Domain Pulse

Fit for Purpose

Table of Contents

# 1 Introduction

1.1 What is Domain Pulse?

Domain Pulse is the ultimate sentiment analysis platform. It gathers and analyses online opinions about any domain, be it a business, a person, or more. With stunning visuals and easy-to-understand statistics, Domain Pulse helps you understand the online presence and sentiment for any domain.

Domain Pulse provides a variety of integrated data sources (namely Google Reviews, Youtube, TripAdvisor, Trustpilot, CSV files, and even a facility to create your own live reviews) with which users can "plug and play". Domain Pulse collects data from the specified source, conducts sentiment analysis on the data, and then presents, visualises, and quantities various useful sentiment metrics in a slick and streamlined fashion.

1.2 The evolution of Domain Pulse

Domain Pulse was originally conceptualised as an exploratory project with the goal of determining whether an application that could leverage sentiment analysis to provide meaningful insight into a generic area of interest to support decision making, could exist and be feasible in principle.

Throughout Domain Pulse's development lifecycle, the project evolved and focussed as we identified areas of potential application, useful and novel features and metrics, and target audiences and use cases. Gradually, Domain Pulse solidified into a stunning, and more importantly, useful, dashboard-style web application which makes insight from sentiment analysis more accessible to a wider variety of users from a number of different contexts - from business, to hospitality and entertainment, to research, and more.
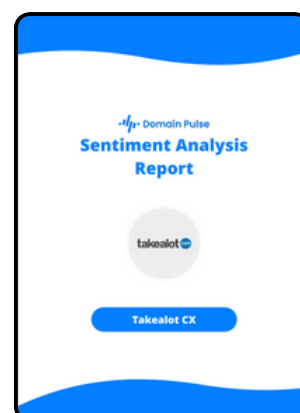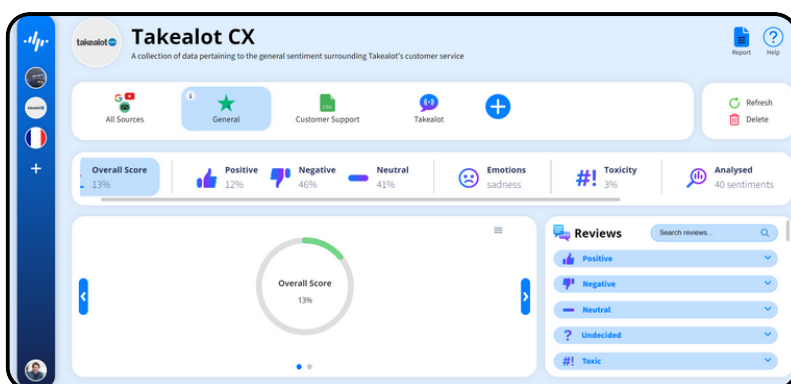
2 Domain Pulse's Purpose

2.1 What is the problem?

The problem that Domain Pulse addresses is as follows: In a Golden Age of data analytics, there is a relentless need to find new and novel ways to extract value from arbitrary data. Generically, this value is typically information to support decision making, identifying areas of strength or weakness, or leverageable insights about the area of interest.

Furthermore, for domains in which public perception can be crucial (such as entertainment, hospitality, business, etc), understanding the sentiment people have towards your operation is a novel, yet substantial advantage. Most of this raw sentiment data is some form of textual review or comment and is more often than not publicly available on platforms such as Google Reviews and Tripadvisor. The challenge is quantifying, measuring, and comparing how "good" or "bad" this feedback is.

Simply reading through all this data is not a viable strategy. The process would be prohibitively slow, and more qualitative than quantitative (which makes measurement difficult). Sentiment analysis is a machine learning technique that is able to autonomously detect and quantify measurable sentiment metrics from this text - for example: prevalent emotions, overall positivity or negativity, and even toxicity (which is a term used to describe or flag text that is nasty, offensive, or otherwise unsavoury).

Being able to conduct sentiment analysis on this feedback is no simple task for most people and organisations. The technical skill required to collect large amounts of data, preprocess it, analyse it, and present it, is often prohibitively high for most individuals, preventing access to the novel and interesting insights that can be gained from sentiment analysis.
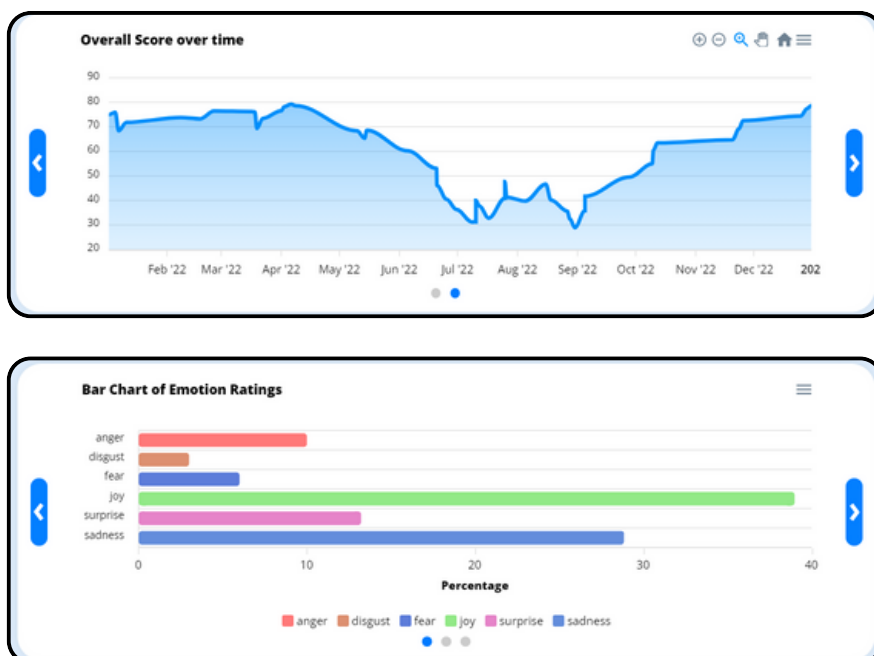
2.2 Target users

Naturally, Domain Pulse's target user base is broad, user groups include:

- Business owners
- Market researchers
- Customer experience managers
- Content creators
- Analytics and insights personnel
- Managers in the hospitality and entertainment industry
- and more.

The key concept tying these groups together is that they all wish to extract value from available review data, with the purposes of supporting decision making, identifying areas of strength or weakness, or acquiring leverageable and measurable insights.

These users do not necessarily possess high levels of preexisting technical expertise when it comes to tech and machine learning, and as such, a novel and easy-to-use analytics tool for sentiment analysis is immensely useful for extracting the aforementioned value from their data.

3 How does Domain Pulse meet its purpose?

3.1 Requirements

In order for Domain Pulse to effectively meet its purpose described above, we deemed it necessary that Domain Pulse exhibits the following five general characteristics, two of which refer to functional requirements, while three pertain to non-functional requirements.

Functional

- **Provide valuable insight and can be quantified and compared.** In order to be useful and insightful, the metrics produced by Domain Pulse need to be interpretable in the sense that they can be compared and measured. Furthermore, Domain Pulse needs to provide different avenues for communicating the results of the analysis.
- **Be robust and generic in its handling of sentiment data.** Domain Pulse is an application that should be applicable to a wide range of problem domains, and limiting it to a more narrow domain would undercut its usefulness and applicability. This involves implementing a robust and sophisticated sentiment analysis pipeline.

Non-Functional

- **Be highly usable and intuitive.** In our goal of making sentiment analysis more accessible, Domain Pulse needs to cater to an audience that is not necessarily technical, hence, the application needs to be as friendly and intuitive as possible - the interface of the application should not prohibit or deter any user from its use.
- **Be performant and scalable.** Domain Pulse handles and analyses large quantities of unstructured text data - a task which can be computationally expensive. If the application is prohibitively slow under stress, as a result, the usability and slickness/usefulness of Domain Pulse will be compromised.
- **Be easily modular and expandable.** Domain Pulse needs to be designed in such a way that the design and addition of new features is a slick process. In particular, adding new data sources and adjusting the nature of the sentiment analysis must be intuitive and quick - such to keep up with evolving user needs.

3.2 Functional

3.2.1 Providing measurable insights

Domain Pulse provides valuable sentiment insight by providing a multidimensional and rich set of metrics from its analysis. These metrics include: an overall "good/bad" score, a summary of prevalent emotions, ratios of positive, negative, and neutral sentiment, as well as a measure/flag for toxicity.

An arrangement of different sentiment analysis models are responsible for computing these different metrics, and this is discussed in the next point. By providing this range of metrics, users are able to gain a richer and deeper understanding of the sentiment prevalent in the data. All of these metrics have numeric representations, meaning that different data sources can be measured and compared directly.
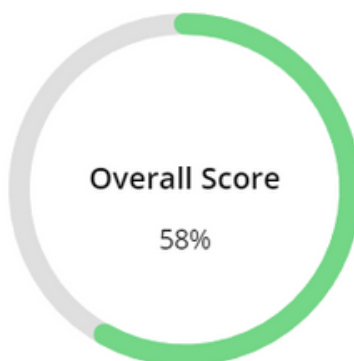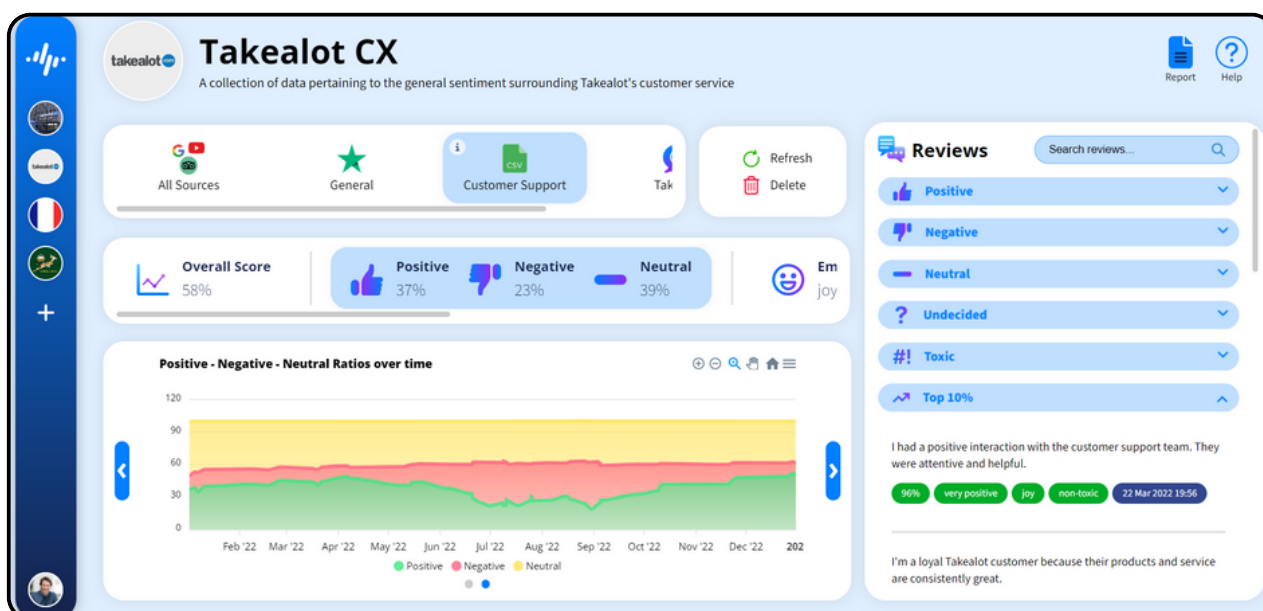
Furthermore, Domain Pulse provides stunning and clear visualisations of each metric to improve interpretability. The "cherry-on-top" is that each metric also has a time-series visualisation (using an exponential moving average), allowing users to identify time periods during which improvements or dips in a particular metric occurred. These graphs can be even further inspected by selecting a specific time period to view the trend for. Cumulatively, this allows users to gain meaningful insight into the sentiment surrounding their domain.

The dashboard of numeric metrics and visualisations is not the only means whereby a user can extract the insight. Domain Pulse includes report-generation functionality, whereby a user can automatically produce, view, and share a PDF report detailing the results presented on the dashboard - this functionality is suitable for users who want to present their findings in a more formal way.
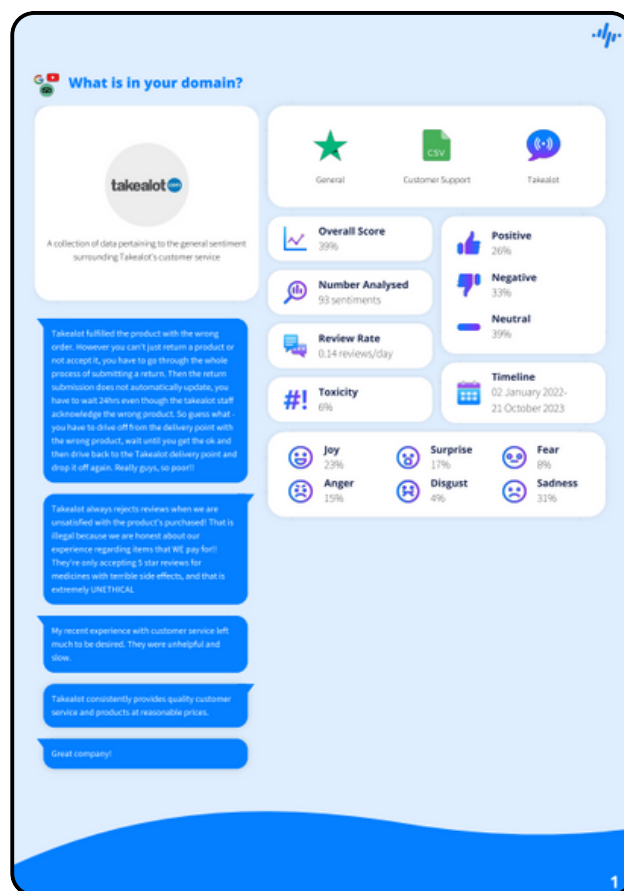
3.2.2 Robust Sentiment data pipeline

The heart and brain of Domain Pulse is its sentiment analysis pipeline, which facilitates the process of data ingestion, preprocessing, analysis, persistence and aggregation.

The first relevant step in this process is to subject incoming data to preprocessing, which conducts a number of standard NLP cleaning steps such to format incoming data consistently before it is fed into the models - reducing structural variation in incoming data from different sources.
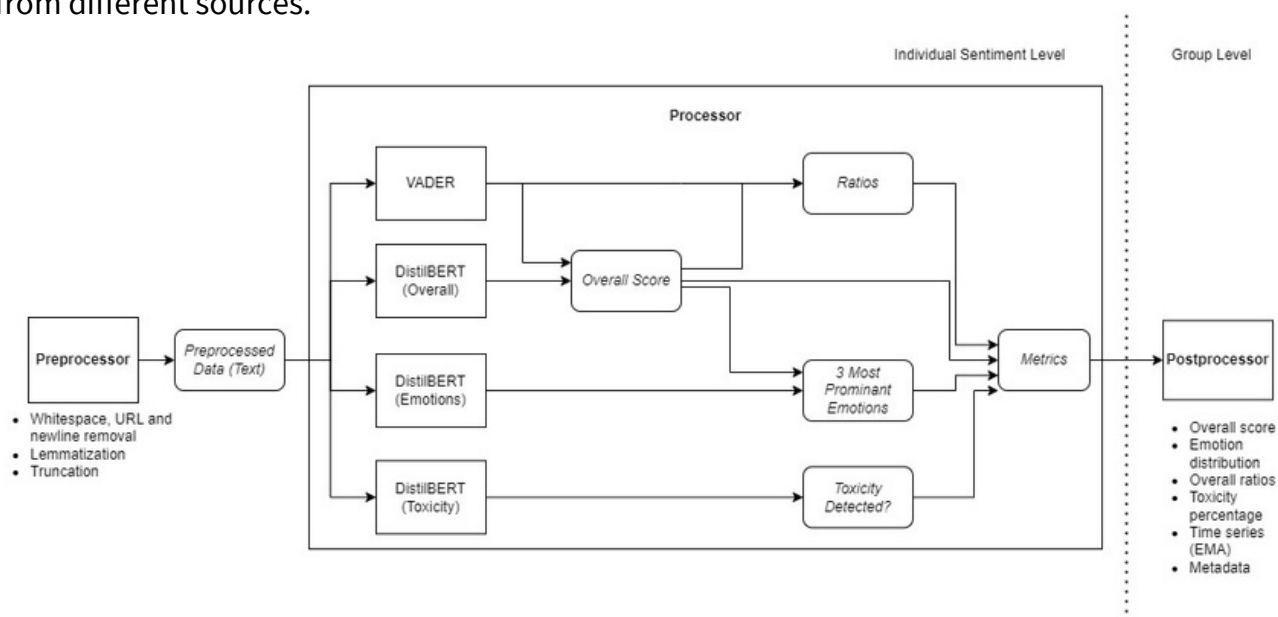
In order to further improve Domain Pulse's generality and applicability, various different sentiment analysis models are used to conduct the analysis. A lexicon-based model (VADER) as well as three neural network models (DistilBERT) work in parallel in order to compute a comprehensive set of metrics for each data record - where each model produces different metrics.

In order to promote consistency and consensus among the metrics, model outputs are not used directly, but rather act as inputs to a 'consensus algorithm' which is responsible for formatting and adjusting the metrics and classifications such that the final output for a piece of sentiment data is more uniform, consistent, and decisive.

Finally, once metrics have been computed for an entire batch of data, our aggregation algorithm systematically produces the overall and final measurements for each metric, including the production of time series data and metadata.

The described process ensures that Domain Pulse's sentiment analysis pipeline is robust and generic such that it can be applied to a wide variety of problem domains.

3.3 Non-Functional

3.3.1 Usability

One of the ways in which we have promoted a smooth and intuitive user experience is by employing recognition over recall. Recognition over recall is an important usability concept and this means that users of the application should be able to easily identify and use the elements and actions performed by the application rather than having to remember them from memory (which is a very daunting task when first using an application).

Another way that our application was created with user experience in mind is by employing a constraint that all actions that the app provides should be able to be completed within 5 clicks, this is especially important as the complexity of the app increases dramatically as the amount of clicks to complete an action increases.
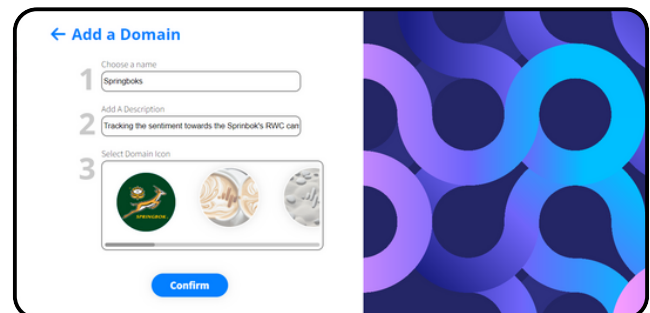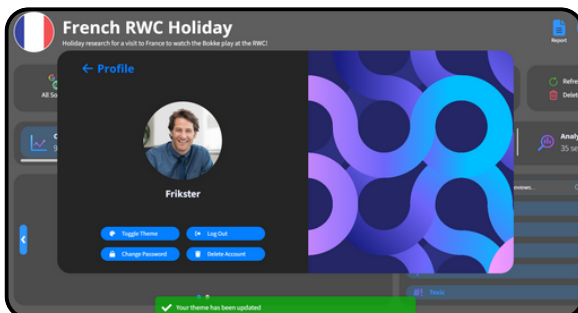
Finally, additional design choices such as providing a light and dark mode, rounding of corners, ensuring symmetry and promoting balance were incorporated to optimise the user as much as possible. These design choices promote the app's user-friendliness by avoiding cognitive overloading (we never display too much information to the user at one time) which aids the user in making easy and quick decisions about what it is that they would like to do/accomplish.

This is balanced using clever iconography that allows us to keep the dashboard dense in terms of functionality (allowing the user maximum productivity and ease of use) while still keeping all possible navigation paths clear implicitly.

From a business and branding perspective, we took extra care to ensure Domain Pulse employed a colour scheme consistent with our branding, such to give users the impression of professionalism promoting trust in our application from users.

3.3.2 Scalability and Performance

Domain Pulse is constructed using a Service-Oriented Architecture. The application consists of five independently deployable subsystems - enabling both vertical and horizontal scaling, aided by automatic load balancing.

The computationally expensive core of our application, the sentiment analysis pipeline, is implemented as a virtual Pip-and-Filter architecture in the sense that the flow of data from collection to analysis is handled by a series of computational stages each responsible for a different task (ex: collection, preprocessing, analysis, persistence, aggregation). These stages/filters are distributed across three of our five independently deployed services - by distributing this data and processing intensive load we promote scalability of this extremely important process while effectively enabling pipelining of sentiment data processing tasks, improving performance. This process is managed using a Facade (our Query Engine).
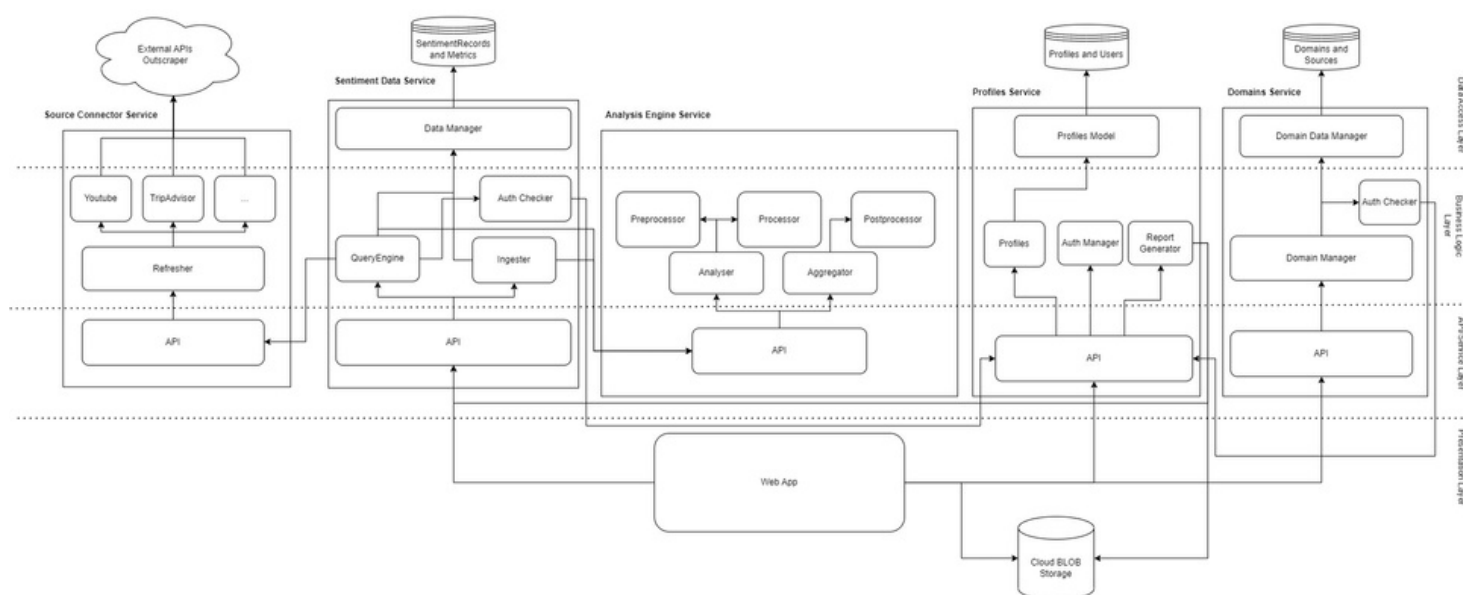
3.3.3 Modularity and Modifiability

The overarching Service-Oriented Architecture is supported by the use of other architectural patterns. Each individual service is structured according to a Layered Architecture, where each service consists of an API, Business Logic, and Data layer. Consequently, not only are the five independent services separated according to function, within each service, clear separation of concerns is achieved. This realises maximal modifiability of application components in both the horizontal and vertical axes.

To further reduce coupling between our services and the databases, we leverage the Repository Pattern - which places an interface between the business logic code and the underlying database system. This allows for the easy 'swap out' of database systems should the need arise, with minimal changes to the rest of the codebase, further emphasising the modifiability of the application.

In particular our Source Connector (which connects to data sources such as Google Reviews, Tripadvisor, etc) is its own service, loosely coupled from the rest of the application, meaning that the addition of new data sources is a simple and streamlined process, requiring minimal adjustment to the rest of the application (ex: Twitter could be easily added as a source were it not for the monetary cost of its external API). Similarly, the Analysis Engine which is responsible for preprocessing, analysing and aggregated sentiment metrics is another one of our independently deployable services - this isolation of this processing from the rest of the application ensures that adjustments to the actual sentiment analysis process are totally independent of the rest of the system.

4 Understanding limitations

It is important to note that each of the models used by Domain Pulse suffers from its own biases and limitations, and that the results of analysis produced by the platform are by no means meant to be regarded as the truth. Domain Pulse is meant to provide general guidance and summary analysis, and does not guarantee true or factual representations of people, places, organisations, or events.

In particular, the models underpinning Domain Pulse may produce biassed or incorrect predictions for input data that reference underrepresented populations or particular identify subgroups, incorrectly flagging these sentiments as negative and/or toxic. Additionally, the models are not aware of all possible contexts of word usage, and benign data may be incorrectly categorised should the text contain words that have alternative meanings in different contexts.

Finally, the sentiment analysis models used by Domain Pulse are not fine-tuned for downstream tasks (ie: classifying specific types of textual data from a specific source). This is a tradeoff that improves the generality of the application (which is a crucial characteristic of the application), at the cost of potentially reduced accuracy. This includes language - Domain Pulse is capable of handling text written in the English language only.

5 Conclusion

In summary, team 'Ctrl Alt Defeat' believes Domain Pulse is an application that has been built with the utmost care and awareness regarding its purpose, while acknowledging and incorporating the evolution and development of this purpose throughout construction.

Domain Pulse's real-world applications are plentiful, and using our technical skills and knowledge, we were able to develop a product that meets both its functional and non-functional requirements in order to be truly fit for purpose.