**Project Reflection**

**Detecting Abrasive Online User Content in Question-and-Answer Forums**

Team 3 – Deepak Rajan, Indranil Pal, Yasser Parambathkandy

**Dr. Lindi Liao**

**George Mason University**

**AIT614 Big Data Analytics**

**04/26/2023**

**Project Reflection**

**In this project reflection document, topics related to the experience working on the term project, technical skills gained from the course, and how that has been applied to the final project, are discussed. Additionally, the lessons learned during the project development, highlighting ways to avoid project risks and streamline issue resolutions are included. Finally, the project management details to complete the project have been submitted.**

**Knowledge and Technical Skills Learned**

**Throughout AIT614, a deep understanding of several core technologies used in Big Data Analytics were obtained. One area that was covered in depth was the differences between RDBMS and NoSQL databases, and when each is best suited for particular use cases. For instance, learning how NoSQL databases like MongoDB can be used for storing unstructured data, including JSON documents, and how they can offer greater flexibility and scalability than traditional RDBMS systems was very informative. In addition to database technology, the course covered data warehousing and data lakes, providing an overview of their architectures and use cases. This included a deep dive into the use of SQL for querying and manipulating structured data, and hands-on experience with MongoDB and PyMongo for working with unstructured data. Another core area covered in the course was the use of Big Data processing technologies like Databricks, Hadoop, HDFS, and HIVE. A comprehensive understanding of the architecture of the Hadoop ecosystem and how to use HDFS for distributed storage and HIVE for distributed querying was gained. Additionally, hands-on experience with the MapReduce programming paradigm and its application to processing large datasets were obtained. This included an understanding of how to use PySpark and Spark for processing large-scale datasets and how to use the Spark MLlib library for machine learning, including data preprocessing, feature engineering, and model building. Finally, the course provided an introduction to Big Data Security, Cloud Computing, and Internet of Things (IoT) applications. This covered the importance of security in Big Data Analytics and the use of Cloud Computing platforms like Microsoft Azure and AWS for processing Big Data and IoT applications. Overall, this course provided a strong foundation in the core technologies and tools used in Big Data Analytics, enabling confidence and expertise in handling Big Data challenges in the future.**

**Using the above technologies, we developed a project that takes a question and predicts if it is sincere or not. The project utilized many of these technologies, including:**

1. **Spark DataFrame: The dataset was loaded into a Spark DataFrame, which is a distributed collection of data organized into named columns. This allowed for easy manipulation and processing of the data in a distributed manner using Spark.**
2. **GloVe and Universal Sentence Encoder: Word embeddings were generated using GloVe, an unsupervised algorithm for generating word embeddings based on the co-occurrence of words in a large corpus of text. The Universal Sentence Encoder, a pre-trained machine learning model, was used to generate sentence embeddings that capture the meaning of a sentence. These embeddings were used to represent the questions in a high-dimensional vector space, allowing them to be used as features in the logistic regression model.**
3. **Spark MLlib: The logistic regression model was trained using Spark MLlib, a library for distributed machine learning in Spark. This allowed for the training of the model on large-scale datasets in a distributed manner, taking advantage of the parallel processing capabilities of Spark.**
4. **Cloud Technologies: The trained model was stored in AWS S3, a cloud-based object storage service that allows for the storage and retrieval of large amounts of data. The project also utilized Databricks, a cloud-based platform for data engineering, data science, and analytics, for running the model training process in the cloud. This allowed for the training of the model on a scalable cloud-based infrastructure, taking advantage of the distributed computing capabilities of Databricks. The use of Databricks demonstrates an understanding of the importance of cloud computing in modern data-driven applications, particularly for training machine learning models on large-scale datasets.**
5. **MongoDB: A prediction service was developed that used the trained model to predict real-time questions and store the results in MongoDB, a NoSQL database optimized for storing and retrieving unstructured data.**

**The project's use of these technologies demonstrates a strong understanding of their capabilities and how they can be applied to real-world data-driven applications. Overall, the project showcases the practical application of the technologies learned in the Big Data Analytics course towards solving real-world problems in a scalable and efficient manner.**

**Project Lessons Learned**

**During the development of the project, we encountered a few issues and challenges that impacted its success. The following lists the issues encountered, the lesson learnt, and how we overcome it.**

1. ***Challenge*: One of the initial challenges was the selection of AWS EMR as a big data platform, instead of using Databricks, which was covered in the course. Although AWS EMR appeared to be a viable option on paper, we quickly discovered that it was not free and would have incurred several hundred dollars in costs.**

***Lesson Learnt*: This oversight taught us the importance of including cost considerations in future planning.**

***Resolution*: Ultimately, we resolved the issue by switching back to Databricks, which we had learned about in the course.**

1. ***Challenge*: Another challenge we faced was the limitation of Databricks in terms of size and time. Our dataset consisted of 1.3 million records and took longer than 2 hours to process, which exceeded the limitations of the Databricks community cluster. Although the documentation indicated that the cluster would only be terminated after 2 hours of idle time, we found that it was actually terminated after 2 hours of use.**

***Lesson Learnt*: This experience taught us the importance of sizing the dataset according to the environment limits.**

***Resolution*: To overcome this issue, we reduced the dataset size to 100K records in the Databricks environment.**

1. ***Challenge:* When selecting AWS Lambda as the prediction service server, we encountered a new challenge. On initialization, SparkNLP downloaded its packages and the pretrained Universal Sentence Encoder and GloVe embeddings. However, AWS Lambda does not allow local file writes, as it is intended to be a function-based processor.**

***Lesson Learnt:* This experience taught us that we need to carefully research the capabilities of a platform and ensure that it is suitable for the solution requirements.**

***Resolution:* We ultimately resolved this issue by switching to an AWS EC2 large instance.**

1. ***Challenge:* Lastly, we found that the use of SparkNLP worked well in a Linux environment but encountered struggles when used in a Windows environment.**

***Lesson Learnt:* This experience taught us the importance of carefully considering the operating system that a package supports, as well as reviewing its installation and execution documentation. *Resolution*: Currently, the code works only in Linux-based systems.**

**These lessons learned will be applied in future projects to avoid risks, streamline issue resolutions, and drive project success.**

**Project Tasks and Timeline**

|  |  |  |  |
| --- | --- | --- | --- |
| ****S.no**** | ****Task**** | ****Assigned To**** | ****Timeline**** |
| **1** | **Finalize project scope and objectives** | **Yasser Parambathkandy**  **Deepak Rajan**  **Indranil Pal** | **3/6/2023 - 3/11/2023** |
| **2** | **Research GloVe and USE** | **Yasser Parambathkandy** | **3/12/2013 - 3/18/2023** |
| **3** | **Research AWS components – EMR, S3, API, cloudformation** | **Deepak Rajan**  **Indranil Pal** | **3/12/2023 - 3/18/2023** |
| **4** | **Data analysis and preprocessing using Spark** | **Deepak Rajan**  **Yasser Parambathkandy**  **Indranil Pal** | **3/19/2023 - 3/25/2023** |
| **5** | **Generate word embeddings and train using logistic regression** | **Yasser Parambathkandy** | **3/26/2023 – 4/8/2023** |
| **6** | **Generate sentence embeddings and train using logistic regression** | **Indranil Pal** | **3/26/2023 – 4/8/2023** |
| **7** | **Develop code to receive real-time question and perform prediction using saved model** | **Deepak Rajan** | **3/26/2023 – 4/8/2023** |
| **8** | **Deploy application in Databricks cluster** | **Yasser Parambathkandy** | **4/9/2023 - 4/15/2023** |
| **9** | **Configure API gateway and S3 and other AWS components** | **Indranil Pal** | **4/9/2023 - 4/15/2023** |
| **10** | **Develop minimal UI for question submission and display predictions** | **Deepak Rajan** | **4/9/2023 - 4/15/2023** |
| **11** | **Testing end-to-end and review** | **Deepak Rajan**  **Yasser Parambathkandy**  **Indranil Pal** | **4/16/2023 – 4/24/2023** |
| **12** | **Project documentation and presentation** | **Deepak Rajan**  **Yasser Parambathkandy**  **Indranil Pal** | **4/16/2023 – 4/24/2023** |

**We w**ould like to express sincere gratitude to the Professors for their valuable guidance and support throughout the project, which greatly contributed to the successful completion of the project.