RiskPACC

INTEGRATING RISK PERCEPTION AND ACTION TO ENHANCE CIVIL PROTECTION-CITIZEN INTERACTION

COMPLETION OF SENTIMENT ANALYSIS TOOLBOX TO MEASURE THE RPAG

Deliverable 5.2

Dissemination Level: Public

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021271

5.2 COMPLETION OF SENTIMENT ANALYSIS TOOLBOX TO MEASURE THE RPAG

Deliverable number:	5.2
Version:	1.0
Delivery date:	07 June 2023
Dissemination level:	PU
Nature:	Report
Main author(s)	Alex Leite (CS) Jesse Manning (CS)
Contributor(s)	Alex Leite (CS) Jesse Manning (CS) Victor Margallo (CS) Tania Sadurschi (CS)
Internal reviewer(s)	Chrysoula Papathanasiou (ICCS), Evangelos Pitidis (UoW)

Document control

Version	Date	Author(s)	Change(s)
0.1	28/04/2023	Alex Leite	Table of contents
0.2	05/05/2023	Alex Leite	First Draft
0.2	05/05/2023	Tania Sadurschi	Functional Design Section
0.3	05/05/2023	Victor Margallo	Data science research section
1.0	23/05/2023	Alex Leite	Addressing feedback
1.1	2/06/2023	Alex Leite	Fixing styles and addressing other feedback

DISCLAIMER AND COPYRIGHT

ABOUT RISKPACC

Increasingly complex and interconnected risks globally highlight the need to enhance individual and collective disaster resilience. While there are initiatives to encourage citizen participation in creating a resilient society, these are typically fragmented, do not reach the most vulnerable members of the communities, and can result in unclear responsibilities for building disaster resilience.

New technologies can also support preparedness and response to disasters, however, there is limited understanding on how to implement them effectively. Awareness of risks and levels of preparedness across Europe remain low, with gaps between the risk perceptions and actions of citizens and between the risk perceptions of citizens and Civil Protection Authorities (CPAs).
The RiskPACC project seeks to further understand and close this Risk Perception Action Gap (RPAG). Through its dedicated co-creation approach, RiskPACC will facilitate interaction between citizens and CPAs to jointly identify their needs and develop potential procedural and technical solutions to build enhanced disaster resilience. RiskPACC will provide an understanding of disaster resilience from the perspective of citizens and CPAs, identifying resilience building initiatives and good practices led by both citizens (bottom-up) and CPAs (top-down). Based on this understanding, RiskPACC will facilitate collaboration between citizens, CPAs, Civil Society Organisations, researchers and developers through its seven (7) case studies, to jointly design and prototype novel solutions.

The “RiskPack” toolbox/package of solutions will include a framework and methodology to understand and close the RPAG; a repository of international best practice; and tooled solutions based on new forms of digital and community-centred data and associated training guidance. RiskPACC consortium comprised of CPAs, NGOs, associated organisations, researchers and technical experts will facilitate knowledge sharing and peer-learning to close the RPAG and build disaster resilience.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021271

TABLE OF CONTENTS

List of Tables

List of Figures

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021271

Executive Summary

This deliverable serves as a comprehensive documentation of the technical engineering processes and design thinking implemented by Crowdsense for Task 5.2, which focuses on the adaptation and further development of the online sentiment analysis toolbox. Within this document, we aim to provide a detailed account of the methodologies employed throughout the task, starting from the initial presentation of use cases and the crucial process of collecting feedback from the Case Study Partners involved in Work Package 3.

The deliverable begins by outlining the essential steps undertaken during the project, emphasizing the iterative nature of the process. We highlight the importance of gathering feedback from the Case Study Partners, as their insights played a vital role in shaping the direction of the sentiment analysis toolbox. By closely collaborating with the partners, we ensured that the final product aligned with their specific requirements and addressed the challenges faced in their respective contexts.

Informed by the principles of Data Science research, we describe the choice, training, and fine-tuning of the multi-sentiment machine learning model. We explain the various techniques and methodologies employed to optimize the model's performance, enabling it to accurately classify sentiment across a diverse range of texts. The research conducted in this phase not only solidified the technical foundation of the sentiment analysis toolbox but also ensured its adaptability and scalability in real-world scenarios.

Moreover, this deliverable sheds light on the product design thinking methodology employed by Crowdsense. We elucidate the systematic approach taken to design the frontend of the sentiment analysis toolbox, leveraging the advancements made in multi-sentiment machine learning. By incorporating user-centered design principles, we prioritized usability, accessibility, and user satisfaction, ultimately resulting in an intuitive and user-friendly interface. The functional design of the frontend is extensively detailed, emphasizing the seamless integration of the machine learning model with the user interface to provide a cohesive and efficient user experience.

In conclusion, we summarize the significance of Task 5.2 in closing the Risk Perception Action Gap (RPAG). By developing and enhancing the online sentiment analysis toolbox, we empower users to effectively gauge and comprehend public sentiment in real-time. This newfound understanding of public sentiment enables informed decision-making, aids in proactive risk management, and bridges the gap between risk perception and action. Task 5.2, through its comprehensive technical engineering processes, design thinking methodologies, and advanced machine learning techniques, serves as a crucial step towards improving risk perception and response strategies in various domains.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021271

Glossary and Acronyms

Term	Definition / Description
ML	Machine Learning
API	Application Programming Interface
PAI	Publicly Available Information
AI	Artificial Intelligence
CPA	Civil Protection Authority
NLP	Natural Language Processing
GPU	Graphics processing unit
GCP	Google Cloud Pages
LLM	Large Language Model
RPAG	Risk Perception Action Gap
EU	European Union
GDPR	General Data Protection Regulation
CSP	Case Study Partner

TABLE 1: GLOSSARY AND ACRONYMS

1 INTRODUCTION

This report fits into the RiskPACC workplan as follows:

Work Package (WP): WP5 Tool Development

Task: Task 5.2. – Sentiment tool box

Deliverable: Deliverable 5.2: Completion of sentiment analysis toolbox to measure the RPAG

1.1 Publicsonar Platform

CrowdSense BV was founded in 2013 and is a spin-off from TU Delft and TNO (Netherlands Organisation for Applied Scientific Research). CrowdSense is an information discovery company, analysing publicly available information (PAI) in real-time to support public service organisations with the best information at the point of decision making.

Crowdsense's niche market is the public safety market, supporting use cases like early warning, situational awareness and incident management. Typical clients are governments, police forces, emergency services, and vital infrastructure organisations, which rely on real-time public information in their daily operation.

The PublicSonar platform is a cloud-based online application offered to these clients. The PublicSonar platform analyses millions of online interactions per day, including data from Twitter, under strict privacy requirements, and delivers real-time insights through its online application.

Among Crowdsense's achievements are: (i) the real-time PAI monitoring and analysis capability within all the real-time intelligence centres of multiple national police forces across Europe, (ii) a fully autonomous earthquake detector for a Dutch region, alerting key stakeholders, and (iii) Crowdsense's solution was used to provide operational support during the refugee crisis in Northern Syria, delivering time-critical information about the safety situation.

CrowdSense is participating as a technology partner to exchange best practices in disaster resilience and collaborate with multiple case study partners to offer its platform and expertise in order to transfer this in utilising crowdsourced data.

1.1.1 TOOL USE CASES

Within a risk framework, the tool can assist in early identification of hazards by scanning social media and monitoring reports from citizens on the ground. This includes tracking the hazard's progression, geographical extent, and impact on the environment.

In a social-political context, the tool can aid in identifying the available resources or deficiencies among individuals, which can inform the communication needs from Civil Protection Authorities (CPAs) to the citizens.

The tool can gather information shared by people on social media, which is widely utilized by a significant portion of the population in a given country.
Moreover, citizens often express themselves online using diverse languages and vocabularies. The tool has the capability to search across multiple languages simultaneously, employing a nearly unlimited set of keywords commonly used online.

Given that risk perception can be highly subjective, the tool can analyze people's positive and negative emotions expressed in different languages through sentiment analysis.

Relating: Although direct citizen engagement is not facilitated by the tool, CPAs can utilize it to identify the queries that citizens pose to CPAs, fellow citizens, or non-citizens via social media.

Building: The tool has the potential to revolutionize CPAs' risk communication processes. While it doesn't offer direct user engagement, it enables them to determine the most effective communication channels based on citizens' preferences.

Additionally, by analysing citizens' raw responses and emotions, the tool can assist in assessing the impact of received communications on individuals.

1.2 Aim and Scope

The focus of WP 5 is the adaptation of existing tools and their further development to the needs of CPAs and citizens as identified in the co-creation lab phases I and II. More specifically, it will respond to the user requirements and experiences of citizens and CPAs in order to bridge the RPAG. Finally, the WP will identify the training needs in terms of technology use and input data validation for the development of training material under WP4. It has the following objectives:

- Adapt existing tools in order to bridge the RPAG making use of the co-creation approach and practitioner perspectives in WP3.
- Advance new tools based on VGI that enable citizens to assist with the local resilience efforts.
- Iterate the development of solutions in response to user needs.
- Develop detailed guidance and training materials to assist tool use.

Task 5.2 encompasses the adaptation and further development of the online sentiment analysis toolbox to make it applicable to measure RPAG.

Sentiment analysis is the process of analyzing digital text to determine if the emotional tone of the message is positive, negative, or neutral. It can be used to measure the

tone and polarity of online discourse and, as such, used to measure engagement or (dis)engagement of the public regarding topics of interest.

The adaptation includes the fine-tuning of the AI-based language models to support the key languages of the case study areas.

To that end, the main features / functionalities we have developed or enhanced for the purpose of this task are:

- Automated sentiment analysis for multiple EU languages;
- Negative sub-sentiment analysis
- Per-messaged sentiment analysis
- Sentiment «word-cloud» visualisation
- Sentiment in reports
- Sentiment dashboards

These functionalities will allow CPAs to measure the Public's emotional perception to a given event or the Authorities response to an event, adding a very relevant data point to which CPAs can act upon, and thus reducing the RPAG.

2 FEATURE DEVELOPMENT

2.1 Use Case

During T3.1 several case study partners participated in the presentation of tool use cases to establish a match between the proposed offering and their needs. The following use cases participated in the workshops:

- Attica, Greece (MRP)
- Brussels, Belgium (IBZ/NCCN)
- Eilat, Israel (MDA, MoE)
- Lancashire Constabulary, United Kingdom (LC)
- MoravianSilesian and Olomouc Regions, Czech Republic (CAFO)
- Municipality of Padova, Italy (CDP/MDP)
- Pandemic (ISAR)

The following use cases for Crowdsense's tool were presented:

Natural Hazards:

There is a storm in the Netherlands/ UK. I would like to monitor the situation in real time. I am interested in building searches for cities, neighbourhoods, streets and areas affected by the disaster. I would like to filter the results on casualties. I would like to

see pictures and videos from the disaster **to be able to estimate the impact and make an informed decision about response.**

Public Safety:

There is a shooting in Germany. I would like to get key insights from the event, which are updated real time. Those include the description perpetrator (if known), status (still on sight, under investigation, caught) and location. I would like to know the number of the victims and their status (shot, dead, alive).

The table below show the results of the matching process between CSP and Tech partner. The number in the table indicate how many case studies thought the proposed functionalities was helpful.

CS	Results
Online tool	5
A tool which allows searches on various sources based on keywords or geographical areas	5
Receive input from social media	4
Analyse and process social media data	4
Sentiment analysis	4
Languages: English	3
Languages: (Dutch/German might be considered)	2

TABLE 2: MATCHING PROCESS RESULTS

From the matching results three Case study partners stood out as the best fit for the development of our tool and were therefore chosen to move forward with Crowdsense: ISAR, CAFO and CDP.

2.2 User Needs and Assumptions

Drawing on Task 3.1, a baseline assessment was completed through the sentiment analysis of both real-time and historical data. The development of a sentiment toolbox matched with nearly all case studies, as long as it supports their native language for both search and sentiment. This has validated Crowdsense's assumptions around building a language agnostic sentiment analysis engine.

From the first discussions with the assigned CSPs, we could derive some assumptions and user needs for our tool:

- create situational awareness during incidents and disasters
- get early warning
- have real time monitoring on sentiment
- get insights on event development
- inform my operational response

Information should include:

- Real time sentiment analysis with a sub-sentiment of negativity
- Message level sentiment analysis

The tool should be:

- Easy to use
- Create complex queries in a user friendly way
- Use boolean operators
- Customisable
- Include the possibility for alerting
- Include the possibility to create reports (incl. automated)
- Have different previews of data (e.g. dashboards)

2.3 Feedback from Case Study Partner

The PublicSonar tool was in state of Production at TLR 9 but certain features were at a lower TLR before the first round of workshops. A unilingual sentiment analysis model was available (TLR 9) before the first round of workshops (see chapter 2.5.1)

Crowdsense held bilateral meetings with CSPs before the first round of workshops. The purpose was to show or existing sentiment analysis and semantic search capabilities and collect requirements. As expected, the main feedback was around the lack of multilingual support for sentiment analysis and key search languages missing from our semantic search engine.

As such Research & development of the multi-sentiment architecture started before the workshops as it was the foundation underpinning the whole 5.2 task. This was by far the most complex step of the project. A detailed account of this process is contained in chapter 2.4.2 and 2.4.3

2.3.1 FIRST ROUND OF WORKSHOP

The tool was tested in the first round of workshops by 3 case study partners: Comune di Padova, ISAR and CAFO.

Comune di Padova tested the tool themselves during the workshop and presented two points of feedback.

1) Italian needed to be added as a search language in order to collect data in Italian. 2) CdP wants to retrieve data from at least as far back as June 2022 to show data relating to natural disaster events.

ISAR also tested the tool themselves during the 1st workshop and they mainly had questions about the interpretation of (analytical) data in the tool and how to act on the insights gained with the tool. Sentiment analysis was tested but was not available for German data.

Crowdsense demonstrated the tool to CAFO during the first workshop and the main remark was not having Czech as a search language.

From the workshops it was also apparent that the case study partners would be interested to share the outcomes of the sentiment analysis internally with their colleagues or externally with peers.

2.3.2 INTERMEDIATE IMPROVEMENT PHASE

The first round of workshops validated the early decision to research a new multi-lingual sentiment paradigm. Crowdsense used the intermediate development phase between workshop 1 and 2 to train the new model in German, Czech and Italian.

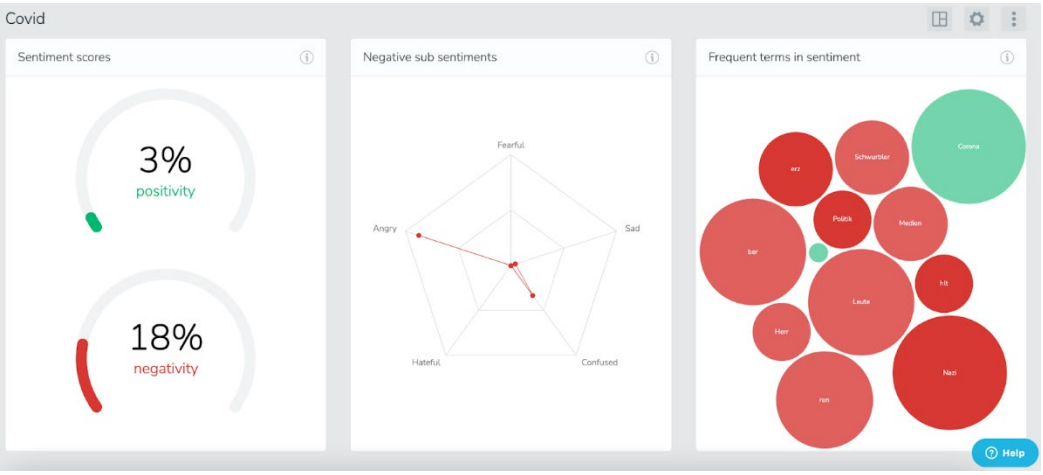

FIGURE 1: SENTIMENT DASHBOARDS

Two features were newly developed between the first and second round of workshops:

- Sentiment widgets in the Dashboard functionality (chapter 2.5.2)
- Sentiment widgets in the Report functionality (chapter 2.5.3)
- Ability to collect data from Twitter in Czech and Italian (chapter 2.5.4)
- Multilingual sentiment analysis in English, German, Italian and Czech (chapters 2.4.2 and 2.4.3)

Crowdsense has developed and / or expanded upon parts of the platform in response to the remarks from the participants in the first round of workshops.

The ability to analyse German data was developed to match the capabilities of the model that was already analysing English data. This includes a positive and negative sentiment score, a sub-sentiment distribution and a wordcloud. Crowdsense also added the ability to use sentiment analysis as part of the 'Dashboard' feature (chapter 2.5.2).

Two new search languages were added in order to collect Tweets in those languages, specifically Czech and Italian. This required an update of our core search engine and associated analytical tooling. (chapter 2.5.2)

2.3.3 SECOND ROUND OF WORKSHOPS

The tool was tested in the second round of workshops by 3 case study partners once again: Comune di Padova, ISAR and CAFO.

Comune di Padova showed various functionalities of the tool to their workshop participants. Among other things, they showed two cases, one for droughts in Venice specifically and one about natural disasters in Italy more generally.

CAFO has demonstrated the tool to their workshop participants by focusing on a specific case with Czech Twitter messages about natural disasters, extreme weather and ukrainian refugees.

CS demonstrated and tested the tool with participants from the ISAR workshop. During this workshop, participants created cases themselves about various topics, such as the Covid pandemic, earthquakes in Turkey, societal unrest and a hostage situation. After testing, participants answered questions about applying the tool in real life scenarios.

The TLR level of the overall software remains at level 9 but the TLR of sentiment analysis improved from TLR 2 / 3 to TLR 5 / 6.

The main feedback received during the second round was the lack of training materials to use the tool. This was addressed as part of T5.4.

2.3.4 GENDER REPRESENTATION

During the workshop testing, organisers strived for good gender representation of workshop participants. This was accomplished across the workshops with 31 females and 40 males overall. This focus on gender balance was to ensure that different perspectives were considered during feedback and development of the tools. This gender focus will be further elaborated on in D3.7, which is due in M24.

2.4 Technical Research & Design

The PublicSonar platform, deployed by CrowdSense, has been developed and customized based on daily experience from several Dutch and EU CPAs over several years. The application of innovative technologies, such as Machine Learning, Deep Learning or NLP, combined with the increase of publicly available data has led CPA users to gain real-time alerts and situational awareness on their domain of expertise. It allows CPAs to recognize at early-stage signs of an emerging disruption and to support operational decisions based a comprehensive overview of an emerging or ongoing event and its effects on the public.

Current semantic search engine functionalities include real-time language-neutral searches that consist of a wide range of word combinations (including slang, names, evolving vocabularies or characteristics) with respect to a (potential) event and can be combined with a variety of location indicators. The user experience is facilitated thanks to functionalities such as AI based on word suggestions; pre-defined building blocks on different kinds of man-made and natural risks and disruptions; as well as analytical dashboards on words, visuals, locations, sources and timelines.

2.4.1 PLATFORM ARCHITECTURE

PublicSonar is a microservice architecture:

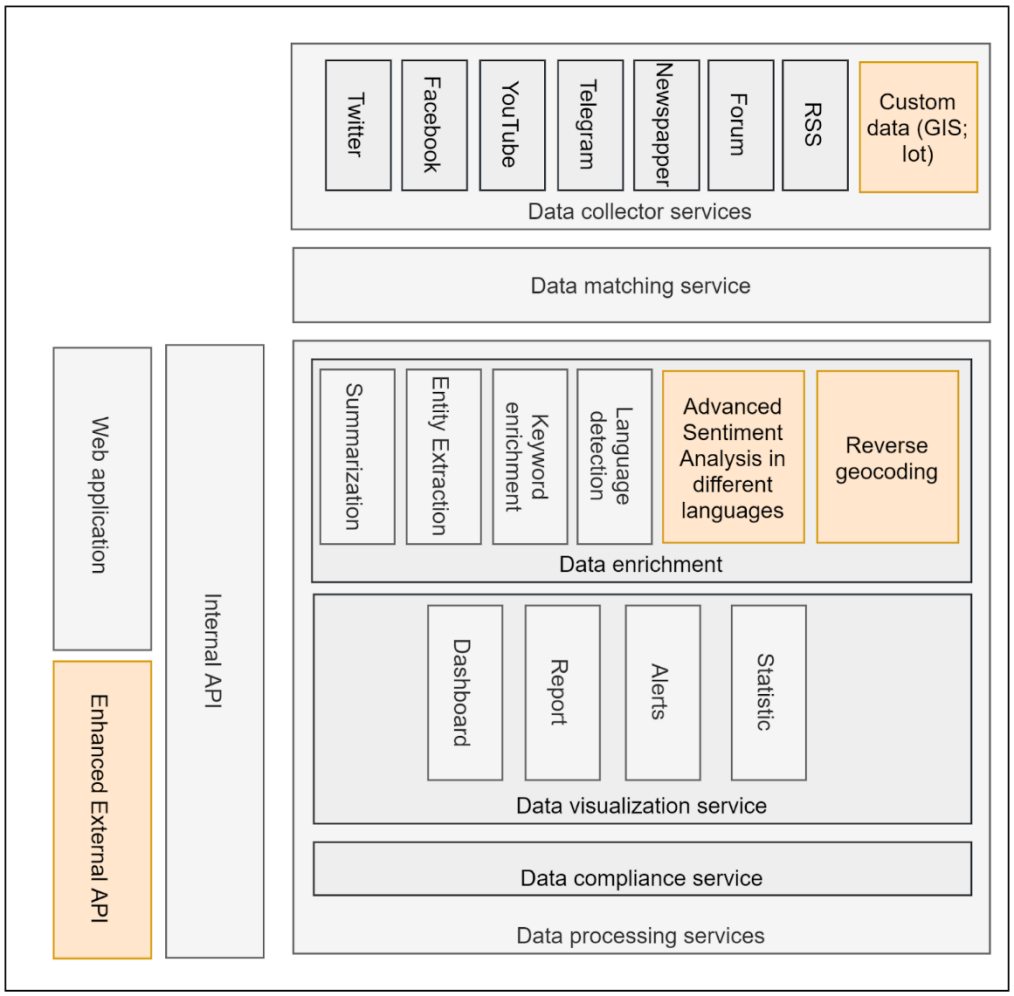

FIGURE 2: LOGICAL ARCHITECTURE DIAGRAM OF PUBLICSONAR

Module	Description
Website	CPA can connect to our website to define which data they need to collect and access to the analysis tools provided by PublicSonar
API	PublicSonar API currently only allows CPA to retrieve data collected on their behalf.

TABLE 3: FRONT END MODULES AND THEIR DESCRIPTIONS FOR THE TOOL

Module	Description
Data collector	CrowdSense connects its PublicSonar application with data provider such as Twitter, under strict privacy requirements, to collect large amounts of data.
Data Macher	The service assigns the correct data with the CPA requests to ensure that no information is missed.

Data enrichment	The services analyse the different raw data collected and enriches them with new information such as sentiment analysis, trends, …The objective of those services is to provide to CPA meta-information that can be useful for their situational analysis.
Data visualization	The services provide data visualization to help the understanding of the information. Information can be used in a report, a dashboard or via automatic alert system.
Data compliance	The compliance service ensures that the data collected comply with GDPR but also with the different T&C of the data providers. To do so, it analyses in real-time the data collected on the behalf of the different CPA and blocks them if they infringe specific regulations.

TABLE 4: BACK-END MODULES AND THEIR DESCRIPTIONS FOR THE TOOL

2.4.2 SENTIMENT ANALYSIS DATA SCIENCE RESEARCH

In order to comprehend the results bellow and the requirements for the model, first and explanation for precision, recall, and the resulting F score will be provided.

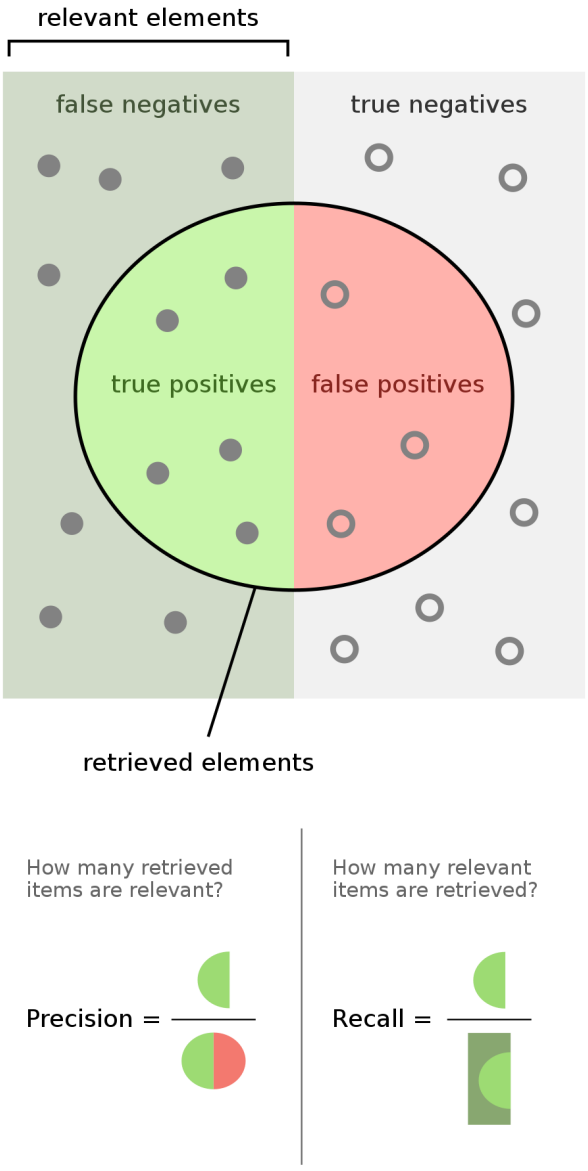

FIGURE 3: PRECISION AND RECALL

In machine learning classification, the **F-score** is a measure of a test's accuracy. It is calculated from the **precision** and **recall**[1] of the test, where the precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive.

As such, the optimal option would be a model with a reasonably high F-score result but also performant both on speed and memory usage.[2]

[1] https://en.wikipedia.org/wiki/Precision_and_recall

[2] https://en.wikipedia.org/wiki/F-score

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021271

The first model tested was the microsoft/Multilingual-MiniLM-L12-H384[3]

	precision	recall	f1-score	support
Pos	0.765	0.802	0.783	12372
Neu	0.651	0.630	0.641	13726
Neg	0.740	0.732	0.736	14512
accuracy			0.719	40610
macro avg	0.719	0.721	0.720	40610
weighted avg	0.718	0.719	0.718	40610

FIGURE 4: MINILM MODEL

Memory consumption 2370MiB and it was able to output 730 msg/s.

Second model was Google's distilbert-base-multilingual-cased[4]

	precision	recall	f1-score	support
Pos	0.765	0.802	0.783	12372
Neu	0.651	0.630	0.641	13726
Neg	0.740	0.732	0.736	14512
accuracy			0.719	40610
macro avg	0.719	0.721	0.720	40610
weighted avg	0.718	0.719	0.718	40610

FIGURE 5: DISTILBERT MODEL

Whilst having a stronger F1 score it faired poorly in both memory usage (3201MiB) and throughput (251 m/s).

[3] https://huggingface.co/microsoft/Multilingual-MiniLM-L12-H384
[4] https://huggingface.co/distilbert-base-multilingual-cased

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021271

Finally, xlm-roberta-base[5] was tested.

	precision	recall	f1-score	support
Pos	0.766	0.807	0.786	12372
Neu	0.658	0.626	0.642	13726
Neg	0.747	0.747	0.747	14512
accuracy			0.725	40610
macro avg	0.724	0.727	0.725	40610
weighted avg	0.723	0.725	0.723	40610

FIGURE 6: ROBERTA MODEL

Once again, results were a bit unconvincing from the performance point of view (3004MiB and 231 m/s).

As the results show, the better option was to move with the MiniLM model due to its capacity to deal with more inputs per second.

An obvious benefit of having sentiment and subsentiment in one ML model is saving on architecture and deployment complexity. We would also save resources as we would merge the two steps in one.

In order to assess whether this is possible without losing accuracy, we have ran the sentiment benchmark while teaching the model to learn the multilabel problem with subsentiment included.

	precision	recall	f1-score	support
Pos	0.765	0.802	0.783	12372
Neu	0.651	0.630	0.641	13726
Neg	0.740	0.732	0.736	14512
accuracy			0.719	40610
macro avg	0.719	0.721	0.720	40610
weighted avg	0.718	0.719	0.718	40610

FIGURE 7: MINILM MODEL

[5] https://huggingface.co/xlm-roberta-base

	precision	recall	f1-score	support
Neg	0.749	0.719	0.734	14512
Neu	0.644	0.652	0.648	13726
Pos	0.767	0.793	0.780	12372
accuracy			0.719	40610
macro avg	0.720	0.721	0.721	40610
weighted avg	0.719	0.719	0.719	40610

FIGURE 8: MINILM MODEL WITH SUBSENTIMENT

Accuracy is kept the same, and thus, we are safe to implement a single model for both sentiment and subsentiment.

2.4.2.1 FINETUNING OF THRESHOLDS

The following figure includes our baseline values before starting the optimization.

	precision	recall	f1-score	support
Neg	0.758	0.773	0.766	3948
Neu	0.654	0.627	0.640	3227
Pos	0.761	0.778	0.769	2495
accuracy			0.726	9670
macro avg	0.724	0.726	0.725	9670
weighted avg	0.724	0.726	0.725	9670
	precision	recall	f1-score	support
Angry	0.675	0.630	0.652	1921
Confusion	0.676	0.747	0.710	1355
Fearful	0.643	0.497	0.561	344
Hateful	0.583	0.460	0.515	1227
Sad	0.588	0.560	0.574	780
micro avg	0.645	0.603	0.623	5627
macro avg	0.633	0.579	0.602	5627
weighted avg	0.641	0.603	0.619	5627
samples avg	0.279	0.273	0.267	5627

FIGURE 9: NO FINETUNING OF THRESHOLDS

The objective is to find metrics that satisfy our business requirements. It was decided to maximize F1 score keeping a precision above 80% for Negative and Positive labels.

```
THS_NEG = 0.78
THS_POS = 0.68
              precision    recall  f1-score   support

         Neg      0.801     0.705     0.750      3948
         Neu      0.597     0.724     0.654      3227
         Pos      0.803     0.734     0.767      2495

    accuracy                          0.719      9670
   macro avg      0.734     0.721     0.724      9670
weighted avg      0.733     0.719     0.722      9670
```

FIGURE 10: INCREASING F1 SCORE

Maximizing the F1 would deliver:

```
THS_NEG = 0.4
THS_POS = 0.48
              precision    recall  f1-score   support

         Neg      0.755     0.786     0.770      3948
         Neu      0.655     0.621     0.638      3227
         Pos      0.770     0.771     0.770      2495

    accuracy                          0.727      9670
   macro avg      0.727     0.726     0.726      9670
weighted avg      0.725     0.727     0.726      9670
```

FIGURE 11: MAXIMIZING F1 SCORE

It's preferable to recall less but being more precise in the sentiments shown. Especially because the interface shows which messages have Negative and Positive labels.

2.4.2.2 CONFIDENCE THRESHOLDS

The precision (being correct when giving a label) in the probabilities above originally looked like this:

```
NEG
0.78  :  0.8006
0.79  :  0.8017
0.8  :  0.8036
0.81  :  0.8052
0.82  :  0.809
0.83  :  0.8104
0.84  :  0.8122
0.85  :  0.813
0.86  :  0.8165
0.87  :  0.8179
0.88  :  0.8212
0.89  :  0.8236
0.9  :  0.8272
0.91  :  0.8346
0.92  :  0.8384
0.93  :  0.8415
0.94  :  0.8461
0.95  :  0.8517
0.96  :  0.8596
0.97  :  0.8658
0.98  :  0.8743
0.99  :  0.8964
```

FIGURE 12: NEGATIVE OUTPUT PRECISION

The precision range is broken in three gaps: 0.8006, 0.8485 and 0.8964 . The thresholds that apply would be 0.78-0.94 low, 0.94-0.99 medium, >0.99 high.

```
NEG            # Samples
0.78-0.94 low       715
0.94-0.99 medium    1053
>0.99     high      1708
```

FIGURE 13: NEGATIVE OUTPUT THRESHOLDS

```
0.68  :  0.8032
0.69  :  0.8052
0.7   :  0.8069
0.71  :  0.808
0.72  :  0.8108
0.73  :  0.8131
0.74  :  0.8136
0.75  :  0.8153
0.76  :  0.8163
0.77  :  0.8188
0.78  :  0.82
0.79  :  0.8217
0.8   :  0.8227
0.81  :  0.8265
0.82  :  0.8265
0.83  :  0.8287
0.84  :  0.8316
0.85  :  0.8357
0.86  :  0.8383
0.87  :  0.8411
0.88  :  0.8433
0.89  :  0.8469
0.9   :  0.8488
0.91  :  0.8523
0.92  :  0.8582
0.93  :  0.8656
0.94  :  0.8689
0.95  :  0.8766
0.96  :  0.8882
0.97  :  0.8964
0.98  :  0.9157
0.99  :  0.953
```

FIGURE 14: POSITIVE OUTPUT PRECISION

Precision range was broken in three gaps 0.8032, 0.8781 and 0.953. The thresholds that apply would be 0.68-0.95 low, 0.95-0.99 medium, >0.99 high.

```
POS           # Samples
0.78-0.95 low       355
0.95-0.99 medium    819
>0.99     high      915
```

FIGURE 15: POSITIVE OUTPUT THRESHOLDS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021271

As for subsentiment the thresholds that were defined are as follows:

```
angry
Best THS: 0.59
PRECISION 0.7
RECALL 0.709
F1 0.705

confusion (including neutral)
Best THS: 0.58
PRECISION 0.702
RECALL 0.699
F1 0.7

fearful
Best THS: 0.74
PRECISION 0.774
RECALL 0.562
F1 0.652

hateful
Best THS: 0.54
PRECISION 0.601
RECALL 0.55
F1 0.574

sad
Best THS: 0.6
PRECISION 0.65
RECALL 0.63
F1 0.64
```

FIGURE 16: NEGATIVE SUBSENTIMENT OUTPUT THRESHOLDS

2.4.2.3 POSITIVE REINFORCEMENT

Testing indicated that the outputs had a negative bias, ie, the model had a tendency to overly label messages as negative. To correct this the models were restrained with a positive reinforcement bias. The results are as follows:

	precision	recall	f1-score	support
Neg	0.748	0.768	0.758	3948
Neu	0.631	0.608	0.619	3227
Pos	0.749	0.753	0.751	2495
accuracy			0.711	9670
macro avg	0.709	0.710	0.709	9670
weighted avg	0.709	0.711	0.710	9670

	precision	recall	f1-score	support
Angry	0.683	0.599	0.638	1921
Confusion	0.680	0.723	0.701	1355
Fearful	0.671	0.439	0.531	344
Hateful	0.627	0.460	0.531	1227
Sad	0.590	0.553	0.570	780
micro avg	0.658	0.582	0.618	5627
macro avg	0.650	0.554	0.594	5627
weighted avg	0.657	0.582	0.614	5627
samples avg	0.271	0.262	0.259	5627

FIGURE 17: POSITIVE REINFORCEMENT TRAINING

2.4.2.4 FINAL ANNOTATION AND TRAINING

It was decided to use artificially generated data from LLMs (ChatGPT and LLaMA) as the input dataset for our final round of training. This proved a very scalable approach with the limited resources at hand, with a negligible loss in annotation quality.

Once again, the focus was kept on positive and neutral messages in order to eliminate the negative bias.

	precision	recall	f1-score	support
Neg	0.772	0.759	0.765	3948
Neu	0.640	0.679	0.659	3227
Pos	0.797	0.756	0.776	2495
accuracy			0.731	9670
macro avg	0.736	0.731	0.733	9670
weighted avg	0.734	0.731	0.732	9670
	precision	recall	f1-score	support
Angry	0.718	0.555	0.626	1921
Confusion	0.663	0.752	0.705	1355
Fearful	0.794	0.369	0.504	344
Hateful	0.622	0.404	0.490	1227
Sad	0.648	0.418	0.508	780
micro avg	0.677	0.539	0.600	5627
macro avg	0.689	0.500	0.567	5627
weighted avg	0.679	0.539	0.592	5627
samples avg	0.250	0.243	0.239	5627

FIGURE 18: FINAL TRAINING RESULTS

2.4.3 MULTISENTIMENT ARCHITECTURE

The previous sentiment architecture was a two-step process. First, the messages are sent to a queue for main sentiment to be analysed (Positive, Negative or Neutral). If the message sentiment is negative or neutral, it is directed to another queue for the subsentiment/s to be extracted. This process happens in three pipelines, one for each language supported (EN, NL & DE).

As this approach is not scalable to many more languages, we decided to implement a multilingual model; a model that is pretrained in many languages so that it can perform a fine-tuning task (in this case sentiment analysis) in any of them. The multilingual model will also condense both analyses (sentiment and subsentiment) in one run.

Running both sentiment and subsentiment in one service will simplify the pipeline but it will also require some changes in the backend.

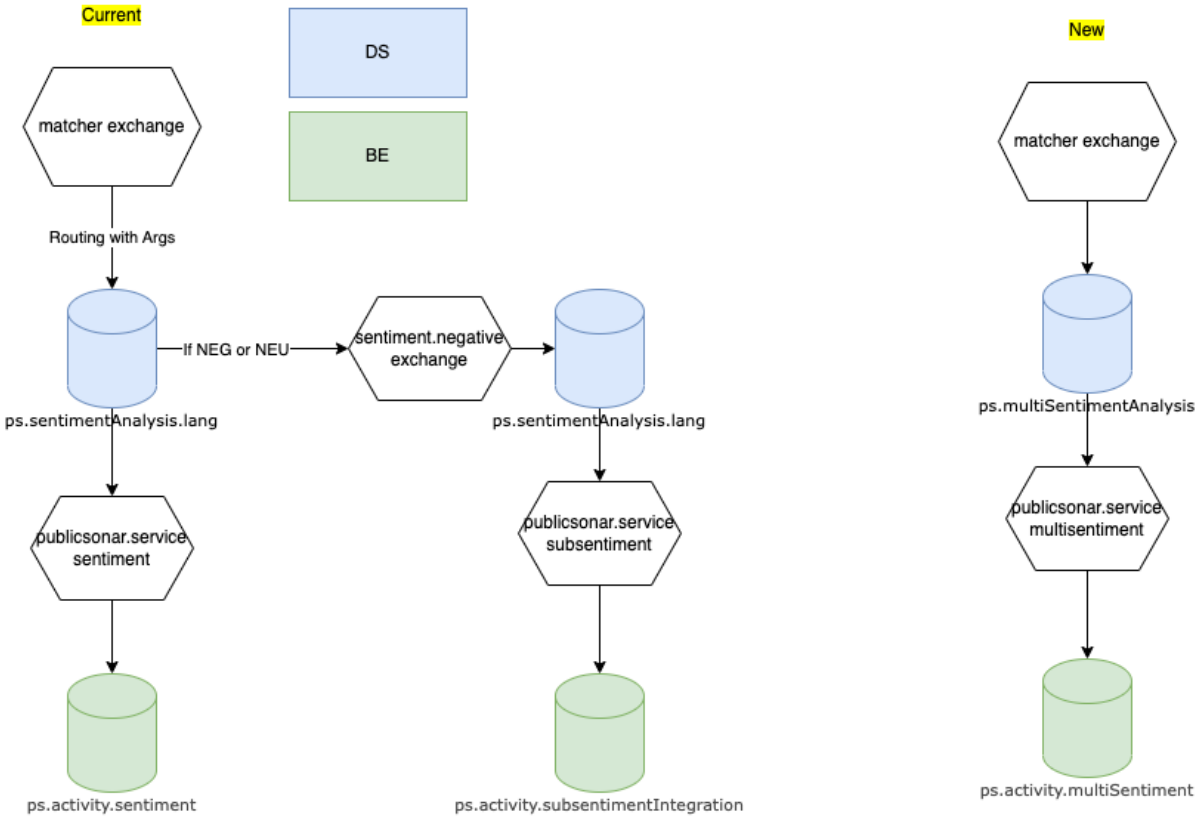

FIGURE 18: MULTISENTIMENT ARCHITECTURE

2.4.3.1 API OUTPUT

The output will stay the same, it will only merge sentiment and subsentiment.

Current output looks like this:

Sentiment

```
{
    'uri': 'message_URI',
    'index': 'case index',
    'sentiment': {
        'class': 'negative' | 'neutral' | 'positive',
        'confidence': 'low' | 'medium' | 'high',
        'probability': prob
        }
}
```

FIGURE 19: SENTIMENT API OUTPUT

SubSentiment

```
{
    'uri': 'message_URI',
    'index': 'case index',
    'subsentiment': {
        'angryScore': prob_angry,
        'confusionScore': prob_confusion,
        'fearfulScore': prob_fearful,
        'hatefulScore': prob_hateful,
        'sadScore': prob_sad
    }
}
```

FIGURE 20: SUBSENTIMENT API OUTPUT

The merged output would look like:

```
{
    'uri': 'message_URI',
    'index': 'case index',
    'sentiment': {
        'class': 'negative' | 'neutral' | 'positive',
        'confidence': 'low' | 'medium' | 'high',
        'probability': prob
    },
    'subsentiment': {
        'angryScore': prob_angry,
        'confusionScore': prob_confusion,
        'fearfulScore': prob_fearful,
        'hatefulScore': prob_hateful,
        'sadScore': prob_sad
    }
}
```

FIGURE 21: SUBSENTIMENT API OUTPUT

2.4.3.2 DEPLOYMENT

Deployment will run in a newly built GPU server architecture. NVIDIA's Triton server was chosen for this application (see appendix 1). This will load balance requests and will accommodate all GPU services.

Each environment will have a Python handler to batch the messages from RabbitMQ and call the triton-server.

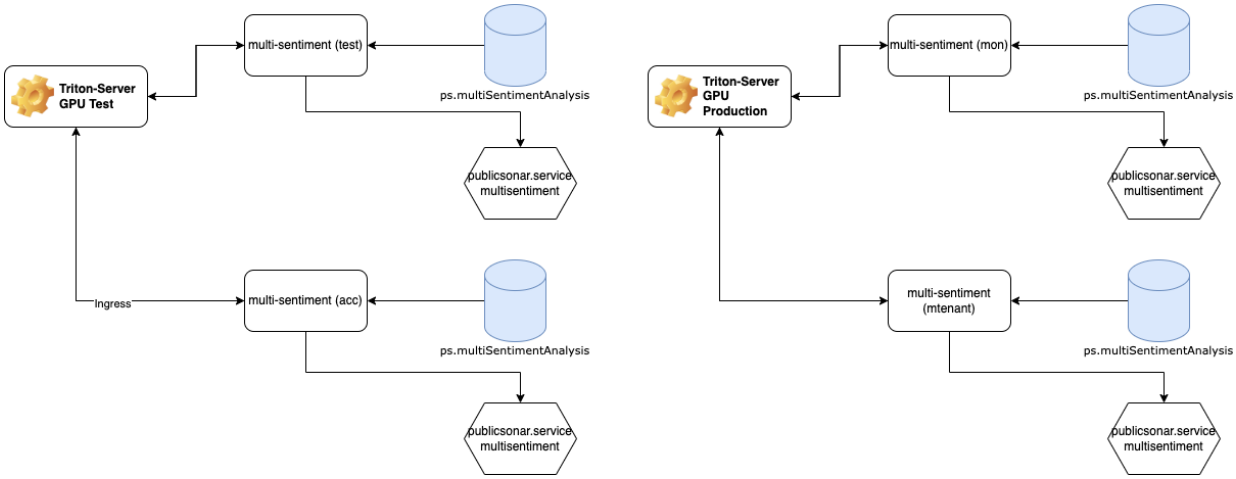

FIGURE 22: MESSAGE QUEUING

The idea is to store models in GCP storage, pull them by triton-server service and deploy them. Thus, deployment of server is independent from deployment of business logic and data workflow. This is not optimal in the sense that business services will have some downtime if one of the models is deployed again. However, GPU balancing is the strong point of NVIDIA Triton and thus it makes sense to rely on it for managing the models (apart from the performance boost in prediction speed).

2.4.3.3 PERFORMANCE

Previous sentiment model has 360 msg/sec throughput with 2 full CPUs and 160 mssg/sec with 1 full CPU.

New GPU model would have at least 1,750 mssg/sec throughput in T4 GPU (not including request overhead).

Besides this, scalling is very easy: resizing to the number of instances of the model we deploy either manually or by setting up an autoscaler based on GPU duty cycle in GKE.

2.5 Functional Design

2.5.1 SENTIMENT ANALYSIS IN CASE PAGE

The design is based on a principle followed by the world of journalism, popularly known as the Inverted Triangle[6] or bottom line up front (BLUF)[7]. As per this principle, the news content is divided into 3 segments on the order of diminishing significance.

[6] https://en.wikipedia.org/wiki/Inverted_pyramid_(journalism)
[7] https://en.wikipedia.org/wiki/BLUF_(communication)

Principle		Applied to sentiment analysis
Insight	Most important information	These are the insights in the sentiment score or the frequent terms graph.
⬇		
Summary	Most important details. Overview of trends about the insight, relevant information	Details like the top messages we are most confident about.
⬇		
Original data	Background information. Finer details that only a few users would want to explore. Allows users to deep dive.	Original messages, visuals and other details.

FIGURE 23: INVERTED TRIANGLE DESIGN APPROACH

The user can move from the insight to the summary and eventually to the original data depending on their needs:

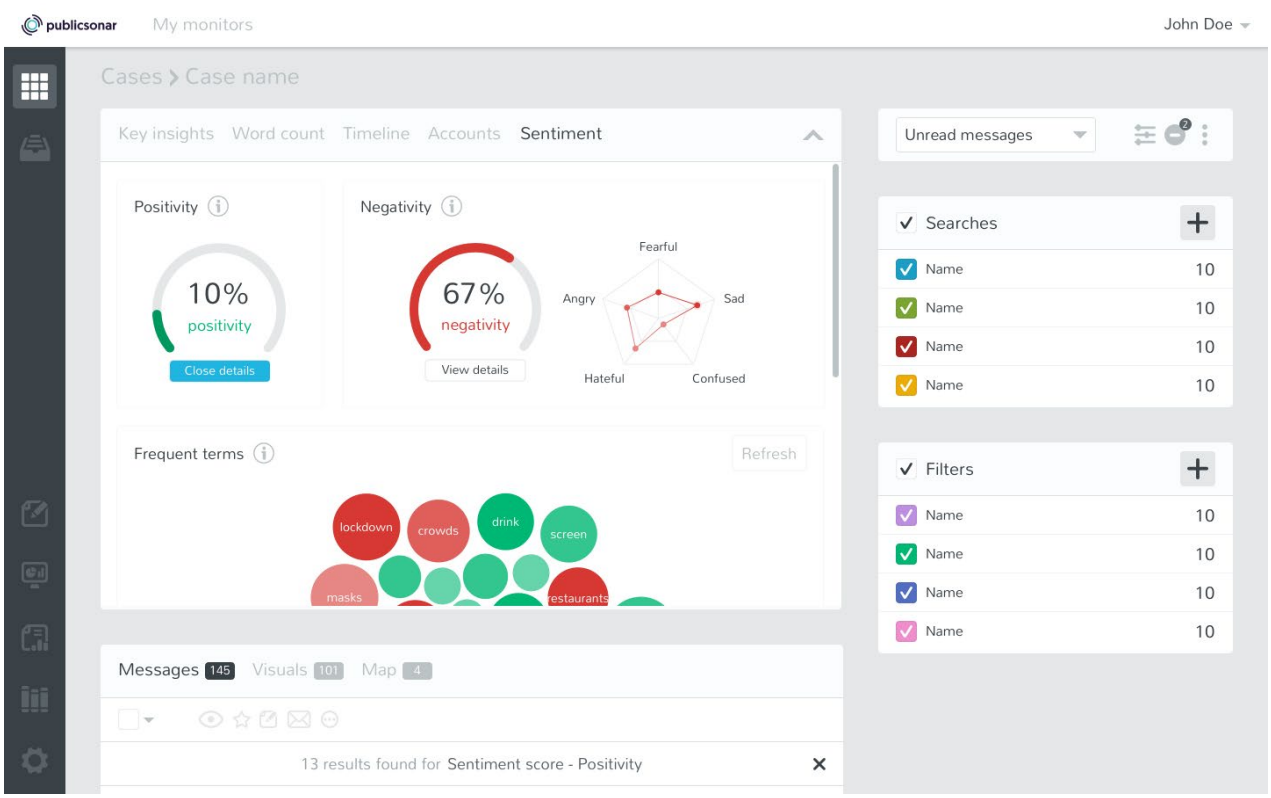

FIGURE 24: INSIGHT

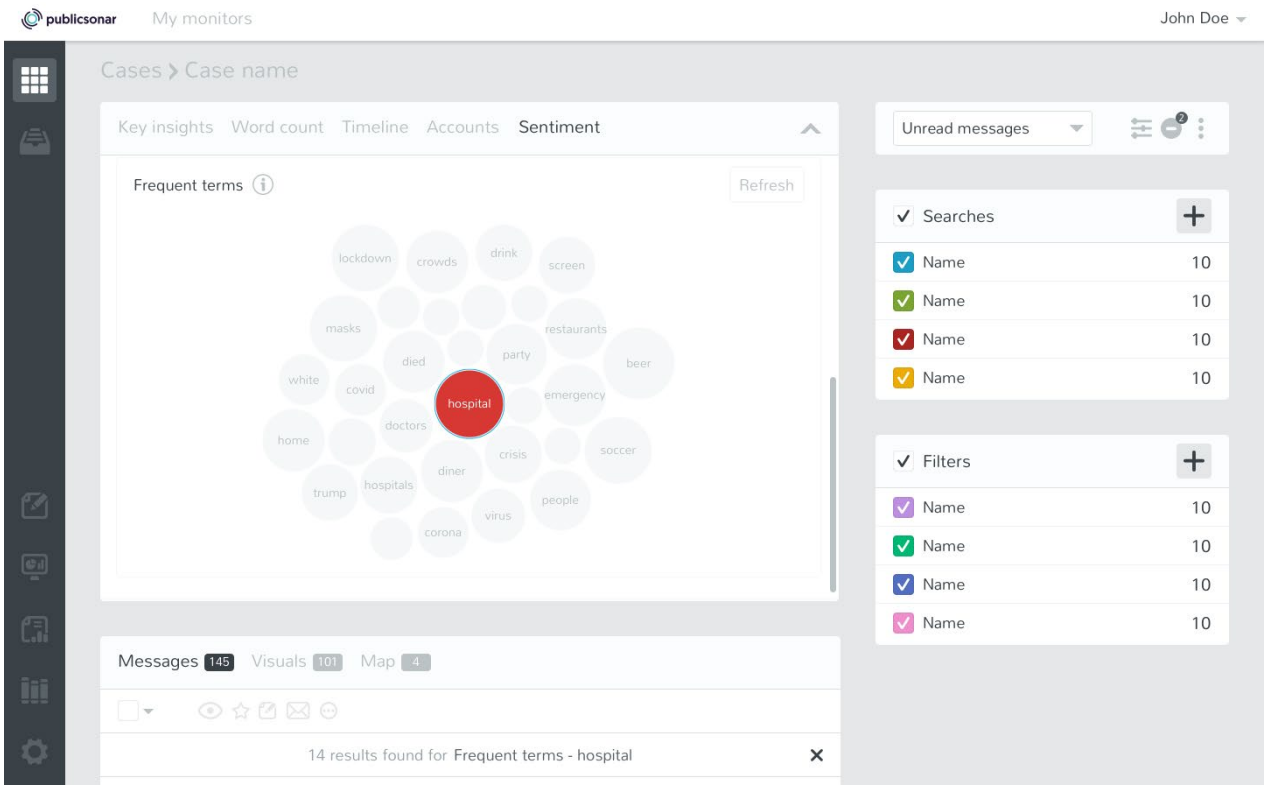

FIGURE 25: INSIGHT SELECTION

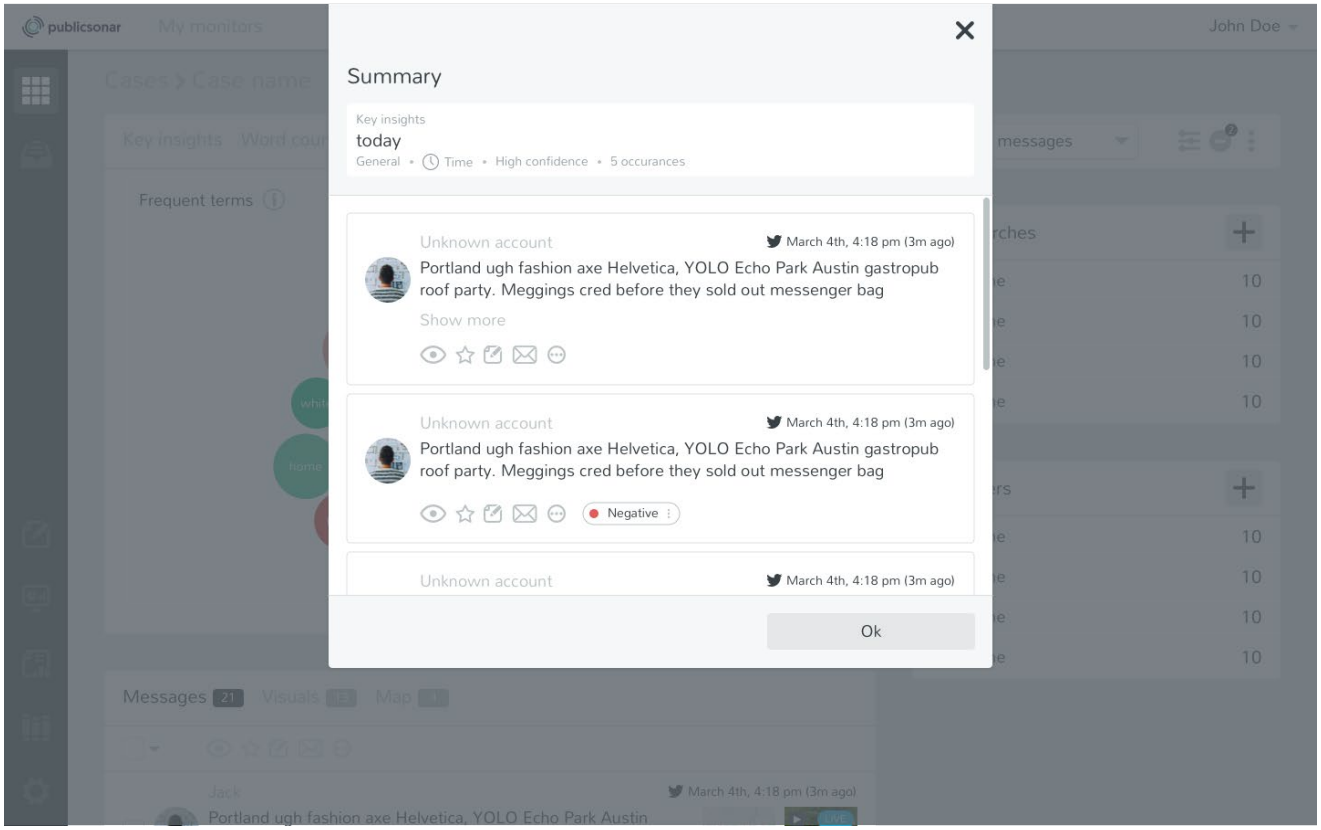

FIGURE 26: SUMMARY & ORIGINAL DATA

From the messages in the selected searches and filters, the top 8 messages with the highest confidence are picked[8].

These 8 messages should be sorted chronologically, with the newest first.

Each message contains the same elements as in the feed in the messages tab and the user should be able to do the same actions[9].

The messages can either show the usual coloured highlights of selected searches / filters in the case, or not.

When there are less than 8 messages, all messages are shown in the modal.

A button 'View all' is placed underneath the last message (also when there are there are less 8 messages).

The user can close the modal by clicking close or 'Ok'.

2.5.1.1 SENTIMENT LABEL ON INDIVIDUAL MESSAGE

Each message with a detected positive of negative sentiment should have a sentiment label. The sentiment label is displayed on the right side of the message actions.

The label shows either 'Negative' or 'Positive', 1 message can only have 1 sentiment. The dot in the label shows the confidence level.

The colour of the dot is Red for 'Negative' and Green for 'Positive'.

Similar to the entity labels, the brightness of the dot corresponds with a low, medium or high confidence level.

[8] The purpose is not to show too few messages and have the user click on 'View all' all the time. The popup will show a sample of messages that we are most confident about in the chosen sentiment selection. Testing indicated that 8 is a good number of message to achieve this goal.

[9] like annotations and clicking on 'show more' or a visual

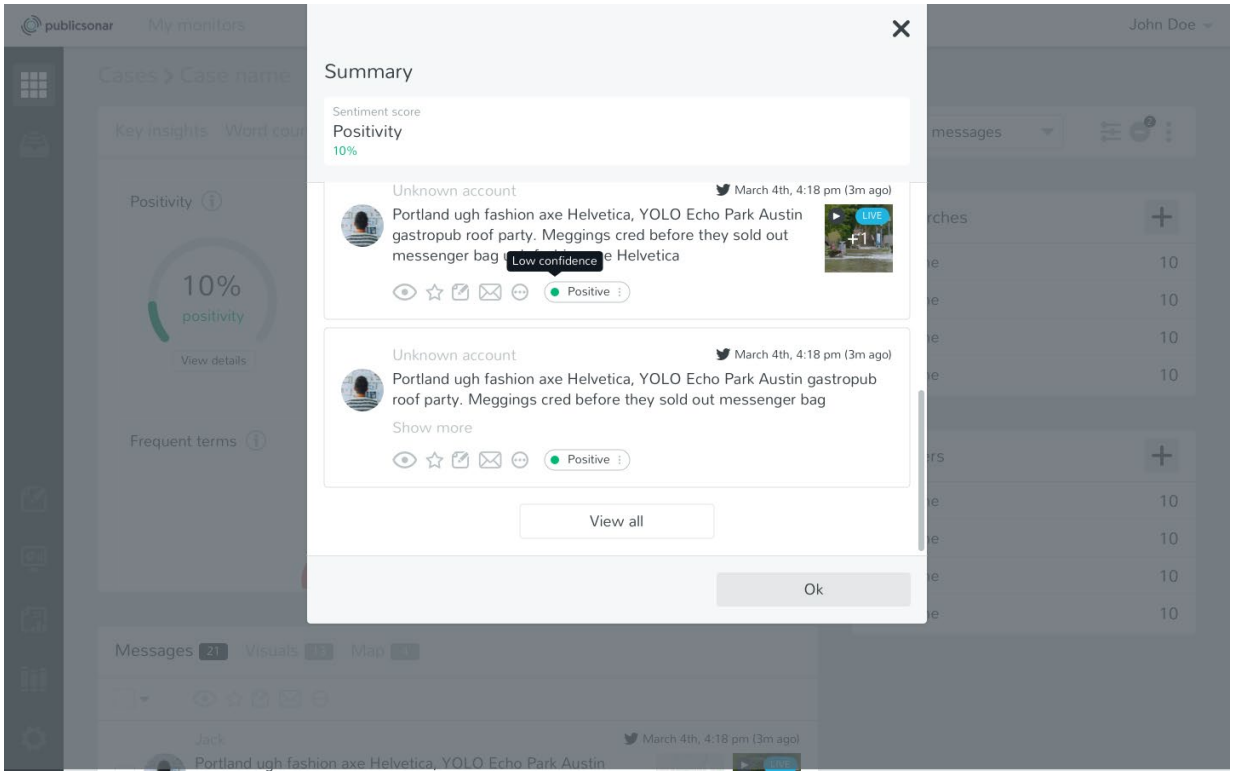

FIGURE 27: MESSAGE LEVEL SENTIMENT

2.5.1.2 USER FEEDBACK

The user can report a message as 'not positive' or 'not negative'.

Similar to the entity labels, the user can click on the 3 dots to open a dropdown. The dropdown shows 1 option to give feedback on the given sentiment.

When reported:

- the label should turn to crossed out state.
- the message will be used by the data annotators to improve the Sentiment models.
- When a sentiment label is in crossed out state, the user can click on the 3 dots to open a dropdown. The dropdown shows 1 option to undo the feedback on the given sentiment.

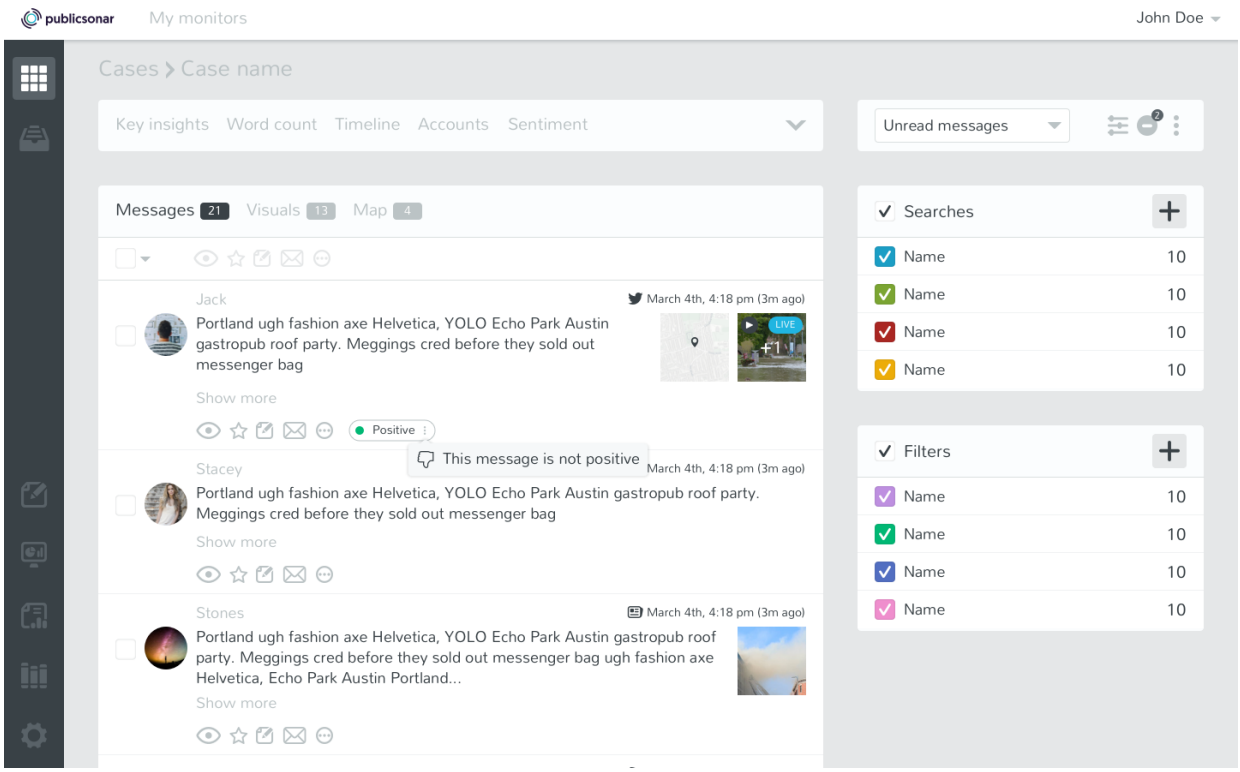

FIGURE 28: USER FEEDBACK

2.5.1.3 RESPONSIVENESS

This feature supports a variety of screen sizes and platforms.

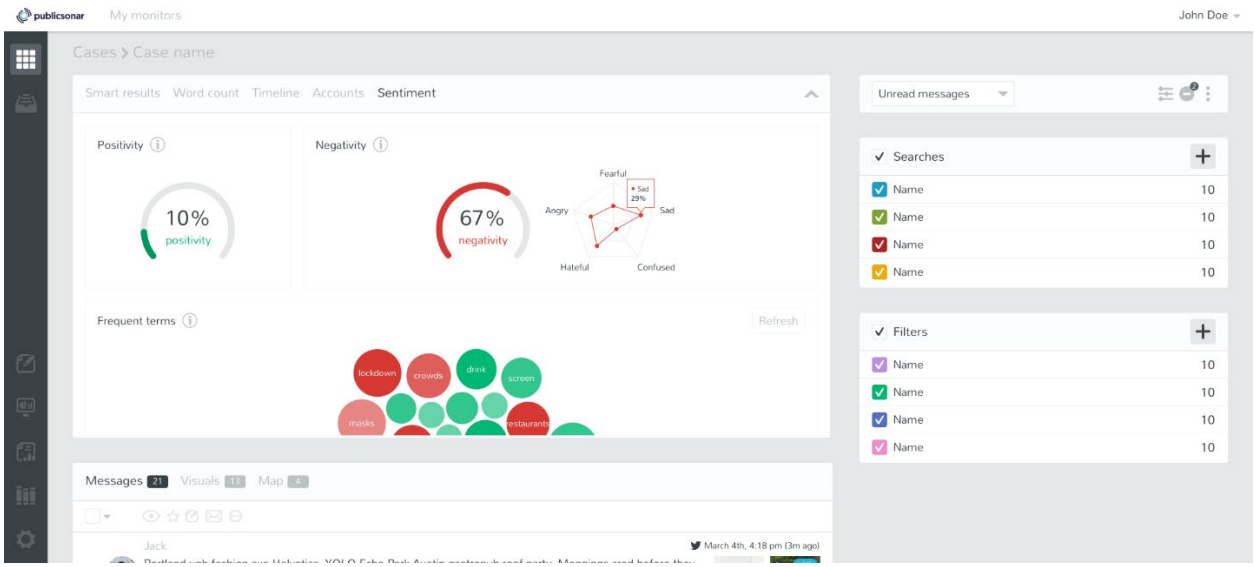

FIGURE 29: DESKTOP LARGE

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021271

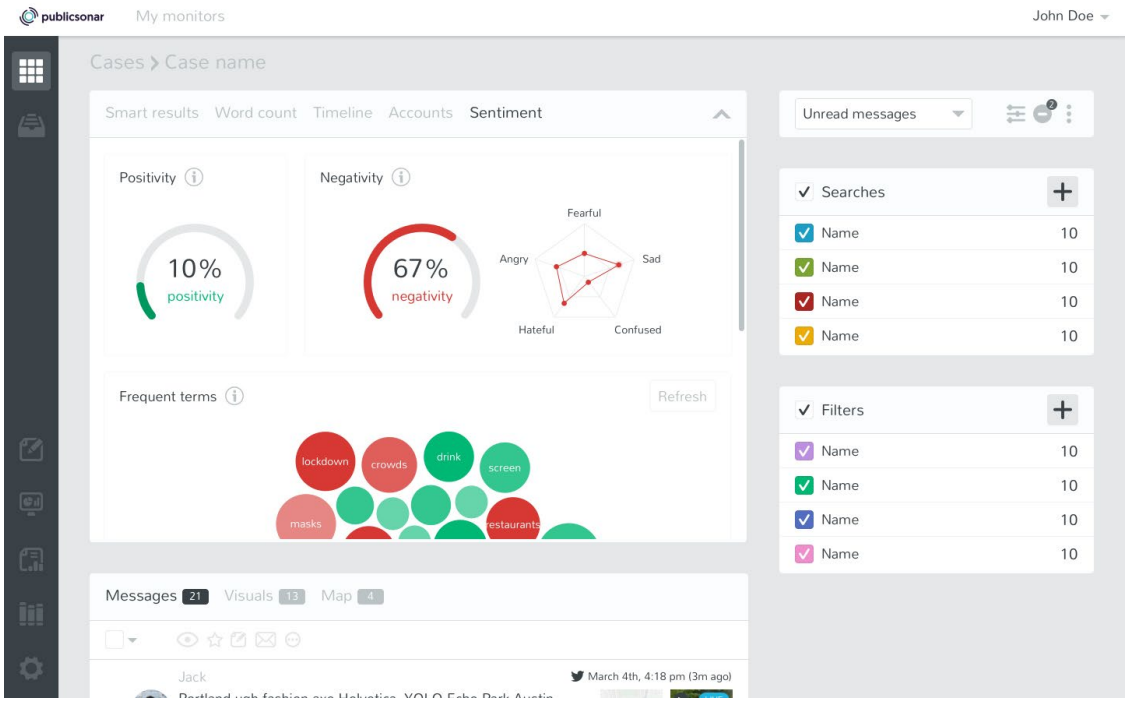

FIGURE 30: DESKTOP SMALL

FIGURE 31: TABLET

2.5.2 SENTIMENT ANALYSIS IN DASHBOARDS

Three new dashboard widgets were created:

- 'Sentiment Scores' – the positive/negative gauges
- 'Negative Sub Sentiments' – sub-sentiment spider web
- 'Frequent Terms in Sentiment' – sentiment word cloud

Unlike the case page, these elements are not clickable.

These widgets interact with other widgets – i.e. if you filter data on another widget it will recalculate sentiment automatically.

We also made responsiveness adjustments to the dashboards so it supports multiple screen sizes and platforms (see Figures 32, 33 and 34).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021271

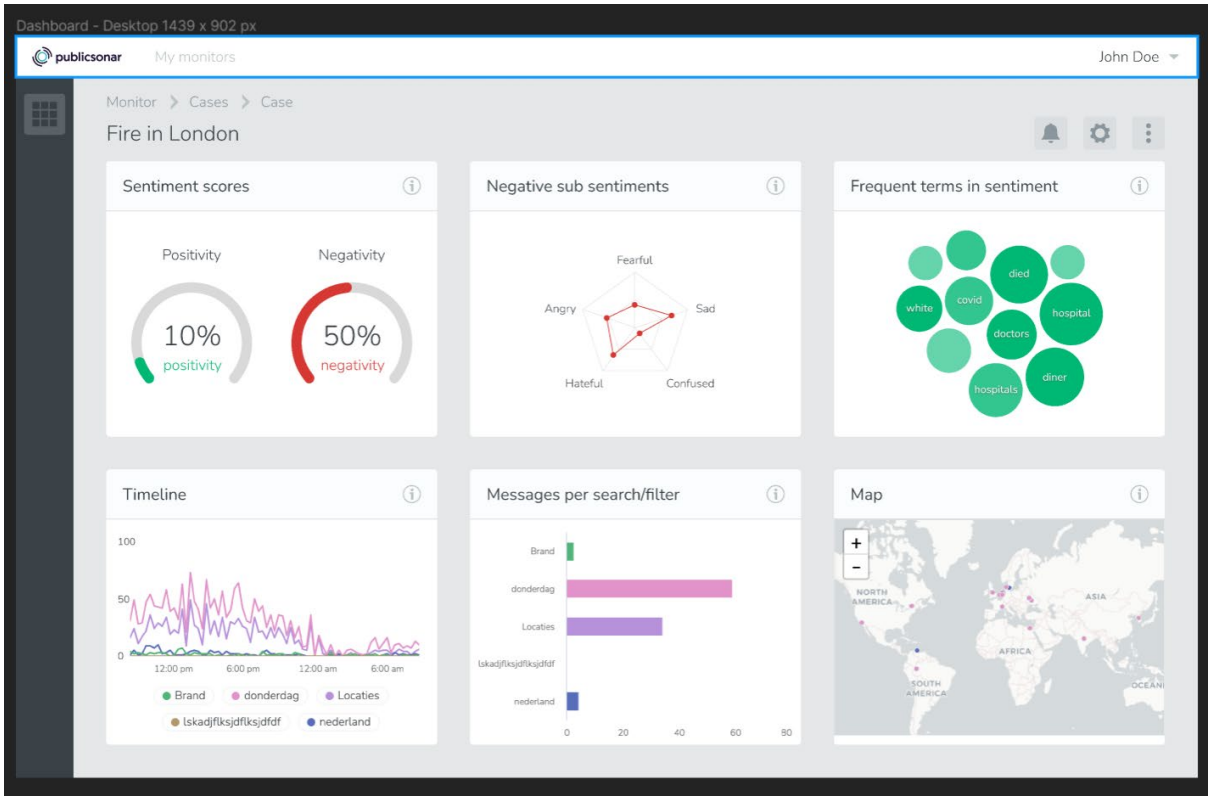

FIGURE 32: DASHBOARD SENTIMENT WIDGETS DESKTOP

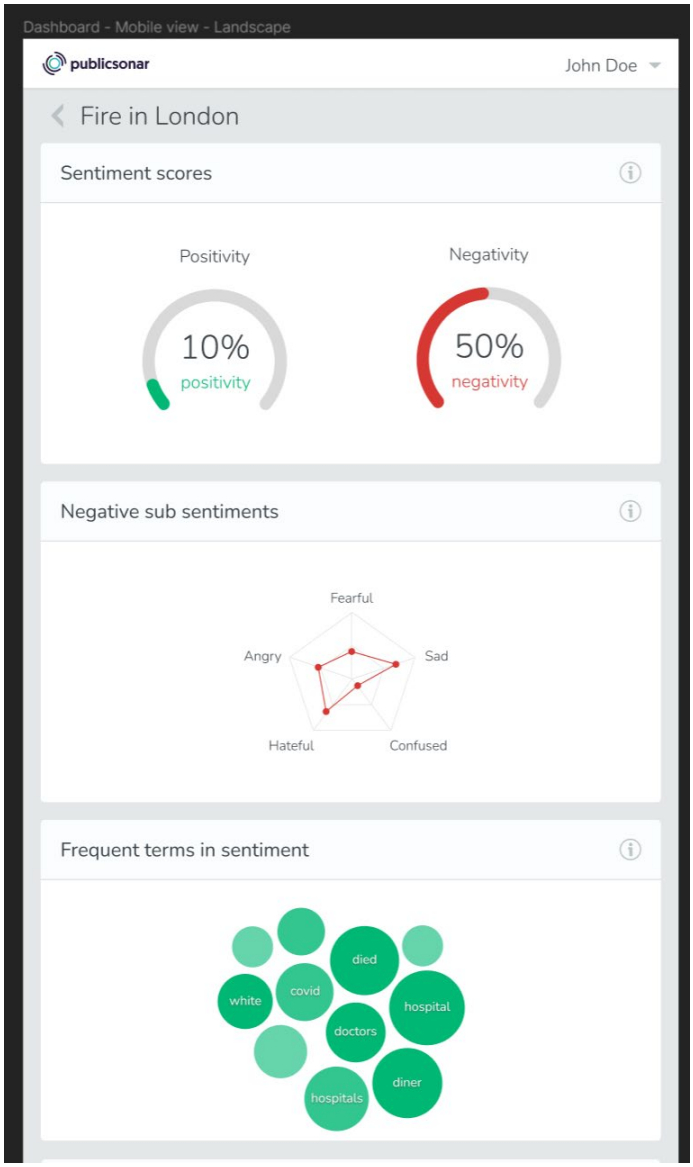

FIGURE 33: DASHBOARD SENTIMENT WIDGETS TABLET

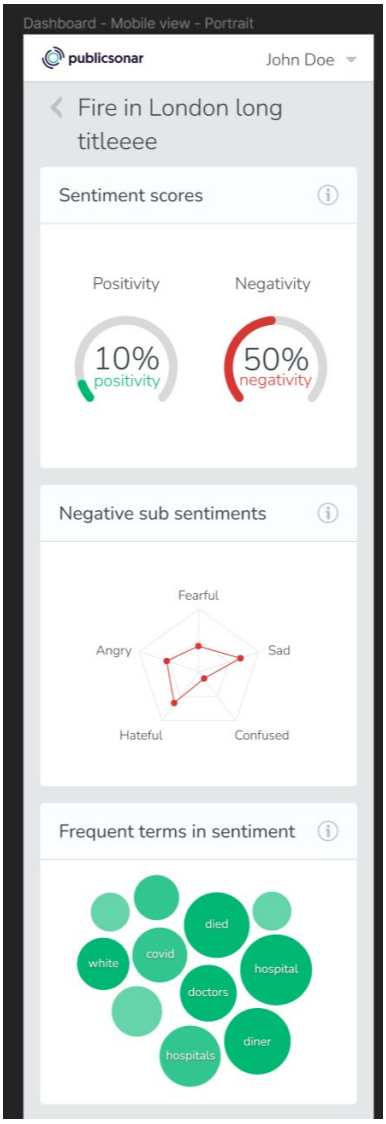

FIGURE 34: DASHBOARD SENTIMENT WIDGETS SMARTPHONE

2.5.3 SENTIMENT ANALYSIS IN REPORTS

In the interface three additional report items were built that the user can add to the report. These items are listed at the end of the existing list of report items and in this order:

- Sentiment scores
- Negative sub sentiments
- Frequent terms in sentiment

These items are not part of the quick report template, instead, user adds the manually.

Per visualization, the user can define the 'Type of messages', 'Searches', and 'Filters' to reflect in the visualization. Same as in other report items.

Each one of the report items has its special image and text.

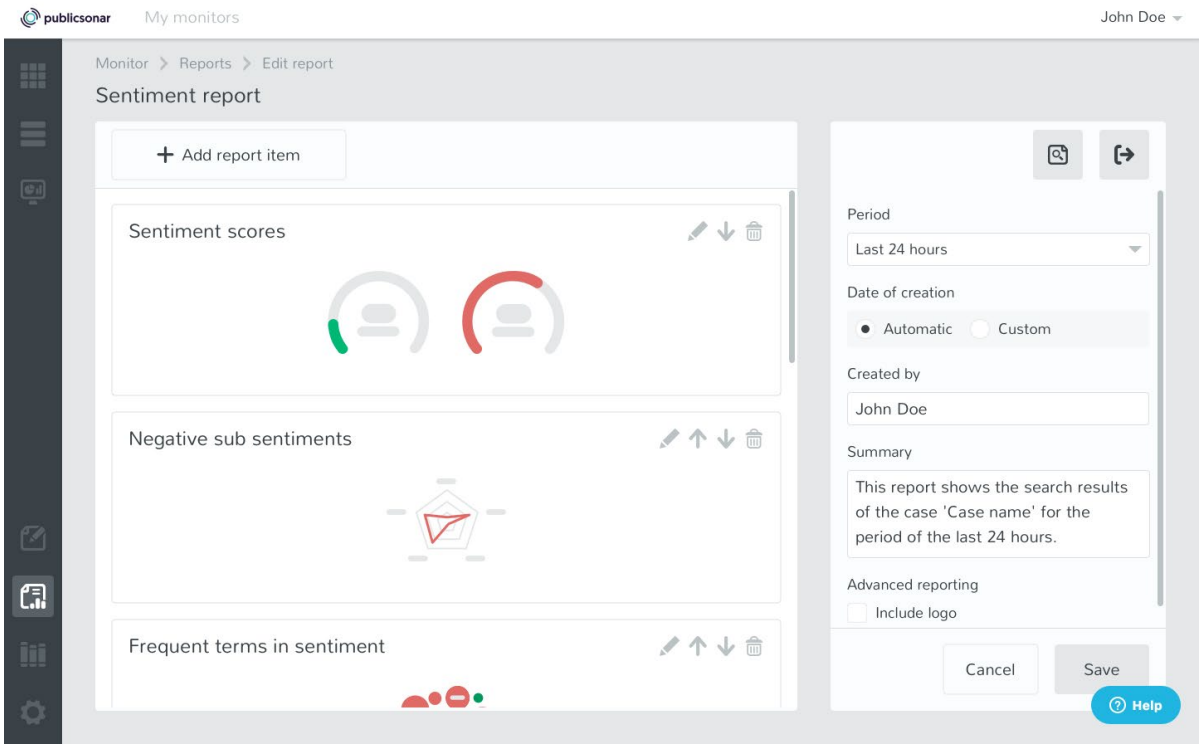

FIGURE 35: SENTIMENT REPORT ITEMS

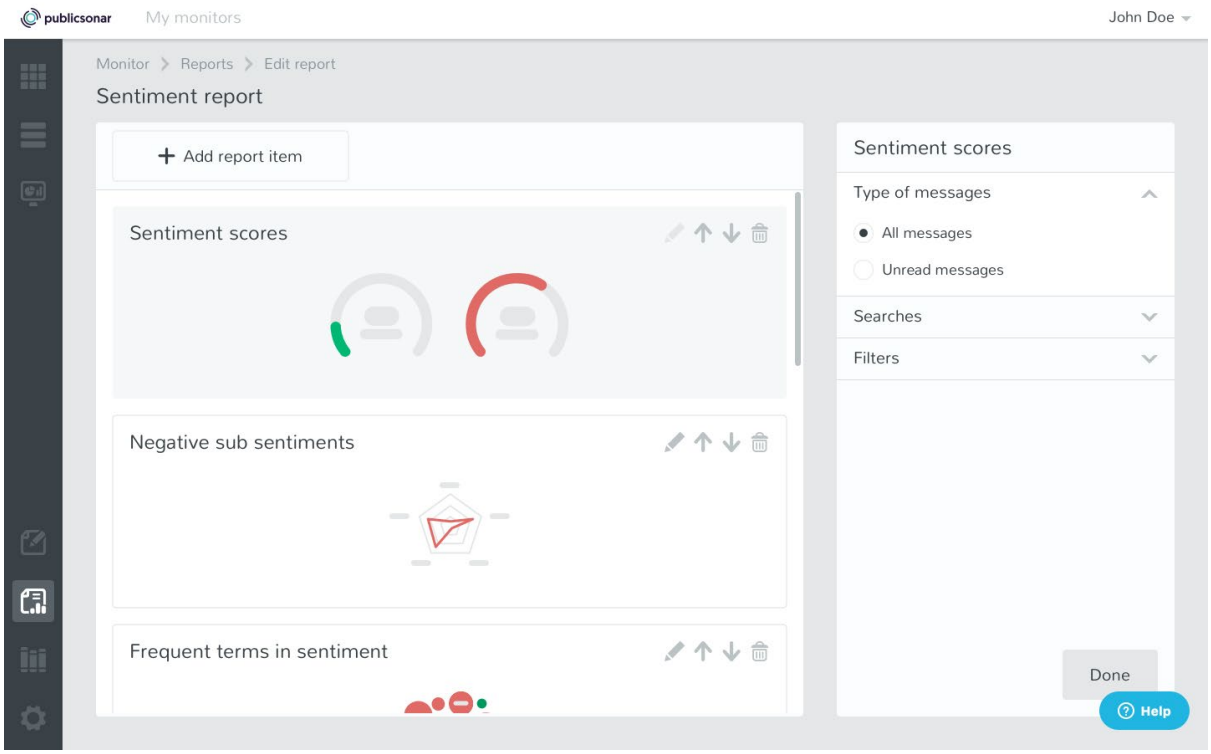

FIGURE 36: SENTIMENT REPORT ITEMS

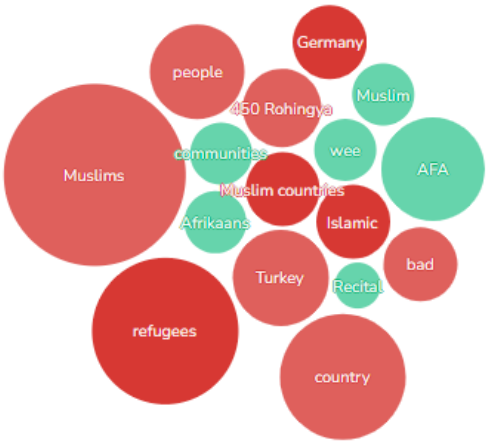

FIGURE 37: SENTIMENT REPORT EXAMPLE

2.5.4 ITALIAN AND CZECH SEARCH LANGUAGES

Publicsonar platform is extensible by design, so adding new search languages, although not trivial, is somewhat a routinely task. These were the steps followed to add Czech and Italian language support:

- Czech and Italian was added to Publicsonar's semantic search engine.
- The ability to set the new languages to monitors was implemented
- Building Block library was updated to support terms as well as offer suggestions in the new languages.
- Text summarization feature was updated to support the new languages
- Search and wordcount "Stopwords"[10] were put in place for the new languages.
- Column names in the export files were extended

2.5.5 FUTURE IMPROVEMENT

One feedback topic mentioned by CdP was the ability to collect historical data and build cases set in the past. Publicsonar is predominantly a realtime tool, but data can be stored and retrieved up to 18 months once the case is created. To research a historical topic, for which no case has been created, an investigation was conducted and several constraints were uncovered:

[10] Stop words are the words in a stop list (or stoplist or negative dictionary) which are filtered out (i.e. stopped) before or after processing of natural language data (text) because they are insignificant.

- Sources: at the moment only Twitter allows for such historical searches. All other sources have a near-realtime focus.
- Cost: Access to Twitter's full-archive comes at a steep price, and was out of scope for the RiskPACC project. Furthermore, historical collection adds to total consumption quotas, thus meaning less quota available for realtime data.
- Complexity: As stated, Publicsonar's ethos is realtime insights, and the design thinking and product positioning reflect this ethos. Adding historical capability would require a complex refactor.

Nevertheless, the value of such capability is obvious, and Crowdsense will continue to work with data sources and CPAs to deliver relevant insights, realtime or otherwise.

3 CONCLUSION

Sentiment analysis plays a crucial role in capturing and understanding perceptions and actions related to risk from both the citizen and Civil Protection Authority (CPA) perspective:

- Sentiment analysis allows for the evaluation of citizens' emotions, opinions, and attitudes expressed in their online interactions and communications. By analyzing sentiment, CPAs can gain valuable insights into how citizens perceive and react to various risks and hazards. This information helps CPAs understand the concerns, fears, and priorities of the public, enabling them to tailor their risk communication strategies accordingly. Additionally, sentiment analysis helps identify emerging trends and shifts in sentiment over time. By continuously monitoring sentiment, CPAs can detect changes in public perception, gauge the effectiveness of their risk communication efforts, and address any misinformation or misconceptions promptly. This promotes two-way communication by enabling CPAs to respond to citizen sentiments, address their needs, and establish a feedback loop for improved risk communication.
- Sentiment analysis is equally valuable for CPAs in understanding the sentiments and perceptions of citizens. By analyzing the sentiments expressed by citizens, CPAs can gauge the effectiveness of their risk communication strategies and identify areas where improvements are needed. Positive sentiments indicate successful communication, while negative sentiments highlight areas of concern or gaps in understanding. Sentiment analysis enables CPAs to identify specific negative sentiments expressed by citizens, providing insights into the underlying issues and concerns. This information helps CPAs adjust their communication approaches, address specific concerns, and provide targeted information to alleviate fears or misconceptions. Moreover, sentiment analysis allows CPAs to monitor the sentiment surrounding their own organization and actions. By analyzing public sentiment towards the CPA, they can gauge the level of trust, credibility, and public perception of their risk management efforts. This knowledge enables CPAs to

enhance their transparency, credibility, and responsiveness, fostering a more effective and trusted relationship with the public.

Overall, sentiment analysis facilitates two-way communication between citizens and CPAs by providing a comprehensive understanding of public perceptions, emotions, and concerns related to risk.

In conclusion, Publicsonar's sentiment technology effectively tackles the primary goal of bridging the Risk Perception Action Gap through multiple avenues:

- Comprehensive sentiment measurement: Publicsonar's technology enables the measurement of prevailing sentiment not only after an incident or relevant topic, but also before and during its occurrence. This comprehensive approach provides valuable insights into the evolving sentiment landscape, allowing organizations to track shifts in public perception over time.
- In-depth analysis of negative sentiments: The platform goes beyond merely capturing sentiment by offering a deep dive into different negative sentiments. This feature enhances the understanding of the underlying issues and challenges, facilitating a more nuanced comprehension of the overall problem at hand. By identifying specific negative sentiments, organizations can tailor their communication strategies and address concerns effectively.
- Flexible user interface for enhanced information sharing: The user-friendly interface allows for the inclusion of sentiment items in reports and dashboards. This flexibility simplifies the process of sharing information across teams and stakeholders, ensuring that sentiment analysis findings are easily accessible and can be integrated into decision-making processes.
- Scalability and adaptability: the solution has been designed with scalability in mind. It offers the capability to incorporate new languages and sources with minimal effort and resources. This adaptability ensures that organizations can expand their monitoring and analysis capabilities as needed, keeping pace with the ever-evolving digital landscape and emerging communication channels.

By encompassing these aspects, Publicsonar's technology empowers CPAs to gain a comprehensive understanding of sentiment dynamics, delve deeper into negative sentiments, facilitate seamless information sharing, and adapt to changing linguistic and data source requirements.

By leveraging sentiment analysis, CPAs can tailor their risk communication strategies, address specific citizen concerns, foster trust, and establish an ongoing dialogue that improves risk perception, action, and overall risk management thus helping to close the Risk Perception Action Gap.

APPENDIX 1: Triton Server deployment

For the deployment of the new multisentiment model we introduced a new server architecture to host the new and existing models. The introduction of the new sentiment model meant we had to find something more performant and scalable to host all of our models.

Triton is a server provided from NVIDIA to run DL models in a very efficient and fast manner. It optimizes ops in the architecture of your running GPU instance. What this means is that the optimizations are tailored to the GPU you select.

We targeted the optimization at 2x-5x improvement versus the previous PyTorch HuggingFace architecture. The first approach was following github.com/ELS-RD/transformer-deploy/ approach. However, the results were rather disappointing. The improvement in speed was not as much as we hoped for and the main problem with the final optimized version was the damage in output compared to the full model:

GPU ELS						
Batch 10						
	Mean	**Std**	**Beams**	**Ratio**	**Mem.**	**Speed**
PyTorch	5.68	1.71			10397.00	
FP16 test-dec-if	2.17	0.76	1.00	**59.40**	7257.00	**2.61**
FP16 test-dec-if	3.00	0.95	2.00	**89.37**		**1.89**
FP16 test-dec-if	3.95	0.97	3.00	**89.54**		**1.44**
FP16 test-dec-if	3.41	0.86	2 Sample	**90.07**		**1.66**
FP16 test-dec-if	2.95	0.92	topk 0	**90.26**		**1.92**
FP16 test-dec-if	3.43	0.90	topk 1	**90.46**		**1.65**
FP16 test-dec-if	3.06	0.91	topk 50	**90.02**		**1.86**

We can see that the maximum speed improvement was 2.61x. Yet the quality of the output was very poor, only 59.4% of the original output. This ratio means how much the final output resembles the original model output. When we optimize models, we

usually move weights to fp16, which decreases the numeric precision. The sweet spot happens when we find a reduction in weights that does not damage our output.

Next shot was to follow the NVIDIA solution to speed up models. It requires, as most optimization solutions, a delicate balance between CUDA versions, drivers, and the optimization tool's versioning. The results are:

						Max mem cache PyTorch	Max mem Triton
GPU Triton						5561MiB	2681.00
Batch 5							
	Mean	**Std**	**Beams**	**Ratio**	**Mem.**	**Speed**	**Mssg/sec**
PyTorch	9.65	3.03			3707.00		
Triton FP 16	0.67	0.14		0.97	2681.00	14.50	**15.02**
Batch 10							
	Mean	Std	Beams	Ratio	Mem.	Speed	Mssg/sec
PyTorch	10.36	3.46			3707.00		
Triton FP 16	0.68	0.14		0.97	2681.00	15.24	**14.71**
Batch 20							
	Mean	Std	Beams	Ratio	Mem.	Speed	Mssg/sec
PyTorch	10.85	3.01			3707.00		
Triton FP 16	0.85	0.07		0.97	2681.00	12.83	**23.64**

With a completion of 97%, most output is equal to the original ones. Those which don't match, are minor differences, sometimes even improving the original output.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021271

The surprising news is that with a batch of 10 we are already at 14.71x speed improvement, being 12.83x when using a batch of 20 messages. This is considerably above expectations. Our maximum tested speed-up was a 20x in a multi-node (2 GPUs).

Note Max mem for Triton is the same as loaded model. This happens because the input is passed as numpy (no cuda device tensors) so no overhead.

The RiskPACC Consortium

FIGURE 38: THE RISKPACC CONSORTIUM