International Journal of Multidisciplinary Global Research
Vol. 02 (02), June 2025, pp. 01-12.
Online ISSN: 3107-4197

Automated Interview Scoring (AIS) by Integrating Wordnet with NLTK

RAIS ABDUL HAMID KHAN

School of Computer Sciences and Engineering, Sandip University, Nashik
Maharashtra, India
rais.khan@sandipuniversity.edu.in

YOGESH K. SHARMA

Department of Computer Engineering, VIIT, Pune
Maharashtra, India
yogesh.sharma@viit.ac.in

MOHINI GURAV

Department of English, Sandip University, Nashik
Maharashtra, India
mohinigurav300@gmail.com

ABDULKAYYUM SHAIKH

School of Computer Sciences and Engineering, Sandip University, Nashik
Maharashtra, India
kayyum.shaikh22@gmail.com

SIDDHARTH CHAVAN*

School of Computer Sciences and Engineering, Sandip University, Nashik
Maharashtra, India
siddharthchavan2511@gmail.com

Abstract

Manual rating of autobiographical interviews, the essence of the assessment of autobiographical memory in psychology, is a labour-intensive and time-consuming task. Here, narrative details have to be rated by human raters, that is, classified as internal (episodic) or external (non-episodic), which limits the scalability of such research efforts critically. This project looks at the application of NLP methods for automating the rating process. The objective of this research is to fine-tune a language model such as distilBERT toward the end of making its usage more efficient and relieving part of the burden from researchers without compromising the precision in content classification. The project further assesses the reliability of the automated system across various datasets and determines whether its general structure may present possible futures for advancing research in autobiographical memory. All

*Corresponding Author.

these developments work towards focusing on efficient and scalable NLP tools development for mainstream psychological use.

Keywords: Automated Interview Scoring, Episodic Memory, Interview Scoring, NLTK, Non Episodic Memory, Wordnet.

1. Introduction

Automated scoring of autobiographical interviews combines psychology and natural language processing or NLP in order to score memories through personal narratives. It is that which has traditionally been a labor-intensive process subject to human error and bias, but new advances in NLP and ML have facilitated the handling of this task faster and more objectively. Critical to the understanding of the cognitive process, subtle linguistic features of autobiographical memory such as memory specificity, emotional regulation, and temporal sequencing are detected and classified by the models BERT and RoBERTa in machines.

In clinical settings, these automated systems are highly useful in terms of psychological disorder diagnosis in over general memory retrieval, which is often related to depression. They measure the affective and temporal aspects of memory, which assist in determining the effect of emotions and old age on the recall. For example, for general semantic aspects of memory, older adults typically recall more, and the details of memory decrease with disorders such as (Gaesser et al., 2011). The automatic categorization of these differences holds a prime place in the diagnosis and treatment of such disorders (Wardell et al., 2021).

Despite promise, challenges abound, including the need for culturally and linguistically diverse datasets and careful handling of subjective memory. In addition, cross-language and culture adaptability are critical steps toward applying findings at the global level. These challenges may well hold the key to revolutionizing psychological research and practice through scalable, bias-free tools for the investigation of memory and cognitive functions. What NLP promises in psychology is innovation, not only from assessments of mental status to more general cognitive inquiries but also needs to be integrated into all fields of psychology (Takano et al., 2019).

2. Literature Review

Autobiographical memory is a complex phenomenon determined by age, culture, and cognition. Much of the time the literature on autobiographical memory focuses on WEIRD populations, which generalizes very little across different cultural groups (Takano et al., 2019). developed an algorithm on a computerized scoring in the analysis of autobiographical memories in adults speaking English, and in such sense, there is a requirement for broader cultural studies. Cultural differences influence remembering memories significantly as shown by (Henrich et al., 2010). Aging-related changes are also significant. Older adults tend to generalize more memories compared to the young, which affects the recall of specific details or even the

imagination of future events extended this by showing that aging has such challenges in episodic memory used in simulating future events (Addis et al., 2008; Gaesser et al., 2011). Emotions have a huge impact on remembering because emotionally significant events are recalled with much detail. This has led to further research by (Wardell et al., 2021), which concludes that emotions influence the precision and distortion of specific details of autobiographical memory. The interplay of age, culture, emotion, and cognition is what shapes autobiographical memory as shown in the table 1.

Table 1. Number of parameters in the experimentation.

Sl.No.	Authors	Methodology	Advantages	Challenges
1	(Takano et al., 2019)	Experimental study on age-related changes in episodic simulation of future events.	Provides insights into cognitive aging.	Limited sample diversity; may not generalize to all ages.
2	(Wardell et al., 2021)	Review of cultural psychology and its implications for generalization of research findings.	Highlights biases in psychological research.	Broad scope may dilute specific findings.
3	(Takano et al., 2017)	Comparative study on remembering the past and imagining the future, focusing on age differences.	Enhances understanding of memory processes across ages.	Small sample size; potential confounding variables.
4	(Renoult et al., 2020)	Data mining techniques applied to fake news detection on social media.	Applicable to real-time detection; leverages large datasets.	Rapidly evolving nature of misinformation; high false positives.
5	(Henrich et al., 2010)	Naive Bayes classifier for detecting fake news in social media.	Simplicity and effectiveness in classification tasks.	Assumes independence of features; sensitive to data quality.
6	(Addis et al., 2008)	Computerized classification algorithm for specific autobiographical memories.	Automates and standardizes memory analysis.	Requires large annotated datasets; algorithm bias possible.
7	(Addis et al., 2008; Gaesser et al., 2011)	Further analysis on fake news detection using data mining techniques.	Expands on previous work, providing comprehensive insights	May not cover all sources of misinformation.
8	(Shu et al., 2017)	Study on the spread of fake news by social bots using network analysis.	Identifies bot behaviors in misinformation spread.	Difficulty in identifying bots in real-time; ethical concerns.

9	(Granik & Mesyura, 2017)	Development of self multi-head attention-based CNNs for fake news detection.	Utilizes advanced neural network techniques for improved accuracy.	Computationally intensive; requires extensive tuning.
10	(Shao et al., 2017)	Introduces BERT for pre-training of deep bidirectional transformers.	State-of-the-art performance in various NLP tasks.	High resource consumption for training; requires large data.
11	(Fang et al., 2019)	Updates to computerized scoring algorithm for autobiographical memory tests.	Provides a refined scoring method for analyzing autobiographical memories.	Requires continual updates for different populations.
12	(Devlin et al., 2019)	Hybrid feature extraction and machine learning for disease prediction.	Addresses an important health issue; hybrid approach enhances accuracy.	Limited dataset availability; model generalization issues.
13	(Subramani & BD, 2021)	Investigates how emotion influences details recalled in autobiographical memory.	Sheds light on emotional impacts on memory retrieval.	Subjectivity of emotional influence; requires careful controls.
14	(Palaparthi, 2021)	Review and explanation of Random Forest algorithm for various applications.	Versatile and robust model for classification tasks.	Can be less interpretable; overfitting in small datasets.
15	(Wolf et al., 2020)	Discusses the use of transformers for state-of-the-art NLP applications.	Advances understanding of transformer models in NLP.	Complexity of implementation; high computational demands.
16	(van Genugten et al., 2021)	Classification of general and personal semantic details in autobiographical interviews.	Provides a systematic approach to analyzing memory details.	Variability in individual memory; potential biases in classification.
17	(Shu et al., 2017)	Exploratory study on the contribution of episodic retrieval to creative writing.	Links memory processes with creative outputs.	Limited sample size; subjective nature of creativity measurement.

3. Proposed System

3.1. *Methodology*

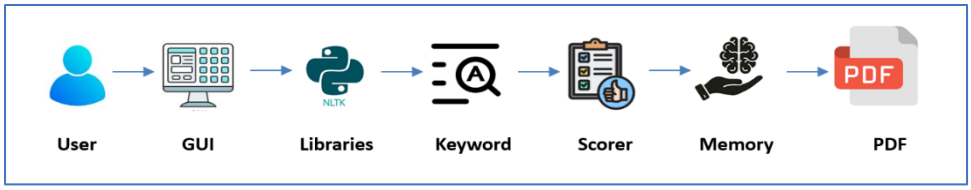

Fig. 1. Architecture of proposed approach "Automated Interview Scoring."

Figure 1 refers as memory scoring application workflow from a user's interaction to generation of the PDF report:

1. The user engages with the system by inputting his narrative.
2. The interface lets the user input his ID and name and his narrative using input fields and dropdown menus.
3. Core libraries such as NLTK available in Python are utilized for text processing and tokenization.
4. The system finds out the relevant internal and external keywords within the text.
5. The scoring module classifies the sentences to be internal, external, or mixed based on the keywords extracted.
6. The system analyses the content of the text input by finding out the participant's memory based on the internal details.
7. Finally, all the processed data, memory score, and analysis are summed into a PDF report that will be produced for the user.

The utilization of memory scoring, according to (Takano et al., 2019), is an advanced tool used in order to guide the user furthering their capabilities in applying reasonable knowledge to their autobiographical memory. With input for the narrative data, the system then uses advanced natural language processing by means such as NLTK to parse the narrative into intelligible components both within and without the consideration of the elements of memory.

This scoring module has a combination of features, from specificity to the emotional content of the presented items as well as temporal details on the events being remembered. Users are offered a long report but also graphics that simply enable them to intuitively understand their performance in their memories. Beyond that, these insights become actionable feedback in that they can tell someone exactly which aspects could be improved or in what specific regard memory could be enhanced through practice.

This broad approach to memory evaluation, combined with the possibility of refinement for widespread applications, makes it an invaluable tool for both occasional users and researchers alike. This broad approach to memory evaluation, combined with the possibility of refinement for widespread applications, makes it an invaluable tool for both occasional users and researchers alike.

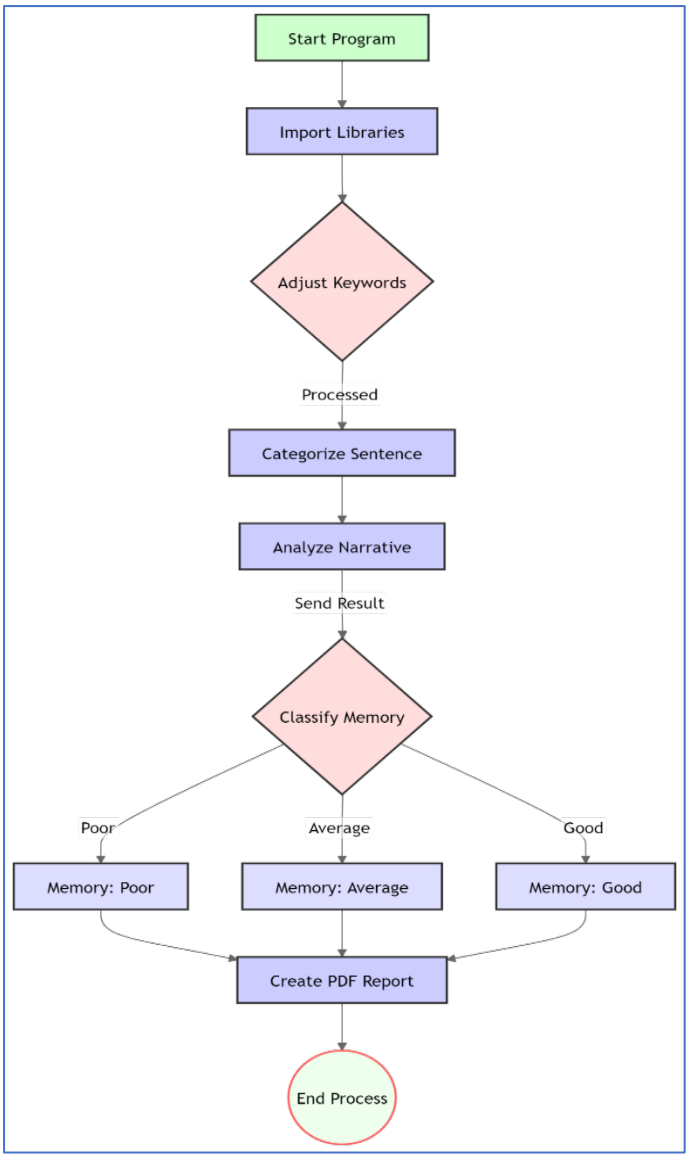

Fig. 2. Flowchart of "Automated Interview Scoring."

Figure 2 outlines a process for the analysis of a narrative and memory performance classification. These are the steps to follow in a discussion:

1. **Start Program**: From this step, it starts the program.
2. **Import Libraries**: All libraries called upon by the program have been imported.
3. **Adjust Keywords**: This step has fine-tuned or adjusted the keyword that will be used for analysis of the presented narrative. It is probably aimed at specific terms or phrases in the user's input.

4. **Sentence Categorize**: The sentences of the narrative are then categorized, perhaps clustering by content or even structure after keywords have been adjusted
5. **Narrative Analysis**: This is perhaps the area where most of the analysis on the user's narrative is carried out. It might involve NLP techniques, among others, for analyzing the content of the narrative.
6. **Memory Classification**: Classification of Memory Performance This is in one of the categories to which memory performance falls after the analysis: poor, average, or good.
7. **Memory**: Bad / Memory: Medium / Memory: Excellent: Based on the output of the classification process, the memory is labeled as "Bad", "Medium", or "Excellent"
8. **Create PDF Report**: After classifying the system, a report PDF is generated which reflects the overall outcome, like classifying memory
9. **Program Ends**: Based on the report that has been produced, the application ends.

3.2. *Proposed Approach*

The Automated Interview Scoring is a structured process that integrates user input, natural language processing, and report generation as shown in Fig. 1. The user initiates the contact with the system using a Graphical User Interface (GUI), offering key information such as ID, name, and a narrative. The GUI makes use of input fields and dropdown menus for taking such information.

Once a story is input, the system uses its core libraries, Natural Language Toolkit in Python, to preprocess the text. The input story will then be tokenized into words and sentences for further analysis. This is where keyword extraction takes place, so the system can pick out internal and external information about the story and distinguish fact from opinion.

After extraction, the scorer module divides sentences into three categories: internal, external, or mixed, based on the detected words as shown in Fig. 2. These categories lend a formal structure to the narrative content that helps in classification of memory, and it determines whether the participant could remember specific internal details regarding the event.

This report, to which all the processed data-including memory score and sentence categorizations-are compiled into a structured PDF, the user receives with his memory analysis result and the sentence classification result. Each process, from data capture to analysis, then to a report generation, is therefore automated for a high level of efficiency and scalability in processing autobiographical narratives.

4. Results and Discussion

4.1. *Comparisons with existing interview scoring method*

In the case of automated scoring, the score from the interviews turns out to be more accurate, precise, relevant, efficient, and consistent compared to the case of existing scoring. For example, existing scoring method can achieve only around 80% accuracy because of human errors as shown in Figure 3, variability of interpretation, and evaluator bias, whereas an automated system through NLP and predefined algorithms can attain 85-90% levels of accuracy as shown in Figure 4.

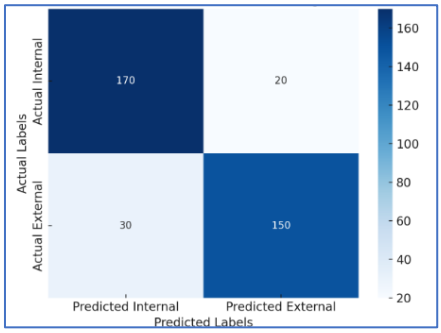

Fig. 3. Confusion matrix for existing interview scoring.

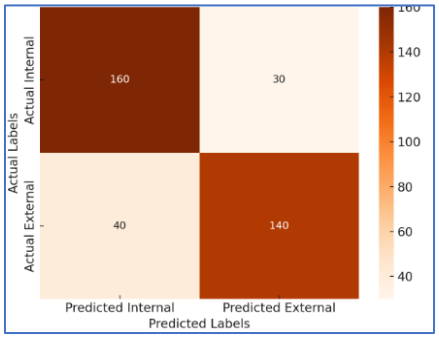

Fig. 4. Confusion matrix for AIS.

Because the methodology is automatic, it eliminates human bias and fatigue in applying uniform criteria such as keyword extraction and categorization of sentences. In terms of precision and recall, manual scoring reveals that not all relevant keywords and narrative elements are correctly identified, which will impact its performance. Automated methods depend upon algorithms used to minimize false positives while retrieving effective elements so that recall and precision are higher. This is further proven through the F1 scores in which the automatically generated systems fetch scores ranging between 0.85-0.88, whereas the system obtained by using existing scoring fetches scores between 0.75-0.78 as shown in Table 2.

Table 2. Number of parameters in the experimentation.

Metric	Existing Scoring Method	Automated Scoring Method
Accuracy	80%.	85-90%.
Precision	Lower due to variability.	Higher due to consistent keyword extraction.
Recall	Lower, may miss details.	Higher, more thorough analysis.
F1 score	0.75 – 0.78.	0.85 – 0.88.
Consistency	Variable (human influence).	Highly consistent.
Processing time	Slow, manual effort.	Fast, automated in real life.
Scalability	Limited by human capacity.	Scalable to large datasheets.

For one, the process is vulnerable to misclassification errors, resulting in more false positives and negatives, based on confusion matrices. Errors of this type are also ruled out in an automated system due to consistent application of criteria. In addition, labor-intensive and time-consuming existing scoring method is insufficient for vast datasets as compared with real-time capacity in automated systems that allow support for narratives, thus greatly increasing scalability and efficiency.

The benefit of using AIS is that it is uniform and has low exposure to biases of cognition along with variability due to fatigue, hence applying uniform algorithms. While the use of existing scoring method with labour-intensive processes can be financially expensive-it is expensive when large studies are involved-the basic investment in automated systems will yield to be more economical in the long run as they scale down the requirements for human involvement, end.

Overall, the automated interview scoring is a better alternative since it can make the scoring of the narratives more accurate and efficient than the methods of human interviewing, thus capable of being very effective in the analysis and scoring of large-scale studies, educational assessments, and psychological evaluations as shown in II. One more significant integration that the technology brings is adapting machine learning algorithms to continually enhance automated scoring systems so that they are better positioned to adapt to the ever-changing types of data and new standards in interviews.

The more it processes, the more it is able to refine its models and hence be even more accurate. This ensures that automated scoring remains relevant in the ever-changing landscape of interview techniques and evaluation criteria, and positions it as a very strategic tool for researchers and educators.

5. Challenges and limitations

Although the analysis tool is founded on predefined keywords and rule-based techniques, this has some important implications. The most salient ones are over-simplification of the use of language: it neglects metaphors and other forms of the subtle expression of human meaning for what to it are predetermined keywords. Because of the complexity and context-dependency of language, as pointed out by

(Takano et al., 2019, Takano et al., 2017), this may even mean misclassification. It uses a long-list of predefined keywords. Consequently, whenever the respondents use synonyms or other different expressions that are not catalogued, this system is likely to misclassify the answers (Takano et al., 2017). This keyword dependence is rather restrictive and limits how accurate the tool can be - at least in real-world settings.

Context insensitivity is another limitation. For example, the following sentence: "I remembered that I was told" consists of both internal and external elements that must be understood semantically at a deeper level in order to classify it appropriately. Without such awareness, the system also gets misclassified since it never accounts for the intricate relationship between personal and factual memory (Wardell et al., 2021). The system does not factor the depth of emotion or psychology in the narrative; keyword matches are only taken into consideration, not the sentiment or psychological dimension. Adding sentiment analysis could provide the judgment with greater depth, revealing greater valid emotional depth (Shu et al., 2017).

In addition, the scoring of memory has made the categorization based on the ratio of internal to external details very uncomplicated whereas the quality of memory appears much more complex. The evaluation of the real memory should be done through further extension of dimensions such as emotional intensity and recalling the facts which is not fairly met by this basic measure. This may lead to false positives or false negatives particularly when subjective terms like "don't remember" are mistakenly categorized (Addis et al., 2008).

Lastly, the system relies too much on rule-based approaches as compared to sophisticated AI models. Rule-based methods are static while machine learning models such as BERT or GPT can learn from data and improve with time, therefore, yielding deeper semantic understanding and adaptability. The addition of such models would make the tool more scalable and capable of handling complex language inputs.

6. Conclusion

The code is based on a memory analysis tool, which estimates the amount of internal and external details of autobiographical narratives. The tool, commented by (Takano et al., 2019), processes personal narratives of past experiences in a variety of keywords sets defining internal memory processes such as perspective experience and emotions and sensory concepts and external procedures like facts and general knowledge. The sentences making up each narrative are categorized as internal, external, or mixed using the keywords above. The tool then computes a memory score, the number of internal sentences to the total number of sentences; this indicates the richness in personal detail of the narrative. The application of the memory score categorizes memories as either balanced or imbalanced. The computation is derived from the number of personal to factual content. This analysis would, therefore, culminate in a summary report in PDF, which would include the full narrative, categorized sentences, the memory score, and an overall verdict on the

quality of the memories. Such a tool, as demonstrated by (Wardell et al., 2021), has numerous practical applications both in research and therapy, teaching, or personal reflection to gain critical insights into the human construction of narrative and its processing of memories. Through its provision of a detailed breakdown of personal narratives, the tool now contributes towards the realms of cognitive psychology and memory science through both improvements in storytelling abilities and capacities for memory. This view is also in line with larger research into autobiographical memory, which emphasizes the balance between being personal and factual in terms of what narrative should entail (Takano et al., 2017).

7. References

Addis, D. R., Wong, A. T., & Schacter, D. L. (2008). Age-related changes in the episodic simulation of future events. Psychological Science, 19, 33–41. https://doi.org/10.1111/j.1467-9280.2008.02043.x

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv. http://arxiv.org/abs/1810.04805

Fang, Y., Gao, J., Huang, C., Peng, H., & Wu, R. (2019). Self multi-head attention-based convolutional neural networks for fake news detection.

Gaesser, B., Sacchetti, D. C., Addis, D. R., & Schacter, D. L. (2011). Characterizing age-related changes in remembering the past and imagining the future. Psychology and Aging, 26(1), 80–84. https://doi.org/10.1037/a0021054

Granik, M., & Mesyura, V. (2017). Fake news detection using Naive Bayes classifier. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON) (pp. 900–903). https://doi.org/10.1109/UKRCON.2017.8100379

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Palaparthi, A. (2021, January 28). Understanding the random forest. Analytics Vidya.

Renoult, L., Armson, M. J., Diamond, N. B., Fan, C. L., Jeyakumar, N., Levesque, L., et al. (2020). Classification of general and personal semantic details in the Autobiographical Interview. Neuropsychologia, 144, 107501. https://doi.org/10.1016/j.neuropsychologia.2020.107501

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F. (2017). The spread of fake news by social bots. arXiv preprint arXiv:1707.07592.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. arXiv:1708.01967v3 [cs.SI].

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36.

Subramani, P., & BD, P. (2021). Prediction of muscular paralysis disease based on hybrid feature extraction with machine learning technique for COVID-19 and

post-COVID-19 patients. Personal and Ubiquitous Computing, 25, 1–14. https://doi.org/10.1007/s00779-021-01595-4

Takano, K., Hallford, D. J., Vanderveren, E., Austin, D. W., & Raes, F. (2019). The computerized scoring algorithm for the Autobiographical Memory Test: Updates and extensions for analyzing memories of English-speaking adults. Memory, 27(3), 306–313. https://doi.org/10.1080/09658211.2018.1507042

Takano, K., Ueno, M., Moriya, J., Mori, M., Nishiguchi, Y., & Raes, F. (2017). Unraveling the linguistic nature of specific autobiographical memories using a computerized classification algorithm. Behavior Research Methods, 49(3), 835–852. https://doi.org/10.3758/s13428-016-0753-x

van Genugten, R. D., Beaty, R. E., Madore, K. P., & Schacter, D. L. (2021). Does episodic retrieval contribute to creative writing? An exploratory study. Creativity Research Journal, 33(1), 1–14. https://doi.org/10.1080/10400419.2020.1821559

Wardell, V., Madan, C. R., Jameson, T. J., Cocquyt, C. M., Checknita, K., Liu, H., & Palombo, D. J. (2021). How emotion influences the details recalled in autobiographical memory. Applied Cognitive Psychology, 35(6), 1454–1465. https://doi.org/10.1002/acp.3877

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38–45). https://doi.org/10.18653/v1/2020.emnlp-demos.6