‚AD-AutoGPT: An Autonomous GPT for Alzheimer's Disease Infodemiology

Section 3 paper 6
Adit Rushil Potta

Summary

The paper presents a tool called AD-AutoGPT which can be used to conduct data collection, processing and analyze health narratives related to Alzheimer's Disease through user's textual prompts. The author's goal with the paper is to show the potential of AI in being able to understand complex health narratives, such as Alzheimer's Disease autonomously, allowing more work similar to this to be brought up in the future.

Alzheimer's Disease is a progressive neurodegenerative disorder, which is one of the most worked on disorders in the health industry. It has a large effect on the social, economic, and health systems throughout the world. To get a public understanding of this disease, professionals have to rely on work intensive methods such as web scraping and complex data pipelines to get information from the news, articles, and other forms of media. As data is ever increasing this can be difficult to manage and plan out into the future. It will also require a lot of technical expertise.

AutoGPT uses the capabilities of LLMS such as GPT-4 to automate and make operations autonomous. It simplifies complex data pipelines, allowing it to be a great tool to be used in this field. In the study the authors modify AutoGPT to public health applications and make AD-AutoGPT, being amongst the pioneers in integrating AutoGPT into the public health domain.

Takeaways

- Advancement of AD-AutoGPT
    - A new LLM-based tool, AD-AutoGPT, is centered around the automation of data collection, processing, and analysis pipelines based on user prompts.

- Overcoming AutoGPT's Limitations: AD-AutoGPT improves efficiency through:
    - Specific prompts for accurate data retrieval
    - Tailored spatiotemporal data extraction.
    - Enhanced text summarization.
    - In-depth summary analysis.

    - Dynamic visualization capabilities.
    - AD-AutoGPT shifts from labor-intensive methods to automated, prompt-based frameworks, enabling efficient analysis of Alzheimer's Disease-related content.

- Case Study and Insights: The tool provides trend analysis, intertopic mapping, and key term identification across four AD-related sources, hence adding further clarity to public health discourse.

With the rise in use of LLMS, originating from transformer based models, models such as BERT, GPT and many others have evolved the field of natural language processing. Going from RNN's to LSTM's to Attention and transformers, a lot of progress has been made. These models have been made to learn the context of the input at a high level, and are being used in various fields such as question answering and information extraction, sentiment analysis, and text generation. Now as of late they are quickly being incorporated into various fields, such as in the medical domain, finance, education, agriculture, etc.

Public health infodemiology is the field that studies the determinants and distribution of information on the internet or within a large group of people, with the objective being informing public health and public policies. This considers how information is being passed along due to trends, news and other sources. Relating search behavior with Alzeihmer's Disease, understanding online behaviors and interests via infodemiology can help improve the public's awareness and education regarding this pressing disease. This can help prevent the spread of misinformation and can inform individuals properly and can aid in the management strategies for the disease.

To build upon this, AutoGPT is used, as it builds on the success of large language models, and takes automation a step further by providing a more user friendly interface for non expert users, allowing it to be more approachable and such. The model can complete complex tasks such as data collection, cleaning, analysis, and even generate human-like text, through simple user prompts. It has also been shown to be able to analyze and understand datasets well.

It has the ability to acquire specialized information quickly and precisely. The authors integrated a specific prompting mechanism to AD-AutoGPT, allowing them to gather more relevant information pertaining to Alzeimer's. It also addresses the challenge AutoGPT faces when it is extracting critical details regarding the time and place of news events from articles accurately. Using web-crawling scripts to extract these timestamps and information regarding the location, AD-AutoGPT overcomes these bottlenecks. AD-AutoGPT also avoids the limitation of AutoGPT that has to provide text in chunks in order to avoid token limits and produce full summaries. The summary study is supported by the extraction of the most important themes by Latent Dirichlet Allocation . Moreover, dynamic visualization tools are included to plot news

trends, locations, and keyword evolution across time. These domain-specific enhancements make AD-AutoGPT faster, more efficient, and quite relevant to public health research.

Method:

- First, AD-AutoGPT breaks down the user's request into small processes, while within the LangChain framework, an interface with the GPT-4 or ChatGPT API is used to perform tasks.
- Find and Save News: used to cover relevant articles from Google API. It has GPT-4 summarizing the major points it extracts from articles.
- Extract Spatio-Temporal Data: Gather spatiotemporal data using geoparsing techniques to extract dates and locations from the news articles.
- Depending on the query, the tools can be executed in parallel or sequentially:
    - The articles get snagged and then cached.
    - GPT-4 generates summaries without a cap on tokens by breaking up the text into chunks
    - Data extracted for further analysis consists of spatiotemporal data.

- Topic identification: AD-AutoGPT employs Latent Dirichlet Allocation (LDA) for topic modeling, identifying emerging themes/keywords across news sources.

- Representations include:
    - Trend graphs that show frequencies of articles over time.
    - Maps of where significant events have occurred.

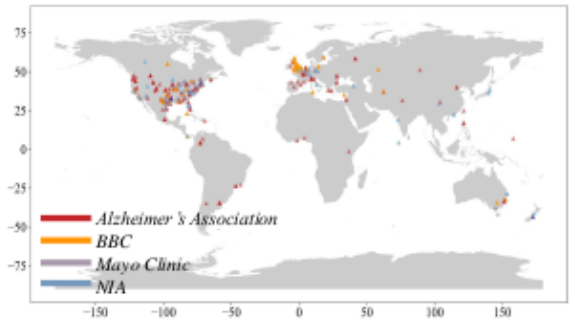
- AD-AutoGPT provides verification of an adequately answered query:
    - It will re-evaluate tasks or tools if any extra refinement is required.

- Final outputs include text summaries, visualizations, and insights, thereby automating the entire workflow to achieve efficient and autonomous results.

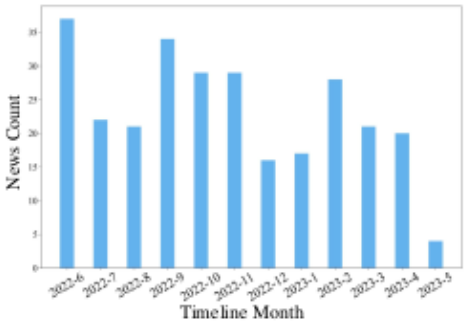
Case Study:

The tool was tested on data provided by authoritative websites reporting ALzheimer's disease, by using the prompt (Can you help me to know something new about Alzheimer's Disease and maybe draw some plots for me?). The authors collected 277 pieces of news in total from websites such as BBC, Alzheimer's Association, National Institute of Aging, and the Mayo Clinic. The time and process, location of the news was also extracted and saved. The tool

automatically creates a pipeline based on the toolsets present in the instruction library without anyone telling it to.

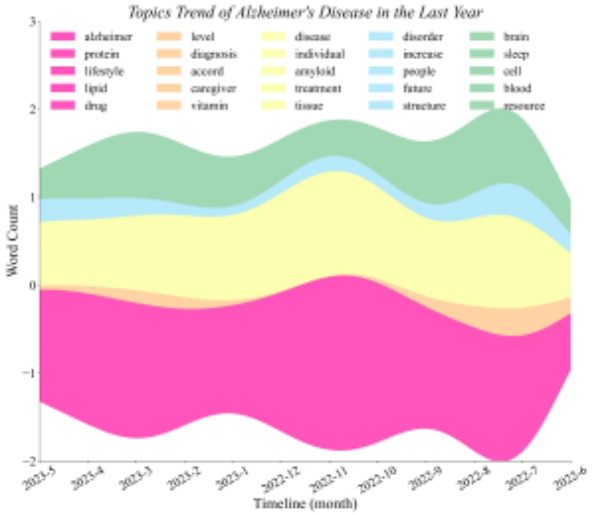


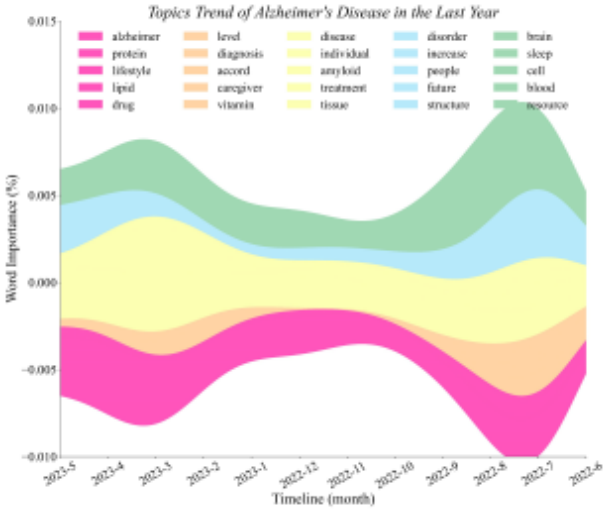
(a) Places where the latest news about Alzheimer's diseases happened.



(b) The number of news collected for each month from June 2022 to May 2023.

The spatial locations of the places extracted from all the news articles were visualized above, and this is automatically generated using AD-AutoGPT, in the first figure. In the second figure, we can see temporal data analysis, where the number of news reports about Alzheimer's Disease in each month of the past year are visualized. The main idea of the purpose of the technology being is to improve the efficiency of researcher's work





Similarly, based on the LDA topic modeling, you can see that hot topic analysis is automatically conducted by AD-AutoGPT. Here the top 5 words with the most occurrences for each of the 5 hot topics are chosen, and it draws stream graphs according to the number of occurrences and word weights of the given words. By doing this we can get the most popular topics in the news and such. The users will no longer need to really extensively on the news.

Scientific Question the paper is trying to solve:

How LLMs and other technologies can help in understanding and awareness of Alzheimer's disease due to its substantial social, economic and health system implications. How it can potentially improve the overall throughput of researchers working in the domain of medical infodemiology, as well to gain insights from the data that is being processed.

In the medical domain, processing data and gathering it is often a very daunting task, requiring a lot of resources and time. For the given modality, in order to gain insight, the internet must be scrapped and analyzed. The amount and quality of the insight that can be gathered also depends on the tools at disposal. There is a gap in the comprehensive data-driven public understanding of complex health narratives. Health professionals would have to API data collection, data post processing, and analysis to get insights, health reports, and other textual sources. As the scale of the global data presents a large challenge, an innovative approach is necessary to streamline these given processes and to extract valuable, actionable insights automatically and efficiently. The lack of technical expertise also poses a problem.

What solution does the paper propose to address this question?

The solution the paper is proposing to mitigate this problem is to build upon AutoGPT which is an open-source application that harnesses the capabilities to automate and optimize many processes. It can understand linguistics and ability to simplify complex data pipelines, it can understand a large amount of data well quickly and efficiently. By modifying this technology for the medical health domain, to analyze news sources pertaining to Alzheimer's. By making an LLM-based tool it can generate data collection, processing and analysis pipelines autonomously based on user prompts. It can improve on the limitations present in AutoGPT, which does not do well on Alzheimer's infodemiology, where it is improved by using specific prompting mechanisms to improve the accuracy of the information that is retrieved. It also uses tailored spatiotemporal information, an improved text summarization ability, and an in depth analysis on the generated text summaries, as well as effective and dynamic visualization. The main idea being able to help make the labor intensive tasks of data collection, processing and analysis into a prompt based automated and simple workflow. On top of that, after collecting and processing this data, being able to provide detailed case studies and trend analysis of various topics, as well as interpreting temporal and spatial information to see how information is spreading and at what rate, which can potentially help in mitigating misinformation.

Potential Improvements:

One of the issues that arise is that these models generate the output based on their given training data, and if it is biased in any manner, it can give outputs that could potentially contain these biases and misinformation. There are ethical concerns regarding artificial intelligence in general as well, which can result in the mistrust of users with this technology as well. Health information is very sensitive information that could be potentially misused, and these tools can potentially leak this private information, which can result in a lot of legal and ethical concerns.

Misuse is also a big factor is one of the issues, as individuals or organizations can use it to target certain regions or certain individuals with the insights to provide the wrong information, or regarding their overall treatment and such. Regulating their appropriate use and managing potential misuse is important.

Potential improvements are to have robust guidelines and moderators who are there to regulate and keep an eye on things. There must be clear user instructions and warnings about the potential pitfalls and such.

Regarding the AD (Alzheimer's Disease) aspect of the improvements, regarding the domain related information where categorizing various pathologies and subtypes of the disease can be challenging, and expanding the dataset can also pose a challenge.

The tool can also be extended to many other modalities so that it can help in many more fields as well. Making it a tool like perplexity would also prove to be potentially useful as it will be in a simple form factor and be more approachable to use. The data visualization aspects can be improved, and much like the search engine perplexity the information that is gathered and generated to the individual can be cited back to the user so that they can see where and how the information came, so that they are more willing to trust the tool overall. Taking in user feedback and having people rate the messages and outputs will also help in improving the model overall. Including more sources of information would also help a lot, as well its integration into various platforms such as social media or any given search engine could help as well.

More features regarding the visualization part can be improved, which would pertain to the generation aspect of the tool. More powerful models can be used. On top of this, user's should also be able to input their own medical information or reports so that they can potentially find help, or more information and help regarding their given situation. The model should also be trained in such a way that the user should be assumed to be an unknowing individual so that no information is kept away from the user and such.