CIRP 79, 2019, pp. 546-550.

[15] E. Hollnagel, “FRAM - the Functional Resonance Analysis Method: Modelling Complex Socio-Technical Systems.” CRC Press, 2012.

[16] J. Lundberg and R. Woltjer, “The Resilience Analysis Matrix (RAM): Visualizing functional dependencies in complex socio-technical systems,” In: Proceedings of the 5th Resilience Engineering Association Symposium, 2013.

[17] R. Patriarca, G. Pinto, G. Di Gravio, and F. Costantino, “FRAM for Systemic Accident Analysis: A Matrix Representation of Functional Resonance,” International Journal of Reliability Quality and Safety Engineering, vol. 25, no. 1, 2018, doi: 10.1142/S0218539318500018

[18] S. Yamamoto, “Comparative Study on Functional Resonance Matrices,” Knowledge-Based Software Engineering: 2022: Proceedings of the 14th International Joint Conference on Knowledge-Based Software Engineering (JCKBSE 2022), M. Virvou, T. Saruwatari, L. Jain *Editors*, Learning and Analytics in Intelligent Systems 30, Springer, 2023, pp. 169-180, doi:org/10.1007/978-3-031-17583-1

[19] M. Sujan and M. Felici, “Combining Failure Mode and Functional Resonance Analyses in Healthcare Settings,” SAFECOMP, LNCS 7612, 2012, pp.364-375.

[20] IEEE, “Standard for Functional Modeling Language - Syntax and Semantics for IDEF0,” IEEE Std. 1320.1-1998.

Improving Minority Stress Detection with Emotions

Jonathan Ivey
Department of Data Science
University of Arkansas
Fayetteville, AR, USA
jwi001@uark.edu

Susan Gauch
Department of Computer Science
University of Arkansas
Fayetteville, AR, USA
sgauch@uark.edu

***Abstract*—Psychological stress detection is an important task for mental healthcare research, but there has been little prior work investigating the effectiveness of psychological stress models on minority individuals, who are especially vulnerable to poor mental health outcomes. In this work, we use the related task of minority stress detection to evaluate the ability of psychological stress models to understand the language of sexual and gender minorities. We find that traditional psychological stress models underperform on minority stress detection, and we propose using emotion-infused models to reduce that performance disparity. We further demonstrate that multi-task psychological stress models outperform the current state-of-the-art for minority stress detection without directly training on minority stress data. We provide explanatory analysis showing that minority communities have different distributions of emotions than the general population and that emotion-infused models improve the performance of stress models on underrepresented groups because of their effectiveness in low-data environments, and we propose that integrating emotions may benefit underrepresented groups in other mental health detection tasks.**

Keywords-stress; emotion recognition; natural language processing; social networking.

I. INTRODUCTION

Psychological stress detection from social media posts has been identified as an important task for mental healthcare research [1], but the datasets for this task may not fairly represent all groups, and little prior work has investigated the effectiveness of psychological stress models on minority individuals.

This issue is especially relevant for Sexual and Gender Minority (SGM) people, who are more vulnerable to poor mental health outcomes than the general population. They are at higher risk of mental illnesses and suicide [2]–[4], and social media is often a place where SGM people find peers, seek help, and cope with prejudice [5]–[8].

One way to evaluate the ability of psychological stress models to understand the language of SGM individuals is through the related task of minority stress detection. Like psychological stress detection, minority stress detection uses Natural Language Processing (NLP) techniques to classify social media posts with whether the poster is experiencing stress [9][10]. However, minority stress is a psychosocial stress specific to minority individuals that they experience due to stigmatized social status [11]. An example of minority stress on social media is provided in Figure 1.

> At school, I have great friends and a good family at home. But I'm a closeted gay. If I ever came out, I know my friends would never talk to me again and my family would disown me. Because of this, I have zero motivation to come out.

> I was kicked out of my online video game squad just for being gay (they said gays are pedophiles). My dad was outside my room listening when this happened so he grabbed me by the throat and kicked me out.

Figure 1. Examples of minority stress disclosure on social media from [9].

This task has an important application in improving the methodology of minority stress studies by circumventing limitations in survey-based self-reporting [12]. The systematic detection of minority stress can also be used to study large-scale health trends on social media that are not feasible to collect survey data on. Additionally, it has applications in automated intervention for those at risk of adverse consequences and screening for comorbid risks, such as cancer, HIV, and reduced cardiovascular health [13].

In this work, we evaluate the effectiveness of psychological stress models at detecting minority stress, and we hypothesize that a lack of diversity in the psychological stress training data causes stress models to overfit and be unable to generalize to minority individuals.

To address this issue, we experiment with the multi-task emotion-infused architectures introduced by [14]. They explored connections between emotions and stress in deep learning models, and they demonstrated that the task of emotion detection, which has more labeled data available, could improve the explainability of stress models.

Initial research found that emotion-infused models did not improve performance on the psychological stress detection task; however, we note that multi-task learning techniques (such as those used for the emotion-infused models) are known to improve generalization [15]. In this work, we explore using multi-task emotion-infused models to improve minority stress detection and highlight their potential for improving the performance of other mental health models on minorities. Our contributions in this work are as follows:

- We conduct experiments to demonstrate that traditional single-task psychological stress detection models under-

 ISBN: 978-1-68558-165-7

perform on minority stress and highlight how this performance difference risks widening preexisting healthcare disparities experienced by minority communities.
- We demonstrate that emotion-infused models reduce the performance gap and exceed State-Of-The-Art (SOTA) performance for the minority stress detection task without training on minority stress data.
- We provide explanatory analysis showing that minority communities have different distributions of emotions than the general population and that emotion-infused models improve the performance of stress models on underrepresented groups because of their effectiveness in low-data environments.

II. Related Work

A. Psychological Stress Detection

Psychological stress detection is best studied with physiological data. Prior work has used audio [16], biological markers [17], neuroimaging [18], thermal imaging [19], or combinations of these signals [20] to achieve the most accurate forms of psychological stress detection. However, [1] demonstrated the value and feasibility of detecting psychological stress purely from social media text.

In [14], the authors introduced the use of emotion-infused models for psychological stress detection. These models improved explainability by integrating emotion with multi-task learning or fine-tuning; however, they did not significantly improve the performance of psychological stress detection.

The authors acknowledged demographic imbalances in the psychological stress dataset and noted a lack of language representing minority groups; however, little previous work has explored the performance of psychological stress models on minorities. In this work, we use the minority stress detection task to highlight the limitations of traditional architectures for detecting stress in minority individuals and explore the benefits of the previously introduced emotion-infused architectures for overcoming those limitations.

B. Minority Stress Detection

In [9], the authors introduced the use of NLP techniques for understanding minority stress. They wrote a codebook for identifying minority stress, created the first dataset of social media posts annotated for disclosure of minority stress, and introduced the first machine learning classifier for minority stress on social media. They experimented with using expertly engineered language features in combination with machine learning models to build a classifier. Their models are the current SOTA for minority stress detection.

Building off that work, [10] introduced a proof-of-concept Bidirectional Long Short-Term Memory (BI-LSTM) model to detect minority stress without expertly engineered features. They were the first to use deep learning for this task; however, they found limited results that did not outperform traditional machine learning models in detecting minority stress.

The authors viewed the minority stress detection task in isolation and trained models directly on the minority stress dataset (which is too small for deep learning models). In this work, we understand minority stress as a subset of psychological stress, and we use this framework to improve minority stress detection by improving models from the related task of psychological stress detection.

C. Pretrained Language Models

Prior work has suggested that domain-specific pretrained language models may benefit mental healthcare tasks. In [21], the authors trained Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pretraining Approach (RoBERTa) models on a corpus of 13,671,785 sentences from mental health-related subreddits. These models, which the authors named MentalBERT and MentalRoBERTa, demonstrated improved performance on psychological stress detection. However, little previous work uses pretrained language models on minority stress detection. In this work, we experiment with four pretrained language models (BERT, RoBERTa, MentalBERT, and MentalRoBERTa) paired with psychological stress models for minority stress detection.

III. Approach

A. Baselines

For our baseline stress model, we use the pretrained BERT language model introduced in [22] followed by an additional dropout layer and dense classification layer. This architecture is the simplest that we evaluate, and it performed the best on the psychological stress detection task when it was introduced in [1]. Later architectures have not provided statistically significant performance improvements. We will refer to this model as Single-Task.

We compare our minority stress models to the current SOTA for minority stress detection established in [9]. This model is a Multilayer Perceptron (MLP) algorithm trained with GloVe word embeddings [23], Linguistic Inquiry and Word Count (LIWC) psycholinguistic categories [24], a gender and orientation hate speech lexicon, n-grams, sentiment classification [25], and stress classification [26]. This series of expertly engineered features draws information from a wide range of data sources that consider lexical and semantic aspects of the text, with a special focus on LGBTQ+ issues.

B. Emotion-Infused Models

We hypothesize that integrating emotions into psychological stress models will account for differences in the emotional expressions of minority individuals and improve the models' ability to generalize to minority stress. To test this claim, we evaluate the emotion-infused models introduced in [14]. When previously evaluated on psychological stress detection, these models did not provide significant performance improvements, but they improved explainability, and they represent key methods for using emotion in mental health tasks. There are three emotion-infused models.

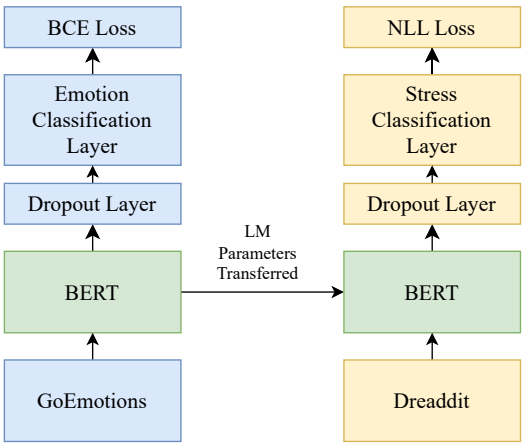

Figure 2. The architecture of Fine-Tune. Components that are used for both tasks are highlighted in green.

1) Fine-Tuning Model (Fine-Tune): A visualization of Fine-Tune is shown in Figure 2. In this architecture, we first fine-tune a single-task model for emotion detection. Because that is a multi-label task, the model trains using Binary Cross-Entropy (BCE) loss. Then the language model parameters from that BERT model are transferred to another single-task model that is further fine-tuned for psychological stress detection. Because that is a single-label task, the model trains using Negative Log-Likelihood (NLL) loss. The rationale for this architecture is that the first task would give the BERT language model a better understanding of emotions, and that understanding would enable a more holistic representation of stress.

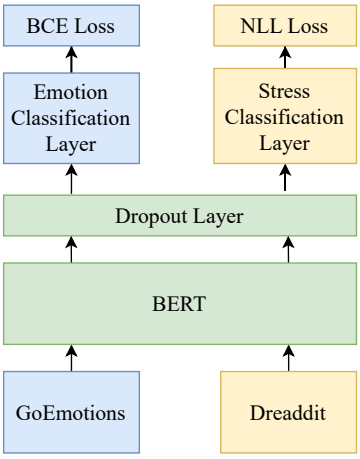

Figure 3. The architecture of Multi^{Alt}. Components that are used for both tasks are highlighted in green.

2) Alternating Multi-Task Model (Multi^{Alt}): A visualization of Multi^{Alt} is shown in Figure 3. It follows a similar rationale as Fine-Tune, but instead of training and then transferring a separate language model, it trains a single, shared language model. During training, it alternates between training for emotion detection and psychological stress detection. Each training batch switches which task it is training for, but these different tasks share the same BERT representation layer. As in the Fine-Tune model, the emotion model trains with BCE loss and the psychological stress model trains with NLL loss.

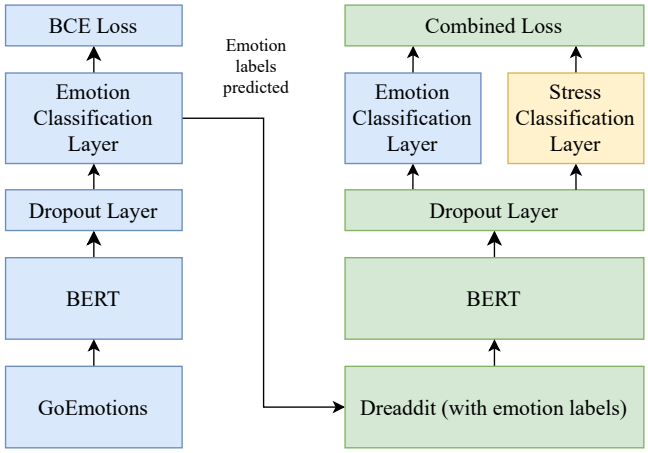

Figure 4. The architecture of Multi. Components that are used for both tasks are highlighted in green.

3) Classical Multi-Task Model (Multi): A visualization of Multi is shown in Figure 4. It differs from the other two models because it uses a classical multi-task architecture that has the same input data for both tasks. However, because the stress data is not labeled with emotions, we first separately train a single-task model for emotion detection and use it to predict emotion labels for the stress data. The multi-task model then uses these emotion labels as targets for training the emotion detection task. In this model the loss is given by $\mathfrak{L} = \lambda\mathfrak{L}_{stress} + (1-\lambda)\mathfrak{L}_{emotion}$ where $\mathfrak{L}_{stress}$ is the NLL loss for psychological stress detection, $\mathfrak{L}_{emotion}$ is the BCE loss for emotion detection, and λ is a weight parameter that we tune during model selection. The rationale for this model follows the traditional understanding that the inductive bias from the emotion detection task would improve its generalization.

C. Pretrained Language Models

Single-Task, Fine-Tune, Multi^{Alt}, and Multi were all originally introduced using BERT, but [21] demonstrated that their domain-specific pretrained language models, MentalBERT and MentalRoBERTa, improved performance on the psychological stress detection task. In this work, we evaluate four pretrained models tested in their work (BERT, RoBERTa, MentalBERT, and MentalRoBERTa) paired with each of the previously mentioned stress models.

IV. Data

For psychological stress detection, we use Dreaddit [1], a dataset of 3,553 segments of Reddit posts from communities where stress is commonly disclosed. Each segment was labeled with whether the poster is expressing stress using

TABLE I
MINORITY STRESS PERFORMANCE

	BERT		RoBERTa		MentalBERT		MentalRoBERTa	
Model	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**
Single-task	69.85	70.88	74.88	**75.09**	73.49	73.56	73.33	73.95
Fine-Tune	69.47	70.31	70.87	71.45	68.71	70.50	72.24	72.61
Multi^{Alt}	70.95	**71.45**	70.60	70.69	73.58	**74.52**	71.88	72.41
Multi	**75.55**	68.58	**75.16**	69.35	**75.58**	72.99	**78.53**	**74.52**

Results of the models evaluated on minority stress detection with different pretrained language models. The best result under each metric is bolded.

TABLE II
PSYCHOLOGICAL STRESS PERFORMANCE

	BERT		RoBERTa		MentalBERT		MentalRoBERTa	
Model	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**
Single-task	77.70	77.95	78.18	78.37	76.03	76.88	79.35	79.39
Fine-Tune	75.85	75.67	77.48	77.53	77.30	77.43	76.37	76.78
Multi^{Alt}	**78.76**	**78.97**	77.00	77.25	**80.80**	**80.89**	79.42	79.59
Multi	77.53	77.36	**79.00**	**79.16**	77.90	78.27	**79.86**	**79.91**

Results of the models evaluated on psychological stress detection with different pretrained language models. The best result under each metric is bolded.

TABLE III
STRESS LABEL DISTRIBUTIONS

Dataset	Split	Stress	Non-Stress
Dreaddit	Training	1,110	1,012
	Development	374	342
	Testing	374	341
MStress	Development	72	103
	Testing	72	103

Label distributions for the training, development, and testing sets of Dreaddit and MStress.

crowdsourced annotation, requiring a majority vote from five annotators. We use this dataset to train the stress models and evaluate their performance on psychological stress detection.

To evaluate how the models generalize to minority stress detection, we use an existing dataset of 350 Reddit posts collected from LGBTQ+ communities by [9]. These posts were manually labeled by the authors with whether they contain the disclosure of minority stress using a codebook built based on Meyer's Minority Stress model [11]. In this paper, we will refer to this dataset as MStress. Table III provides more information about the label distributions of Dreaddit and MStress.

Finally, to train the emotion-infused models, we use both Dreaddit and the GoEmotions dataset [27]. GoEmotions consists of 58,009 Reddit comments labeled by crowd workers with one of more than 27 emotions (or neutral). Based on the findings of [14] we use a relabeling of this dataset created with agglomerative clustering to cluster the original labels into the Ekman 6 basic emotions (anger, disgust, fear, joy, sadness, surprise, neutral) [28].

V. EXPERIMENTAL SETUP

Due to the scarcity of minority stress data, we do not train our models directly on MStress. Instead, all the models are trained on Dreaddit and GoEmotions (for the emotion-infused models) with minibatch gradient descent using the Adam optimizer [29].

We test on Dreaddit to evaluate their psychological stress detection, and we test on MStress to evaluate their minority stress detection. For Dreaddit and GoEmotions, we use 60% of the data for training, 20% for hyperparameter tuning, and 20% for testing. We choose F1 score to evaluate the performance of our models because of its ability to account for class imbalances.

We replicate [14]'s training and hyperparameter tuning processes with the same parameter ranges; however, for our primary tests, we use 50% of MStress for hyperparameter tuning instead of using Dreaddit. We make this change to find the peak performance of the models for minority stress detection and highlight the associated reductions in psychological stress detection. We run each of these experiments three times with three different random seeds, and we report the mean of the three runs.

VI. RESULTS

We report the results of our primary tests when evaluated on minority stress in Table I, and we report the results of our primary tests when evaluated on psychological stress in Table II.

A. Single-Task Models

We find that the traditional models underperform on minority stress detection. The Single-Task models achieve F1 scores between 2.54 and 7.85 points lower on minority stress detection than on psychological stress detection. The MentalRoBERTa model performs best on psychological stress detection out of the Single-Task models with an F1 of 79.35, but that score drops to 73.33 when evaluated on minority stress detection.

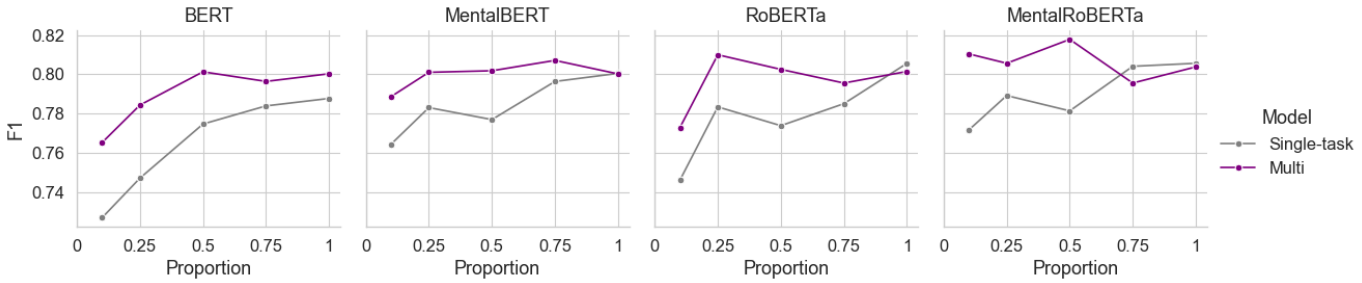

Figure 5. Performance of the Single Task and Multi models trained with different proportions (10%, 25%, 50%, 75%, and 100%) of the original training set and evaluated on the psychological stress data.

B. Emotion-Infused Models

Though Fine-Tune and Multi^{Alt} do not provide significant performance improvements, we find that the Multi models improve minority stress detection performance over baselines in all cases. Multi combined with MentalRoBERTa achieves an F1 of 78.53 on minority stress detection. While this result is still lower than the best psychological stress models, it closes the gap significantly and outperforms 13 out of the 16 psychological stress models. This result demonstrates that using the Multi architecture improves the ability of psychological stress models to generalize to minority stress detection.

Importantly, our best Multi models are not trained directly on minority stress data, but they outperform the current SOTA MLP for minority stress detection. The MLP proposed by [9], which is trained directly on minority stress data, achieved an F1 of 75, and our best Multi model achieves an F1 score of 78.53.

C. Pretrained Language Models

We find that in most cases, domain-specific language models such as MentalBERT and MentalRoBERTa provide marginal improvements in both psychological stress detection and minority stress detection compared to the standard BERT and RoBERTa models. This result confirms prior work demonstrating that the MentalBERT and MentalRoBERTa language models perform better on a variety of mental healthcare tasks [21].

D. Discussion

These findings have important implications for the use of stress models in research and healthcare applications. First, traditional Single-Task models perform worse on minority stress detection than psychological stress detection and risk reinforcing preexisting mental healthcare disparities for SGM individuals. Second, the Multi architecture creates models that can generalize well to minority stress detection and significantly reduce the performance gap. Finally, minority stress researchers can benefit from using psychological stress detection data to surpass the current SOTA without directly training on minority stress data.

In the next section, we support these conclusions with an analysis of the Single-Task and Multi models for psychological stress detection with reduced training sets.

VII. Analysis

A. Data Reduction Analysis

We propose that the disparity in performance of the baseline models between minority stress and psychological stress is due to overfitting on the psychological stress data. The Single-Task models gain too much sample-specific information and, as a result, are struggling to perform well on out-of-sample stress disclosures like minority stress.

Multi-task learning techniques improve generalization by using domain information contained in related tasks as an inductive bias [15]. We hypothesize that this improvement in generalization explains why the Multi models have improved performance on minority stress detection compared to baselines.

To further support this reasoning, we experiment with reducing the size of the Dreaddit training set for the psychological stress detection task. This reduction in the training set simulates the data scarcity that is present for minority stress detection.

We perform the same experimental setup as described in Section V, but we use psychological stress data for our hyperparameter tuning and change the size of the training set to be either 10%, 25%, 50%, 75%, or 100% of the original training set. We perform these experiments with the Single-Task and Multi models paired with each of the four language models (BERT, RoBERTa, MentalBERT, and MentalRoBERTa).

We report our results in Figure 5. We find that while Single-Task and Multi achieve equivalent performance with the full training set, reduced training sets reduce the performance of Single-Task models much more significantly than Multi models. We see that at 100% the Single-Task models all have F1 scores near 80, but at 50% they drop to be between 77.38 and 78.13. By comparison, the Multi models have F1 scores of at least 80 with only 50% of the training data, but they do not significantly improve as the training size increases.

This finding demonstrates that the improved generalizability of the multi-task architecture of the models makes them more effective in low-data environments. It consequently explains why Multi models are more effective at minority stress detection: the training set has a limited amount of minority stress, so detecting it is a low-data environment.

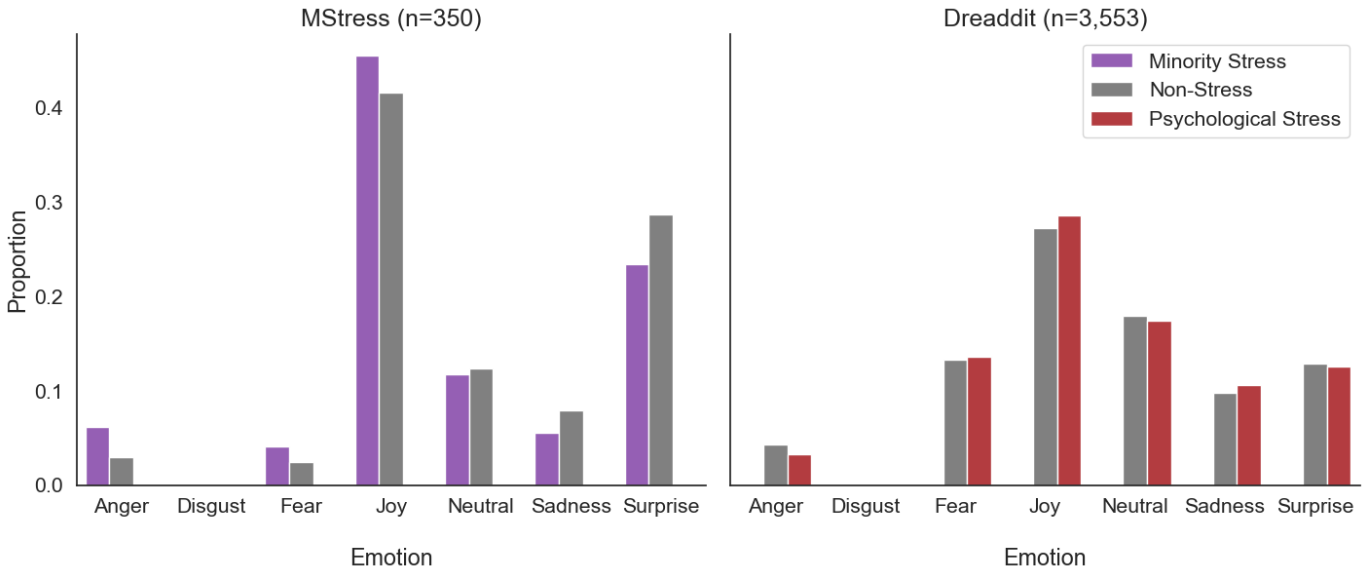

Figure 6. Distributions of predicted emotion labels in MStress and Dreaddit. Note that posts can have multiple emotion labels.

This finding also suggests that multi-task emotion-infused architectures may improve stress detection for other underrepresented groups, and further work should be done to explore using emotions to create equitable mental health models.

B. Emotion Distributions

To provide additional support for the importance of emotion analysis for supporting underrepresented groups, we examine the predicted emotion distributions of both MStress and Dreaddit (shown in Figure 6). These emotion labels were created using a single-task MentalRoBERTa model with a macro F1 of 61.13.

From these distributions, we see that emotions do not significantly vary based on stress status. Posts in MStress marked as minority stress have a similar emotion distribution as posts marked as non-stress, and posts in Dreaddit marked as psychological stress have a similar emotion distribution as posts marked as non-stress.

However, we see a significant difference in the emotion distributions of posts from MStress compared to Dreaddit. This suggests that posts from minority communities exhibit different emotions, and that difference may affect single-task models' ability to understand mental health conditions in minorities. Other work has found that minority stress mediates emotion regulation, leading to dysregulation and emotion suppression [30], which supports this finding.

This difference provides further explanation for the underperformance of single-task stress models on minorities. They are trained on a distribution of emotional expressions that are not representative of minority communities. This finding further suggests that multi-task, emotion-infused architectures may make more equitable models for other mental health tasks.

VIII. Conclusion

In this work, we find that traditional single-task stress models underperform on minority stress detection and are at risk of widening the healthcare gap for SGM individuals. We also find that risk can be reduced with the use of a multi-task architecture that integrates the task of emotion detection. Our experiments show that architecture performs well on psychological stress detection and outperforms the SOTA for minority stress detection without training on minority stress data. Finally, we provide explanatory analysis demonstrating the Multi model's superior performance in low-data environments, and we highlight how differences in emotion expressions in minority communities make them vulnerable to reduced effectiveness in mental health modeling. Our analysis suggests that integrating emotions may be effective for improving the performance of mental health models on underrepresented groups, and future work should explore using emotions to create equitable models for other mental health tasks.

IX. Limitations

Our data was collected from Reddit, which is not necessarily representative of the general population. While our work focuses on generalizing stress models to work effectively on minority stress, it does not evaluate their ability to generalize to other social media platforms or beyond social media.

Additionally, our models focus on detecting minority stress for sexual and gender minorities, but we did not explore the detection of minority stress for other underrepresented groups such as racial and ethnic minorities. Other minority groups may disclose minority stress differently, and future work should evaluate the performance of current minority stress models on other minority groups.

 ISBN: 978-1-68558-165-7

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Award # OIA-1946391, "Data Analytics that are Robust and Trusted (DART)".

REFERENCES

[1] E. Turcan and K. McKeown, "Dreaddit: A Reddit dataset for stress analysis in social media," in *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 97–107. DOI: 10.18653/v1/D19-6213.

[2] R. M. Díaz, G. Ayala, E. Bein, J. Henne, and B. V. Marin, "The impact of homophobia, poverty, and racism on the mental health of gay and bisexual Latino men: Findings from 3 US cities.," *American Journal of Public Health*, vol. 91, no. 6, pp. 927–932, Jun. 2001, ISSN: 0090-0036.

[3] G. Remafedi, S. French, M. Story, M. D. Resnick, and R. Blum, "The relationship between suicide risk and sexual orientation: Results of a population-based study.," *American Journal of Public Health*, vol. 88, no. 1, pp. 57–60, Jan. 1998, ISSN: 0090-0036.

[4] E. A. Tebbe and B. Moradi, "Suicide risk in trans populations: An application of minority stress theory," eng, *Journal of Counseling Psychology*, vol. 63, no. 5, pp. 520–533, Oct. 2016, ISSN: 0022-0167. DOI: 10.1037/cou0000152.

[5] E. Formby, *Exploring LGBT Spaces and Communities: Contrasting Identities, Belongings and Wellbeing*, en. Routledge, 2017, Google-Books-ID: GEccswEACAAJ, ISBN: 978-1-138-81400-4.

[6] S. Tropiano, "'A Safe and Supportive Environment': LGBTQ Youth and Social Media," en, in *Queer Youth and Media Cultures*, C. Pullen, Ed., London: Palgrave Macmillan UK, 2014, pp. 46–62, ISBN: 978-1-349-48056-2 978-1-137-38355-6. DOI: 10.1057/9781137383556_4.

[7] L. B. McInroy and S. L. Craig, ""It's like a safe haven fantasy world": Online fandom communities and the identity development activities of sexual and gender minority youth," *Psychology of Popular Media*, vol. 9, pp. 236–246, 2020, Place: US Publisher: Educational Publishing Foundation, ISSN: 2689-6575. DOI: 10.1037/ppm0000234.

[8] N. Woznicki, A. S. Arriaga, N. A. Caporale-Berkowitz, and M. C. Parent, "Parasocial relationships and depression among LGBQ emerging adults living with their parents during COVID-19: The potential for online support," *Psychology of Sexual Orientation and Gender Diversity*, vol. 8, pp. 228–237, 2021, Place: US Publisher: Educational Publishing Foundation, ISSN: 2329-0390. DOI: 10.1037/sgd0000458.

[9] K. Saha *et al.*, "The Language of LGBTQ+ Minority Stress Experiences on Social Media," *Proceedings of the ACM on human-computer interaction*, vol. 3, no. CSCW, p. 89, Nov. 2019, ISSN: 2573-0142. DOI: 10.1145/3361108.

[10] C. J. Cascalheira *et al.*, "Classifying Minority Stress Disclosure on Social Media with Bidirectional Long Short-Term Memory," en, *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, pp. 1373–1377, May 2022, ISSN: 2334-0770. DOI: 10.1609/icwsm.v16i1.19390.

[11] I. H. Meyer, "Prejudice, Social Stress, and Mental Health in Lesbian, Gay, and Bisexual Populations: Conceptual Issues and Research Evidence," *Psychological bulletin*, vol. 129, no. 5, pp. 674–697, Sep. 2003, ISSN: 0033-2909. DOI: 10.1037/0033-2909.129.5.674.

[12] P. P. Heppner, B. E. Wampold, J. Owen, K. T. Wang, and M. N. Thompson, *Research design in counseling*. Boston, MA: Cengage Learning, 2016, ISBN: 978-1-305-08731-6.

[13] A. Flentje, N. C. Heck, J. M. Brennan, and I. H. Meyer, "The relationship between minority stress and biological outcomes: A systematic review," eng, *Journal of Behavioral Medicine*, vol. 43, no. 5, pp. 673–694, Oct. 2020, ISSN: 1573-3521. DOI: 10.1007/s10865-019-00120-6.

[14] E. Turcan, S. Muresan, and K. McKeown, "Emotion-infused models for explainable psychological stress detection," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 2895–2909. DOI: 10.18653/v1/2021.naacl-main.230.

[15] R. Caruana, "Multitask Learning," en, *Machine Learning*, vol. 28, no. 1, pp. 41–75, Jul. 1997, ISSN: 1573-0565. DOI: 10.1023/A:1007379606734.

[16] X. Zuo, T. Li, and P. Fung, "A multilingual natural stress emotion database," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 1174–1178.

[17] A. P. Allen, P. J. Kennedy, J. F. Cryan, T. G. Dinan, and G. Clarke, "Biological and psychological markers of stress in humans: Focus on the Trier Social Stress Test," eng, *Neuroscience and Biobehavioral Reviews*, vol. 38, pp. 94–124, Jan. 2014, ISSN: 1873-7528. DOI: 10.1016/j.neubiorev.2013.11.005.

[18] F. Al-Shargie *et al.*, "Mental stress assessment using simultaneous measurement of EEG and fNIRS," *Biomedical Optics Express*, vol. 7, no. 10, pp. 3882–3898, Sep. 2016, ISSN: 2156-7085. DOI: 10.1364/BOE.7.003882.

[19] S. Kumar *et al.*, *StressNet: Detecting Stress in Thermal Videos*, arXiv:2011.09540 [cs, eess], Nov. 2020. DOI: 10.48550/arXiv.2011.09540.

[20] M. Jaiswal *et al.*, "MuSE: A multimodal dataset of stressed emotion," English, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, May 2020, pp. 1499–1510, ISBN: 979-10-95546-34-4.

[21] S. Ji *et al.*, "MentalBERT: Publicly available pretrained language models for mental healthcare," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, Jun. 2022, pp. 7184–7190.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[23] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

[24] Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," en, *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, Mar. 2010, ISSN: 0261-927X, 1552-6526. DOI: 10.1177/0261927X09351676.

[25] C. Manning *et al.*, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 55–60. DOI: 10.3115/v1/P14-5010.

[26] K. Saha and M. De Choudhury, "Modeling Stress with Social Media Around Incidents of Gun Violence on College Campuses," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, 92:1–92:27, Dec. 2017. DOI: 10.1145/3134727.

[27] D. Demszky *et al.*, "GoEmotions: A dataset of fine-grained emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 4040–4054. DOI: 10.18653/v1/2020.acl-main.372.

[28] P. Ekman, "Are there basic emotions?" eng, *Psychological Review*, vol. 99, no. 3, pp. 550–553, Jul. 1992, ISSN: 0033-295X. DOI: 10.1037/0033-295x.99.3.550.

[29] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. arXiv: 1412.6980 [cs.LG].

[30] A. Singh, A. Dandona, V. Sharma, and S. Z. H. Zaidi, "Minority stress in emotion suppression and mental distress among sexual and gender minorities: A systematic review," en, *Ann. Neurosci.*, vol. 30, no. 1, pp. 54–69, Jan. 2023.

 ISBN: 978-1-68558-165-7