





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

Shubhanshu Mishra<sup>1</sup>, Rezvaneh Rezapour<sup>2</sup>, and Jana Diesner<sup>3</sup>

<sup>1</sup> Twitter, Inc., Chicago, USA  
mishra@shubhanshu.com

<sup>2</sup> Drexel's College of Computing and Informatics, Philadelphia, USA  
shadi.rezapour@drexel.edu

<sup>3</sup> University of Illinois at Urbana-Champaign, Champaign, USA  
jdiesner@illinois.edu  
<https://shubhanshu.com/>

**Abstract.** Information extraction (IE) is a common sub-area of natural language processing that focuses on identifying structured data from unstructured data. The community of Information Retrieval (IR) relies on accurate and high-performance IE to be able to retrieve high quality results from massive datasets. One example of IE is to identify named entities in a text, e.g., “Barack Obama served as the president of the USA”. Here, Barack Obama and USA are named entities of types of PERSON and LOCATION, respectively. Another example is to identify sentiment expressed in a text, e.g., “This movie was awesome”. Here, the sentiment expressed is positive. Finally, identifying various linguistic aspects of a text, e.g., part of speech tags, 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 reproducibility of research. Participants will learn and practice various semantic and syntactic IE techniques that are commonly used for analyzing tweets. Additionally, participants will be familiarized with the landscape of publicly available tweet data, and methods for collecting and preparing them for analysis. Finally, participants will be trained to use a suite of open source tools (SAIL for active learning, TwitterNER for named entity recognition<sup>3</sup>, 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 context can be integrated in Information Extraction systems to make them better. 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/>.

**Keywords:** Information extraction · Multi-task learning · Natural language processing · Social media data · Twitter · Machine learning bias

## 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, and (vi) detecting and prioritizing needs during crisis events (e.g., COVID19). (vii) 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 upon our previous tutorials on this topic at ACM Hypertext 2019, IC2S2 2020, WWW 2021.

### 1.2 Scope and Benefit to the ECIR 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 when dealing with social media corpora, which tends to exhibit very different linguistic patterns compared to the other corpora. It has also been found that publicly available tools for IE,

which are trained on news and academic corpora do not tend to perform very 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.

Scholars in 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, especially the knowledge of its difference from regular newswire text. This tutorial will help them learn state-of-the-art methods for processing social media text which can help them improve their information retrieval systems on social media text. They will also learn how social media text has a social context, which can be included as part of the analysis.

### 1.3 Presenter Bios

*Shubhanshu Mishra*, Twitter, Inc. Shubhanshu Mishra is a 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. He has finished his Integrated Bachelor’s and Master’s degree in Mathematics and Computing from the Indian Institute of Technology, Kharagpur in 2012.

*Rezvaneh (Shadi) Rezapour*, Department of Information Science at Drexel’s College of Computing and Informatics, 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 Societal Computing from the School of Computer Science at Carnegie Mellon University.

## 2 Tutorial Details

- **Duration of the tutorial:** 1 day (full day)
- **Interaction Style:** Hands-on-tutorial with live coding session.
- **Target audience:** We expect the participants to have familiarity with python programming and social media platforms like Twitter and Facebook.

**Setup and Introduction (1 h)** (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 h)** (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, and (ix) Detecting COVID-19 Misinformation in Videos on YouTube.

**Collecting and distributing social media data (30 min)** (i) Overview on available annotated tweet datasets, (ii) Respecting API terms and user privacy considerations for collecting & sharing social media data, (iii) Demo on collecting data from a few social media APIs, such as Twitter and Reddit.

**Break 30 min**

**Improving IE on social media data via Machine Learning (2 h 30 min)** (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](#) (another version pySAIL to be released soon), (iv) Finetuning transformer models for monolingual and multi-lingual social media NLP tasks. (v) Biases in social media NER. (vi) Utilizing Social Context for improving NLP Models.

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

## References

1. Addawood, A., Rezapour, R., Mishra, S., Schneider, J., Diesner, J.: Developing an information source lexicon. In: *Prioritising Online Content Workshop Co-located at NIPS (2017)*
2. Collier, D., Mishra, S., Houston, D., Hensley, B., Mitchell, S., Hartlep, N.: Who is most likely to oppose federal tuition-free college policies? Investigating variable interactions of sentiments to America's college promise. *SSRN Electron. J.* (2019). <https://doi.org/10.2139/ssrn.3423054>
3. Collier, D.A., Mishra, S., Houston, D.A., Hensley, B.O., Hartlep, N.D.: Americans 'support' the idea of tuition-free college: an exploration of sentiment and political identity signals otherwise. *J. Furth. High. Educ.* **43**(3), 347–362 (2019). <https://doi.org/10.1080/0309877X.2017.1361516>
4. Diesner, J., Carley, K.M.: Relation extraction from texts (in German: Extraktion relationaler Daten aus Texten). In: Stegbauer, C., Häußling, R. (eds.) *Handbook network research (Handbuch Netzwerkforschung)*, pp. 507–521. VS Verlag (2010)
5. Diesner, J., Kumaraguru, P., Carley, K.M.: Mental models of data privacy and security extracted from interviews with Indians. In: *Proceedings of 55th Annual Conference of International Communication Association (ICA)*. New York, NY (2005)
6. Diesner, J., Chin, C.L.: Usable ethics: practical considerations for responsibly conducting research with social trace data. In: *Proceedings of Beyond IRBs: Ethical Review Processes for Big Data Research (2015)*
7. Diesner, J., Chin, C.L.: 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 (2016)*
8. Eisenstein, J.: 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*, pp. 359–369. Association for Computational Linguistics, Atlanta, Georgia (June 2013)
9. Han, K., Yang, P., Mishra, S., Diesner, J.: WikiCSSH: extracting computer science subject headings from Wikipedia. In: *Workshop on Scientific Knowledge Graphs (SKG 2020)* (2020)
10. Hutto, C.J., Gilbert, E.: 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 (2014)
11. Kaplan, A.M., Haenlein, M.: Users of the world, unite! The challenges and opportunities of social media. *Bus. Horiz.* **53**(1), 59–68 (2010). <https://doi.org/10.1016/j.bushor.2009.09.003>
12. Kosinski, M., Matz, S.C., Gosling, S.D., Popov, V., Stillwell, D.: Facebook as a research tool for the social sciences: opportunities, challenges, ethical considerations, and practical guidelines. *Am. Psychol.* **70**(6), 543–556 (2015). <https://doi.org/10.1037/a0039210>
13. Kulkarni, V., Mishra, S., Haghghi, A.: LMSOC: an approach for socially sensitive pretraining. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2967–2975. Association for Computational Linguistics, Stroudsburg, PA, USA (November 2021). <https://doi.org/10.18653/v1/2021.findings-emnlp.254>

14. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web—WWW 2010, p. 591. ACM Press, New York, New York, USA (April 2010). <https://doi.org/10.1145/1772690.1772751>
15. Mishra, S.: SCTG: social communications temporal graph - a novel approach to visualize temporal communication graphs from social data. In: UIUC Data Science Day (October 2017)
16. Mishra, S.: 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 2019, pp. 283–284. ACM Press, New York, New York, USA (2019). <https://doi.org/10.1145/3342220.3344929>
17. Mishra, S.: Information extraction from digital social trace data with applications to social media and scholarly communication data. ACM SIGIR Forum **54**(1), 1–2 (2020). <https://doi.org/10.1145/3451964.3451981>
18. Mishra, S.: Information Extraction from Digital Social Trace Data with Applications to Social Media and Scholarly Communication Data. Ph.D. thesis, University of Illinois at Urbana-Champaign (2020)
19. Mishra, S.: Non-neural structured prediction for event detection from news in Indian languages. In: Mehta, P., Mandl, T., Majumder, P., Mitra, M. (eds.) Working Notes of FIRE 2020—Forum for Information Retrieval Evaluation. CEUR Workshop Proceedings, CEUR-WS.org, Hyderabad, India (2020)
20. Mishra, S., Agarwal, S., Guo, J., Phelps, K., Picco, J., Diesner, J.: Enthusiasm and support: alternative sentiment classification for social movements on social media. In: Proceedings of the 2014 ACM conference on Web science - WebSci 2014, pp. 261–262. ACM Press, Bloomington, Indiana, USA (June 2014). <https://doi.org/10.1145/2615569.2615667>
21. Mishra, S., Collier, D.: A framework for generating annotated social media corpora with demographics, stance, civility, and topicality. SSRN Electron. J. (2020). <https://doi.org/10.2139/ssrn.3757554>
22. Mishra, S., Diesner, J.: Semi-supervised named entity recognition in noisy-text. In: Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT), pp. 203–212. The COLING 2016 Organizing Committee, Osaka, Japan (2016)
23. Mishra, S., Diesner, J.: 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 2018, pp. 2–10. ACM Press, New York, New York, USA (2018). <https://doi.org/10.1145/3209542.3209562>
24. Mishra, S., Diesner, J.: Capturing signals of enthusiasm and support towards social issues from Twitter. In: Proceedings of the 5th International Workshop on Social Media World Sensors - SIdEWayS 2019, pp. 19–24. ACM Press, New York, New York, USA (2019). <https://doi.org/10.1145/3345645.3351104>
25. Mishra, S., Diesner, J., Byrne, J., Surbeck, E.: 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 2015, pp. 323–325. ACM Press, New York, New York, USA (2015). <https://doi.org/10.1145/2700171.2791022>
26. Mishra, S., Haghighi, A.: 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), pp. 381–388. Association for Computational Linguistics, Stroudsburg, PA, USA (November 2021). <https://doi.org/10.18653/v1/2021.wnut-1.42>

27. Mishra, S., He, S., Belli, L.: Assessing demographic bias in named entity recognition. In: *Bias in Automatic Knowledge Graph Construction—A Workshop at AKBC 2020* (August 2020)
28. Mishra, S., Mishra, S.: 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*, pp. 208–213. Kolkata, India (2019)
29. Mishra, S., Mishra, S.: Scubed at 3C task a—a simple baseline for citation context purpose classification. In: *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pp. 59–64. Association for Computational Linguistics, Wuhan, China (2020)
30. Mishra, S., Mishra, S.: Scubed at 3C task b—a simple baseline for citation context influence classification. In: *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pp. 65–70. Association for Computational Linguistics, Wuhan, China (2020)
31. Mishra, S., Prasad, S., Mishra, S.: 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*, pp. 120–125. European Language Resources Association (ELRA), Marseille, France (2020)
32. Mishra, S., Prasad, S., Mishra, S.: Exploring multi-task multi-lingual learning of transformer models for hate speech and offensive speech identification in social media. *SN Comput. Sci.* **2**(2), 1–19 (2021). <https://doi.org/10.1007/s42979-021-00455-5>
33. Mohammad, S.M., Kiritchenko, S., Zhu, X.: NRC-Canada: building the state-of-the-art in sentiment analysis of tweets. In: *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, vol. 2, pp. 321–327. Association for Computational Linguistics, Atlanta, Georgia, USA (2013)
34. Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Found. Trends® Inf. Retr.* **2**(1–2), 1–135 (2008). <https://doi.org/10.1561/1500000011>
35. Rezapour, R., Dinh, L., Diesner, J.: 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*, pp. 177–188. ACM, New York, NY, USA (August 2021). <https://doi.org/10.1145/3465336.3475112>
36. Rezapour, R., Shah, S.H., Diesner, J.: 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*, pp. 35–45. Association for Computational Linguistics, Stroudsburg, PA, USA (2019). <https://doi.org/10.18653/v1/W19-1305>
37. Rezapour, R., Wang, L., Abdar, O., Diesner, J.: 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)*, pp. 93–96. IEEE (2017). <https://doi.org/10.1109/ICSC.2017.92>
38. Sarawagi, S.: Information extraction. *Found. Trends® Databases* **1**(3), 261–377 (2007). <https://doi.org/10.1561/1900000003>
39. Sarol, M.J., Dinh, L., Rezapour, R., Chin, C.L., Yang, P., Diesner, J.: An empirical methodology for detecting and prioritizing needs during crisis events. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4102–4107. Association for Computational Linguistics, Stroudsburg, PA, USA (2020). <https://doi.org/10.18653/v1/2020.findings-emnlp.366>

40. Schwartz, H.A., et al.: Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS ONE* **8**(9), e73791 (2013). <https://doi.org/10.1371/journal.pone.0073791>
41. Yee, K., Tantipongpipat, U., Mishra, S.: Image cropping on twitter: fairness metrics, their limitations, and the importance of representation, design, and agency. *Proc. ACM Hum. Comput. Interact.* **5**(CSCW2), 1–24 (2021). <https://doi.org/10.1145/3479594>