ProtectMe

Elaboration Phase

Context

In the current days, the various forms of social media that exist have become a very impactful part of the common person's life, and day by day, we can see a rapidly rising user base for every social media platform.

And with this fast growth in usage, there is also a growth in users who try to take advantage of these social media platforms to spread malicious content and deceitful information.

Although there have been a lot of attempts to monitor the spread of malicious content in social medias, most of them rely on detecting spam attacks and detection of improper content in a social context, whilst the ProtectMe project aims to be able to detect malicious content through the analysis of all the content made available by a given post (image, video and text)

SOA

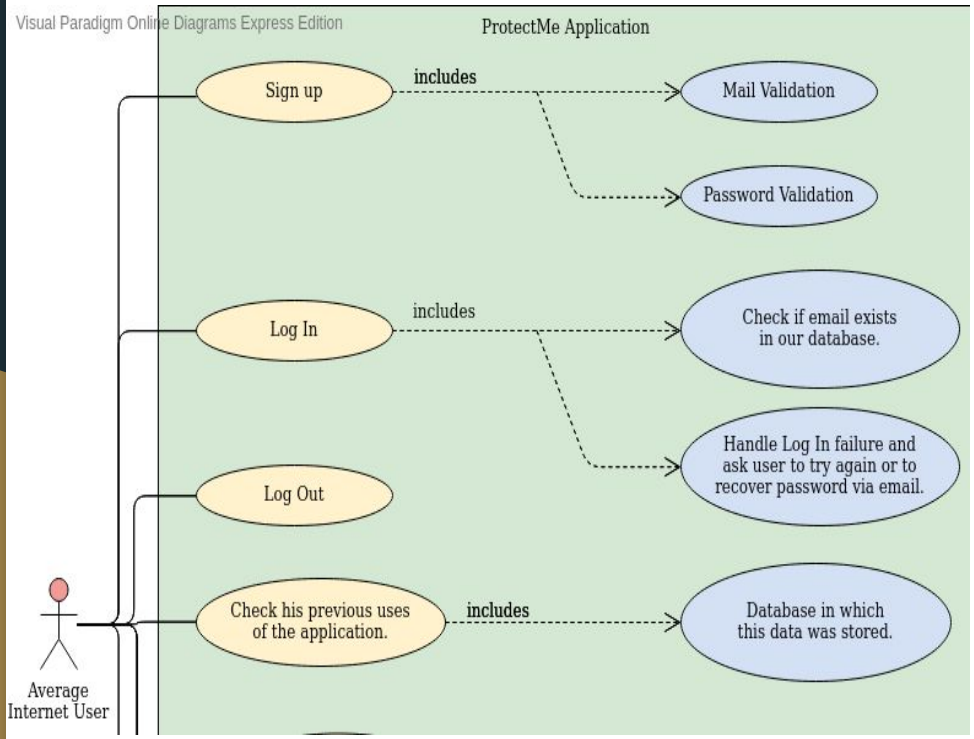
In the context we are trying to approach,although there isn't much work done there is notable done made by Johnny Wales (Data Scientist at Wales Data Technology,LLC) (https://towardsdatascience.com/full-pipeline-project-python-ai-for-detecting-fake-news-with-nlp-bbb1eec4936d)

In this article, Wales talks about a project he made where he used NLP and Machine Learning in order to detect Fake News, and got results of up to 65% accuracy for one given model (SVM).

Actors

The target users for the application are people that wish to know if the social media posts that they read and see contains some sort of manipulation and if that content has malicious intent behind it.
There isn’t a need for a special skill set to use the application, it should be relatively easy and intuitive to anyone who uses a browser to navigate the web on a regular basis.
It’s important to note that the users can be totally anonymous in our system using it without doing any type of log in. They can also sign up and log in if they wish to be able to access their previous iterations with the application but the only registered information saved will be their email and a password, without keeping any other kind of personal information about them.
So, there will be just one actor which can be described as a regular Internet user that wishes to know more about the veracity and truthfulness of the contents he/she sees on his/her media feeds.

Use cases



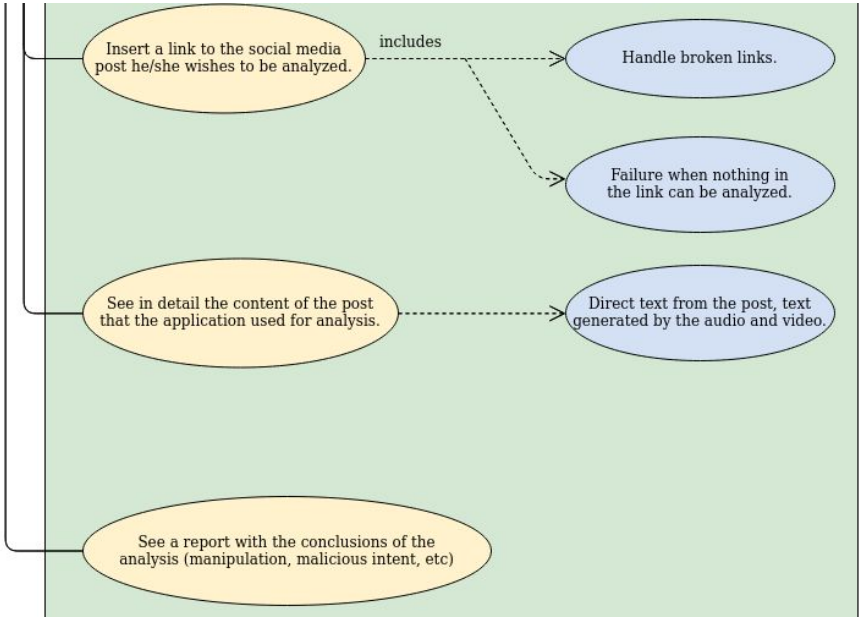
Currently defined use cases:

->Sign up

->Login/Logout

->Get User search History

Use Cases



Currently Defined Use Cases:

->Analyze a social media post

->Check content extracted from the social media post

->See results of a social media post analysis

Non Functional Requirements

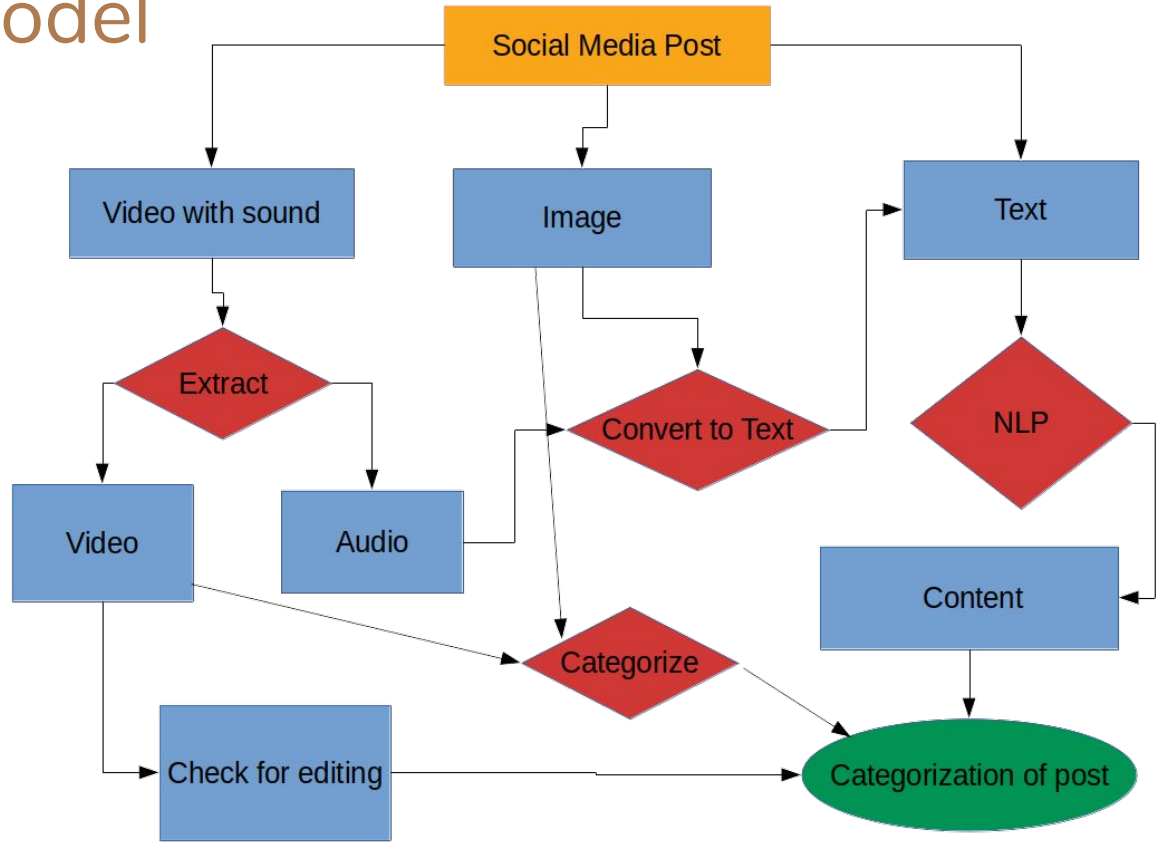
We want our application to be responsible and with maximum availability as possible. It should be intuitive and easy to use. We want to show the user the best results we can manage, even if it compromises the system's speed a bit. We want the system to be as reliable as possible, however, we cannot guarantee that the results shown by the application can be completely reliable, so instead of showing a "malicious/harmless" result, we expect to show the user a given probability of how harmless a given post is.

When it comes to security, the system by itself can be used in a completely anonymous manner, with no tracking of user activity whatsoever. We also want the user to be given the chance to leave a track of his searches, and so, we will also allow the users to register an account, but we won't store any personal information besides the user's email and a password.

Functional Requirements

The system must be able to create a new user account and store information relevant to the user (minimal amounts of user data and the user post verification history). It must also be able to receive a query to verify if a given social media post is “malicious” or not as well as give a score of how likely it is to be malicious.

Domain Model

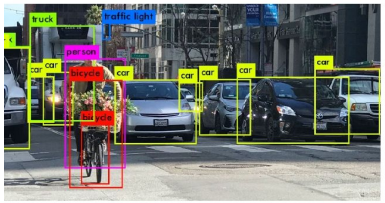


Mockup Interface

ProtectMe!

Lorem Ipsum (https://loreipsum.dolor/sit.amet)

Extracted Content:

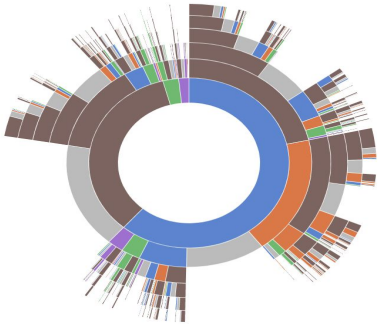


Text Analysis

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non arcu ac nunc placerat efficitur. Donec mauris neque, pretium a ligula id, efficitur mattis lacus. Duis id finibus felis. Phasellus cursus nec ligula a viverra.

78%

Result: Possibly Harmless



Additional Resources

In order to better understand and ascertain the viability of the project we used the following resource:

https://youtu.be/RuI0DK7qiwU