Assignment for 19.11.2022
Submission from News Summary Team:
Seida Basha, Lily Djami, Eric J. Herschbach, and Samuel Khodak

For this assignment, we watched a video lecture by Prof. Kaestner on working in an interdisciplinary team, as well as the paper “Data Scientists in Software Teams”.

In the video lecture, Prof. Kaestner mentioned some of the problems that could occur when working together in a team, such as group thinking, social loafing, and lack of motivation. Group thinking is when everyone in the group agrees on a decision without further discussion. This has a negative impact on the group’s quality of work due to lack of diverse ideas within the group. If ideas are not questioned it could lead to implementation of problematic ideas that for example do not fit together with other ideas. Social loafing is the decrease in team member’s motivation due to the mentality that there are others in the group that can pick up the work. First, this leads to an imbalance of workload. Furthermore, it could lead to the team not being able to finish their project because the work will be too much for the other team members. This would be very problematic in smaller teams which usually define roles at the beginning.

These are the challenges that we face as a team. To tackle these challenges, we have come up with a few strategies. To avoid group thinking, we have more than one person in a role, which allows for discussions in smaller teams. In addition, each team member has made different experiences which leads to many diverse ways of thinking within the group. Assigning roles is also a way to prevent social loafing. Since the roles are shared by mostly two people there is a higher risk of failure if one of the two people does not do anything. It is important that both take their responsibilities seriously. As another way to improve accountability and motivation in the team, we plan to have regular and frequent meetings so we can give updates on our progress to each other, as well as check if anyone needs help. This also helps against social loafing.

Something else that was mentioned in the video was the concept of a “unicorn”. This concept describes a person that is experienced and capable with everything the project needs. This can lead to problems. Since they know so much, it might be different for such people to think outside the box. Inexperienced members might have ideas that someone who is experienced would not think about. We prevent this again by assigning roles and not only having one person assigned to these roles. This gives the opportunity to talk about different ideas and have a better chance to work out a thought-out plan.

The paper “Data Scientists in Software Teams” discusses the way data scientists work in software projects, the challenges they face, as well as best practices. One of the challenges mentioned in the paper is the difficulty of obtaining data with good enough quality for the project. Scalability of machine learning models may also be expensive when processing large datasets. Lack of understanding of the dataset due to bad documentation or naming inconsistency may also impact the data analysis.

These issues will also affect our project. To address the issue of bad data quality, we would need to thoroughly research the dataset that we want to use for our project. There are existing news article datasets from CNN and BBC that are available online. These datasets have already been used in similar projects, so there is some guarantee to their quality. To tackle the

issue of scalability, we can use services such as Kaggle or Google Collab to train our model on better hardware.

Another issue mentioned in the paper was data availability. It was described that the data can be missing or incomplete or contain too much useless information to filter. As it was mentioned above, there is a certain amount of datasets, which have been used before and therefore are “proved” to be qualitative. Also, from previous research, we have realized that there are plenty of datasets available, which we can then choose for our project. Picking the right dataset for this project, would also help with filtering the needed information as well, e.g. if a dataset was used before, that means that there is little or in best case no work at all to be done in filtering the info. Moreover, the filtering process might also be available, considering the dataset was used before.

Integration of data from multiple sources might be another important challenge as keeping data of good quality in this project. It is necessary that the data is understandable, especially by data scientists. To avoid these challenges, this data has to be checked by data preparers and analysts, so the data are on the same level and comparable. In this case, news articles have to be understood, so that the quality of the summary can be checked by the quality assurance team. Furthermore, the source should be verified by checking the authors’ or publishers' seriousness and reputation. Our team has to keep a documentation regarding the data recentness, label the data correctly and, most importantly, talk to stakeholders, who work with these data to prevent misunderstandings and space for interpretation. The data must be easy to understand and clarified.

The paper also mentions several best practices in data science that we can implement in our team work. These best practices include:

- Having a good understanding of statistics and machine learning
- Setting a clear goal for the project
- Using homogenous tools and standardized data
- Understanding the data and making sure that the analysis makes sense in the context of the problem.

As our team consists not only of students with a background in machine learning, we make sure that the team has good understanding in machine learning and statistics by working together with a member of the machine learning group, who will act as a machine learning consultant for our group. We have also discussed what our goal is as a team: a web application that gives a summary of news of the day. To further detail our goal, a requirement document can be written. To make sure that we are using homogenous tools and standardized data in the group, we have also had preliminary discussions on what tools we want to use for our project (i.e. Git, Visual Studio Code).