Information Extraction from Social Media: A Hands-on Tutorial on Tasks, Data, and Open Source Tools

Shubhanshu Mishra
Twitter, Inc.
Chicago, Illinois, USA
mishra@shubhanshu.com

Rezvaneh Rezapour
Drexel University
Philadelphia, Pennsylvania, USA
shadi.rezapour@drexel.edu

Jana Diesner
University of Illinois at Urbana-Champaign
Champaign, Illinois, USA
jdiesner@illinois.edu

ABSTRACT

Information extraction (IE) is a common sub-area of natural language processing that focuses on identifying structured data from unstructured data. One application domain of IE is Information Retrieval (IR), which relies on accurate and high-performance IE to retrieve high quality results from massive datasets. Another example of IE is to identify named entities in a text. For example, in the the sentence "Katy Perry lives in the USA", *Katy Perry* and *USA* are named entities of types of PERSON and LOCATION, respectively. Also, identify the sentiment expressed in a text is another instance of IE: in the sentence, "This movie was awesome", the expressed sentiment is positive. Finally, IE is concerned with identifying various linguistic aspects of text data, e.g., part of speech of words, noun phrases, dependency parses, etc., which can serve as features for additional IE tasks. This tutorial introduces participants to a) the usage of Python based, open-source tools that support IE from social media data (mainly Twitter), and b) best practices for ensuring the responsible use of IE and research data. Participants will learn and practice various lexical, semantic, and syntactic IE techniques that are commonly used for analyzing tweets. Participants will also be familiarized with the landscape of publicly available social media data (including popular NLP and IE benchmarks) and methods for collecting and preparing them for analysis. Furthermore, participants will be trained to use a suite of open source tools (SAIL for active learning, TwitterNER for named entity recognition, TweetNLP for transformer based NLP, and SocialMediaIE for multi task learning), which utilize advanced machine learning techniques (e.g., deep learning, active learning with human-in-the-loop, multi-lingual, and multi-task learning) to perform IE on their own or existing datasets. Participants will also learn how social contexts of text production and usage of results can be integrated into IE systems to improve these systems and to consider the role of time in improving social media IE quality. Finally, participants will learn about the governance of social media data for research purposes. The tools introduced in the tutorial will focus on the three main stages of IE, namely, collection of data (including annotation), data processing and analytics, and visualization of the extracted information. More details can be found at: https://socialmediaie.github.io/tutorials/

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557503

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Multi-task learning**; • **Software and its engineering** → *Software libraries and repositories*; • **Human-centered computing** → Social media.

KEYWORDS

Social media, Twitter, Information extraction, Multitask learning, Deep learning, Machine learning, Named entity recognition, Part of speech tagging, Chunking, Supersense tagging, Open source tool, Text Classification, Open data, Natural Language Processing, Machine Learning Bias, Data governance

ACM Reference Format:
Shubhanshu Mishra, Rezvaneh Rezapour, and Jana Diesner. 2022. Information Extraction from Social Media: A Hands-on Tutorial on Tasks, Data, and Open Source Tools. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3511808.3557503

1 INTRODUCTION

1.1 Aims and Learning Objectives

In this hands-on tutorial (details and material at: https://socialmediaie.github.io/tutorials/), we introduce the participants to working with social media data, which are an example of Digital Social Trace Data (DSTD). The DSTD abstraction allows us to model social media data with rich information associated with social media text, such as authors, topics, and time stamps. We introduce the participants to several Python-based, open-source tools for performing Information Extraction (IE) on social media data. Furthermore, the participants will be familiarized with a catalogue of more than 30 publicly available social media corpora for various IE tasks such as named entity recognition (NER), part of speech (POS) tagging, chunking, super sense tagging, entity linking, sentiment classification, and hate speech identification. We will also show how these approaches can be expanded to word in a multi-lingual setting. Finally, the participants will be introduced to the following applications of extracted information: (i) combining network analysis and text-based signals to rank accounts, and (ii) correlation between sentiment and user-level attributes in existing corpora. The tutorial aims to serve the following use cases for social media researchers: (iii) high accuracy IE on social media text via multi-task and semi–supervised learning, including the recent transformer-based tools

which work across languages, (iv) rapid annotation of new data for text classification via active human-in-the-loop learning, (v) temporal visualization of the communication structure in social media corpora via social communication temporal graph visualization technique, (vi) detecting and prioritizing needs during crisis events (e.g, COVID19), and (vii) responsible collection and use of social media data. (viii) Furthermore, the participants will be familiarized with a catalogue of more than 30 publicly available social media corpora for various IE tasks, e.g., named entity recognition (NER), part of speech (POS) tagging, chunking, super sense tagging, entity linking, sentiment classification, and hate speech identification. We propose a full day tutorial session using Python based open-source tools. This tutorial builds on our past tutorials at ACM Hypertext 2019, IC2S2 2020, WWW 2021, ECIR 2022, LREC 2022.

Novelty over previous versions. The tutorial will feature demos of recently released tools for doing social media NLP like TweetNLP[1], discussion on temporal degradation of models trained on social media data as suggested in TimeLMs and Temporal NER, and an updated catalogue of social media datasets for information extraction covering platforms like Twitter, YouTube, Reddit, etc.

1.2 Scope and benefit to the CIKM Community

Information extraction (IE) is a common sub-area of natural language processing that focuses on identifying structured data from unstructured data. While many open source tools are available for performing IE on newswire and academic publication corpora, there is a lack of such tool for working with social media corpora. These data tend to exhibit different linguistic patterns than other genres of corpora. It has also been found that publicly available tools for IE, which are trained on news and academic corpora, might not perform well on social media corpora. Topics of interest include: (i) Machine learning for social media IE (ii) Generating annotated text classification data using active human-in-the-loop learning (iii) Public corpora for social media IE (iv) Open source tools for social media IE (v) Visualizing social media corpora (vi) Bias in social media IE systems (vii) Responsible computing with social media data Scholars in the Information Retrieval community who work with social media text can benefit from the recent machine learning advances in information extraction and retrieval in this domain, e.g., knowledge in how social media differs from newswire and literary data. This tutorial will help attendees to learn state-of-the-art methods for processing social media text and to improve information retrieval systems for social media data. They will learn presence and usage of social context in social media text.

1.3 Presenter Bios

***Shubhanshu Mishra**, Twitter, Inc.* Shubhanshu Mishra is a Senior Machine Learning Researcher at Twitter. He earned his Ph.D. in Information Sciences from the University of Illinois at Urbana-Champaign in 2020. His thesis was titled "Information extraction from digital social trace data: applications in social media and scholarly data analysis". His current work is at the intersection of machine learning, information extraction, social network analysis, and visualizations. His research has led to the development of open source tools of open source information extraction solutions from large scale social media and scholarly data.

***Rezvaneh (Shadi) Rezapour**, College of Computing and Informatics, Drexel University, USA.* Shadi is an Assistant Professor in the Department of Information Science at Drexel's College of Computing and Informatics. Her research interests lie at the intersection of Computational Social Science and Natural Language Processing (NLP). More specifically, she is interested in bringing computational models and social science theories together, to analyze texts and better understand and explain real-world behaviors, attitudes, and cultures. Her research goal is to develop "socially-aware" NLP models that bring social and cultural contexts in analyzing (human) language to better capture attributes, such as social identities, stances, morals, and power from language, and understand real-world communication. Shadi completed her Ph.D. in Information Sciences at University of Illinois at Urbana-Champaign (UIUC) where she was advised by Dr. Jana Diesner.

***Jana Diesner**, The iSchool at University of Illinois Urbana Champaign, USA.* Jana is an Associate Professor at the School of Information Sciences (the iSchool) at the University of Illinois at Urbana-Champaign, where she leads the Social Computing Lab. Her research in social computing and human-centered data science combines methods from natural language processing, social network analysis, and machine learning with theories from the social sciences to advance knowledge and discovery about interaction-based and information-based systems. Jana got her PhD (2012) in from the School of Computer Science at Carnegie Mellon University.

2 TUTORIAL DETAILS

(i) **Duration of the tutorial:** 6 hrs (full day) (ii) **Interaction Style:** Hands-on live coding session. (iii) **Target audience:** We expect the participants to have familiarity with Python programming and social media platforms like Twitter, Reddit, Facebook, etc.

2.1 Tutorial Outline

Setup and Introduction (1 hr). (i) Introducing the differences between social media data versus newswire and academic data, (ii) Digital Social Trace Data abstraction for social media data, (iii) Introduction to information extraction tasks for social media data, e.g., sequence tagging (named entity, part of speech tagging, chunking, and super-sense tagging), and text classification (sentiment prediction, sarcasm detection, and abusive content detection)

Applications of information extraction (1 hr). (i) Indexing social media corpora in database, (ii) Network construction from text corpora, (iii) Visualizing temporal trends in social media corpora using social communication temporal graphs, (iv) Aggregating text-based signals at the user-level, (v) Improving text classification using user-level attributes, (vi) Analyzing social debate using sentiment and political identity signals otherwise, (vii) Detecting and Prioritizing Needs during Crisis Events (e.g., COVID19), (viii) Mining and Analyzing Public Opinion Related to COVID-19, (ix) Detecting COVID-19 Misinformation in Videos on YouTube.

Collecting and distributing social media data (30 mins). (i) Overview on available annotated social media datasets (Twitter,

[1] https://tweetnlp.org/

Reddit, Youtube, etc.), (ii) Respecting API terms and user privacy considerations for collecting & sharing social media data, (iii) Demo on collecting data from social media APIs, e.g. Twitter and Reddit.

Break 30 mins.

Improving IE on social media data via Machine Learning (2 hr 30 mins). (i) Semi-supervised learning for Twitter NER, (ii) Multi-task learning for social media IE, (iii) Active learning for annotating social media data for text classification via SAIL, (iv) Pre-trained transformer models for Tweets via TweetNLP and HuggingFace Model Hub, (v) Finetuning monolingual and multi-lingual language models for social media NLP tasks. (vi) Biases in social media NER. (vii) Utilizing Social Context for improving NLP Models. (viii) Role of time in the quality of NLP Models.

Conclusion and future directions (10 mins). (i) Open questions in social media IE, (ii) Tutorial feedback and questions.

REFERENCES

[1] Aseel Addawood, Rezvaneh Rezapour, Shubhanshu Mishra, Jodi Schneider, and Jana Diesner. 2017. Developing an Information Source Lexicon. In *Prioritising Online Content workshop co-located at NIPS*.

[2] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration. *Epidemiologia* 2, 3 (2021), 315–324. https://doi.org/10.3390/epidemiologia2030024

[3] Francesco Barbieri, Luis Espinosa-Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of LREC*.

[4] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[5] Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* 15, 5 (2012), 662–679.

[6] Jose Camacho-Collados, Yerai Doval, Eugenio Martínez-Cámara, Luis Espinosa-Anke, Francesco Barbieri, and Steven Schockaert. 2020. Learning Cross-lingual Embeddings from Twitter via Distant Supervision. In *Proceedings of ICWSM* (Atlanta, United States).

[7] Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara, et al. 2022. TweetNLP: Cutting-Edge Natural Language Processing for Social Media. *arXiv preprint arXiv:2206.14774* (2022).

[8] Kathleen M Carley, Jana Diesner, Jeffrey Reminga, and Maksim Tsvetovat. 2004. An integrated approach to the collection and analysis of network data. In *IN PROC OF THE (NAACSOS) 2004 CONFERENCE*. Citeseer.

[9] Daniel Collier, Shubhanshu Mishra, Derek Houston, Brandon Hensley, Scott Mitchell, and Nicholas Hartlep. 2019. Who is Most Likely to Oppose Federal Tuition-Free College Policies? Investigating Variable Interactions of Sentiments to America's College Promise. *SSRN Electronic Journal* (2019). https://doi.org/10.2139/ssrn.3423054

[10] Daniel A. Collier, Shubhanshu Mishra, Derek A. Houston, Brandon O. Hensley, and Nicholas D. Hartlep. 2019. Americans 'support' the idea of tuition-free college: an exploration of sentiment and political identity signals otherwise. *Journal of Further and Higher Education* 43, 3 (mar 2019), 347–362. https://doi.org/10.1080/0309877X.2017.1361516

[11] Laura Dabbish, Ben Towne, Jana Diesner, and James Herbsleb. 2011. Construction of association networks from communication in teams working on complex projects. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 4, 5 (2011), 547–563. https://doi.org/10.1002/sam.10135

[12] Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 141–152.

[13] Jana Diesner. 2015. Small decisions with big impact on data analytics. *Big Data & Society* 2, 2 (2015). https://doi.org/10.1177/2053951715617185

[14] Jana Diesner. 2015. *Words and Networks: How Reliable Are Network Data Constructed from Text Data?* Springer International Publishing, Cham, 81–89. https://doi.org/10.1007/978-3-319-05467-4_5

[15] Jana Diesner and Kathleen M Carley. 2008. Conditional random fields for entity extraction and ontological text coding. *Computational and Mathematical Organization Theory* 14, 3 (2008), 248–262. https://doi.org/10.1007/s10588-008-9029-z

[16] Jana Diesner and Kathleen M Carley. 2008. *Looking Under the Hood of Stochastic Machine Learning Algorithms for Parts of Speech Tagging (CMU-ISR-08-131).* Technical Report. Pittsburgh, PA: Carnegie Mellon University, School of Computer Science, Institute for Software Research.

[17] Jana Diesner and Kathleen M Carley. 2009. He says, she says. Pat says, Tricia says. How much reference resolution matters for entity extraction, relation extraction, and social network analysis. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*. IEEE, Ottawa, Canada, 1–8.

[18] Jana Diesner and Kathleen M. Carley. 2010. Extraktion relationaler Daten aus Texten [Relation extraction from texts]. In *Handbuch Netzwerkforschung [Handbook network research]*, Christian Stegbauer and Roger Häußling (Eds.). VS Verlag für Sozialwissenschaften, 507–521. https://doi.org/10.1007/978-3-531-92575-2_44

[19] Jana Diesner and Kathleen M Carley. 2010. A methodology for integrating network theory and topic modeling and its application to innovation diffusion. In *2010 IEEE Second International Conference on Social Computing*. IEEE, Minneapolis, MN, 687–692. https://doi.org/10.1109/SocialCom.2010.106

[20] Jana Diesner and Kathleen M Carley. 2011. Words and Networks. In *Encyclopedia of social networks*, George A Barnett (Ed.). Sage Publications, 958–961.

[21] Jana Diesner and Chieh-Li Chin. 2015. Usable ethics: practical considerations for responsibly conducting research with social trace data. *Proceedings of Beyond IRBs: Ethical Review Processes for Big Data Research* (2015).

[22] Jana Diesner and Chieh-Li Chin. 2016. Gratis, libre, or something else? Regulations and misassumptions related to working with publicly available text data. In *ETHI-CA² Workshop (ETHics in Corpus Collection, Annotation & Application), 10th Language Resources and Evaluation Conference (LREC), Portoroz, Slovenia.*

[23] Jana Diesner and Chieh-Li Chin. 2016. Seeing the forest for the trees: Considering applicable types of regulation for the responsible collection and analysis of human centered data. In *Human-Centered Data Science (HCDS) Workshop at 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing.*

[24] Jana Diesner and Craig S Evans. 2015. Little bad concerns: Using sentiment analysis to assess structural balance in communication networks. In *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 342–348. https://doi.org/10.1145/2808797.2809403

[25] Jana Diesner, Ponnurangam Kumaraguru, and Kathleen M Carley. 2005. Mental models of data privacy and security extracted from interviews with indians. In *55th Annual Conference of the International Communication Association (ICA), New York, NY.*

[26] Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, 359–369.

[27] Ahmed El-Kishky, Thomas Markovich, Serim Park, Chetan Verma, Baekjin Kim, Ramy Eskander, Yury Malkov, Frank Portman, Sofía Samaniego, Ying Xiao, and Aria Haghighi. 2022. TwHIN: Embedding the Twitter Heterogeneous Information Network for Personalized Recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) *(KDD '22)*. Association for Computing Machinery, New York, NY, USA, 2842–2850. https://doi.org/10.1145/3534678.3539080

[28] Ramy Eskander, Peter Martigny, and Shubhanshu Mishra. 2020. Multilingual Named Entity Recognition in Tweets using Wikidata. In *The fourth annual WeCNLP (West Coast NLP) Summit (WeCNLP)* (virtual). Zenodo. https://doi.org/10.5281/zenodo.7014432

[29] Casey Fiesler, Nathan Beard, and Brian C Keegan. 2020. No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 187–196.

[30] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[31] Kanyao Han, Pingjing Yang, Shubhanshu Mishra, and Jana Diesner. 2020. WikiCSSH: Extracting Computer Science Subject Headings from Wikipedia. In *Workshop on Scientific Knowledge Graphs (SKG 2020).*

[32] Liam Peter Hebert, Raheleh Makki, Yuval Merhav, Hamidreza Saghir, and Shubhanshu Mishra. 2022. Robust Candidate Generation for Entity Linking on Short Social Media Texts. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (WNUT).*

[33] Martin Hilbert, George Barnett, Joshua Blumenstock, Noshir Contractor, Jana Diesner, Seth Frey, Sandra González-Bailón, PJ Lamberson, Jennifer Pan, Tai-Quan Peng, Cuihua (Cindy) Shen, Paul E. Smaldino, Wouter van Atteveldt, Annie Waldherr, Jingwen Zhang, and Jonathan J. H. Zhu. 2019. Computational Communication Science: A Methodological Catalyzer for a Maturing Discipline. *International Journal of Communication* 13, 0 (2019), 3912–3934.

[34] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *International AAAI Conference on Web and Social Media*. Ann Arbor, Michigan, USA.

[35] Andreas M. Kaplan and Michael Haenlein. 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons* 53, 1 (jan 2010), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

[36] Michal Kosinski, Sandra C. Matz, Samuel D. Gosling, Vesselin Popov, and David Stillwell. 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist* 70, 6 (sep 2015), 543–556. https://doi.org/10.1037/a0039210

[37] Michal Kosinski, Sandra C Matz, Samuel D Gosling, Vesselin Popov, and David Stillwell. 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American psychologist* 70, 6 (2015), 543–556.

[38] Vivek Kulkarni, Shubhanshu Mishra, and Aria Haghighi. 2021. LMSOC: An Approach for Socially Sensitive Pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, 2967–2975. https://doi.org/10.18653/v1/2021.findings-emnlp.254

[39] Jinning Li, Shubhanshu Mishra, Ahmed El-Kishki, Sneha Mehta, and Vivek Kulkarni. 2022. Enriching Social Media Text Representations with Non-Textual Units. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (WNUT)*.

[40] Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic Language Models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Dublin, Ireland, 251–260. https://doi.org/10.18653/v1/2022.acl-demo.25

[41] Shubhanshu Mishra. 2017. SCTG: Social Communications Temporal Graph – A novel approach to visualize temporal communication graphs from social data. In *UIUC Data Science Day*.

[42] Shubhanshu Mishra. 2019. *Multi-Dataset Multi-Task Learning Benchmark for Social Media Information Extraction*. https://doi.org/10.5281/zenodo.5867160

[43] Shubhanshu Mishra. 2019. Multi-dataset-multi-task Neural Sequence Tagging for Information Extraction from Tweets. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media - HT '19*. ACM Press, New York, New York, USA, 283–284. https://doi.org/10.1145/3342220.3344929

[44] Shubhanshu Mishra. 2020. Improving Social Media Information Extraction using Multitask Multidataset Learning. In *The fourth annual WeCNLP (West Coast NLP) Summit (WeCNLP)* (virtual). Zenodo. https://doi.org/10.5281/zenodo.7014470

[45] Shubhanshu Mishra. 2020. Information Extraction from Digital Social Trace Data with Applications to Social Media and Scholarly Communication Data. *ACM SIGIR Forum* 54, 1 (2020).

[46] Shubhanshu Mishra. 2020. *Information Extraction from Digital Social Trace Data with Applications to Social Media and Scholarly Communication Data*. Ph. D. Dissertation. University of Illinois at Urbana-Champaign.

[47] Shubhanshu Mishra. 2020. *Information extraction from digital social trace data with applications to social media and scholarly communication data*. Ph. D. Dissertation. University of Illinois at Urbana-Champaign. https://shubhanshu.com/phd_thesis/

[48] Shubhanshu Mishra. 2020. Non-neural Structured Prediction for Event Detection from News in Indian Languages. In *Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation*, P Mehta, T Mandl, P Majumder, and M Mitra (Eds.). CEUR Workshop Proceedings, CEUR-WS.org, Hyderabad, India.

[49] Shubhanshu Mishra, Sneha Agarwal, Jinlong Guo, Kirstin Phelps, Johna Picco, and Jana Diesner. 2014. Enthusiasm and support: alternative sentiment classification for social movements on social media. In *Proceedings of the 2014 ACM conference on Web science - WebSci '14*. ACM Press, Bloomington, Indiana, USA, 261–262. https://doi.org/10.1145/2615569.2615667

[50] Shubhanshu Mishra and Daniel Collier. 2020. A Framework for Generating Annotated Social Media Corpora with Demographics, Stance, Civility, and Topicality. *SSRN Electronic Journal* (2020). https://doi.org/10.2139/ssrn.3757554

[51] Shubhanshu Mishra and Jana Diesner. 2016. Semi-supervised Named Entity Recognition in noisy-text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*. The COLING 2016 Organizing Committee, Osaka, Japan.

[52] Shubhanshu Mishra and Jana Diesner. 2018. Detecting the Correlation between Sentiment and User-level as well as Text-Level Meta-data from Benchmark Corpora. In *Proceedings of the 29th on Hypertext and Social Media - HT '18*. ACM Press, New York, New York, USA, 2–10. https://doi.org/10.1145/3209542.3209562

[53] Shubhanshu Mishra and Jana Diesner. 2019. Capturing Signals of Enthusiasm and Support Towards Social Issues from Twitter. In *Proceedings of the 5th International Workshop on Social Media World Sensors - SIdEWayS'19*. ACM Press, New York, New York, USA, 19–24. https://doi.org/10.1145/3345645.3351104

[54] Shubhanshu Mishra, Jana Diesner, Jason Byrne, and Elizabeth Surbeck. 2015. Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*. ACM Press, New York, New York, USA, 323–325. https://doi.org/10.1145/2700171.2791022

[55] Shubhanshu Mishra and Aria Haghighi. 2021. Improved Multilingual Language Model Pretraining for Social Media Text via Translation Pair Prediction. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*. 381–388. https://doi.org/10.18653/v1/2021.wnut-1.42

[56] Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing Demographic Bias in Named Entity Recognition. In *Bias in Automatic Knowledge Graph Construction - A Workshop at AKBC 2020*. arXiv:2008.03415

[57] Shubhanshu Mishra and Sudhanshu Mishra. 2019. 3Idiots at HASOC 2019: Fine-tuning Transformer Neural Networks for Hate Speech Identification in Indo-European Languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*. Kolkata, India, 208–213.

[58] Sudhanshu Mishra, Shivangi Prasad, and Shubhanshu Mishra. 2020. Multilingual Joint Fine-tuning of Transformer models for identifying Trolling, Aggression and Cyberbullying at TRAC 2020. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. European Language Resources Association (ELRA), Marseille, France, 120—-125.

[59] Sudhanshu Mishra, Shivangi Prasad, and Shubhanshu Mishra. 2021. Exploring Multi-Task Multi-Lingual Learning of Transformer Models for Hate Speech and Offensive Speech Identification in Social Media. *SN Computer Science* 2, 2 (apr 2021), 72. https://doi.org/10.1007/s42979-021-00455-5

[60] Shubhanshu Mishra, Aman Saini, Raheleh Makki, Sneha Mehta, Aria Haghighi, and Ali Mollahosseini. 2022. *TweetNERD - End to End Entity Linking Benchmark for Tweets*. https://doi.org/10.5281/zenodo.6617192

[61] Shubhanshu Mishra, Aman Saini, Raheleh Makki, Sneha Mehta, Aria Haghighi, and Ali Mollahosseini. 2022. TweetNERD-End to End Entity Linking Benchmark for Tweets. (2022).

[62] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[63] Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska De Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics* 42, 3 (2016).

[64] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.

[65] Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2 (2008), 1–135. https://doi.org/10.1561/1500000011

[66] Rezvaneh Rezapour, Ly Dinh, and Jana Diesner. 2021. Incorporating the Measurement of Moral Foundations Theory into Analyzing Stances on Controversial Topics. In *Proceedings of the 32st ACM Conference on Hypertext and Social Media*. ACM, New York, NY, USA, 177–188. https://doi.org/10.1145/3465336.3475112

[67] Rezvaneh Rezapour, Saumil H. Shah, and Jana Diesner. 2019. Enhancing the Measurement of Social Effects by Capturing Morality. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Stroudsburg, PA, USA, 35–45. https://doi.org/10.18653/v1/W19-1305

[68] Rezvaneh Rezapour, Lufan Wang, Omid Abdar, and Jana Diesner. 2017. Identifying the Overlap between Election Result and Candidates' Ranking Based on Hashtag-Enhanced, Lexicon-Based Sentiment Analysis. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*. IEEE, 93–96. https://doi.org/10.1109/ICSC.2017.92

[69] Sunita Sarawagi. 2007. Information Extraction. *Foundations and Trends® in Databases* 1, 3 (mar 2007), 261–377. https://doi.org/10.1561/1900000003

[70] M Janina Sarol, Ly Dinh, and Jana Diesner. 2021. Variation in Situational Awareness Information due to Selection of Data Source, Summarization Method, and Method Implementation. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 597–608.

[71] M. Janina Sarol, Ly Dinh, Rezvaneh Rezapour, Chieh-Li Chin, Pingjing Yang, and Jana Diesner. 2020. An Empirical Methodology for Detecting and Prioritizing Needs during Crisis Events. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Stroudsburg, PA, USA, 4102–4107. https://doi.org/10.18653/v1/2020.findings-emnlp.366

[72] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E P Seligman, and Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE* 8, 9 (jan 2013), e73791. https://doi.org/10.1371/journal.pone.0073791

[73] Indira Sen, Fabian Flöck, Katrin Weller, Bernd Weiß, and Claudia Wagner. 2021. A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly* 85, S1 (2021), 399–422.

[74] Shawn A Weil, Pacey Foster, Jared Freeman, Kathleen Carley, Jana Diesner, Terrill Franz, Nancy J Cooke, Steve Shope, and Jamie C Gorman. 2017. Converging approaches to automated communications-based assessment of team situation awareness. In *Macrocognition in Teams*. CRC Press, 276–304.

[75] Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra. 2021. Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (oct 2021), 1–24. https://doi.org/10.1145/3479594

[76] Michael Zimmer. 2020. "But the data is already public": on the ethics of research in Facebook. In *The Ethics of Information Technologies*. Routledge, 229–241.