Background

This project arises from the business need of LiveWorld[1] to add, to the current content moderation pipeline offered to its customers, a software solution that automatically detects posts/messages, on customers' platforms, containing sexists or racists statements (hate speech statements), providing as extension a "mitigating answer" tool. Such a tool returns, for an identified hate speech post, an answer that can be used as a discouraging comment, therefore a comment that can prevent possible supportive replies. This software is positioned as solution for communication through Internet for companies.

Currently, Internet plays an important role in the communication field and it is always more and more important to apply the different companies' terms and conditions, making sure the content that does not respect them is identified as soon as possible.

For this reason, automated content moderation is a rising application domain, also thanks to the recent advancements in the applicable AI techniques. Moreover, studies[2] show that negative answers to racists/sexists posts ("mitigating answers") can reduce their impact limiting subsequent following messages and interactions. Consequently, we can see the benefits of an automated content moderation tool that can also take a more active role in limiting the effect of negative posts and consequent reactions.

Emerging customer needs cannot be satisfied, at the moment, by commercial dominant software solutions in the market, specifically they do not offer answering modules able to react automatically to negative posts, as can be seen in the following table, comparing the main competitors:

Product/Company Name	Reference URL	Short description
Besedo	https://besedo.com/product/customized-ai-moderation-models/	AI – tool based on training data obtained by the customer, possibility to keep the algorithm up to date using continuous learning.
Utopia AI Moderator	https://utopiaanalytics.com/utopia-ai-moderator/	Fully automated, requires a 2-weeks training on the company's data. It also exploits continuous learning.
Dialogfeed	https://www.dialogfeed.com/our-ai-automated-moderation-tool-for-social-media/	Focused on social media, allows the connection of different accounts and feeds and returns the filtered content via a widget.

Scope

At the completion of the project, the developed software solution will offer:

a) A Machine Learning model able to distinguish "clean" posts from the racists and sexists ones, classifying so them properly. The model works after a training on a data set. For this reason, a dataset

[1] https://www.liveworld.com/social-media-content-moderation/
[2] https://arstechnica.com/science/2016/11/twitter-bots-can-reduce-racist-slurs-if-people-think-the-bots-are-white/ and https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf

with labelled data, extracted from the LiveWorld databases, will be used to train the associated Machine Learning model (as described in the data architecture section).

b) An extension to the current content moderation pipeline, providing an automatically generated "mitigating answer". The release of this feature is dependent on the level of performances shown in the detecting phase (as denoted in the metrics section). In case the performances will be satisfactory, the model will be paired with an "answering module" which, given a classified "negative" post and possibly some keywords, returns automatically a customised answer to the post.
On the other hand, if the performances shown in the detecting phase will not be considered satisfactory, the classification of the post and the suggested answer will be notified to a human content moderator, who will then revise them before deciding on their approval.

The software solution will be deployed in the existing LiveWorld content moderation pipeline, using a web server hosted by LiveWorld, that offers to the customer's technicians access to all the different functionalities through a web service. The software solution will then work on real time input data, coming from external platforms of LiveWorld's customers.

More in detail, the project will start with a data acquisition phase, which will grant access, to the data scientists working on the project, to the data available in LiveWorld's databases. In this phase the virtual machines provided by the customer will also be configured.

Following, the data will be analysed and cleaned, making so possible to understand their properties, narrowing so the scope of the subsequent data pre-processing phase, which will prepare the data to the usage in the training of a Machine Learning model.

Next, the two modules will be developed concurrently, using the respective training data for building the models and optimising the associated parameters and, after analysing their performances, they will then be integrated in a single software solution.

Thereafter, a Web Service, that will offer the access to the different functionalities to LiveWorld's technicians, will be developed and deployed on a server hosted by the customer. A phase of testing and integration with the existing systems will then conclude the project.

Goals

The goals of the project, previously described, target to deploy the following services:

- Detection and distinction of racists/sexists posts from the "clean" ones, applying a Machine Learning model
- Provision, for each negative post, of an appropriate "mitigating answer"

The two goals imply:

- Reduction of the number of "negative" interactions on the platform on which the smart moderation software has been applied, thanks to the effect provided by the mitigating answers.

In the Proof Of Concept (POC), the capabilities of the models can be tested. It will also implement a basic definition of the answering module.

Metrics

Below, we report the metrics used in this project, distinguishing the ones related directly to the tool (TOOL) from those related with the platform on which it is applied (PLATFORM)

1. TOOL: Detect at least the 95% of the real "negative" posts analysed (accuracy metric)
2. PLATFORM: Reduce the number of negative posts on the single platform by 10%. Note: the average number of negative posts per day will be measured on a 30-days timeframe, prior to the deployment of the tool and then measured again after the deployment.

Personnel involved

In the development of this project, the following professional roles will be present:

Us

- Project Lead and Project Manager
- Four Data Scientist: analyse the data, build and tune the needed models
- Two Developers: implement the model and the interfaces used to communicate with it
- A System administrator: deployment and management of the resulting API

The customer

- Business contact
- Data administrator
- System administrator

Key Stakeholder

Client	LiveWorld
Sponsor	LiveWorld
Project Manager	Davide Sbetti
Project team members	Piers Ford, Homer Greer, Barrett Daniels, Scott Hicks, Bert Austin, Ernest Steel and Jamie Hodson

Project Milestones

The table below reports milestones, with time required in working days, since the start of the project. The project will start on the first Monday after the signature of the present document.

When?	What?
Day 0	Project start - Access to the customer's data and architecture is granted
Day 13	Preliminary Data Analysis completed
Day 34	Predictive Model and Answering Module finished – First invoice
Day 45	Models Integration and Web Service Development finished – Second invoice
Day 64	Deployment and Testing Phase finished – Third Invoice
Day 65	Project End

Project Budget

Considering the previously outlined team (8 people including the project manager), their salaries and the planned number of business days needed to complete the project, we can calculate an estimated project budget of about €120,000.

Data Architecture

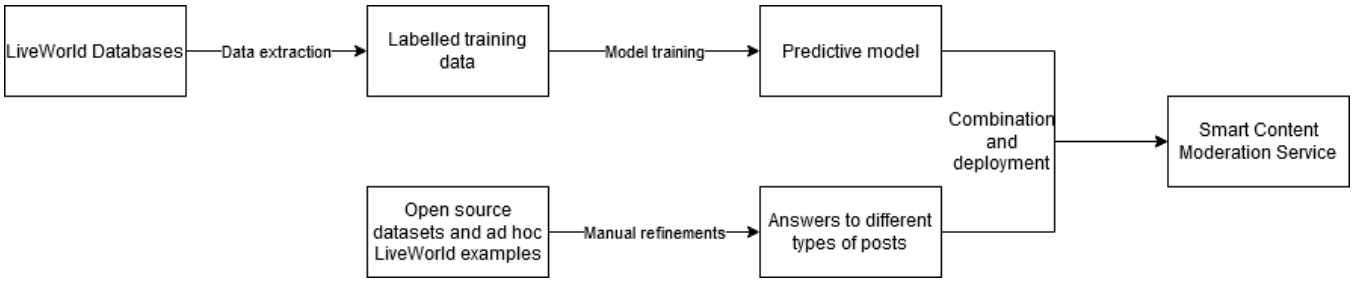
The data, used mainly for training the predictive module, is extracted from LiveWorld Postgres databases, to which access will be granted. Regarding the answering modules, data provided by LiveWorld, which is collecting answering data from its customers, is used to define the most used replies in the different contexts. Possibly, additional ad-hoc examples, provided by the customer or coming from open source datasets, could be processed. In all cases, data shall be processed on virtual machines made available by LiveWorld.

For the Proof of Concept provided along with the project, only open source datasets[3] are used

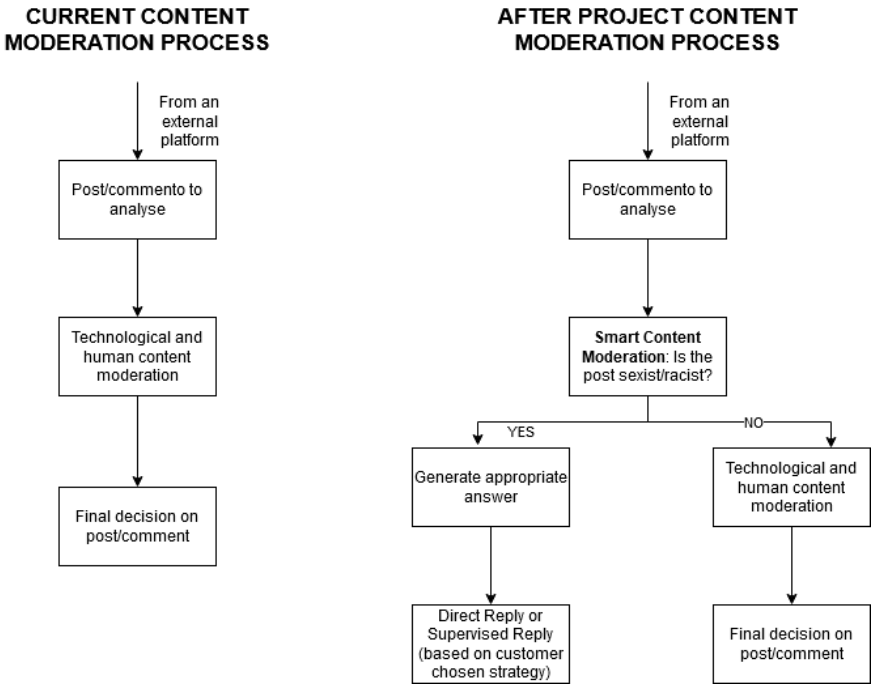
For what concerns the tools and technique adopted, all software related solutions (libraries and frameworks) and possible additional data will come **only** from open sources granting a free commercial use.

We report below a schema of the data extraction and processing:

[3] https://gombru.github.io/2019/10/09/MMHS/



A comparison of the current architecture with the final one is here depicted:



Constraints, Assumptions, Risk and Dependencies

Type	Description
Constraints	The customer will grant access to the training labelled data already available from their past moderation activity through a set of 3 virtual machines (equipped with an NVidea GPU conformed to at least the Volta architecture), used then in all phases of the project. An NDA agreement, concerning the data provided by the customer, will be signed to ensure non-disclosure. The personnel provided by the customer's side will be available to develop the project in an agile form, with multiple interactions.
Assumptions	The customer is able to provide the requested items (data/resources) and no technical outbreaks to the provided resources prevent the project from progressing.

Risks and Dependencies	The main risks are related to the quality of the provided data. Machine learning models heavily rely on the quality of the training data (although this can be augmented) and the real-world performances can be influenced by these factors.

Approval Signatures

Project Client	Project Sponsor	Project Manager