Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
April 24, 2025
A Regional Conference of the Society for Industrial and Systems Engineering

Maximizing the Usability of Publicly Available Information

Scott Belveal, Dalton Harkins, Charles O'Donnell, Sarah Platt, and Donald Koban

Department of Systems Engineering, United States Military Academy, West Point, New York 10996

Corresponding author's Email: donald.koban@westpoint.edu

Author Note: This work was supported by USSOCOM under support agreement No. USMA25004. The views expressed herein are those of the authors and do not reflect the position of the United States Military Academy, the Department of the Army, or the Department of Defense.

Abstract: As modern warfare increasingly relies on data-driven intelligence, leveraging Publicly Available Information (PAI) has become essential for military operations. To maximize its usability, PAI must be structured into databases that support large-scale queries and integration with other intelligence sources. This process requires filtering U.S. Person Information (USPI) to ensure compliance with collection, storage, and usage regulations. Although automated approaches for USPI filtering exist, few studies have systematically evaluated their performance. This study evaluates a DoD-developed Regex model and explores the feasibility of using Large Language Models (LLMs) for USPI filtering. Compared to the Regex model, an open-source LLM, Llama 3.2, demonstrated significantly higher classification performance and an ability to generate coherent explanations. However, these gains come with increased processing time. Our findings suggest that, given sufficient computational resources, LLMs may outperform current USPI filtering methods. Future research should explore efficiency improvements and evaluate performance across broader data sets.

Keywords: Publicly Available Information, Large Language Models, Text Classification, U.S. Person Information

1. Introduction

The Department of Defense's (DoD's) role in protecting national security increasingly depends on its ability to process and analyze vast amounts of publicly available information (PAI) from sources such as social media, traditional news, and blogs (Browne et al., 2024). PAI is valuable because it can satisfy information requirements without specialized collection methods and can also tip and cue other intelligence sources. Additionally, organizing PAI into databases enhances its usability, enabling large-scale queries and integration with other intelligence data, which improves operational efficiency. However, legal and ethical constraints—particularly those protecting the privacy of U.S. persons—limit how such data can be collected, stored, and analyzed.

To ensure compliance with federal laws and DoD policies, intelligence personnel must carefully screen PAI for U.S. Person Information (USPI). The Army defines PAI as publicly available information from sources such as online content, broadcast media, public events, and subscription-based publications (Department of Defense, 2019). Open-source intelligence (OSINT) refers to the collection, analysis, and dissemination of intelligence derived from PAI to support decision-making (109th Congress, 2006). While OSINT is helpful for maintaining situational awareness, practitioners must ensure that data collection adheres to intelligence oversight guidelines. If information is copied, saved, stored, or otherwise preserved in any manner, it is "collected" and must be evaluated in accordance with Public Law 109-163, DODM 5250.01, and AR 381-10.

Given the complexities of how individuals represent themselves on social media, screening PAI for USPI can be time-consuming, complicated, and subjective. Consequently, manually reviewing data is impractical for large datasets (Edwards et al., 2015), and creating rules that account for every possible scenario is infeasible. Nonetheless, some units have attempted to perform USPI screening using Regex-based filtering, which detects predefined keywords and patterns to flag potential USPI. An example of Regex filtering can be seen in spam detection. Early spam detection methods primarily relied on rules-based approaches, including Regex filters and blacklists, to flag messages containing predefined spam indicators like "free money" or "click here" (Xia, 2020). Although effective at capturing straightforward spam patterns, these models suffer from high false positive rates and require constant updates as spammers continuously modify their phrasing. Additionally, applying Regex approaches to new data or tasks is challenging, as it requires significant time and adjustments (Sykes et al., 2021).

To address these challenges, transformer-based architecture, a class of deep learning designed to process and understand natural language with greater contextual awareness, have been applied to natural language processing tasks (Brown et al., 2020). Unlike rules-based approaches, transformers utilize self-attention mechanisms to weigh the importance of different words within a given text, allowing them to capture nuanced relationships and infer meaning beyond keyword matching. Hence,

transformer-based models don't require users to maintain a predefined set of key words. These capabilities make transformer-based models well-suited for tasks like USPI classification, where context plays a crucial role in differentiating between relevant and irrelevant information (Raschka, Sebastian, 2023). Additionally, transformer models can account for misspellings, abbreviations, and acronyms that may be missed or extremely time consuming to include in Regex models.

Bidirectional Encoder Representations from Transformers (BERT) offer a deep contextual understanding of language, making it well-suited for text classification. However, a notable limitation of BERT is its lack of localized explanations for its predictions, often leading to its characterization as a black-box model (Van Aken et al., 2019). Black-box models are generally less trusted than explainable models that provide traceability. In contrast, LLMs excel at generating explanations, enhancing interpretability. Although LLMs are prone to hallucination—where the model generates inaccurate or misleading information—we opted to examine LLMs due to their ability to provide coherent and contextualized explanations, which are essential for tasks requiring interpretability and trust.

In this study, we systematically compare a DoD-developed Regex model with an open-source LLM developed by Meta, Llama 3.2. We hypothesize that Llama 3.2 will outperform the Regex method due to its ability to incorporate contextual clues and generate coherent explanations. To test this, we analyzed verified X (formerly known as Twitter) profiles available on the Internet Archive website (*Internet Archive*, 2023). We first manually annotated each account as "American" or "Not American" using a three-person majority vote. We then compared the classification performance of the Regex and LLM-based methods against the manually annotated data, which served as ground truth. Additionally, we reviewed the explanations generated by the LLM to assess whether it produced coherent justifications. To our knowledge, no prior studies have systematically compared Regex models and LLMs for identifying USPI. Our findings suggest that LLMs could enhance current classification systems by making context-informed predictions and providing more coherent explanations.

2. Methodology

The following section provides a description of our research methodology shown in Figure 1.

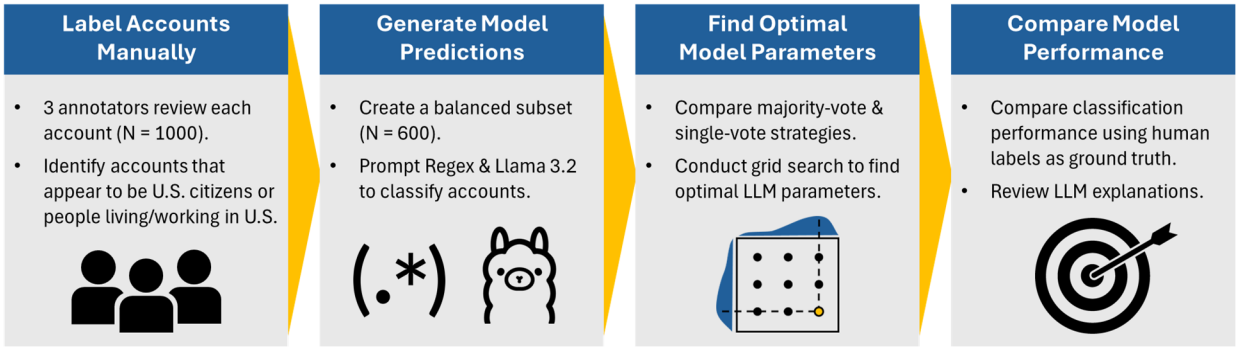

Figure 1. Research methodology for comparing Regex- and LLM-based classification performance in identifying X profiles likely belonging to U.S. citizens or individuals residing in the U.S.

2.1 Label Accounts Manually

To evaluate performance, we first established ground truth labels by manually annotating the X profiles. Although DODM 5240.01 defines a "U.S. person" to include U.S. citizens, permanent resident aliens, and corporations in or owned by U.S. entities, our annotations focused mainly on determining whether an entity was an U.S. citizen or resided or worked in the United States (Department of Defense, 2016). Because we were working with profile information with varying levels of credibility, we attempted to adhere to the principles discussed in US military guidelines for assessing confidence (Irwin & Mandel, 2019). The annotators conducted surface-level research which included tasks such as searching for names on Google or Wikipedia, viewing webpages included on the profile, and inspecting the user's recent posting activity. Three independent annotators reviewed each profile, assigning a label of "Y" (American) or "N" (Non-American). The final classification was determined by a majority vote—if the first two annotators disagreed, a third annotator cast the deciding vote. If a profile lacked clear indicators of U.S. affiliation, the account was assumed to be American.

2.2 Generate Model Predictions

To maintain data privacy, we implemented our approach using Ollama, an open-source tool that allows users to run LLMs locally on a computer. We performed all experiments using a Nvidia RTX 6000 Ada Generation GPU. We used the 3B-parameter Llama 3.2 model because it was the most recent model available at the time of the study and was developed by Meta, a well-established company with expertise in large-scale AI models. Unlike our manual annotation process, our automated classification approaches were based solely on metadata found in each X profile. The metadata included the following fields: 1) name, 2) screen name, 3) self-reported locations, 4) URL, 5) description, 6) email, and 7) profile banner URL. This makes our study fundamentally different from many geo-inference research efforts, which often rely on analyzing a user's posting activity, linguistic patterns across tweets, or social network connections to infer geographic location (Cheng et al., 2010; Jurgens et al., 2021). In contrast, our approach is constrained to a small amount of static text, which presents unique challenges in classification. We used the following classification prompt:

"Your job is to classify this Twitter account as belonging to a U.S. Person or not. A U.S. Person is defined as: "A U.S. citizen, a known permanent resident alien, a U.S.-based corporation not controlled by a foreign government, or an unincorporated group primarily composed of such individuals". Respond with 'Yes' or 'No' and provide an explanation for your prediction. Format your response as: 'American: (Yes or No), Explanation: (optional).' Here is the profile information:"

By restricting input to profile metadata, we aimed to test how well LLMs infer nationality from minimal data, rather than relying on behavioral signals. This distinction is critical when comparing our results to prior work in geo-inference, where richer datasets typically yield higher accuracy. Furthermore, since it is often impractical for large-scale, analytic workflows to incorporate external data beyond metadata, our approach is more applicable to real-world implementations.

2.3 Find Optimal Model Parameters

To address the highly imbalanced data we employed stratified sampling. Using balanced data was essential for model tuning and hyperparameter optimization, ensuring fair representation and improved generalization in the classification task. We varied the model's temperature parameter between 0 and 1.0 to explore its impact on classification performance. Temperature controls response randomness—a low setting (0) makes the model highly deterministic, ensuring consistent classifications across runs, while a high setting (1.0) increases variability, allowing for more creative responses. We conducted a grid search to identify the optimal temperature setting and classification strategy. To mitigate the effect of hallucinations and account for temperature-driven variations in model response, we implemented an ensemble approach, running each account through the model multiple times (1–15 runs). Temperature-driven variations occur due to the inherent randomness introduced by the model's sampling mechanism, which can lead to different outputs for the same input. By running multiple iterations, we aimed to stabilize the predictions and minimize the impact of hallucinations. We evaluated two labeling strategies:

- Majority-Vote Approach: The label was assigned based on the most frequent classification across multiple runs. This approach assumes that hallucinations are rare events and that the model is generally more accurate than erroneous. It prioritizes accuracy by minimizing the impact of occasional hallucinations.

- Single-Vote Approach: If any single run classified an account as American, the account was labeled as American. This method assumes that, with enough attempts, the model is more likely to correctly identify American accounts that exhibit less definitive indicators. It prioritizes recall but may reduce precision.

2.4 Compare Model Performance

To benchmark performance, we compared LLM classifications with those from a Regex-based screening tool used by the study sponsor. The Regex method was designed to detect keywords in categories such as U.S. state and city names, email domains, and colloquialisms referring to well-known locations. The sponsor provided the research team with a spreadsheet of terms flagged as USPI, along with their occurrences in X profiles and the associated Regex rule that triggered each flag. To ensure accuracy, annotators manually reviewed the Regex results to verify correct mapping to each profile. Both classification methods were then compared against the ground truth labels to assess accuracy, precision, recall, and F1 scores. Additionally, we recorded computation times to compare time requirements for each approach. To assess the effect of model input, we conducted a second set of experiments in which users' self-reported locations were masked from the LLM. Finally, to assess explanation quality, we manually reviewed LLM predictions to identify common patterns. We selected specific examples to highlight where the model performed well and where it struggled, providing insights into its strengths and limitations.

3. Results

Our study sample consisted of 1,000 accounts that we sampled from a spreadsheet of X profiles found on the internet archive website. All accounts were created between 2006 and 2008, had complete information in all meta-data fields, and included emojis. Two teams, each consisting of two annotators, demonstrated substantial agreement in classifying the accounts. Team 1 achieved an inter-rater reliability (Cohen's kappa) of 0.822 with 93.6% agreement, while Team 2 had a kappa of 0.854 with 93.4% agreement. After disputes were resolved by a 3rd annotator, we assessed that 75% of the X profiles belonged to a U.S. citizen or someone who lived or worked in the United States. A balanced sample of 600 accounts was selected from the set of manually annotated data; this balance was achieved through stratified sampling to ensure diversity in classification labels.

Model tuning analysis showed that F1 scores remained consistently high at temperatures of 0.2 or less for both voting strategies (Figures 2a and 2b), indicating that LLMs performed better when configured to generate less creative responses. As expected, the majority-voting strategy outperformed single runs, achieving higher F1 scores with 15 predictions (Figure 2a). In contrast, the single-vote strategy produced lower F1 scores under the same conditions (Figure 2b), largely due to higher false positive rates outweighing gains in recall. Both classification strategies were sensitive to model input. Masking the location information led to significant drops in F1 scores (Figure 2c, 2d). This suggests that the model heavily relies on location data for accurate classification, raising potential concerns about generalizability and robustness when location data is unavailable or misleading.

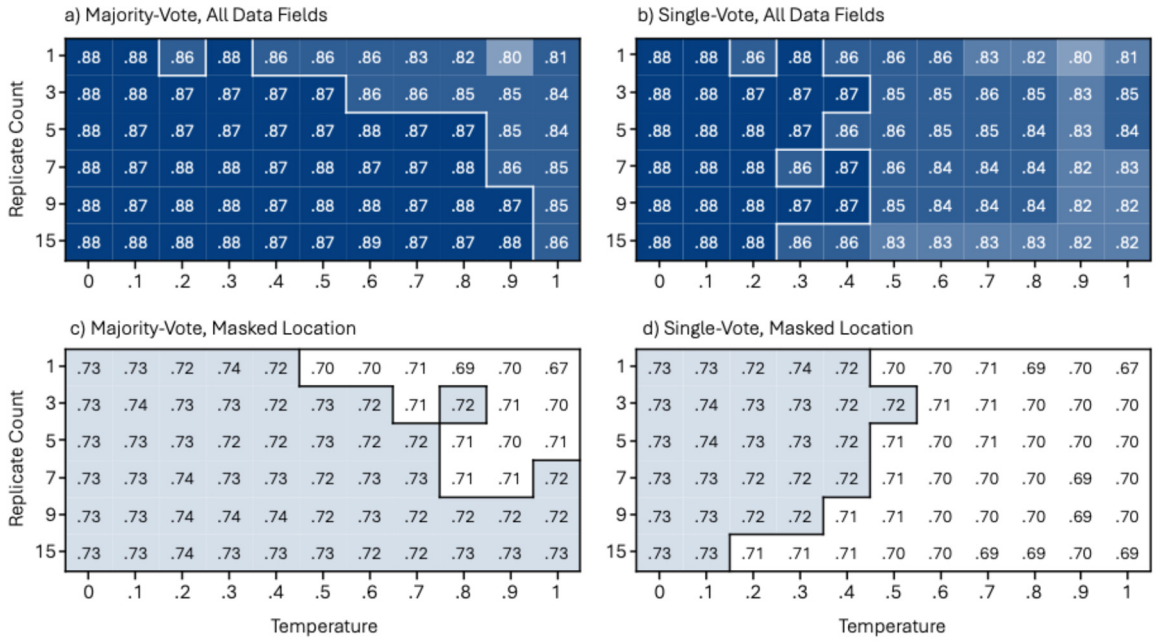

Figure 2. Left to right: F1 score heat maps for the majority-vote approach (a, c) versus the single-vote approach (b, d). Top to bottom: F1 score heat maps using all available metadata fields (a, b) versus masking the location (c, d).

Our analysis revealed that using a single run of the LLM with a temperature of 0 outperformed the Regex model by a substantial margin, achieving an F1 score of 0.88 compared to 0.48 (Table 1). As expected, applying majority-vote and single-vote strategies across multiple LLM runs further improved accuracy and recall, respectively. However, these gains come with a trade-off of increased computation time. Processing 600 accounts with a GPU took 1 minute and 15 seconds, leading to an estimated runtime of approximately 3.7 hours for 100,000 accounts. Running the process across 3 replicates would extend the total runtime to approximately 11.1 hours. Scaling to a million accounts would require multiple days, posing a challenge for large-scale applications.

By analyzing the explanations provided by the LLM for classifying X accounts, we identified key themes about how the model was making its predictions. The model determined account classifications based primarily on factors such as location, language, content, and affiliations. Accounts were classified as USPI if they referenced U.S. locations, used American slang, mentioned with U.S.-based accounts, or discussed U.S. political and cultural topics. In contrast, non-USPI accounts tended to avoid U.S.-specific references, used other languages, and mentioned primarily non-U.S. entities. The model also considers

hashtags, keywords, and profile bios to assess whether an account has a connection to the U.S. Overall, the LLM relied on a combination of linguistic patterns, locational references, and affiliations to make its classifications.

Table 1. Classification performance on labeling 600 X profiles. The DoD-developed Regex model labeled accounts as American if any pattern matched. LLM predictions were generated using a single run and two voting strategies (single-vote and majority-vote), as described in section 3. Temperature was set to 0 for the single run and 0.6 for multiple runs.

Model	Temperature	Accuracy	Precision	Recall	F1	Time (min)
Regex Model (Baseline)	N/A	.65	.88	.33	.48	0:05
Single Run LLM	0	.88	.91	.84	.88	**1:15***
Majority-Vote LLM, 3 replicates	.6	.87	.90	.83	.86	2:30
Majority-Vote LLM, 15 replicates	.6	**.90***	**.94***	.85	**.89***	12:30
Single-Vote LLM, 3 replicates	.6	.84	.82	.88	.85	2:30
Single-Vote LLM, 15 replicates	.6	.80	.74	**.95***	.83	12:30

Note: Asterick with bold text indicates the maximum or minimum value within each column

4. Discussion

A key tradeoff in this study is the balance between precision, recall, and processing time. Using multiple LLM runs can improve precision through a majority-vote strategy or enhance recall via a single-vote strategy. However, these ensemble-style approaches significantly increase processing time, posing practical challenges for large-scale applications where time efficiency is crucial. Organizations adopting LLM-based classification must weigh the benefits of enhanced performance against the computational costs and resource demands associated with using multiple runs. Notably, our results showed that models performed better at lower temperature settings (0.2 or below), which is significant given that Llama 3.2's default temperature is 0.7. Additionally, the LLM's reliance on location data raises concerns about its ability to generalize to accounts with missing, misleading, or inaccurate location descriptions.

Another important consideration is the role of explainable AI (XAI) in classification tasks (Phillips et al., 2021). We found that the LLM's explanations were generally coherent and presented in a manner that humans could easily understand. This enhances transparency and builds user trust. However, a key limitation is that the LLM's factually inaccurate explanations can be difficult to recognize because they are often presented confidently and appear plausible. In contrast, the Regex-based approach provided precise indications of specific words and text patterns that triggered classifications. Although this precision aids traceability and accountability, non-coding-savvy users often struggle to understand Regex patterns. This trade-off highlights the need for explainability solutions that balance human interpretability with technical accuracy. Future work should explore methods to enhance model explainability while maintaining classification accuracy and efficiency.

Privacy and cost considerations also influence model implementation. A strength of our approach is that Ollama enables local deployment of LLMs and does not require a subscription fee, as it is open source. Additionally, since this workflow is designed for databasing rather than exploiting information, the risks associated with incidental data retention are lower. However, implementing post-processing steps to filter or anonymize collected data can further mitigate privacy concerns while preserving classification integrity. Future work should establish best practices for data management that strike a balance between data minimization and maintaining necessary classification information.

4.1 Limitations

While the LLM significantly outperformed the Regex model, several limitations constrain the broader applicability of this approach. One key limitation is sampling bias within our dataset, as we tested the model using only verified accounts. Since account verification was reserved for well-known figures with confirmed identities when the accounts were created, it is uncertain how the model would perform when classifying non-verified accounts, which may contain a higher proportion of bots, false information, or misleading content. This limitation also raises concerns about the model's generalizability across different social media platforms and non-English-speaking users. Another limitation is that we tested only one LLM, Llama 3.2, despite research suggesting that other models may achieve higher performance (Gao et al., 2025). Furthermore, while our results show that Llama 3.2 outperforms the Regex model, it is possible that a more advanced Regex model could bridge the performance gap. Finally, our accuracy estimates rely on ground truth labels derived from a manual annotation process that is inherently imperfect. These limitations highlight the need for further research into dataset selection, LLM model comparisons, and hybrid approaches that combine Regex-based precision with LLM flexibility to optimize performance.

5. Conclusion

In summary, our study demonstrates that LLMs have the potential to outperform Regex models in filtering USPI from X profiles and ultimately enhance the usability of PAI by facilitating its structured storage and retrieval. Lama 3.2 consistently outperformed a DoD-developed Regex model, especially at low, non-default temperature settings. Its highest precision and recall were achieved using majority-vote and single-vote classification methods over multiple runs with a temperature setting of 0.6, respectively. However, despite its strong performance, the LLM required significantly more processing time than the Regex model and relied heavily on location data from X profiles. Future research should focus on reducing processing time and expanding the study's scope. Exploring batch prompting, alternative LLMs, other transformer models, or hybrid approaches that combine Regex precision with LLM flexibility may improve efficiency. Additionally, testing the model on other social media platforms, non-English profiles, and data sources beyond metadata could yield broader insights into its generalizability.

5.1 Acknowledgements

The authors used ChatGPT to assist with editing this article by asking ChatGPT “Can you review this sentence or paragraph for clarity and conciseness?” ChatGPT offered recommendations that were incorporated into the final manuscript.

6. References

109th Congress. (2006). *National Defense Authorization Act For Fiscal Year 2006*. Senate and House of Representatives.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Browne, T. O., Abedin, M., & Chowdhury, M. J. M. (2024). A systematic review on research utilising artificial intelligence for open source intelligence (OSINT) applications. *International Journal of Information Security*, *23*(4), 2911–2938. https://doi.org/10.1007/s10207-024-00868-2

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: A content-based approach to geo-locating twitter users. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 759–768. https://doi.org/10.1145/1871437.1871535

Department of Defense. (2016). *Procedures Governing the Conduct of DoD Intelligence Activities (DoD Manual 5240.01)*. Department of Defense.

Department of Defense. (2019). *Open-Source Intelligence (ATP 2-22.9)*. Department of Defense.

Edwards, M., Rashid, A., & Rayson, P. (2015). A Systematic Survey of Online Data Mining Technology Intended for Law Enforcement. *ACM Computing Surveys*, *48*(1), 1–54. https://doi.org/10.1145/2811403

Gao, T., Jin, J., Ke, Z. T., & Moryoussef, G. (2025). *A Comparison of DeepSeek and Other LLMs* (arXiv:2502.03688). arXiv. https://doi.org/10.48550/arXiv.2502.03688

Internet Archive. (2023, January 6). Twitter-Databases Directory Listing. https://archive.org/download/twitter-databases

Irwin, D., & Mandel, D. (2019). How Intelligence Organizations Communicate Confidence (Unclearly). *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3441302

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y., & Ruths, D. (2021). Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice. *Proceedings of the International AAAI Conference on Web and Social Media*, *9*(1), 188–197. https://doi.org/10.1609/icwsm.v9i1.14627

Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). *Four principles of explainable artificial intelligence* (NIST IR 8312; p. NIST IR 8312). National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/NIST.IR.8312

Raschka, Sebastian. (2023). *Self Attention*. https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

Sykes, D., Grivas, A., Grover, C., Tobin, R., Sudlow, C., Whiteley, W., Mcintosh, A., Whalley, H., & Alex, B. (2021). Comparison of rule-based and neural network models for negation detection in radiology reports. *Natural Language Engineering*, *27*(2), 203–224. https://doi.org/10.1017/S1351324920000509

Van Aken, B., Winter, B., Löser, A., & Gers, F. A. (2019). How Does BERT Answer Questions?: A Layer-Wise Analysis of Transformer Representations. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1823–1832. https://doi.org/10.1145/3357384.3358028

Xia, T. (2020). A Constant Time Complexity Spam Detection Algorithm for Boosting Throughput on Rule-Based Filtering Systems. *IEEE Access*, *8*, 82653–82661. https://doi.org/10.1109/ACCESS.2020.2991328