Factoring in Context for the Automatic Detection of Misrepresentation

Bruna Paz Schmid[1,*], Annette Hautli-Janisz[2] and Steve Oswald[1]

[1]University of Fribourg, Avenue de l'Europe 20, 1700 Fribourg, Switzerland
[2]University of Passau, 94030 Passau, Germany

Abstract

The aim of the paper is to show how a solid theoretical pragmatic underpinning informs an automatic approach to identifying and classifying misrepresentation in social media. To that end we present a dataset that encodes misrepresentation as well as the source that is misrepresented, building on a set of pragmatically informed annotation guidelines. The performance of standard statistic classifiers for misrepresentation detection is promising. We also perform a fine-grained manual error analysis. The paper closes with a longitudinal analysis of misrepresentation in our dataset and shows that items labelled as misrepresentation increase in years that coincide with political campaigns.

Keywords

Misrepresentation, pragmatics, natural language processing

1. Introduction

A key feature of Trump's political campaign and one-term presidency, which started in early 2017 and ended four years later in 2021, was the strategical use of social media. In October and November 2023 alone, CNN has fact-checked twelve speeches, concluding that his "fall remarks were teeming with false claims - a staggering quantity of misrepresentations, exaggerations and outright lies that made sheer wrongness a central feature of each of his addresses" [1]. Through social media, misrepresentations occur faster and in a more targeted way than in traditional media outlets where political messages are usually assessed in terms of their factual content.

Identifying *misrepresentations, i.e., a metarepresentation that is not similar enough to the original representation at the inferential level given a certain context*, in a systematic manner is one building block for helping voters assess the political strategies and worldviews of potential future leaders. But tackling problematic content that is spread in connection to political campaigns is not simply an issue of quickly sifting through large quantities of data. What makes it especially challenging is the quality of the content. As the results of CNN's fact-checking indicate, "sheer wrongness" comes in various forms, one of which is misrepresentation – a notion that may be understood as a form of misinformation, that is, false information.

The aim of the paper is to show how a solid theoretical pragmatic underpinning informs an automatic approach to identifying and classifying misrepresentation in social media. To this end, we present a dataset that encodes misrepresentation as well as the source that is misrepresented, building on a set of pragmatically informed annotation guidelines. The performance of standard statistic classifiers for misrepresentation detection is promising. We also perform a fine-grained manual error analysis. The paper closes with a longitudinal analysis of misrepresentation in our dataset and shows that items labelled as misrepresentation increase in years that coincide with political campaigns.

Proceedings of the 1st Workshop on COuntering Disinformation with Artificial Intelligence (CODAI), co-located with the 27th European Conference on Artificial Intelligence (ECAI), pages 11–18, October 20, 2024, Santiago de Compostela, Spain
*Corresponding author.
bruna.pazschmid@unifr.ch (B. Paz Schmid); annette.hautli-janisz@uni-passau.de (A. Hautli-Janisz); steve.oswald@unifr.ch (S. Oswald)
https://www.fim.uni-passau.de/en/cornlp (A. Hautli-Janisz); https://www.steveoswald.ch (S. Oswald)
0009-0009-7734-0699 (B. Paz Schmid); 0000-0002-5901-9633 (A. Hautli-Janisz); 0000-0002-5946-1691 (S. Oswald)

The paper proceeds as follows: Section 2 discusses related work in automatically identifying misrepresentation. Section 3 presents the theoretical pragmatic underpinning, followed by a description of the annotation guidelines in Section 4. Section 5 includes information regarding data preprocessing steps, classification, model performance and error analysis. Section 6 discusses longitudinal aspects of the study, while section 7 discusses its results and Section 8 its limitations.

2. Related work

Research and governmental-driven efforts alike are trying to contain the effects of misrepresentation by finding ways to automatically identify information online that is misrepresenting the original or is outright false. For instance, the European Commission has sponsored projects aimed at developing AI-tools, such as Monitio, a media monitoring platform that includes fact-checking [2]. The fact-checking system is evidence-based in that it works by retrieving documents from an available collection of news articles which then serve as evidence for the predictions.

However, these approaches to fact-checking and misinformation tend not to differentiate between the various forms of false information. Instead, identification often occurs with the help of vocabularies and datasets labelled based on stance detection [3], truthfulness [4, 5, 6], topic-matching [7], or linguistic features associated with fake news [8, 9].

Misrepresentation itself has rarely been a topic of research in NLP. One exception is Michael Yeomans' study about partisan misrepresentation of political opponents through straw man arguments [10]. In an experimental setting, participants were tasked with articulating their own together with their opponents' positions. They were instructed to write down open-ended responses about the Affordable Care Act. Responses were then labelled depending on whether they were the participants' genuine or imitated positions. At the computational level, texts were scored with sentiment analysis and a lexicon of morally charged words. This was followed by a machine learning model – a logistic LASSO regression - trained to distinguish between texts from opponents and supporters. However, it is difficult to conclude that the study was about the straw man argument specifically since the theoretical underpinning remains unspecified in the paper. As such, the scope of the study appears to have been limited to the analysis of partisan incendiary language accompanying misrepresentation.

In the next section, we will elaborate on our theoretical underpinning. Our approach being theory-driven is what differentiates it from others: Every step is guided by pragmatic theory because, next to identifying misrepresentation, the aim of our study was also to understand the phenomenon from a linguistic and political perspective.

3. Pragmatic theory and misrepresentation

The core theoretical underpinning of the present paper is an observation from pragmatics, namely that people tend to show regularity in their language use due to the social aspects of communication [11, pp. 4-6], for instance if people intend to misrepresent or discredit the original. This means that certain patterns in language depend on the context they are embedded in. Therefore, by defining and describing the relevant contexts, we can link patterns of language use to certain pragmatic phenomena such as misrepresentation.

In this study, we build on theories from political discourse analysis [12, 13], pragmatics [14, 15] and philosophy of language [16] in defining misrepresentation as a metarepresentation that is not similar enough to the original representation at the inferential level given a certain context. Example (1) shows a tweet posted by Donald J. Trump on 27 July 2017 in which he claims that the New York Times (belonging to the left-wing spectrum) asserts that 'Fox and Friends' (right-wing spectrum) is the most powerful TV show in America.

(1) **Misrepresentation**
Wow, the Failing @nytimes said about @foxandfriends "....the most powerful T.V. show in America."
(ID: 890524387773997056)

The original text from an opinion piece in the NYT from 19 July 2017 is shown in (2) and includes some of the text that precedes and follows the quotation:

(2) **Original**
For years, it was a nontaxing mix of news, lifestyle and conservative couch gab, a warm-up before Fox's day of politics and commentary. Suddenly, for no other reason than its No. 1 fan, it is the most powerful TV show in America. (It's also easily the most-watched cable news morning show, averaging 1.6 million viewers in the year's second quarter, following a post-Trump ratings boost.) [...] [17]

Based on the theoretical pragmatic underpinning of this paper, a mispresentation M in the political context needs to meet the following criteria C:

C1: M is a metarepresentation in terms of intentionality.

C2: There are perceivable structural or componential differences between the original and its metarepresentation.

C3: There is noticeable change in the metarepresentation.

C4: The difference between the original representation and the metarepresentation results in a difference in comprehension.

C5: The difference in comprehension is politically relevant.

Based on *C1*, Example (1) is a metarepresentation, because it contains representative content discernible on a verbatim quotation, on the reported speech verb "said," and on the use of "[w]ow" to express a psychological state through the positive evaluation of the content of the verbatim quotation. The tweet is also significantly shorter than the overall article, satisfying *C2*.

Regarding *C3*, the quotation is isolated from the article and as a result, there is an emphasis on the content of the quotation which is evaluated positively with "[w]ow." Criteria *C4* is met for a variety of reasons: For one, the quotation was taken from an opinion article and therefore represents an individual's opinion (and not necessarily NYT's). Secondly, Fox & Friends being described as "the most powerful TV show in America" is surprising and concerning given the adverb "[s]uddenly", which marks unexpectedness. Removing the beginning of the sentence thus changes the overall sentiment in the tweet to a positive one that is absent in the original representation. Thirdly, the author of the original does not appear to believe that the show deserves its new status, since the latter is said to result from "no other reason than its No. 1 fan", which undermines the inherent quality of the show. Finally, from the first to the second sentence in the original, Fox & Friends develops from "conservative couch gab" to "the most powerful TV show in America", i.e., the author implies that the world is turned upside down as a result of the former president's relationship with Fox & Friends.

In terms of C5, a politically relevant difference emerges between the original and the misrepresentation: The New York Times' original article achieves relevance by being identified as a critical piece which deems the show's new importance to be undeserved and perhaps even dangerous, considering that it is the result of the influence of the president of the United States. From this point of view, it may be a warning against the manipulation of the media for political purposes given the media's role as a check on governmental power and democracy. This criticism is especially strong since it originates in the media itself. Strikingly, Trump's metarepresentational rendition achieves political relevance by concluding the opposite, with important contextual implications: Even the New York Times, which is biased and left-wing and is known to be critical of Trump, recognizes how important his favorite show, in which he often participates, is. Additionally, given Trump's relationship with the

show, presenting it in a positive light may be an attempt to promote himself as well. Thus, it would represent an instance of positive self-presentation. Whereas the original representation is likely an attempt to revise available assumptions and thus change the current state of the world where Fox & Friends is portrayed as undeservedly powerful, the metarepresentation is likely to function as an attempt to strengthen those same assumptions and thus to protect the current state of the world.

4. The dataset

The basis of the investigation is the Trump Twitter Archive, a database that contains most tweets posted from Trump's personal account, @realDonaldTrump, between 2009 and 2021. The site was launched in 2016 and includes 56,571 tweets [18]. The maximum character count of the tweets ranges between 140 and 280. In the following, we discuss the steps taken to prune the dataset in order to be able to model pragmatic theory and misrepresentation in a meaningful way. The dataset and classification code are available at https://github.com/runastef/auto-ident-trump-misrep.git.

Filtering Two filters are applied consecutively with the aim of increasing the likelihood that the resulting array contains misrepresentation. The first filter extracted tweets containing quotation marks, which usually signal the presence of representative content for instance in the form of reported speech. Therefore, the presence of quotation marks is more likely to be used to comment on an original representation. We exclude tweets predating Trump's presidential campaign announcement from the selection as well as retweets. The same was attempted for quoted replies by excluding tweets containing the handle @realDonaldTrump. The intention behind excluding retweets and quoted replies was to limit the conversational context of the tweets to the relevant original representations by reducing the amount of representative content. This reduces the contextual complexity of the tweets so as to strengthen the relationship between the utterances and the pragmatic phenomenon of misrepresentation. Eventually, the filter excluded the expressions 'Nobody', 'establishment', 'Washington', 'elite', and 'Congress'. The resulting pre-annotation dataset, which combines both selections, contained a total of 1,737 tweets.

Annotation study The annotation of the selected tweets was done by two annotators after instruction, one of them being a co-author of the paper. The annotation guidelines reflect the criteria for misrepresentation discussed in Section 3 in that they are the deciding factors when a tweet is judged as being a misrepresentation of an original. The decision is binary, i.e., 'misrepresentation' versus 'not-misrepresentation'. Inter-annotator agreement with Cohen's Kappa was 0.765 over the whole dataset, which signals substantive agreement. To increase the quality of the dataset only tweets that both annotators agreed on were included. The resulting dataset has 214 items, 107 are labelled as being an instance of misrepresentation and 107 not being considered misrepresentations.

5. Predicting misrepresentation

5.1. Preprocessing

In preparation for the application of the text classification algorithms, we normalize, remove noise and anticipate and prevent issues connected to expressions such as URLs. Stop words were removed. Tokenization was done with TweetTokenizer from the NLTK library, which takes into account the specific linguistic structures prevalent in social media.

The list of stop words is updated to reflect Trump's language use. Since Trump's language use lacks complexity, removing frequent words may result in the removal of a significant amount of meaning because Trump's vocabulary contains many elements that would normally count as stop words. As a result, removing such words could influence the classification in a negative way as important patterns linked to his language use might be lost. This might cause an issue with the Naïve Bayes classifier, for example, since it is a probabilistic model that bases its decisions on the frequency with which the

different tokens are present in a certain label during the training phase. Thus, to avoid losing potentially important information, most verbs and conjunctions are kept while pronouns were removed from the list of stop words.

5.2. Classification

In the next step, the TfidfVectorizer (Term-Frequency Inverse Document Frequency Vectorizer) from scikit-learn was used for vectorization, i.e., a bag of words representation of all tweets was created, containing the tf-idf values for each word across all tweets. Normally, this vectorizer does its own tokenization, i.e., a library-internal module splits the running text into tokens. For the purpose of this paper, we overwrite this module since tweets have a unique format that can be challenging for tokenization. Since scikit-learn still requires tokenization for internal reasons, we follow the method introduced in David Batista's blog and pass a dummy tokenizer and preprocessor that returns the same input without changing it in scikit-learn [19].

5.3. Results

The text classification algorithms employed in this study are: Naïve Bayes, Support Vector Classifier (with a linear kernel) and Random Forest, all imported from the scikit-learn library. All classifiers were left on their standard configurations for learning purposes. On average, all three classifiers performed at around 70% based on mean accuracy (Naïve Bayes: 0.71, SVC: 0.73, Random Forest: 0.68). The accuracy scores ranged between 71% and 72% for Naïve Bayes, between 72% and 74% for SVC, and between 67% and 69% for Random Forest based on a 95% confidence interval. The scores are promising considering that a larger sample size may well improve the performance of the classifiers given that the content of the training data is expected to be more balanced with a larger sample size. The results are promising even when compared with related work: Miranda et al.'s [2] evidence-based automated fact checking platform presents predictions maintained to be correct 58% of the time, and Pérez-Rosas et al.'s [9] fake news detector reportedly presents accuracies between 50% and 76% depending on the domain associated with the dataset that is used. The results also suggest that the Support Vector Machine classifier performed slightly better than the other two.

5.4. Error analysis

The performances of Naïve Bayes and SVC are slightly reduced in testing sets containing larger tweets. That is, the performance seems to decrease the higher the number of tokens in the testing set is. This could be due to an imbalance between the training set and the testing set. Random Forest was probably less affected by this because it relies on the decision of multiple classifiers that reach their individual decisions based on a significantly smaller sample size than SVC and Naïve Bayes during the same run.

Incorrect predictions often arise when the language in the tweet is uncharacteristically complex, for instance when a supporter's tweet is copied and posted through Trump's account and contains language that is more complex than Trump's typical writing style. Here is such an example:

(3) **Copied and posted tweet**
"@racheljoycowley: I'm done with Macy's. Apparently, they follow the trend of trying to force legislation on every American freedom. Done!" (ID: 616646192609558528)

In an arbitrary run, this tweet was incorrectly predicted as not-misrepresentation by all three classifiers.

6. Tracking misrepresentation over time

Having annotated the tweets to be either misrepresentation or not, counting them over the years on the x-axis allows for a longitudinal view of the tweeting behavior of Donald J. Trump.

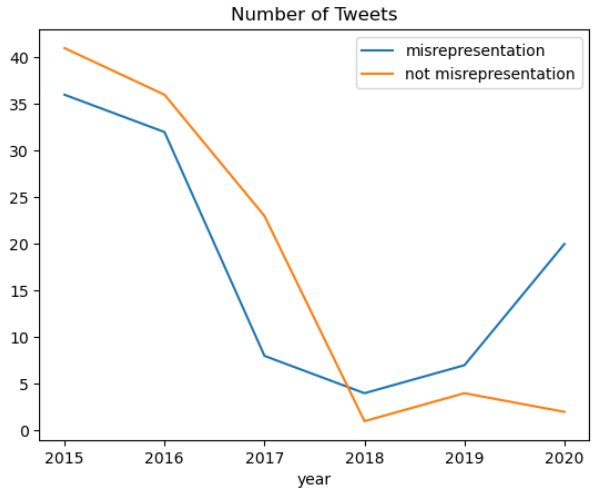

Figure 1: Number of Misrepresentations over time

Figure 1 hints at a possible correlation between the number of tweets containing misrepresentations that Trump posted and the years in which he was active in presidential election campaigns. The overall trend towards less tweets was already evident in the pre-annotation dataset. Thus, it is not surprising. What is interesting is the way in which the trend appears to change abruptly in 2018 and then again in 2019. Both categories experience an increase between 2018 and 2019. However, whereas the instances of not-misrepresentation begin to decrease from 2019 on, the instances of misrepresentation rise sharply.

It is also worth pointing out that given the connection to campaign years, the graph may be interpreted in terms of the reliability of the created dataset since misrepresentation is probably more likely to occur during political campaigns due to the nature of politics.

7. Discussion and summary

Future work could include a general analysis of Trump's language use, which might help to improve the preprocessing. This could take the form of a linguistic features engineering study to determine his writing style. The findings could then be used to create a stop words list that is better able to reflect Trump's language use although, perhaps, a more generalized approach based on country and political affiliation may be more helpful for the study of misrepresentation itself because it would be easier to generalize. It may also prove to be more practical in terms of implementation.

The process of evaluating the performances of the classifiers could be simplified and improved with an explainer. The LIME (Local Interpretable Model-Agnostic Explanations) Text Explainer was difficult to implement even after replacing the LinearSVC classifier with an SVC classifier with a linear kernel. Initially, the LinearSVC classifier was used, but it was changed into an SVC classifier with a linear kernel. It should produce similar results seeing as LinearSVC is an implementation of SVC. The change was necessary because LinearSVC does not support the function predict_proba, which calculates the probabilities for each class prediction. It would have been difficult to evaluate the performance of the LinearSVC classifier without predict_proba. LIME was chosen because, in theory, it should work well with all three classifiers as long as the models are able to "predict the probabilities of the categories." According to Albrecht et al., LIME "works locally by taking a look at each prediction separately. This is achieved by modifying the input vector to find the local components that the predictions are sensitive to" [20, p. 195]. Then, "[f]rom the behavior in the vicinity of the vector, it will draw conclusions about which components are more or less important" and "visualize the contributions and explain the decision mechanism of the algorithm for individual documents" [20, p. 195]. However, LIME's implementation and interpretation was challenging. As such, in the end, it was not taken into consideration in the error analysis. And yet, evaluating the performance of the classifiers would probably have been significantly

more straightforward with such an explainer. In its absence, the process is considerably slowed down. Improving explainability would have practical implications for future research, for instance, although we favor a theory-driven approach, together with the misrepresentation dataset the findings could be used to improve or expand available vocabularies employed in current fact-checking systems.

To summarize, this study contributes to the research of pragmatically relevant phenomena with computational linguistic methods by discussing how to account for various aspects of the context at different stages of the research process. Specifically, efforts were made to retain contextual information related to the social and political contexts. To this end, a framework was developed for the pragmatic analysis of political misrepresentation with computational methods. Based on the framework, annotation guidelines were written to enable the creation of a misrepresentation dataset, which was then employed to train supervised machine learning algorithms used for text classification. The three algorithms performed at around 70% with SVC performing slightly better than the other two algorithms.

8. Limitations

The findings of this study may be limited by the small size of the dataset, the data selection process, data source, and the format of the text data.

The study relied on tweets posted from Trump's Twitter account. Consequently, the methods applied in this study might yield different results on the discourses of other individuals especially if one considers Trump's unique language use and political affiliation. Widening the scope of the study to include political discourse from a larger number of politicians is likely to lead to better insight into political misrepresentation. To this end, our study will hopefully provide a basis for further research into a topic that has not received a lot of attention so far.

The relative novelty of pragmatic research into misrepresentation with computational linguistic methods also explains the chosen format. The smaller number of tokens present in tweets were expected to facilitate computation given the theoretical underpinning. Although the small format may limit the study's generalizability, it facilitates the qualitative analysis of the results which will help us to widen the scope of the analysis in future studies on this topic.

References

[1] D. Dale, Trump's avalanche of dishonesty: Fact-checking 102 of his false claims from this fall, 2023. URL: https://edition.cnn.com/2023/12/01/politics/trump-dishonesty-avalanche-102-fall-false-claims/index.html.

[2] S. Miranda, D. Nogueira, A. Mendes, A. Vlachos, A. Secker, R. Garrett, J. Mitchel, Z. Marinho, Automated fact checking in the news room, in: L. Liu, R. White (Eds.), The World Wide Web Conference, ACM, New York, NY, USA, 2019, pp. 3579–3583. doi:10.1145/3308558.3314135.

[3] W. Ferreira, A. Vlachos, Emergent: a novel data-set for stance classification, in: K. Knight, A. Nenkova, O. Rambow (Eds.), Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, CA, USA, 2016, pp. 1163–1168. doi:10.18653/v1/N16-1138.

[4] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2931–2937. doi:10.18653/v1/D17-1317.

[5] W. Y. Wang, "liar, liar pants on fire": A new benchmark dataset for fake news detection, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Short Papers, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 422–426. URL: https://aclanthology.org/P17-2067. doi:10.18653/v1/P17-2067.

[6] Y. Qiao, D. Wiechmann, E. Kerz, A language-based approach to fake news detection through interpretable features and brnn, in: A. Aker, A. Zubiaga (Eds.), Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM), Association for Computational Linguistics, Barcelona, Spain, 2020, pp. 14–31. URL: https://aclanthology.org/2020.rdsm-1.2.

[7] A. Miani, T. Hills, A. Bangerter, Loco: The 88-million-word language of conspiracy corpus, Behavior research methods 54 (2022) 1794–1817. doi:10.3758/s13428-021-01698-z.

[8] G. Kuzmin, Larionov, Daniil, D. Pisarevskaya, I. Smirnov, Fake news detection for the russian language, in: A. Aker, A. Zubiaga (Eds.), Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM), Association for Computational Linguistics, Barcelona, Spain, 2020, pp. 45–57. URL: https://aclanthology.org/2020.rdsm-1.5/.

[9] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in: E. M. Bender, L. Derczynski, P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, NM, USA, 2018, pp. 3391–3401. URL: https://aclanthology.org/C18-1287.

[10] M. Yeomans, The straw man effect: Partisan misrepresentation in natural language, Group Processes & Intergroup Relations 25 (2022) 1905–1924. doi:10.1177/13684302211014582.

[11] G. Yule, Pragmatics, Oxford introductions to language study, Oxford University Press, Oxford, UK, 2011.

[12] T. A. van Dijk, Ideology and discourse analysis, Journal of Political Ideologies 11 (2006) 115–140. doi:10.1080/13569310600687908.

[13] J. Wilson, Political discourse, in: D. Tannen, H. E. Hamilton, D. Schiffrin (Eds.), The Handbook of Discourse Analysis, Blackwell Handbooks in Linguistics, Wiley Blackwell, Malden and Oxford, 2015, pp. 775–794.

[14] E.-J. Noh, Metarepresentation: A Relevance-Theory Approach, volume 69, John Benjamins Publishing Company, Amsterdam, 2000.

[15] D. Wilson, D. Sperber, Relevance theory, in: L. R. Horn, G. Ward (Eds.), The Handbook of Pragmatics, Wiley, Malden, MA, 2006, pp. 607–632.

[16] J. R. Searle, Intentionality: An Essay in the Philosophy of Mind, Cambridge University Press, Cambridge, UK, 2012. doi:10.1017/CBO9781139173452.

[17] J. Poniewozik, Watching 'fox & friends,' trump sees a two-way mirror, 2017. URL: https://www.nytimes.com/2017/07/19/arts/television/donald-trump-fox-friends.html.

[18] B. Brown, Trump twitter archive, 2016. URL: https://www.thetrumparchive.com/.

[19] D. Batista, Applying scikit-learn tfidfvectorizer on tokenized text, 2018. URL: https://www.davidsbatista.net/blog/2018/02/28/TfidfVectorizer/.

[20] J. Albrecht, S. Ramachandran, C. Winkler, Blueprints for Text Analytics Using Python: Machine Learning-Based Solutions for Common Real World (NLP) Applications, O'Reilly Media, Inc., Sebastopol, CA, 2021.