OPEN ACCESS

International Journal Of Modern Engineering Research (IJMER)

Natural Language Processing and Artificial Intelligence: A Survey

Riyazuddin Khan[1], Sudhir Kumar Das[2], Rosalin Samantasinghar[3], Deepak Kumar Biswal[4]

[1.]Department of Computer Science and Engineering, Einstein Academy of Technology & Management, Bhubaneswar

[2.]Department of Computer Science and Engineering, Einstein Academy of Technology & Management, Bhubaneswar

[3.]Department of Computer Science and Engineering, Einstein Academy of Technology & Management, Bhubaneswar

[4.]Department of Computer Science and Engineering, Einstein Academy of Technology & Management, Bhubaneswar

Abstract

Modern models have significantly improved in quality in recent years; however, this has resulted in models losing some of their interpretability. An overview of Explainable AI (XAI) as it is currently understood within the field of Natural Language Processing (NLP) is provided by this survey. We discuss how explanations are categorized, how they can be arrived at, and how they might be visualized. As a service to the community of model developers, we describe in detail the operations and explainability strategies that are currently available for producing explanations for NLP model predictions. In conclusion, we highlight the existing deficiencies and propose future paths for this significant research field.

***Keywords:** - Local, Global, XAI, NLP*

I. Introduction

Traditionally, Natural Language Processing (NLP) systems have been mostly based on techniques that are inherently explainable. Examples of such ap- proaches, often referred to as *white box* techniques, include rules, decision trees, hidden Markov mod- els, logistic regressions, and others. Recent years, though, have brought the advent and popularity of *black box* techniques, such as deep learning mod- els and the use of language embeddings as features. While these methods in many cases substantially advance model quality, they come at the expense of models becoming less interpretable. This ob- fuscation of the process by which a model arrives at its results can be problematic, as it may erode trust in the many AI systems humans interact with daily (e.g., chatbots, recommendation systems, in- formation retrieval algorithms, and many others). In the broader AI community, this growing under- standing of the importance of explainability has cre- ated an emerging field called Explainable AI (XAI). However, just as tasks in different fields are more amenable to particular approaches, explainability must also be considered within the context of each discipline. We therefore focus this survey on XAI works in the domain of NLP, as represented in the main NLP conferences in the last seven years. This is, to the best of our knowledge, the first XAI sur- vey focusing on the NLP domain.

As will become clear in this survey, explainabil- ity is in itself a term that requires an explanation. While explainability may generally serve many purposes (see, e.g., Lertvittayakumjorn and Toni, 2019), our focus is on explainability from the per- spective of an end user whose goal is to understand how a model arrives at its result, also referred to as the *outcome explanation problem* (Guidotti et al., 2018). In this regard, explanations can help users of NLP-based AI systems build trust in these sys- tems' predictions. Additionally, understanding the model's operation may also allow users to provide useful feedback, which in turn can help developers improve model quality (Adadi and Berrada, 2018).

Explanations of model predictions have previ- ously been categorized in a fairly simple way that differentiates between (1) whether the explanation is for each prediction individually or the model's prediction process as a whole, and (2) determin- ing whether generating the explanation requires post-processing or not (see Section 3). However, although rarely studied, there are many additional characterizations of explanations, the most impor- tant being the techniques used to either generate or visualize explanations. In this survey, we ana- lyze the NLP literature with respect to both these dimensions and identify the most commonly used *explainability and visualization techniques*, in ad- dition to *operations* used to generate explanations (Sections

4.1-Section 4.3). We briefly describe each technique and point to representative papers adopting it. Finally, we discuss the common *evalu- ation techniques* used to measure the quality of ex- planations (Section 5), and conclude with a discus- sion of gaps and challenges in developing successful explainability approaches in the NLP domain (Section 6).

Related Surveys: Earlier surveys on XAI in- clude Adadi and Berrada (2018) and Guidotti et al. (2018). While Adadi and Berrada provide a com- prehensive review of basic terminology and fun- damental concepts relevant to XAI in general, our goal is to survey more recent works in NLP in an effort to understand how these achieve XAI and how well they achieve it. Guidotti et al. adopt a four-dimensional classification scheme to rate var- ious approaches. Crucially, they differentiate be- tween the "explanator" and the black-box model it explains. This makes most sense when a surrogate model is used to explain a black-box model. As we shall subsequently see, such a distinction applies less well to the majority of NLP works published in the past few years where the same neural network (NN) can be used not only to make predictions but also to derive explanations. In a series of tutorials, Lecue et al. (2020) discuss fairness and trust in ma- chine learning (ML) that are clearly related to XAI but not the focus of this survey. Finally, we adapt some nomenclature from Arya et al. (2019) which presents a software toolkit that can help users lend explainability to their models and ML pipelines.

Our goal for this survey is to: (1) provide the reader with a better understanding of the state of XAI in NLP, (2) point developers interested in building explainable NLP models to currently avail- able techniques, and (3) bring to the attention of the research community the gaps that exist; mainly a lack of formal definitions and evaluation for ex- plainability. We have also built an interactive web- site providing interested readers with all relevant aspects for every paper covered in this survey. [1]

II. Methodology

We identified relevant papers (see Appendix A) and classified them based on the aspects defined in Sections 3 and 4. To ensure a consistent classification, each paper was individually analyzed by at least two reviewers, consulting additional reviewers in the case of disagreement. For simplicity of presentation, we label each paper with its main applicable category for each aspect, though some papers may span multiple categories (usually with varying degrees of emphasis.) All relevant aspects for every paper covered in this survey can be found at the aforementioned website; to enable readers of this survey to discover interesting explainability techniques and ideas, even if they have not been fully developed in the respective publications.

III. Categorization of Explanations

Explanations are often categorized along two main aspects (Guidotti et al., 2018; Adadi and Berrada, 2018). The first distinguishes whether the expla- nation is for an individual prediction (*local*) or the model's prediction process as a whole (*global*). The second differentiates between the explanation emerging directly from the prediction process (*self- explaining*) versus requiring post-processing (*post- hoc*). We next describe both of these aspects in de- tail, and provide a summary of the four categories they induce in Table 1.

3.1 Local vs Global

A *local* explanation provides information or justifi- cation for the model's prediction on a specific in- put; 46 of the 50 papers fall into this category.

A *global* explanation provides similar justifica- tion by revealing how the model's predictive pro- cess works, independently of any particular input. This category holds the remaining 4 papers cov- ered by this survey. This low number is not surprising given the focus of this survey being on explana- tions that justify predictions, as opposed to explanations that help understand a model's behavior in general (which lie outside the scope of this survey).

3.2 Self-Explaining vs post-hoc

Regardless of whether the explanation is local or global, explanations differ on whether they arise as part of the prediction process, or whether their generation requires post-processing following the model making a prediction. A *self-explaining* approach, which may also be referred to as directly interpretable (Arya et al., 2019), generates the ex- planation at the same time as the prediction, us- ing information emitted by the model as a result of the process of making that prediction. Decision trees and rule-based models are examples of global self-explaining models, while feature saliency ap- proaches such as attention are examples of local self-explaining models.

In contrast, a post-hoc approach requires that an additional operation is performed after the pre- dictions are made. LIME (Ribeiro et al., 2016) is an example of producing a local explanation us- ing a surrogate model applied following the predic- tor's operation. A paper might also be considered to span both categories – for example, (Sydorova et al., 2019) actually presents both self-explaining and post-hoc explanation techniques.

Local Post-Hoc	Explain a single prediction by performing additional operations (*after* the model has emitted a prediction)
Local Self-Explaining	Explain a single prediction using the model itself (calculated from information made available from the model *as part of* making the prediction)
Global Post-Hoc	Perform additional operations to explain the entire model's predictive reasoning
Global Self-Explaining	Use the predictive model itself to explain the entire model's predictive reasoning (*a.k.a.* directly interpretable model)

Table 1: Overview of the high-level categories of expla- nations (Section 3).

IV. Aspects of Explanations

While the previous categorization serves as a con- venient high-level classification of explanations, it does not cover other important characteristics. We now introduce two additional aspects of explana- tions: (1) techniques for deriving the explanation and (2) presentation to the end user. We discuss the most commonly used explainability techniques, along with basic operations that enable explainabil- ity, as well as the visualization techniques com- monly used to present the output of associated ex- plainability techniques. We identify the most com- mon combinations of explainability techniques, op- erations, and visualization techniques for each of the four high-level categories of explanations pre- sented above, and summarize them, together with representative papers, in Table 2.

Although explainability techniques and visual- izations are often intermixed, there are fundamental differences between them that motivated us to treat them separately. Concretely, explanation derivation - typically done by AI scientists and engineers - fo- cuses on mathematically motivated justifications of models' output, leveraging various explainabil- ity techniques to produce "raw explanations" (such as attention scores). On the other hand, explana- tion presentation - ideally done by UX engineers - focuses on how these "raw explanations" are best presented to the end users using suitable visualiza- tion techniques (such as saliency heatmaps).

4.1 Explainability Techniques

In the papers surveyed, we identified five major explainability techniques that differ in the mecha- nisms they adopt to generate the raw mathematical justifications that lead to the final explanation pre- sented to the end users.

Feature importance. The main idea is to derive explanation by investigating the importance scores of different features used to output the final pre- diction. Such approaches can be built on differ- ent types of features, such as manual features ob- tained from feature engineering (e.g., Voskarides et al., 2015), lexical features including word/tokens and n-gram (e.g., Godin et al., 2018; Mullenbach et al., 2018), or latent features learned by NNs (e.g., Xie et al., 2017). Attention mechanism (Bahdanau et al., 2015) and first-derivative saliency (Li et al., 2015) are two widely used operations to enable feature importance-based explanations. Text-based features are inherently more interpretable by hu- mans than general features, which may explain the widespread use of attention-based approaches in the NLP domain.

Surrogate model. Model predictions are ex- plained by learning a second, usually more explain- able model, as a proxy. One well-known example is LIME (Ribeiro et al., 2016), which learns sur- rogate models using an operation called input per- turbation. Surrogate model-based approaches are model-agnostic and can be used to achieve either local (e.g., Alvarez-Melis and Jaakkola, 2017) or global (e.g., Liu et al., 2018) explanations. How- ever, the learned surrogate models and the original models may have completely different mechanisms to make predictions, leading to concerns about the fidelity of surrogate model-based approaches.

Example-driven. Such approaches explain the prediction of an input instance by identifying and presenting other instances, usually from available labeled data, that are semantically similar to the input instance. They are similar in spirit to nearest neighbor-based approaches (Dudani, 1976), and have been applied to different NLP tasks such as text classification (Croce et al., 2019) and question answering (Abujabal et al., 2017).

Provenance-based. Explanations are provided by illustrating some or all of the prediction deriva- tion process, which is an intuitive and effective ex- plainability technique when the final prediction is the result of a series of reasoning steps. We observe several question answering papers adopt such approaches (Abujabal et al., 2017; Zhou et al., 2018; Amini et al., 2019).

Declarative induction. Human-readable repre- sentations, such as rules (Pro¨ llochs et al., 2019), trees (Voskarides et al., 2015), and programs (Ling et al., 2017) are induced as explanations.

As shown in Table 2, feature importance-based and surrogate model-based approaches have been in frequent use (accounting for 29 and 8, respec- tively, of the 50 papers reviewed). This should not come as a surprise, as features serve as building blocks for machine learning models (explaining the proliferation of feature importance-based ap- proaches) and most recent NLP papers employ NN- based models, which are generally black-box mod- els (explaining the popularity of surrogate model- based approaches). Finally note that a complex NLP approach consisting of different

Component say employ more than one of these explainabil- ity techniques. A representative example is the QA system QUINT (Abujabal et al., 2017), which dis- plays the query template that best matches the user input query (example-driven) as well as the instan- tiated knowledge-base entities (provenance).

4.2 Operations to Enable Explainability

We now present the most common set of operations encountered in our literature review that are used to enable explainability, in conjunction with relevant work employing each one.

First-derivative saliency. Gradient-based ex- planations estimate the contribution of input *i* to- wards output *o* by computing the partial derivative of *o* with respect to *i*. This is closely related to older concepts such as sensitivity (Saltelli et al., 2008). First-derivative saliency is particularly convenient for NN-based models because these can be computed for any layer using a single call to auto-differentiation, which most deep learning en- gines provide out-of-the-box. Recent work has also proposed improvements to first-derivative saliency (Sundararajan et al., 2017). As suggested by its name and definition, first-derivative saliency can be used to enable feature importance explainability, es- pecially on word/token-level features (Aubakirova and Bansal, 2016; Karlekar et al., 2018).

Layer-wise relevance propagation. This is an- other way to attribute relevance to features com- puted in any intermediate layer of an NN. Defini- tions are available for most common NN layers in- cluding fully connected layers, convolution layers and recurrent layers. Layer-wise relevance propa- gation has been used to, for example, enable feature importance explainability (Poerner et al., 2018) and example-driven explainability (Croce et al., 2018).

Input perturbations. Pioneered by LIME (Ribeiro et al., 2016), input perturbations can ex- plain the output for input **x** by generating ran- dom perturbations of **x** and training an explainable model (usually a linear model). They are mainly used to enable surrogate models (e.g., Ribeiro et al., 2016; Alvarez-Melis and Jaakkola, 2017).

Attention (Bahdanau et al., 2015; Vaswani et al., 2017). Less an operation and more of a strategy to enable the NN to explain predictions, attention lay- ers can be added to most NN architectures and, be- cause they appeal to human intuition, can help indi- cate where the NN model is “focusing”. While pre- vious work has widely used attention layers (Luo et al., 2018; Xie et al., 2017; Mullenbach et al., 2018) to enable feature importance explainability, the jury is still out as to how much explainability at- tention provides (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019).

LSTM gating signals. Given the sequential na- ture of language, recurrent layers, in particular LSTMs (Hochreiter and Schmidhuber, 1997), are commonplace. While it is common to mine the out- puts of LSTM cells to explain outputs, there may also be information present in the outputs of the gates produced within the cells. It is possible to uti- lize (and even combine) other operations presented here to interpret gating signals to aid feature impor- tance explainability (Ghaeini et al., 2018).

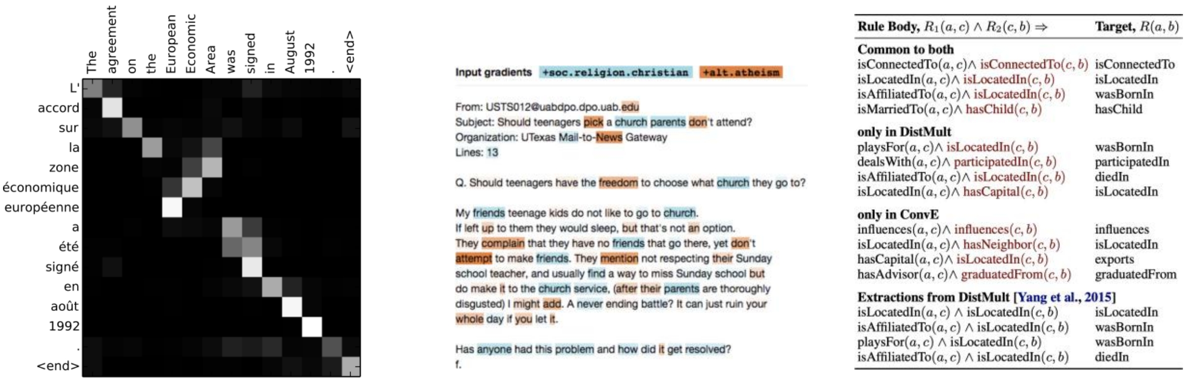
Explainability-aware architecture design. One way to exploit the flexibility of deep learning is to devise an NN architecture that mimics the process humans employ to arrive at a solution. This makes the learned model (partially) interpretable since the architecture contains human-recognizable compo- nents. Implementing such a model architecture can be used to enable the induction of human-readable programs for solving math problems (Amini et al., 2019; Ling et al., 2017) or sentence simplification problems (Dong et al., 2019). This design may also be applied to surrogate models that generate expla- nations for predictions (Rajani et al., 2019a; Liu et al., 2019).

Previous works have also attempted to compare these operations in terms of efficacy with respect to specific NLP tasks (Poerner et al., 2018). Oper- ations outside of this list exist and are popular for particular categories of explanations. Table 2 men- tions some of these. For instance, Pro¨ llochs et al. (2019) use reinforcement learning to learn simple negation rules, Liu et al. (2018) learns a taxonomy post-hoc to better interpret network embeddings, and Pryzant et al. (2018b) uses gradient reversal (Ganin et al., 2016) to deconfound lexicons.

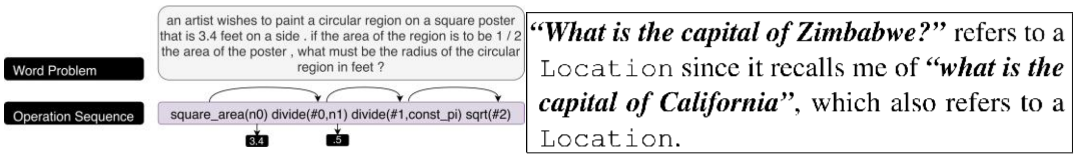
4.3 Visualization Techniques

An explanation may be presented in different ways to the end user, and making the appropriate choice is crucial for the overall success of an XAI ap- proach. For example, the widely used attention mechanism, which learns the importance scores of a set of features, can be visualized as raw at- tention scores or as a saliency heatmap (see Fig- ure 1a). Although the former is acceptable, the lat- ter is more user-friendly and has become the stan- dard way to visualize attention-based approaches. We now present the major visualization techniques identified in our literature review.

Saliency. This has been primarily used to visu- alize the importance scores of different types of elements in XAI learning systems, such as show- ing input-output word alignment (Bahdanau et al., 2015) (Figure 1a), highlighting words in input text (Mullenbach et al., 2018) (Figure 1b) or displaying extracted relations (Xie et al., 2017). We observe a strong correspondence between feature importance- based explainability and saliency-based visualiza- tions; namely, all papers using feature importance to generate explanations also chose saliency-based visualization techniques. Saliency-based visualiza- tions are popular because they present visually per- ceptive explanations and can be easily understood by different types of end users. They are there-



(a)Saliency heatmap (Bahdanau et al., 2015) (b)Saliency highlighting (Mullenbach et al., 2018) (c)Raw declarative rules (Pezeshkpour et al., 2019b)



(d)Raw declarative program (Amini et al., 2019) (e) Raw examples (Croce et al., 2019)

Figure 1: Examples of different visualization techniques

fore frequently seen across different AI domains (e.g., computer vision (Simonyan et al., 2013) and speech (Aldeneh and Provost, 2017)). As shown in Table 2, saliency is the most dominant visualization technique among the papers covered by this survey.

Raw declarative representations. As suggested by its name, this visualization technique directly presents the learned declarative representations, such as logic rules, trees, and programs (Figure 1c and 1d). Such techniques assume that end users can understand specific representations, such as first- order logic rules (Pezeshkpour et al., 2019a) and reasoning trees (Liang et al., 2016), and therefore may implicitly target more advanced users.

Natural language explanation. The explanation is verbalized in human-comprehensible natural lan- guage (Figure 2). The natural language can be generated using sophisticated deep learning mod- els, e.g., by training a language model with human natural language explanations and coupling with a deep generative model (Rajani et al., 2019a). It can also be generated by using simple template- based approaches (Abujabal et al., 2017). In fact, many declarative induction-based techniques can use template-based natural language generation (Reiter and Dale, 1997) to turn rules and programs into human-comprehensible language, and this mi- nor extension can potentially make the explanation more accessible to lay users.

Table 2 references some additional visualiza-

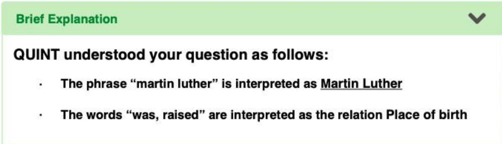


Figure 2: Template-based natural language explanation for a QA system (Abujabal et al., 2017).

tion techniques, such as using *raw examples* to present example-driven approaches (Jiang et al., 2019; Croce et al., 2019) (e.g., Figure 1e), and de- pendency parse trees to represent input questions (Abujabal et al., 2017).

V. CONCLUSION

Finally, it is interesting to note that we found only four papers that fall into the global explana- tions category. This might seem surprising given that white box models, which have been fundamen- tal in NLP, are explainable in the global sense. We believe this stems from the fact that because white box models are clearly explainable, the focus of the explicit XAI field is in explaining black box models, which comprise mostly local explanations. White box models, like rule based models and de- cision trees, while still in use, are less frequently framed as explainable or interpretable, and are hence not the main thrust of where the field is going. We think that this may be an oversight of the field since white box models can be a great test bed for studying techniques for evaluating explanations.

References

[1]. Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2017. Quint: Inter- pretable question answering over knowledge bases. In Proceedings of the 2017 Conference on Empiri- cal Methods in Natural Language Processing: Sys- tem Demonstrations, pages 61–66.

A. Adadi and M. Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelli- gence (xai). IEEE Access, 6:52138–52160.

[2]. Zakaria Aldeneh and Emily Mower Provost. 2017. Us- ing regional saliency for speech emotion recognition. In 2017 IEEE International Conference on Acous- tics, Speech and Signal Processing (ICASSP), pages 2741–2745. IEEE.

[3]. David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In Pro- ceedings of the 2017 Conference on Empirical Meth- ods in Natural Language Processing, pages 412–421, Copenhagen, Denmark. Association for Com- putational Linguistics.

[4]. Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Ha- jishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. 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), pages 2357–2367, Minneapolis, Minnesota. Associ- ation for Computational Linguistics.

[5]. Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Alek- sandra Mojsilovic, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John T. Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yi Zhang. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. ArXiv, abs/1909.03012.

[6]. M. Aubakirova and M. Bansal. 2016. Interpreting neu- ral networks to improve politeness comprehension. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (Austin, Texas, 2016), page 2035–2041.

[7]. AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU- MOSEI dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol- ume 1: Long Papers), pages 2236–2246, Melbourne, Australia. Association for Computational Linguis- tics.

[8]. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Ben- gio. 2015. Neural machine translation by jointly learning to align and translate. In ICLR.

[9]. Francesco Barbieri, Luis Espinosa-Anke, Jose Camacho-Collados, Steven Schockaert, and Hora- cio Saggion. 2018. Interpretable emoji prediction via label-wise attention LSTMs. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4766–4771, Brussels, Belgium. Association for Computational Linguistics.

[10]. Dimitris Bertsimas, Arthur Delarue, Patrick Jaillet, and Se´bastien Martin. 2019. The price of interpretability. ArXiv, abs/1907.03419.

[11]. Nikita Bhutani, Kun Qian, Yunyao Li, H. V. Jagadish, Mauricio Hernandez, and Mitesh Vasa. 2018. Ex- ploiting structure in representation of named entities using active learning. In Proceedings of the 27th In- ternational Conference on Computational Linguis- tics, pages 687–699, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[12]. Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2018. Extractive adversarial networks: High-recall expla- nations for identifying personal attacks in social me- dia posts. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Process-

ing, pages 3497–3507, Brussels, Belgium. Associa- tion for Computational Linguistics.

[13]. Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics, 8(8):832. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[14]. Danilo Croce, Daniele Rossini, and Roberto Basili. 2018. Explaining non-linear classifier decisions within kernel-based deep architectures. In Proceed- ings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 16–24, Brussels, Belgium. Association for Computational Linguistics.

[15]. Danilo Croce, Daniele Rossini, and Roberto Basili. 2019. Auditing deep learning processes through kernel-based explanatory models. In Proceedings of the 2019 Conference on Empirical Methods in Nat- ural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4037–4046, Hong Kong,

[16]. China. Association for Computational Linguistics.

[17]. Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplifi- cation through explicit editing. In Proceedings of the 57th Annual Meeting of the Association for Com- putational Linguistics, pages 3393–3402, Florence, Italy. Association for Computational Linguistics.

[18]. Sahibsingh A Dudani. 1976. The distance-weighted k-nearest-neighbor rule. IEEE Transactions on Sys- tems, Man, and Cybernetics, (4):325–327.

[19]. Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

[20]. Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Franc¸ois Lavi- olette, Mario Marchand, , and Victor Lempitsky. 2016. Domain-adversarial training of neural net- works. JMLR.

[21]. Nicolas Garneau, Jean-Samuel Leboeuf, and Luc Lam- ontagne. 2018. Predicting and interpreting embed- dings for out of vocabulary words in downstream tasks. In Proceedings of the 2018 EMNLP Work- shop BlackboxNLP: Analyzing and Interpreting Neu- ral Networks for NLP, pages 331–333, Brussels, Bel- gium. Association for Computational Linguistics.

[22]. Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language infer- ence. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4952–4957, Brussels, Belgium. Association for Computational Linguistics.

[23]. Fre´deric Godin, Kris Demuynck, Joni Dambre, Wesley De Neve, and Thomas Demeester. 2018. Explaining character-aware neural networks for word-level pre- diction: Do they discover linguistic rules? In Pro- ceedings of the 2018 Conference on Empirical Meth- ods in Natural Language Processing, pages 3275– 3284, Brussels, Belgium. Association for Computa- tional Linguistics.

[24]. Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. ACM Comput. Surv., 51(5).

[25]. Pankaj Gupta and Hinrich Schu¨tze. 2018. LISA: Ex- plaining recurrent neural network judgments via layer-wIse semantic accumulation and example to pattern transformation. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and In- terpreting Neural Networks for NLP, pages 154–164, Brussels, Belgium. Association for Computational Linguistics.

[26]. Joseph Y. Halpern. 2016. Actual Causality. MIT Press.

[27]. David Harbecke, Robert Schwarzenberg, and Christoph Alt. 2018. Learning explanations from language data. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpret- ing Neural Networks for NLP, pages 316–318, Brus- sels, Belgium. Association for Computational Lin- guistics.

[28]. Sepp Hochreiter and Ju¨rgen Schmidhuber. 1997. Long short-term memory. Neural Computation.

[29]. Shiou Tian Hsu, Changsung Moon, Paul Jones, and Na- giza Samatova. 2018. An interpretable generative adversarial approach to classification of latent entity relations in unstructured sentences. In AAAI Confer- ence on Artificial Intelligence.

[30]. Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In Proceedings of the 2019 Con- ference of the North American Chapter of the Asso- ciation for Computational Linguistics: Human Lan- guage Technologies, Volume 1 (Long and Short Pa- pers), pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

[31]. Yichen Jiang and Mohit Bansal. 2019. Self-assembling modular networks for interpretable multi-hop rea- soning. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natu- ral Language Processing (EMNLP-IJCNLP), pages 4474–4484, Hong Kong, China. Association for Computational Linguistics.

[32]. Yichen Jiang, Nitish Joshi, Yen-Chun Chen, and Mohit Bansal. 2019. Explore, propose, and assemble: An interpretable model for multi-hop reading compre- hension. In Proceedings of the 57th Annual Meet- ing of the Association for Computational Linguis- tics, pages 2714–2725, Florence, Italy. Association for Computational Linguistics.

[33]. Dongyeop Kang, Varun Gangal, Ang Lu, Zheng Chen, and Eduard Hovy. 2017. Detecting and explaining causes from text for a time series event. In Pro- ceedings of the 2017 Conference on Empirical Meth- ods in Natural Language Processing, pages 2758–2767, Copenhagen, Denmark. Association for Com- putational Linguistics.

[34]. Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018. Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models. In Proceed- ings of the 2018 Conference of the North American Chapter of the Association for Computational Lin- guistics: Human Language Technologies, Volume 2 (Short Papers) (New Orleans, Louisiana, Jun. 2018), page 701–707.

[35]. Shun Kiyono, Sho Takase, Jun Suzuki, Naoaki Okazaki, Kentaro Inui, and Masaaki Nagata. 2018. Unsupervised token-wise alignment to improve in- terpretation of encoder-decoder models. In Proceed- ings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 74–81, Brussels, Belgium. Association for Computational Linguistics.

[36]. Freddy Lecue, Krishna Gade, Sahin Cem Geyik, Krish- naram Kenthapadi, Varun Mithal, Ankur Taly, Ric- cardo Guidotti, and Pasquale Minervini. 2020. Ex- plainable ai: Foundations, industrial applications, practical challenges, and lessons learned. In AAAI

[37]. Conference on Artificial Intelligence. Association for Computational Linguistics.

[38]. Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded evaluations of explanation meth- ods for text classification. In Proceedings of the 2019 Conference on Empirical Methods in Natu- ral Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5195–5205, Hong Kong,

[39]. China. Association for Computational Linguistics.

[40]. Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2015. Visualizing and understanding neural models in nlp. arXiv preprint arXiv:1506.01066.

[41]. Qiuchi Li, Benyou Wang, and Massimo Melucci. 2019. CNM: An interpretable complex-valued network for matching. 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), pages 4139–4148, Minneapolis, Minnesota. Associ- ation for Computational Linguistics.

[42]. Chao-Chun Liang, Shih-Hong Tsai, Ting-Yun Chang, Yi-Chung Lin, and Keh-Yih Su. 2016. A meaning- based English math word problem solver with under- standing, reasoning and explanation. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstra- tions, pages 151–155, Osaka, Japan. The COLING 2016 Organizing Committee.

[43]. Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blun- som. 2017. Program induction by rationale genera- tion: Learning to solve and explain algebraic word problems. In Proceedings of the 55th Annual Meet- ing of the Association for Computational Linguistics (Volume 1: Long Papers), pages 158–167, Vancou- ver, Canada. Association for Computational Linguis- tics.

[44]. Hui Liu, Qingyu Yin, and William Yang Wang. 2019. Towards explainable NLP: A generative explanation framework for text classification. In Proceedings of the 57th Annual Meeting of the Association for Com- putational Linguistics, pages 5570–5581, Florence, Italy. Association for Computational Linguistics.

[45]. Ninghao Liu, Xiao Huang, Jundong Li, and Xia Hu. 2018. On interpretation of network embedding via taxonomy induction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowl- edge Discovery & Data Mining, KDD ’18, page 1812–1820, New York, NY, USA. Association for Computing Machinery.

[46]. Junyu Lu, Chenbin Zhang, Zeying Xie, Guang Ling, Tom Chao Zhou, and Zenglin Xu. 2019. Construct- ing interpretive spatio-temporal features for multi- turn responses selection. In Proceedings of the 57th Annual Meeting of the Association for Computa- tional Linguistics, pages 44–50, Florence, Italy. As- sociation for Computational Linguistics.

[47]. Ling Luo, Xiang Ao, Feiyang Pan, Jin Wang, Tong Zhao, Ningzi Yu, and Qing He. 2018. Beyond polar- ity: Interpretable financial sentiment analysis with hierarchical query-driven attention.

[48]. Seungwhan Moon, Pararth Shah, Anuj Kumar, and Ra- jen Subba. 2019. OpenDialKG: Explainable conver- sational reasoning with attention-based walks over knowledge graphs. In Proceedings of the 57th An- nual Meeting of the Association for Computational Linguistics, pages 845–854, Florence, Italy. Associ- ation for Computational Linguistics.

[49]. James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable pre- diction of medical codes from clinical text. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computa- tional Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

[50]. An Nguyen, Aditya Kharosekar, Matthew Lease, and Byron Wallace. 2018. An interpretable joint graph- ical model for fact-checking from crowds. In AAAI Conference on Artificial Intelligence.

[51]. Alexander Panchenko, Fide Marten, Eugen Ruppert, Stefano Faralli, Dmitry Ustalov, Simone Paolo Ponzetto, and Chris Biemann. 2017. Unsupervised, knowledge-free, and interpretable word sense dis- ambiguation. In Proceedings of the 2017 Confer- ence on Empirical Methods in Natural Language Processing: System Demonstrations, pages 91–96, Copenhagen, Denmark. Association for Computa- tional Linguistics.

[52]. Nikolaos Pappas and Andrei Popescu-Belis. 2014. Ex- plaining the stars: Weighted multiple-instance learn- ing for aspect-based sentiment analysis. In Proceed- ings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 455–466, Doha, Qatar. Association for Computa- tional Linguistics.

[53]. Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019a. Investigating robustness and interpretability of link prediction via adversarial modifications. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Compu- tational Linguistics: Human Language Technolo- gies, Volume 1 (Long and Short Papers), pages 3336–3347, Minneapolis, Minnesota. Association for Computational Linguistics.

[54]. Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019b. Investigating robustness and interpretability of link prediction via adversarial modifications. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computa- tional Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3336– 3347.

[55]. Nina Poerner, Hinrich Schu¨tze, and Benjamin Roth. 2018. Evaluating neural network explanation meth- ods using hybrid documents and morphosyntactic agreement. In Proceedings of the 56th Annual Meet- ing of the Association for Computational Linguis- tics (Volume 1: Long Papers), pages 340–350, Mel- bourne, Australia. Association for Computational Linguistics.

[56]. Nicolas Pro¨llochs, Stefan Feuerriegel, and Dirk Neu- mann. 2019. Learning interpretable negation rules via weak supervision at document level: A reinforce- ment learning approach. 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), pages 407–413, Minneapolis, Minnesota. Association for Computational Linguistics.

[57]. Reid Pryzant, Sugato Basu, and Kazoo Sone. 2018a. Interpretable neural architectures for attributing an ad’s performance to its writing style. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: An- alyzing and Interpreting Neural Networks for NLP, pages 125–135, Brussels, Belgium. Association for Computational Linguistics.

[58]. Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. 2018b. Deconfounded lexicon induction for interpretable social science. In Proceedings of the 2018 Conference of the North American Chap- ter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Pa- pers), pages 1615–1625, New Orleans, Louisiana. Association for Computational Linguistics.

[59]. Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019a. Explain your- self! leveraging language models for commonsense reasoning. In Proceedings of the 57th Annual Meet- ing of the Association for Computational Linguis- tics, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

[60]. Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019b. Explain your- self! leveraging language models for commonsense reasoning. arXiv preprint arXiv:1906.02361.

[61]. Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. Natural Lan- guage Engineering, 3(1):57–87.

[62]. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?”: Explain- ing the predictions of any classifier. In Proceed- ings of the 22Nd ACM SIGKDD International Con- ference on Knowledge Discovery and Data Mining (New York, NY, USA, 2016), page 1135–1144.

[63]. Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelli- gence,

IJCAI-17, pages 2662–2670.

[64]. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Taran- tola. 2008. Global Sensitivity Analysis: The Primer. John Wiley & Sons.

[65]. Robert Schwarzenberg, David Harbecke, Vivien Mack- etanz, Eleftherios Avramidis, and Sebastian Mo¨ller. 2019. Train, sort, explain: Learning to diagnose translation models. In Proceedings of the 2019 Con- ference of the North American Chapter of the Asso- ciation for Computational Linguistics (Demonstra- tions), pages 29–34, Minneapolis, Minnesota. Asso- ciation for Computational Linguistics.

[66]. Prithviraj Sen, Yunyao Li, Eser Kandogan, Yiwei Yang, and Walter Lasecki. 2019. HEIDL: Learning linguis- tic expressions with deep learning and human-in-the- loop. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 135–140, Florence, Italy. Association for Computational Linguistics.

[67]. Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In Proceedings of the 57th Annual Meeting of the Association for Computational Lin- guistics, pages 2931–2951, Florence, Italy. Associa- tion for Computational Linguistics.

[68]. Karen Simonyan, Andrea Vedaldi, and Andrew Zisser- man. 2013. Deep inside convolutional networks: Vi- sualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

[69]. Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In Inter- national Conference on Machine Learning, Sydney, Australia.

[70]. Alona Sydorova, Nina Poerner, and Benjamin Roth. 2019. Interpretable question answering on knowl- edge bases and text. In Proceedings of the 57th An- nual Meeting of the Association for Computational Linguistics, pages 4943–4951, Florence, Italy. Association for Computational Linguistics.

[71]. James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Gener- ating token-level explanations for natural language inference. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech- nologies, Volume 1 (Long and Short Papers), pages 963–969, Minneapolis, Minnesota. Association for Computational Linguistics.

[72]. Martin Tutek and Jan Šnajder. 2018. Iterative recur- sive attention model for interpretable sequence clas- sification. In Proceedings of the 2018 EMNLP Work- shop BlackboxNLP: Analyzing and Interpreting Neu- ral Networks for NLP, pages 249–257, Brussels, Belgium. Association for Computational Linguistics.

[73]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In NeurIPS.