PROJECT PROPOSAL

TEXT SUMMARIZATION FOR WEATHER FORECASTING USING MACHINE LEARNING

A graduate project proposed in partial fulfillment of the requirements

For the degree of Master of Science in Software Engineering

By

Sai Venkata Shreyas Bandaru

Student ID : 203140710

E-mail: sai-venkata-shreyas.bandaru.785@my.csun.edu

Katya Mkrtchyan, Committee Chair

Date 12th September 2023

Background of the study

Our everyday decisions, activities, and security are all profoundly impacted by weather forecasts. Reliable forecasts are essential for determining when it is best to leave on a vacation, what to pack for the trip, or how to best prepare for a weather-related emergency (Xu et al., 2021). However, the large amounts of data and specialized vocabulary used in weather forecasting provide substantial obstacles for those looking to break into the area. The average person has struggled greatly to make sense of something so intricate. Accurate weather predictions are only possible because of the wide variety of instruments, satellites, and sensors used by scientists (Schultz et al., 2021). The air's temperature, humidity, kinetic energy, and pressure are all considered. Mostajabi et al.'s (2019) findings can aid in the forecasting of a wide range of weather phenomena, including precipitation, temperature trends, storm systems, and more. Weather predictions sometimes employ scientific language and specialized vocabulary. Words like "low-pressure system," "frontal boundary," and "isobar" may seem foreign to listeners who aren't meteorologists, according to Hartnett's (2019) research. There is a great potential for misunderstanding when it comes to meteorological data because of its complexity and the requirement for specialized jargon. Inadequate knowledge of weather forecasts may lead to poor decision making, such as skipping precautions during extreme weather events or failing to account for weather-related dangers while organizing a trip (Bruine et al., 2021). Access to reliable weather forecasts is critical to the public's health and safety. Hurricanes, tornadoes, floods, and wildfires are just some of the natural calamities that they help individuals and communities prepare for and respond to. Providing accurate and timely information, say Ye et al. (2020), will help reduce casualties and property damage. This information access issue is especially pressing for the economically disadvantaged, who may lack both the means and the

education to make sense of complex weather forecasts. Disseminating accessible and clear weather information is crucial for promoting inclusion and protecting the public's welfare. There is a disconnect between the large amount of sophisticated meteorological data available to meteorologists and the public's capacity to interpret and effectively utilize this knowledge, despite substantial breakthroughs in weather forecasting. This study intends to fill this void by creating a text summary instrument that can simplify detailed weather forecasts. The ultimate goal is to improve public safety, increase accessibility, and increase knowledge of weather trends.

2.1 Aim

To help people make educated decisions based on weather forecasts, this project aims to develop a powerful text summary tool that can take complex meteorological data and provide easy-to understand summaries.

2.2 Objectives

The information we collect will serve as the backbone of our investigation. Careful attention to detail will be given to the task of compiling a database containing both current and historical meteorological data. Using a sizable data set, machine learning models will be trained and fine tuned. Now that we have all of the information we need, we can begin preparing the text. The raw weather data needs to be cleaned, formatted, and organized before it can be analyzed. Throughout the model's development, the data will be cleaned and organized to maximize its use by machine learning algorithms. Then, we'll get down to brass tacks and focus on the meat of the job at hand: creating and honing machine learning models tailored to the task of text summarization. The goal is to develop algorithms that can efficiently glean and summarize

crucial meteorological data from massive and complex amounts of data. The success of our initiative relies heavily on the accuracy of these models. Following the development of the summary models, our focus will shift to the design of an intuitive interface. The proposed user interface would streamline the process of submitting weather-related inquiries or data and providing clear, concise responses. The interface's intuitiveness is crucial in determining whether or not it will be widely used. Further, testing is required to ensure the device's reliability and precision. We plan to use extensive testing and crowdsourced feedback to determine how efficient our summary generator is. Important for making changes to the models and UI based on actual user feedback. In conclusion, widespread adoption of the summary tool is anticipated. The third step, "deployment," will involve making the technology accessible to the general public. Our mission to improve the usability and accessibility of weather forecasts will be complete when the tool is available to users across a variety of web and mobile platforms. By simplifying difficult meteorological data and scientific terminology, the aforementioned goals will ultimately aid individuals and communities in making better decisions in response to weather conditions.

3. Problem Statement

At the heart of this effort is the intricacy of weather forecasting and the challenges it presents to the general public. Today's weather forecasts might be tough to decipher due to technical jargon, excessive detail, and abstract scientific concepts (Halegoua, 2020). Many factors contribute to its inaccessibility. The crux of this complexity is that it could result in erroneous interpretations of vital meteorological data. People may make bad decisions with unanticipated effects if they struggle to interpret complex weather predictions. To give just one example, people may be unprepared for extreme weather like thunderstorms, floods, and heat waves if they don't fully understand the dangers involved. Similarly, they can wreak havoc on travel plans, outdoor

activities, and other events. Poor decision-making can have far-reaching consequences for businesses and industries that rely on weather-sensitive operations, such as agriculture, transportation, and tourism. Our project's goal is to develop a text summary instrument to address this complex issue. The intention of this software is to simplify complex weather forecasts for the general public. This allows people to make better informed decisions, which boosts their safety, health, and quality of life. The inaccessibility of weather forecasts and their potential effect on public decision-making and outcomes drives our work.

4. Technical Approach

Data Collection: Our technical strategy begins with gathering relevant information. In order to train our machine learning models accurately, it is essential to acknowledge the importance of a complete and varied dataset. To do this, we will compile meteorological data from a wide variety of reputable sources, such as national weather services, private weather forecasting firms, and archives of past weather records. Expect to see descriptions of the weather, numbers representing things like the temperature and humidity, and other location-specific information, like latitude and longitude, in these reports. Our summarization tool's efficacy in dealing with a wide variety of weather-related data relies heavily on the depth and breadth of our dataset. Text Preprocessing: After the information has been collected, the text will be carefully prepared. In order to conduct analysis and machine learning procedures, raw data must first be prepared. In text preprocessing, "data cleaning" refers to the act of removing any unnecessary or irrelevant information from the text. Tokenization will be performed on the text so that individual words can be studied more closely. For the sake of summarization, we will also do feature extraction to help us get at the meat of the data. At this point, you should focus on creating a properly annotated training dataset. Human-generated summaries are needed to improve the raw data

collected from weather reports. Using the available training data, the machine learning algorithms used in this study will learn how to synthesize weather reports. Machine Learning Models: Building machine learning models with the ability to effectively summarize weather data is the main focus of our work. Machine learning (ML) algorithms and natural language processing (NLP) techniques will be employed to complete the task. The proposed methods will employ both extractive and abstractive summarizing techniques. Selecting and reducing the most crucial sentences or phrases from the original text is what extractive summarization is all about. Aiming to provide concise summaries presented in a form that is more akin to human language, abstractive summarizing may involve paraphrasing and rephrasing the original material. We will also investigate the potential applications of deep learning models, which have demonstrated significant progress in NLP problems. We will also investigate the possibility of using transfer learning from pre-trained language models in order to make advantage of the information contained within the huge language models developed by the natural language processing (NLP) research community. In order to improve the accuracy of summarization, this study will compare and contrast state-of-the-art natural language processing (NLP) models with more traditional machine learning methods including Logistic Regression, Random Forest, and Support Vector Machines (SVM).

5. Project Risk

Data Quality: We have serious concerns about the accuracy of the weather data we collect (Mahanti, 2019). Having access to high-quality data is crucial to the performance of our machine learning models. Wrong inferences could be derived from faulty or inconsistent data. To mitigate this, we will use extensive data validation and cleaning procedures. Integrity in data collection also requires the use of trustworthy historical data and partnerships with reputable

meteorological agencies. Model Complexity: Kouris et al. (2022) note how difficult it might be to train a machine learning model to effectively summarize text. It is crucial to strike a balance between the complexity and usefulness of a model. Although complex models may be able to produce more reliable summaries, they can be time-consuming and resource-intensive to train. However, overly simplified models may not do justice to the complexities included in weather data. We want to use well-established methods from the field of natural language processing to conduct extensive experiments on a variety of techniques and architectures to lower this risk. In order to reach our performance goals, we will systematically evaluate and tweak our models. User Adoption: The project will only be successful if people actually use it. Lack of user-friendly presentation of meteorological data or operational complexity of the instrument could impede the tool's adoption among its intended users (Shirish & Batuekueno, 2021). By making user experience (UX) design a top priority right from the start of development, we can lessen the impact of this risk. Incorporating user feedback and implementing user testing at various stages of development helps improve the interface and the summary result. It will be crucial to gather user feedback in order to make sure the tool meets the needs and expectations of the public.

6. Significance of the Project

Increasing access to weather reports is the primary focus of this effort. It aims to broaden people's access to vital weather data by making complicated data more digestible and offering succinct summaries (Schintler et al. 2022). The language and terminology used in weather forecasts can be a significant barrier to understanding for people who are not meteorologists. This initiative addresses that need, expanding access to accurate weather forecasts. This software ensures that everyone, from farmers planning crop upkeep to commuters planning a trip and families getting ready for outdoor activities, can easily access and understand weather forecasts.

The public safety benefits of this initiative go far beyond its easily accessible features. Communities and individuals need easily digestible weather reports to protect themselves and their property. Having access to timely and reliable information is essential during natural disasters such as hurricanes, floods, and heat waves (Ponce & Spataru 2022). The program will help its users be better prepared for extreme weather, whether they choose to stay put or must evacuate quickly. Therefore, it lessens the impact of natural disasters and improves safety in general. It's possible that this project could be put to good use in the classroom (Alam, 2022). Easy-to understand data visualizations can help the general public learn more about weather trends and predictions. This effort to educate goes well beyond making people feel more confident in their ability to make climate-related decisions on their own. It has the potential to pique people's interest and educate them on the interconnectedness of our planet's weather and climate. This deeper comprehension of climatic and environmental concerns may lead to increased environmental consciousness and the advocacy of sustainability. The effort has ramifications for concerns of universal access, public health, and youth development, far beyond its stated purpose of making meteorological data easier to grasp and use. The general public becomes more weather-literate and environmentally conscious, while also empowering individuals to make better decisions in their daily lives and increasing public safety during weather extremes.

Schedule

Dates	Task	Estimated Duration	Output
September 15 - 20	Project Kickoff and Planning	1 week	Project plan and milestones defined
September 21 - 30	Data Collection and Acquisition	10 days	Comprehensive weather dataset
October 1 - 10	Data Preprocessing and Cleaning	10 days	Cleaned and structured dataset
October 11 - 25	Machine Learning Model Development and Training	15 days	Trained summarization models
October 26 - 31	User Interface Design and Development	6 days	User-friendly interface prototype
November 1 - 10	Initial Model Testing and Evaluation	10 days	Model performance assessment
November 11 - 20	User Interface Testing and Refinement	10 days	Improved user interface
November 21 - 27	Final Testing, Feedback Incorporation, and Deployment	7 days	Deployed tool for public access

References

Xu, X., Rioux, T. P., Gonzalez, J., Hansen, E. O., Castellani, J. W., Santee, W. R., ... & Potter, A. W. (2021). A digital tool for prevention and management of cold weather injuries—Cold Weather Ensemble Decision Aid (CoWEDA). International Journal of Biometeorology, 65, 1415-1426.

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., ... & Stadtler, S. (2021). Can deep learning beat numerical weather prediction?. Philosophical Transactions of the Royal Society A, 379(2194), 20200097

Halegoua, G. (2020). Smart cities. MIT press.

Mahanti, R. (2019). Data Quality. Quality Press.

Shirish, A., & Batuekueno, L. (2021). Technology renewal, user resistance, user adoption: status quo bias theory revisited. Journal of Organizational Change Management, 34(5), 874-893. Schintler, L. A., & McNeely, C. L. (Eds.). (2022). Encyclopedia of big data. Cham: Springer International Publishing. Ponce-López, V., & Spataru, C. (2022). Behavior in social media for floods and heat waves in disaster response via Artificial Intelligence. arXiv preprint arXiv:2203.08753.