L'interpretabilità dei modelli di machine learning come parametro di qualità: L'esperienza Datalake Giustizia

Roberto Bizzoni[1], Maria Borriello[1], Andrea Favalli[1], Cristina Giannone[1,*], David Preti[1], Raniero Romagnoli[1] and Federico Wolenski[1]

[1]Almawave S.p.A, via Casal Boccone 188/190, Roma, 00137, Italy

Abstract

I dati sono essenziali per l'apprendimento e il riconoscimento del linguaggio naturale da parte delle macchine, ma devono essere di alta qualità per garantire l'addestramento di modelli affidabili. Tuttavia, nonostante una cura nella qualità dei dati, i modelli addestrati, a causa della loro complessità, possono avere dei comportamenti non attesi o addirittura dannosi. L'articolo riporta l'esperienza del progetto Datalake Giustizia, concentrandosi sulla gestione della qualità nei processi di apprendimento automatico e sull'importanza di rendere i modelli appresi non opachi tramite tecniche di explainability come l'algoritmo LIME.

Keywords

Natural Language Processing, Explanable AI, Qualità dei dati, LIME, Information Extraction, Datalake Giustizia

1. Introduzione

La cura dei dati e delle informazioni passa attraverso l'analisi delle qualità che tali dati devono possedere per garantire agli utenti una corretta fruizione degli stessi e una efficace "data governance". In ambito Machine Learning (ML) e Natural Language Processing (NLP), i dati svolgono un ruolo centrale: Per dar vita al processo di apprendimento e al susseguente riconoscimento del linguaggio naturale da parte delle macchine è necessario disporre, oltre che degli algoritmi, di dati di alta qualità.

I dati costituiscono gli esempi da cui la macchina apprende lo svolgimento di un task (come, ad esempio, l'estrazione di informazioni, la classificazione di testi e immagini, la traduzione, ecc) e, una volta appreso, può essere riprodotto su nuovi dati tramite inferenza del modello.

I risultanti modelli addestrati vengono spesso impiegati in processi decisionali in ambiti sensibili quali, ad esempio, quello sanitario o giuridico, permettendo di automatizzare processi molto onerosi come l'analisi di grandi moli di documenti e abilitando un accesso alle informazioni estratte.

L'interpretabilità costituisce un parametro fondamentale per definire la qualità dei processi decisionali, che siano completamente automatizzati da modelli di IA o affianchino l'utente nella decisione. Tuttavia, i modelli di Machine Learning e quelli di Deep Learning da questo punto di vista sono definiti delle "black box" in quanto non è possibile decomporre l'output di un modello nei fattori costituenti che sono alla base delle sue decisioni finali.

Questo articolo descrive l'esperienza nella gestione della qualità nei processi di apprendimento automatico nel progetto DataLake Giustizia[1], focalizzandosi in particolare sulla necessità di rendere i modelli appresi non *opachi*, ovvero poter ricostruire le decisioni prese dal modello tramite tecniche algoritimiche che vanno sotto il nome di explainability[1].

L'articolo è strutturato nelle seguenti sezioni: La sezione 2 descrive come l'interpretabilità dei modelli di ML sia un aspetto di qualità fondamentale, in particolar modo nel contesto Giustizia, nella sezione 3 si introduce l'explanable AI, l'area di ricerca che si occupa di definire algoritmi per l'interpretazione dei modelli di Machine Learning, e infine la sezione 4 descrive l'algoritmo LIME[2] l'algoritmo di explainability utilizzato nel progetto Datalake.

2. Intelligenza Artificiale in ambito giuridico

I sistemi di intelligenza artificiale, ormai impiegati nei più disparati settori, si prestano ad essere utilizzati anche come strumenti di ausilio per gli operatori giuridici. Strumenti sempre più sofisticati, migliorano la ricerca di informazioni [3, 4], automatizzando la redazione o il controllo di atti e documenti [5], fino ad abilitare valutazioni automatiche di natura tecnica (giustizia predittiva)[6].

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.

r.bizzoni@almawave.it (R. Bizzoni); m.borriello@almawave.it (M. Borriello); a.favalli@almawave.it (A. Favalli); c.giannone@almawave.it (C. Giannone); d.preti@almawave.it (D. Preti); r.romagnoli@almawave.it (R. Romagnoli); f.wolenski@almawave.it (F. Wolenski)

[1]https://www.giustizia.it /cmsresources/cms/documents /studio_dgsia_ricognizione_digitalizzazione_febbraio2021.pdf

L'adozione di strumenti del genere può fornire grandi vantaggi, contribuendo a migliorare l'efficienza e la qualità della giustizia. Allo stesso tempo la loro adozione, se non controllata e affiancata a strumenti di analisi e verifica dei risultati, potrebbe causare diverse problematiche.

Il progetto DataLake Giustizia si pone l'obiettivo di individuare le modalità più idonee per ottenere il massimo risultato nello sfruttamento del patrimonio informativo di cui il Ministero di Giustizia dispone, e della relativa conoscenza estraibile dai dati che in esso confluiscono, curandone la qualità e mettendola al servizio per il miglioramento e l'efficientamento delle procedure legate alle indagini del processo penale e del processo civile.

La grande quantità di informazioni a disposizione del Datalake pone in primo piano l'esigenza di controllarne e valutarne l'effettiva qualità, con conseguente variazione dell'approccio operativo con l'introduzione, nel processo di estrazione e aggregazione delle informazioni, di algoritmi di intelligenza artificiale basati sul Natural Language Processing.

Nel sistema informativo di Giustizia e, quindi, nel Datalake, una quantità rilevante di informazioni proviene da dati non strutturati (generalmente testi) o parzialmente strutturati, di origine interna o esterna all'Amministrazione e conservata in differenti formati e modalità. I soli documenti prodotti e valutati nel corso di un procedimento penale sono molto eterogenei tra loro, sia che essi siano in forma testuale - verbali di interrogatorio o di testimonianza - sia in forma audio o video - nel caso di intercettazioni telefoniche o ambientali che vengono archiviate, scambiate e valutate giornalmente. Su questi documenti si basano sentenze o decisioni importanti come, ad esempio i filoni di indagine da aprire o proseguire, o ancora la stessa identificazione delle parti, dei testimoni e dei fatti accaduti.

L'uso di tali documenti per l'addestramento di algoritmi di Natural Language Processing (NLP) rende necessario un focus su aspetti che devono essere attenzionati nel processamento automatico dei testi attraverso NLP.

Al fine di consentire la creazione di modelli validi, è necessario porre attenzione ad alcuni aspetti qualitativi a cui devono rispondere i dati[7] per una corretta costruzione del dataset di addestramento. Nel contesto giuridico, se non correttamente addestrati, i modelli di ML possono portare ad un trattamento iniquo e discriminatorio, come nel caso dell'algoritmo COMPAS[8]. Il rischio di distorsioni e di discriminazioni non intenzionali verso sottoinsiemi della popolazione è molto legato a fenomeni di bias [9] nei dati. Tali fenomeni sono insidiosi e possono verificarsi anche con dati di buona qualità e ben etichettati.

Al fine di gestire e mitigare i rischi di bias e unfairness è necessario affiancare i sistemi di IA adottati con strumenti di analisi e interpretazione dei risultati che permettono agli utenti di comprendere l'inferenza effettuata dal modello e, ove necessario, poter intervenire sui dati di addestramento per eliminare i comportamenti distorti.

3. Trasparenza nel Machine Learning

I metodi di Machine Learning (ML) basano la propria efficacia su un *modello*, ovvero "qualcosa in grado di riconoscere (tramite un opportuno apprendimento) determinati *pattern*". In generale distinguiamo modelli *parametrici* (ovvero dipendenti da una serie di coefficienti, comunemente detti "pesi" per le reti neurali) o *non-parametrici* (ad esempio alberi decisionali e algoritmi di clustering). Focalizzandosi sui primi, ed in particolare su modelli basati su reti neurali profonde, è evidente come al crescere della complessità, in relazione diretta con il numero dei parametri del modello, diventi sempre più difficile riuscire ad "interpretare" o in qualche caso "giustificare" determinati comportamenti del modello che possono apparire poco intuitivi da un punto di vista umano. Data questa esigenza si origina il campo dell'Explainable AI (XAI), che ha come scopo quello di rendere più trasparenti le "motivazioni" dietro determinati comportamenti dei Recentilli. In questo ambito, generalmente i termini "Explainability" o "Interpretability" vengono erroneamente utilizzati come sinonimi. Tuttavia, mentre il primo si riferisce alla capacità di spiegare dal punto di vista umano una meccanica interna al modello, il secondo si riferisce alla capacità di comprendere la relazione causa-effetto osservata all'interno di un modello. Chiaramente entrambi gli aspetti sono di fondamentale importanza per poter instaurare un rapporto di *fiducia* nei confronti delle predizioni del modello. È tuttavia importante notare come l'investimento verso la creazione di modelli "spiegabili" sia in forte competizione rispetto all'incredibile crescita dei modelli in termini di complessità (e di performances). Recenti studi [10, 11] sulle leggi di scala di performance in relazione alla complessità del modello e alla dimensione e qualità del dataset suggeriscono che modelli molto grandi non siano solo correlati ad alte performances, ma siano capaci di risolvere una serie di "nuovi" compiti grazie a delle dinamiche emergenti [12]. Questo confronto attualmente sembrerebbe essere vinto da parte della complessità [13], tuttavia, senza necessariamente citare problemi legati alla fattualità nei modelli generativi [14, 15, 16], in domini particolarmente sensibili quale quello riguardante la Giustizia continua a rimanere di fondamentale importanza la capacità di controllare ed avere una comprensione delle ragioni dietro le predizioni di un modello.

I metodi di XAI costituiscono un panorama molto complesso e frastagliato, tuttavia possiamo distinguere tra metodi basati su *gradienti* oppure basati su *perturbazioni*. I metodi basati su gradienti sono stati i primi ad essere

sviluppati e consistono nell'interpretare la rete neurale tramite i gradienti dei suoi pesi. Di questa tipologia distinguiamo

- `Gradient with respect to the input` [17]: ovvero gradienti degli output calcolati a partire da variazioni degli input, utilizzati in congiunzione con una *saliency map* per aumentarne il contrasto.
- `Integrated Gradients` [18]: similmente al precedente, in cui però i gradienti sono calcolati a partire da input riscalati rispetto ad una baseline e poi mediati tra loro.
- `Layer-wise relevance propagation (LRP)` [19]: Introdotta originariamente per explanations "pixel-wise" si basa su regole di ridistribuzione di uno score di rilevanza all'interno di ogni layer della rete neurale. Anche in questo caso la rilevanza e' proporzionale al gradiente indicando se dato un input, questo supporta o meno una data predizione.
- `DeepLIFT` [20]: simile al precedente, ma con l'utilizzo di Integrated Gradients invece di gradienti calcolati sugli input originali.

I metodi *perturbation-based* invece si basano su (piccole) perturbazioni degli esempi sui quali si vuole ottenere una explanation e misurano la variazione nelle predizioni del modello localmente attorno all'esempio perturbato. Tipicamente questi metodi sono agnostici rispetto al modello sul quale fornire una explanation. In questo ambito è fondamentale citare

- `LIME` [2]: un metodo che fornisce explanations attraverso un'approssimazione locale del modello di cui si vogliono ottenere informazioni attorno ad un determinato esempio. La località e' definita tramite una misura di prossimità ed il modello viene approssimato linearmente.
- `ANCHORS` [21]: Questo metodo costituisce un'estensione del precedente introducendo il concetto di *coverage*, ovvero la regione per la quale una data explanation può essere applicata.
- `SHAP` [22]: questo metodo si basa sul concetto di *Shapley Values* introdotto per la prima volta nella teoria dei giochi cooperativi [23], ovvero la media del contributo marginale di una determinata feature calcolata su tutte le possibili coalizioni.

Data la complessità delle features e la necessità di un metodo di XAI in grado di fornire explanations semplicemente interpretabili, per questo lavoro è stato utilizzato LIME [2], con le opportune modifiche per l'utilizzo nel caso di un modello di Named Entity Recognition (NER).

4. Interpretabilità Locale

Come citato nella sezione precedente, LIME [2] acronimo di "Local Interpretable Model-agnostic Explanations" è un metodo di interpretazione model-agnostic per modelli di classificazione. Lo scopo fondamentale di questo approccio è quello di fornire informazioni altamente interpretabili sul comportamento di un modello. Le explanations *locali* fornite da LIME si basano su un'approssimazione *lineare* del modello originale attorno ad un esempio, a differenza di modelli di explanation *globali* che tentano di approssimare il modello per ogni esempio. Quest'ultima è molto più complessa ed è soggetta alla generazione di informazioni di difficile interpretazione. Inoltre, informazioni globali sul modello tendono a rappresentare comportamenti medi, spesso di marginale interesse rispetto allo specifico esempio. Da un punto di vista più formale il modello che fornisce le explanations, l'*explainer*, è un modello intepretabile $g \in G$ dove G è l'insieme dei modelli potenzialmente interpretabili. Per poter avere delle explanations interpretabili, le features di un esempio di input devono condividere tra loro la stessa rappresentazione interpretabile, anche se il modello internamente utilizza una rappresentazione differente. Una rappresentazione interpretabile è ad esempio un vettore binario che sta ad indicare la presenza o l'assenza di una parola in un determinato input testuale, mentre il modello internamente può rappresentare ogni parola con una rappresentazione più complessa. Quindi, per un dato esempio $x \in \mathbb{R}^d$, consideriamo la sua rappresentazione interpretabile come il vettore binario $x' \in \mathbb{R}^{d'}$ ed il dominio $\{0,1\}^{d'}$ del modello g. Possiamo inoltre definire una misura di complessità $\Omega(g)$ per il modello *explainer*. Dato un modello $f : \mathbb{R}^d \to \mathbb{R}$ sul quale fornire explanations, definiamo l'intorno di un dato esempio x come $\pi_x(z)$ in base ad una misura di prossimità. LIME utilizza un kernel esponenziale definito su una distanza angolare per input testuali. Possiamo introdurre una misura della bontà dell'approssimazione dell'*explainer* g dato f nell'intorno di x come $\mathscr{L}(f, g, \pi_x)$. Per tener conto del bilanciamento tra complessità ed errore di approssimazione la quantità da minimizzare è quindi definita come

$$\epsilon_\pi^f(x) = \text{argmin}_{g \in G} \mathscr{L}(f, g, \pi_x) + \Omega(g) \qquad (1)$$

dove $\mathscr{L}$ è calcolata in via approssimata estraendo casualmente esempi attorno ad x con probabilità π_x. Un esempio estratto è ottenuto estraendo elementi non-zero di x' uniformemente per ottenere un esempio perturbato $z' \in \mathbb{R}^{d'}$. L'esempio viene poi classificato dal modello come $f(z)$ ottenuto facendo la predizione del modello originale su $z \in \mathbb{R}^d$ ovvero la vera rappresentazione di z'. Queste istanze perturbate costituiscono un nuovo dataset $\mathscr{Z}$ sul quale viene calcolato $\epsilon(x)$. Nello specifico l'errore

di approssimazione del modello è quadratico

$$\mathcal{L}(f, g, \pi_x) = \sum_{z,z' \in \mathcal{Z}} \pi_x(z)(f(z) - g(z'))^2 . \qquad (2)$$

Come menzionato precedentemente, per mantenere un alto grado di interpretabilità, g è una funzione lineare nei parametri $g(z') = \sum_{i=1}^{d'} w_i z'_i$. Per classificazioni testuali la rappresentazione interpretabile è data da un semplice "Bag Of Words" (BOW) con un limite K sul numero massimo di parole. Successivamente, per semplificare la penalità dovuta alla complessità $\Omega(g)$ vengono selezionate le features più rilevanti ed i pesi w_i vengono calcolati con il metodo dei minimi quadrati.

4.1. Interpretabilità nella predizione di Entità Nominali

Il caso dell'estrazione di annotazioni di sequenze di tokens è differente rispetto ad un modello di classificazione di frasi, in quanto la classificazione avviene a livello di ogni token e non dell'intera frase. Modelli di questo genere vengono tipicamente utilizzati per il task di Named Entity Recognition (NER), un'attività che consiste nell'identificazione delle Entità Nominali espresse all'interno di un input testuale. Prima di poter discutere dell'applicazione del metodo LIME a questo genere di modello, bisogna fare un'importante precisazione: per ogni istanza di entità all'interno della frase, LIME deve produrre un diverso modello lineare. Questo perché i coefficienti del modello lineare prodotto da LIME sono dipendenti sia dall'esempio che dall'entità. Va inoltre tenuto conto del fatto che un'entità può essere rappresentata da più di un token. Nel tipico schema *BIO* introdotto a CoNLL nel 2002,2003 [24, 25], per una ipotetica entità "NOME" si annotano con $B-NOME$ ed $I-NOME$ rispettivamente il token iniziale e tutti i token successivi rappresentanti l'entità. Nel caso di queste entità multi-token nella nostra implementazione, a differenza di [26], le explanation non sono direttamente legate alle entità, ma alla loro annotazione BIO. In altri termini, nel caso di un'entità multi-token avremo explanations per le label predette $B-$, ed $I-$ in modo indipendente, senza effettuare alcuna aggregazione tra i token che la compongono.
Il modello di NER utilizzato nella nostra implementazione si basa su tecnologia Transformer, in particolare su Bert [27]. Tale modello, grazie a meccanismi di "self-attention" che rendono così performanti i Transformers [28], è in grado di produrre rappresentazioni contestuali per ogni singolo token appartenente al testo. Una particolarità di questo genere di modelli è quello dell'utilizzo di frammenti di parole come tokens, generalmente chiamati "word-piece"[2]. Il concetto di rappresentazione contestuale è particolarmente centrale in questo lavoro perché consente di ottenere una rappresentazione vettoriale di ogni token in cui i token che lo circondano forniscono un contributo importante [28]. Un'ulteriore differenza rispetto a [26] consiste nel modo in cui vengono applicate le perturbazioni agli esempi. Mentre letteratura i token appartenenti ad un'entità sono mantenuti inalterati durante le perturbazioni per preservare l'entità ed evitare che tutti o parte dei token che la compongono vengano mascherati. Nella nostra implementazione, invece, abbiamo scelto di non inserire questa policy. Questa scelta è motivata dal fatto che mascherare i token che compongono l'entità stessa non limita la valutazione ai contributi dei soli token di contesto. Infatti, analizzare i contributi dei token che compongono l'entità può risultare interessante. Ad esempio, nel caso in cui la forma superficiale compaia sempre associata ad una stessa entità nel dataset su cui è stato addestrato il modello, si osserva che il contributo dei token che la compongono è cruciale per la classificazione di tale entità. Diversamente, se una forma superficiale compare associata a più entità nel dataset di addestramento, si osserva che i token di contesto hanno un peso maggiore per la classificazione. Entrambi i fenomi sono facilmente osservabili tramite la nostra implementazione di LIME che riesce bene a rappresentare il contributo delle features tramite i coefficienti delle funzioni lineari locali.

[2]Per semplicità in questo lavoro ci si riferisce genericamente a "tokens" senza specificare il tipo di tokenizzazione.

5. Conclusioni

La sempre più centrale rilevanza che stanno assumendo gli algoritmi di IA in ambiti sensibili ha reso necessaria la definizione di parametri di qualità di tali artefatti. Questo articolo riporta l'esperienza effettuata nel progetto Datalake Giustizia, in cui sono state esplorate tecniche di Explanability AI (XAI) per l'interpretazione dei risultati prodotti da modelli di NLP. In particolare l'esperienza dell'applicazione di un algoritmo di explainability ad un modello per l'identificazione di Entita Nominali (NE), apre le porte all'adozione di tali approcci anche per modelli su task più complessi, che verranno sviluppati in futuro all'interno progetto.

6. Ringraziamenti

Ringraziamo il professor Carlo Batini per il coordinamento del progetto Datalake Giustizia e per i sui preziosi contributi sul tema qualità dei dati e interpretazione dei modelli di Machine Learning.

References

[1] P. Linardatos, e. a. Papastefanopoulos, Explainable ai: A review of machine learning interpretability methods, Entropy 23 (2021). URL: https://www.mdpi.com/1099-4300/23/1/18. doi:10.3390/e23010018.

[2] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, 2016. arXiv:1602.04938.

[3] B. Liu, Y. Hu, Y. Wu, et al., Investigating conversational agent action in legal case retrieval, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part I, volume 13980 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 622–635. URL: https://doi.org/10.1007/978-3-031-28244-7_39. doi:10.1007/978-3-031-28244-7_39.

[4] R. Zhang, Q. Ai, Y. Wu, et al., Diverse legal case search, CoRR abs/2301.12504 (2023). URL: https://doi.org/10.48550/arXiv.2301.12504. doi:10.48550/arXiv.2301.12504. arXiv:2301.12504.

[5] A. Farzindar, G. Lapalme, Legal text summarization by exploration of the thematic structures and argumentative roles (2011).

[6] F. Romeo, Giustizia e predittività. un percorso dal machine learning al concetto di diritto, Rivista di filosofia del diritto, Journal of Legal Philosophy (2020) 107–124. URL: https://www.rivisteweb.it/doi/10.4477/97023. doi:10.4477/97023.

[7] C. Batini, Enciclopedia dei dati digitali,Volume terzo: L'etica dei dati digitali: l'Equità, Milano : in proprio., 2022.

[8] S. M. e. a. Julia Angwin, Jeff Larson, Machine bias, 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[9] K.-W. Chang, V. Prabhakaran, V. Ordonez, Bias and fairness in natural language processing, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts, Association for Computational Linguistics, Hong Kong, China, 2019. URL: https://aclanthology.org/D19-2004.

[10] J. Kaplan, S. McCandlish, T. Henighan, et al., Scaling laws for neural language models, 2020. arXiv:2001.08361.

[11] J. S. Rosenfeld, Scaling laws for deep learning, 2021. arXiv:2108.07686.

[12] A. Srivastava, A. Rastogi, A. Rao, et al., Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2022. arXiv:2206.04615.

[13] A. Sarkar, Is explainable ai a race against model complexity?, 2022. arXiv:2205.10119.

[14] A. Devaraj, W. Sheffield, B. C. Wallace, et al., Evaluating factuality in text simplification, 2022. arXiv:2204.07562.

[15] R. Weng, H. Yu, X. Wei, et al., Towards enhancing faithfulness for neural machine translation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 2675–2684. URL: https://aclanthology.org/2020.emnlp-main.212. doi:10.18653/v1/2020.emnlp-main.212.

[16] J. Maynez, S. Narayan, B. Bohnet, et al., On faithfulness and factuality in abstractive summarization, 2020. arXiv:2005.00661.

[17] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. arXiv:1312.6034.

[18] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, CoRR abs/1703.01365 (2017). URL: http://arxiv.org/abs/1703.01365. arXiv:1703.01365.

[19] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, 2017. arXiv:1708.08296.

[20] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, 2019. arXiv:1704.02685.

[21] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, Proceedings of the AAAI Conference on Artificial Intelligence 32 (2018). URL: https://ojs.aaai.org/index.php/AAAI/article/view/11491. doi:10.1609/aaai.v32i1.11491.

[22] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, 2017. arXiv:1705.07874.

[23] L. S. Shapley, A Value for N-Person Games, RAND Corporation, Santa Monica, CA, 1952. doi:10.7249/P0295.

[24] E. F. T. K. Sang, Introduction to the conll-2002 shared task: Language-independent named entity recognition, 2002. arXiv:cs/0209010.

[25] E. F. T. K. Sang, F. D. Meulder, Introduction to the conll-2003 shared task: Language-independent named entity recognition, 2003. arXiv:cs/0306050.

[26] S. H. Villarroya, Z. Akata, Interpretability in sequence tagging models for named entity recognition by sofia herrero villarroya, 2018.

[27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for

language understanding, 2019. arXiv:1810.04805.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.