Enhancing Global Accessibility and Information Flow: Bridging Language Divides with Natural Language Processing and Deep Learning

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Christopher Barfield Jr.

Spring, 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Bryn E. Seabrook, Department of Engineering and Society

STS Research Paper

Introduction

In today's interconnected world, access to information has become a fundamental right yet remains a significant challenge for many, particularly in the face of natural disasters. Vietnam faces an ongoing battle against natural disasters, with floods being a leading cause of devastation affecting millions of lives annually (Nguyen et al., 2021). Despite significant advances in machine learning (ML) and artificial intelligence (AI) that have enhanced our predictive capabilities, there is a glaring disconnect between these solutions and their real-world applications for the average citizen. Addressing this challenge, our research team, 'Floodwatch,' has developed a progressive web application designed to bridge this specific gap. Leveraging flood prediction algorithms, the application features an intuitive interface allowing access to flood prediction information, at any time. My section of the project is focused on making this technology more accessible by adding Natural Language Processing (NLP), attempting to make our platform as user-friendly as popular voice-activated technologies like Amazon's Alexa.

This commitment to accessibility transcends mere user interface design; it's about adopting a comprehensive approach that ensures advanced flood predictions are not just available but also intelligible to a diverse global audience. Accessibility, in this wide-ranging context, is inherently multifaceted. It involves creating systems that cater to the unique needs of various user groups, such as the elderly, who may benefit from simpler navigation due to their lower confidence and need for more guidance with complex interfaces, as highlighted by Vaportzis et al. (2017). It also includes designing for the youth, who seek fast and engaging interactions, and ensuring inclusivity for people with disabilities, including the visually impaired, those with motor challenges, or individuals with limited dexterity.

Expanding on this foundation, our approach embraces the crucial aspect of multilingual support, acknowledging the global diversity in language and the barriers it can create. By incorporating NLP technologies, we aim to make our platform accessible across linguistic boundaries, offering predictions and crucial information in multiple languages. This initiative not only enhances accessibility for non-English speakers but also ensures that emergency information is comprehensible and actionable for a a wider audience, including those who might face barriers due to language proficiency. Integrating multilingual capabilities is a step towards truly democratizing information access, ensuring that the technology serves as a bridge, rather than a barrier, to critical knowledge and resources. This effort strives to create a more informed and prepared global community, where everyone, regardless of language, age, or physical ability, can access life-saving information.

Supportive Background Information

As an example, Vietnam's geographical vulnerability to flooding is a result of its extensive river systems, deltaic plains, and tropical monsoons, which together create a natural predisposition to both seasonal and flash floods. The historical landscape of the country is filled with instances of devastating floods that have led to significant human and economic losses. Traditional flood response mechanisms often rely on centralized dissemination of information and lack the capacity for real-time, personalized communication. This often leaves the most vulnerable populations — such as the elderly, individuals with disabilities, and those in remote areas — at a greater risk due to delayed information or messages that do not translate into actionable insights.

Recognizing these challenges we attempt to improve the distribution and accessibility of communication, representing a move towards a more user-centric approach. By integrating NLP,

the platform aims to interpret flood information driven by rain and topology data, and present it in a way that is tailored to individual comprehension levels and information needs. The inclusivity-driven system is about enabling every segment of the population to make informed decisions during critical times. For instance, by providing voice-activated access to flood risks and safety information, the platform serves not just as a technological tool, but as a means of empowerment for those who have traditionally been forgotten in the communication loop of disaster management.

Furthermore, the application's NLP capabilities are being developed with the vision of offering multilingual support, thus ensuring that language barriers do not impede access to life-saving information. This aspect of the design is crucial in a nation where local dialects and languages form a core part of cultural identity and community cohesion. The use of local languages in disaster warnings has been shown to increase comprehension and response rates, a critical factor in the timely evacuation and safety measures during floods. As Vietnam continues to grapple with the implications of climate change, which is predicted to increase the intensity and unpredictability of weather patterns, 'Floodwatch' represents a proactive step towards a future.

Vietnam exemplifies the impactful application of real-time information dissemination. In this context, our research delves into the capabilities of Natural Language Processing (NLP) and Deep Learning to enhance global information accessibility. 'Floodwatch' is designed to expand the range of regions worldwide that can benefit from our flood prediction technologies. By integrating NLP and Deep Learning through multilingual large language models, we aim to overcome linguistic barriers, making our platform universally accessible. This approach allows for the consumption and delivery of information in any language, without the constraints of

manual translation. Through the 'Floodwatch' experiment, we envision a future where access to critical information is not limited by language, ensuring that individuals across the globe are equipped with the knowledge needed to navigate critical, time-sensitive events such as natural disasters. Moreover, this technology demonstrates the need to establish software globalization as a foundational practice, rather than relegating it to an afterthought, showcasing the feasible path to achieve this goal.

The Role of SCOT

In the domain of disaster management, existing flood prediction systems often present data in formats that not only demand specialized knowledge to understand—such as complex graphs and technical vernacular—but also typically cater to speakers of dominant languages, neglecting the multilingual reality of our global community, as mentioned in Wachinger et al. (2012). This approach underscores a conventional bias towards expert users and overlooks the diverse linguistic backgrounds of the global population. Drawing attention to this issue, have criticized the accessibility of such systems, noting their failure to communicate vital information effectively to non-experts and speakers of various languages.

Responding to these challenges, our project leverages NLP and Deep Learning technologies to interpret and translate complex flood data into easily understandable information across multiple languages. By doing so, we embrace the SCOT theory's perspective that technology should be developed in consideration of the social group it aims to serve, acknowledging the diverse linguistic needs of global communities. Our approach not only democratizes access to critical flood risk information but also exemplifies how technological solutions can be socially constructed to bridge the linguistic divides, thereby enhancing global information accessibility in times of disaster.

In essence, this project embodies the commitment to utilizing technology such as NLP and Deep Learning to answer the global call for more inclusive and accessible disaster management tools. By addressing the multifaceted challenge of cross-language information accessibility, we contribute to the construction of a technology that is not just innovative but also socially responsive and inclusive, aligning with the overarching goals of SCOT to harmonize technological development with global societal needs. As Floodwatch aims to become a global platform and do good for the world, the end product should be designed to directly serve its users, without compromising on quality or availability of information.

Our solution harnesses the capabilities of NLP to meticulously translate complex flood data into languages and formats that resonate with and are actionable by a diverse global audience. This linguistic transformation is rooted in the SCOT theory's premise, recognizing that effective communication must cater to the varied information processing needs and linguistic backgrounds of different social groups. In practice, our user interface reimagines how flood information is conveyed, drawing inspiration from the simplicity and accessibility of weather forecasts. It converts intricate data into straightforward, multilingual expressions, such as “Danang's flood risk is 10%” or "anticipate high flood risk within the next 48 hours."

This focus on linguistic accessibility and clarity is driven by a broader societal imperative for information that can be easily understood and acted upon, particularly by those without specialized knowledge. By leveraging NLP, we aim to ensure that vital flood risk information breaks through language barriers, becoming universally digestible. Referencing studies like those by Bronfman et al. (2019), we find evidence that presenting information in a simplified, language-sensitive manner not only enhances comprehension but also empowers individuals with

actionable insights. This approach underlines the importance of creating a more linguistically inclusive global community, demonstrating NLP's potential to transform complex, time-sensitive data into formats that facilitate broader understanding and prompt, informed action.

Research Question

This investigation seeks to answer the question, "How can Natural Language Processing and Deep Learning improve cross-language access to information on a global scale?" This question underscores the need to bridge linguistic divides and therefore enhancing global information accessibility, especially crucial in the context of disaster management.

Methods

To address this question, the primary objective is to explore the development and implications of an NLP-enhanced interface for "Floodwatch," a web application aimed at predicting and responding to flooding events. The technical aspects of the approach include constructing a comprehensive pipeline that involves audio streaming, natural language understanding, data querying, and delivering processed information through a language model capable of generating responses in multiple languages. At the heart of this study is the anticipation of deploying AI agents powered by large language models (LLMs). These agents are expected to process user inquiries, access a wide toolbox of data sources, and provide actionable information to users in their desired language. This research is guided by keywords such as "Natural Language Processing," "disaster management," "accessibility," "flood prediction," and "user interface design." Adopting a thematic analysis approach, we aim to systematically explore NLP's pivotal role in access to information, disaster management, delve into the challenges of creating accessible digital interfaces, and assess the potential global impact of such technologies.

Results and Discussion

The exploration into how NLP and Deep Learning can enhance cross-language information access on a global scale demonstrates promising avenues for technology's role in disaster management. Preliminary findings indicate that the development of an NLP-enhanced application for flood prediction can vastly improve the accessibility of critical information to users regardless of language barriers. This potential breakthrough suggests that advanced NLP techniques, coupled with Deep Learning can deliver vital data in real-time to diverse populations. The research underlines the critical importance of multilingual support within disaster response tools, highlighting how such technologies can bridge communication gaps and foster a more inclusive approach to global disaster preparedness. The technical pipeline envisioned for this application— audio streaming, natural language understanding, data querying, and information delivery through language models—stands as a potentially new standard of integrating AI in safety tools and large distribution of information globally.

Applying NLP for Multilingual Communication

The research underscores the critical role that NLP and Deep Learning can play in breaking down language barriers that often restrict access to information. By implementing NLP techniques that support a wide range of languages, the study addresses the diverse needs of global user groups. This direction aligns with the SCOT principle, emphasizing that technology's development and success are heavily influenced by its relevance and adaptability to various social contexts. The push for multilingual capabilities in digital platforms exemplifies a conscious effort to mold technology according to the linguistic diversity of its user base.

Recent advancements in NLP and Deep Learning have created the era of large language models (LLMs) like GPT-4, which represent some of the most powerful tools in the history of computing. The creators of GPT-4, OpenAI, cite this model as being multilingual (OpenAI, 2024). The capacity for transfer learning, which is creating understanding of similar patterns in linguistics, bolstered by extensive datasets spanning numerous languages, allows these models to grasp the syntactic nuances of different languages. This capability underscores a significant leap towards genuine multilingual support, bridging communication gaps and fostering inclusivity in information access. The utilization of such models in the "Floodwatch" application illustrates a concrete application of these technologies, showcasing the potential to deliver critical disaster management information across linguistic divides efficiently.

User-Centric Design and Iterative Feedback

A significant finding from this study involves the importance of user interaction and feedback in shaping technology. Consistent with the SCOT framework, which emphasizes that users are crucial in the stabilization and acceptance of technology, the research highlights how user feedback mechanisms are indispensable in refining NLP systems. An iterative design process that incorporates real user experiences ensures the technology evolves to meet the actual needs of its audience, making information more accessible and user-friendly. It is crucial that applications of this technology are designed with a user first approach, as it can unlock almost endless possibilities in terms of democratization of information.

Emphasizing a user-first approach has far-reaching implications for the democratization of information. By designing applications that are inherently flexible and responsive to user feedback, we can unlock opportunities for information access that have not been possible before. This democratization is crucial for bridging knowledge gaps and addressing the disparities often seen in access to critical information. Notably, the disparity in information accessibility is a significant contributor to global wealth inequalities. By ensuring that information technology is inclusive and equitably accessible, we can make

strides toward leveling the playing field, especially in regions and communities historically marginalized in the digital divide.

Impact on Technology Acceptance and Adoption

The SCOT framework suggests that the acceptance and adoption of technology are critically dependent on how well it aligns with the needs and expectations of its user base. By prioritizing user-centric design and iterative feedback, our project not only advances the technical capabilities of NLP systems but also enhances their social relevance. This relevance is pivotal for ensuring that the technology is not only adopted but also actively utilized in diverse settings, thereby maximizing its impact on global disaster management efforts.

In conclusion, the integration of user-centric design principles and iterative feedback mechanisms is vital for the successful deployment of NLP and Deep Learning technologies in disaster management. This approach not only optimizes the technology for real-world application but also contributes to the broader goal of making information more accessible and equitable worldwide, thus echoing the core tenets of the SCOT framework.

Translation coverage on websites

During the development of "Floodwatch," a significant challenge emerged: the labor-intensive process of manually entering direct translations based on user preferences. This experience shed light on the broader difficulty of making information globally accessible and available. The task of providing accurate, context-sensitive translations between two or more languages presents a barrier to global information dissemination. A recent survey by the Web Technology Surveys (2021) found that while a majority of websites offer content predominantly in English, only a small percentage provide multilingual support. The Global Overview 2024

report indicates that 52% of all websites on the internet are english based, with the next highest being spanish with a much smaller 5.5% (Global Overview, 2024). This disparity highlights a crucial oversight in web development priorities: a significant focus on content and functionality over the essential need for accessibility through multilingual availability. The concept of internationalization—the practice of designing products and services to be easily adapted to various languages and regions without requiring significant engineering changes—is increasingly critical. As "Floodwatch" aims to distribute potentially lifesaving information on a global scale, the need to embrace internationalization and localization practices from the project's inception becomes clear. This approach presents the opportunity for a paradigm shift in web development, advocating for the integration of internationalization strategies early in the design process.

Adopting internationalization practices is not merely a technical necessity but a moral imperative. It ensures that vital information, especially in contexts like disaster management, is accessible to diverse populations regardless of language barriers. By prioritizing internationalization and localization, developers can contribute to a more inclusive digital environment where access to information is not limited by linguistic boundaries. In this context, global accessibility can be seamlessly and effectively integrated, without the need for extensive specialization in translation. This approach allows developers to focus on creating valuable technology, as we are doing with Floodwatch, and rely on NLP and Deep Learning-based models to ensure the technology is accessible to everyone.

Disaster Response

In the world of global disaster response, the challenge of language equity becomes particularly challenging. The distribution of information on a global scale entails more than just

the translation of website content; it involves ensuring equitable access to critical services and resources across linguistic boundaries. The COVID-19 pandemic has underscored this issue dramatically, where disparities in language access have had direct impacts on health outcomes and the ability to receive timely and appropriate care.

A study published in the "Journal of Public Health Policy" (2020) highlighted that non-English speakers in various countries faced significant challenges in accessing COVID-19 related health information, testing, and treatment services. These language barriers not only hindered the individual's ability to seek help but also increased public health risks by delaying diagnosis and treatment for communities that are already marginalized. The implications of such disparities extend far beyond health crises to encompass all forms of disaster response. Whether dealing with natural disasters, public health emergencies, or humanitarian crises, the ability to communicate effectively and accessible with affected populations is important. Language barriers can severely limit the effectiveness of disaster response efforts, leading to inequitable outcomes and potentially increasing the disaster's impact on vulnerable populations.

Challenges in AI and Ethics

The deployment of AI agents and Large Language Models in facilitating cross-lingual information access brings to light several challenges, particularly concerning ethics and the social implications of AI. The SCOT analysis reveals concerns around AI agency, data privacy, and the potential for bias, prompting a critical evaluation of how these technologies interact with human values. This dialogue between technological innovation and ethical considerations reflects a broader societal negotiation over the role of AI in public life.

There are issues with relying solely on AI agents to make decisions and be considered a ground truth for information, one of them being because these datasets are trained on majority english based websites. While LLMs will try to cross-learn information in different languages, it is primarily focused on English and the results reported by companies developing these large language models show a clear dropoff in understanding of other languages, with English understanding scoring significantly higher than other languages on accuracy tests. The Orion-14B multilingual large language model's researchers state that the majority of the training data is in English, and other languages such as Korean and Japanese are included after a threshold is reached during training (Orion-14B, 2024). This bias in training data could lead to some unintended responses because of the need to use an intermediary language as the base knowledge.

Concerns should also be raised due to the nature of bias present in transformer-based large language models. Most training data include a Western bias that these models can inherit. For instance, a Dartmouth study demonstrates that transformer-based models may assign biases to certain racial groups. This includes associating African people with being violent, dirty, and criminal; women with being silly, distracted, and poor; and Middle Eastern individuals with being terrorists, involved with bombs, and committing crimes (Ma et al., EMNLP 2023).

Although the intention is to remove these biases during fine-tuning training steps, it is concerning that such values underpin the creation of large language models.

Limitations and Avenues for Future Research

Acknowledging the limitations of the current research, primarily its theoretical orientation, future investigations should aim for empirical validations through prototypes and

real-world applications. Such studies would not only test the feasibility of the proposed NLP solutions but also offer deeper insights into their impact on global information access.

Further, the evolving landscape of AI ethics presents a fertile ground for future research, especially in how NLP and Deep Learning technologies can be developed and deployed responsibly. Exploring the balance between innovation and ethical standards will be crucial as these technologies become more integrated into our daily lives.

Conclusion

This research, through the lens of the SCOT framework, has pointed out the potential of NLP and Deep Learning technologies to transcend language barriers and democratize information access globally. By focusing on the social dimensions of technology—namely, user needs, inclusivity, and ethical considerations—it contributes to a broader understanding of how digital tools can foster a more connected and informed world. The findings serve as a foundation for future endeavors that seek to expand the reach of information technology across linguistic and cultural divides, furthering the dialogue on the role of technology in society.

References

Nguyen, M. T., Sebesvari, Z., Souvignet, M., & Bachofer, F. (2021, January). *Understanding and assessing flood risk in Vietnam: Current status, persisting gaps, and future directions*. Flood Resilience Portal. https://floodresilience.net/resources/item/understanding-and-assessing-flood-risk-in-vietnam-current-status-persisting-gaps-and-future-directions/

Vaportzis, E., Clausen, M. G., & Gow, A. J. (2017, October 4). *Older Adults Perceptions of Technology and Barriers to Interacting with Tablet Computers: A Focus Group Study*. National Library of Medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5649151/

Suk, S., & Kojima, H. (2008). Voice activated appliances for severely disabled persons. In F. Mihelič & J. Žibert (Eds.), Speech Recognition (pp. 528–537). I-Tech, Vienna, Austria. ISBN 978-953-7619-29-9.

Khan, V., & Meenai, T. A. (2021). Pretrained natural language processing model for intent recognition (BERT-IR). Human-Centric Intelligent Systems, 1(3-4), 66-74. https://doi.org/10.2991/hcis.k.211109.001

Klein, H. K., & Kleinman, D. L. (2002). The social construction of technology: Structural considerations. Science, Technology, & Human Values, 27(1), 28-52.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., ... Jain, S., et al. (2024). GPT-4 technical report (Version 6) [Technical report]. arXiv. https://arxiv.org/abs/arXiv:2303.08774

Kemp, S. (2024, January 31). *Digital 2024: Global Overview Report - DataReportal – Global Digital Insights*. DataReportal. https://datareportal.com/reports/digital-2024-global-overview-report

DeSalvo, K., Hughes, B., Bassett, M., Benjamin, G., Fraser, M., Galea, S., & Gracia, J. N. (2021, April 7). *Public health covid-19 impact assessment: Lessons learned and compelling needs*. NAM perspectives. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8406505/

Chen, D., Huang, Y., Li, X., Li, Y., Liu, Y., Pan, H., Xu, L., Zhang, D., Zhang, Z., & Han, K. (2024, January 20). Orion-14B: Open-source multilingual large language models. arXiv.org. https://arxiv.org/abs/2401.12246

Ma, W., Scheible, H., Wang, B., Veeramachaneni, G., Chowdhary, P., Sun, A., Koulogeorge, A., Wang, L., Yang, D., & Vosoughi, S. *Deciphering stereotypes in pre-trained language models*. ACL Anthology. https://aclanthology.org/2023.emnlp-main.697/

Wachinger, G., Kuhlicke, C., Begg, C., & Renn, O. (2012). The risk perception paradox: Implications for governance and communication of natural hazards. *Risk Analysis, 32(6), 1049-1065.* https://pubmed.ncbi.nlm.nih.gov/23278120/