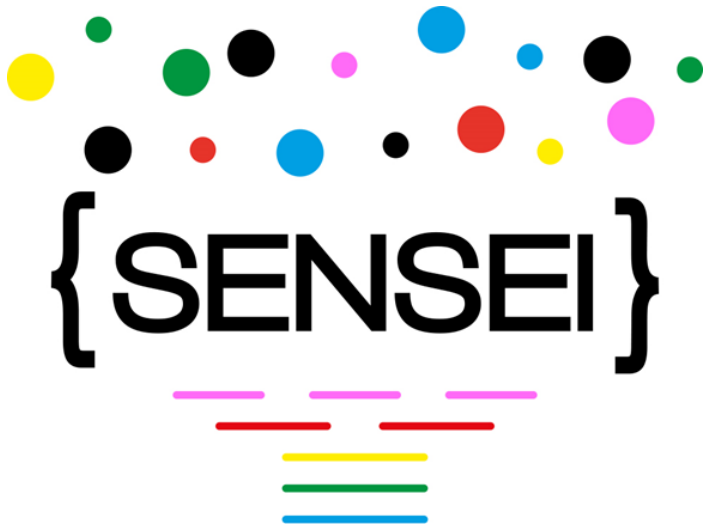


D1.3 – Report on Intermediate Evaluation

Document Number	D1.3
Document Title	Report on Intermediate Evaluation
Version	2.0
Status	Draft
Work Package	WP1
Deliverable Type	Report
Contractual Date of Delivery	10.31.2015
Actual Date of Delivery	10.30.2015
Responsible Unit	UNITN
Keyword List	Evaluation, Task Based Evaluation, Speech Use Case, Social Media Use Case, User Requirements
Dissemination level	PU

Editors

Morena Danieli (University of Trento, UNITN)

Emma Barker (University of Sheffield, USFD)

Contributors

Giuseppe Riccardi (University of Trento, UNITN)

Carmelo Ferrante (University of Trento, UNITN)

Benoit Favre (University of Marseille, AMU)

Balamurali A R (University of Marseille, AMU)

Rob Gaizauskas (University of Sheffield, USFD)

Emina Kurtic (University of Sheffield, USFD)

Monica Lestari Paramita (University of Sheffield, USFD)

Letizia Molinari (Teleperformance Italy, TP)

Adele Palumbo (Teleperformance Italy, TP)

Cosima Caramia (Teleperformance Italy, TP)

Vincenzo Lanzolla (Teleperformance Italy, TP)

SENSEI Coordinator

Prof. Giuseppe Riccardi

Department of Information Engineering and Computer Science

University of Trento, Italy

giuseppe.riccardi@unitn.it

Document change record

Version	Date	Status	Author (Unit)	Description
0.1	07/07/2015	Draft	Morena Danieli, Giuseppe Riccardi (UNITN)	Table of Content
0.2	07/22/2015	Draft	Morena Danieli (UNITN)	Added names of contributors and content description to sections
0.3	08/01/2015	Draft	Morena Danieli (UNITN)	Modifications to ToC on the basis of partners' contribution to v0.2
0.4	08/05/2015	Draft	Morena Danieli (UNITN)	Authorship of social media evaluation sections
0.5	08/12/2015	Draft	Morena Danieli (UNITN)	Edited Introduction, 1.1 and 1.1.2.1
0.6	08/31/2015	Draft	Benoit Favre (AMU)	Add ACOF and synopsis reliability
0.7	09/01/2015	Draft	Morena Danieli (UNITN)	Edited Section 1 and 2; completed section 1.1.2.2
0.8	09/01/2015	Draft	Emma Barker, Emina Kurtic, Monica Lestari Paramita, Rob Gaizauskas (USFD)	Added to sections:1.1.2.3; 2.2.Added Appendices 1-6.
0.8.1	09/15/2015	Draft	Emma Barker, Monica Lestari Paramita, Rob Gaizauskas (USFD)	Added to sections 1.2.2 and 2.2.
0.9	09/15/2015	Draft	Morena Danieli (UNITN)	Added content to most of the speech sections, restructured the table of content, writing of the Introduction
0.10	09/16/2015	Draft	Morena Danieli, Carmelo Ferrante (UNITN)	Integration of versions 0.8.1 and 0.9; due to the new structure of the index, old section 1.2.2 is now 2.2, and old section 2.2 is

				now 3.3.
0.11	09/29/2015	Draft	Morena Danieli (UNITN), Letizia Molinari, Cosima Caramia, Adele Palumbo, Vincenzo Lanzolla (Teleperformance)	Added to Section 1, Introduction; added content and results to the speech evaluation sections
012	09/30/2015	Draft	Morena Danieli (UNITN)	Executive summary added
0.13	10/01/2015	Draft	Morena Danieli (UNITN)	Added the acronym table, revised and completed Section 1, and section 2 (up to 2.3 excluded, also the missing references)
0.14	10/01/2015	Draft	Morena Danieli (UNITN)	Completed Sections 2.3; (speech parts with complete description of the evaluation tasks); updated the list of contributors; added references
1.0	10/03/2015	Draft	Emma Barker, Rob Gaizauskas (USFD), Morena Danieli (UNITN)	Final version of Sections 2.2.3, 3.1, 3.2, 3.3; 3.3.5. drafted Section 4. Further revision of Executive Summary, Introduction, Section 1, References
1.1	10/12/2015	Draft	Elisa Chiarani (UNITN)	Quality check completed. Some remarks and comments
1.2	10/18/2015	Draft	Morena Danieli (UNITN)	Reviewed after quality and scientific checks
1.3	10/24/2015	Draft	Giuseppe Riccardi (UNITN)	Final review and approval
2.0	10/28/2015	Final	Morena Danieli	Final version ready for submission

Executive summary

The SENSEI Deliverable D1.3 reports on the evaluation experiments done with the SENSEI prototype during Period 2 of the project. The evaluation experiments we illustrate and discuss here are based on the principle of evaluating the SENSEI prototype with users in realistic task settings. The goal is gathering feedback on the present state of development of the prototype and providing insights for improving it. The intermediate SENSEI evaluation effort has the further aim of setting up the different components of the evaluation protocol, including the new methodologies designed for the social media scenarios, and the joint use of qualitative and quantitative methods in the assessment of the speech tasks. The assessment of the evaluation protocol is important due either to the intrinsic difficulty of setting baselines for automatically generated summaries, and to the need of fixing in advance possible issues of the evaluation protocol in view of the larger scale trials planned for the third year of the project.

D1.3 deliverable focuses on extrinsic evaluation, i.e. on evaluation tasks based on activities typically carried out by the SENSEI potential users. In this document we report the details of the task-oriented experiments, and their results. We recruited users with real experience in the tasks selected for the evaluation. For the speech use cases the users were professionals of a call centre company, for the social media use case they were graduates with experience of using online news and reader comments.

The prototype we have been evaluating generates different types of summaries, including short summaries of call centre calls (synopses), summaries of reader comments, (including text summaries, linked to clusters of comments and graphical summaries, where clusters of comments are represented in a pie-chart), and filled questionnaires used to summarise some aspects of call centre agents' communication behaviour when they interact with their customers.

For the speech scenario we implemented three experiments, including the evaluation of the reliability and the evaluation of the accuracy of the automatically generated questionnaires, and the collection of insights and feedback based on users' experience in using the call summaries (in form of synopses of the calls) within a task of call centre agent supervising. The evaluation task designed and implemented for the social media scenario was based on assessing the quality of user outputs and gathering the experience of the users after having performed a set of tasks, with and without the contribution of the SENSEI prototype. In addition, we also report in this deliverable the results of the assessment of the reliability of the annotated speech corpus.

In general, the results of the evaluation highlighted that the evaluation protocol and the tasks are realistic, potentially accepted by the users, and feasible on a larger scale. Summaries at very low compression rate (7% of the original conversation) have been judged useful by the users for potentially reducing decision-making time in their job. Some possible improvements of the experimental settings have been identified, they are mainly about the prototype interface and they will be presented to the technical SENSEI workpackages. As for the prototype system, from the participants' comments we could appreciate both their preferences for some types of summaries among the ones proposed, both in speech and

social media scenarios, and useful insights for improving the underlying technologies, whose critical aspect seems to be related with the accuracy requirement.

Table of Contents

List of acronyms and names

Acronym	Meaning
ACOF	**A**gent **C**onversation **O**bservation **F**orm
RATP-DECODA or DECODA	Call centre human-human spoken conversation corpus
ICC	**I**ntraclass **C**orrelation **C**oefficient
LUNA	Spoken **L**anguage **UN**derstanding in Multilingu**A**l Communication Systems (human-human spoken conversation corpus)
SUMMAC	Text Summarization Evaluation Conference
QA	**Q**uality **A**ssurance

1. Introduction

Evaluating the quality of automatic generated summaries is difficult, principally because humans also tend to agree only approximatively when they are asked to judge about the quality of a summary with respects to the original text [Radev, Howy & McKeown, 2002]. In the literature it has been showed that using a task as motivation could help for approaching the inherent subjectivity of this type of evaluation. For example, the Summarisation Evaluation Conference (SUMMAC) included three task-oriented assignments for single news articles summarisation: the categorization task, the *ad hoc* task, and the question task. For the kind of summaries generated by the SENSEI prototype the adoption of task-oriented evaluation and human perspective is crucial.

The advances in summarisation technologies developed in SENSEI enable the creation of extractive summaries that provide a more efficient way of navigating huge archives of social media and call centre conversations. The implementation of the SENSEI prototype takes into consideration user preferences and related requirements. For instance, at present the speech use case attains “user defined” summaries of call centre conversation, based on the requirements abstracted from the analysis of the Agent Conversation Observation Forms (ACOFs henceforth) used by call centre quality assurance supervisors. We believe that qualitative task-oriented evaluation may help in verifying the appropriateness of the method, and the usefulness and quality of the summaries, i.e. to answer to the question: how well do summaries help a user carry out a task? So the evaluation of the summaries generated by the system needs to be based both on intrinsic evaluation metrics, the ones commonly used for assessing the results of the machine learning techniques, and on extensive extrinsic evaluation of qualitative user experience and/or outputs created by a user with a SENSEI system, in a task-oriented context.

In this document we present the application of the SENSEI evaluation model introduced in the SENSEI deliverable D1.2. The details here presented include the set of evaluation tasks and experimental settings for speech and social media, the results of such experiments, and the discussion of their implications for future activities. The evaluation framework is organized into three different levels. The first level relates to the different technology-oriented, intrinsic evaluation measures. This intrinsic evaluation is performed within the three technological work packages of the project, and the results are discussed in D3.2, D4.2, and D5.2.

The goal of the second level, task-based extrinsic evaluation, is to assess if, and at what extent, the different types of summaries generated by the SENSEI prototype are usable and effective for potential users.

The third level of evaluation, that we name “insight-oriented evaluation”, is focused on users’ perception of task success. In this document we will focus on the extrinsic evaluation tasks that have been used for this intermediate evaluation campaign of the SENSEI prototype.

In WP1 we have evaluated the reliability of the annotated speech corpora used for training the machine learning algorithms1. Here we describe the principles, the goals, and the results of inter-annotator agreement, and test-retest assessment performed on the annotated speech corpora. In the past year we also created a novel, Gold Standard Corpus of manually written summaries and related annotations (e.g. clusters) of social media comment sets.[2]

In the following parts of this Introduction we describe the follow up in WP1 to project Period 1 activities, and we explore in further details the motivations for summary evaluation. The rest of the document is organized as follows: Section 2 reports the results of the assessment of corpora annotation reliability, and the evaluation tasks for speech and social media. Section 3 describes metrics and results of the evaluation, and Section 4 reports result discussion and suggestions for further work.

1.1. Follow up to Period 1 activities

During Period 1 of the project, SENSEI WP1 activities were focused on

1. the identification of the users of SENSEI technologies,
2. the definition of an initial set of speech and social media use cases,
3. the setup of the evaluation framework to be applied during the second year of the project,
4. the definition of the annotation guidelines to be used for the speech corpora,
5. the design of user requirements: this activity had impact on the three technological work packages (WP3, WP4, and WP5), and on WP6.

During the second year of the project the main goals for WP1 have been related with the improvement of the evaluation scenarios, with the choice of appropriated metrics, and with the evaluation of the SENSEI prototype. We focused the set of evaluation tasks described in Section 2. The selection was based on the large set of tasks designed during Period 1. The selected tasks were implemented, in terms of both protocols and system interface, for the evaluation campaign reported in this document.

For the extrinsic evaluation of the speech scenario we chose to evaluate the tasks based on the generation of two types of summaries of the call centre conversations, i.e. the generation of very short summaries of the calls, and the automatic filling of questionnaires focused on summarising the agents' communicative behaviour as showed in the conversations. From now on we will refer to the first type of speech summary with the term "synopsis", and to the second type with the acronym ACOF, that stands for Agent Conversation Observation Form. Both synopses and ACOFs are described in details in D1.2, but for ease of reading we report here their definitions.

[1] The details of the speech corpora annotation are described in Section 2.1 of the SENSEI deliverable D2.3.

[2] This work is reported in detail in D 5.2, together with results from the initial intrinsic evaluation of SENSEI social media clustering and summarization technologies, which compared system outputs against manual summaries and cluster annotations.

The ACOFs are questionnaires modelled on the ones actually used in call centres for assessing how the call centre agents communicate with their customers during the phone conversations. By scoring the items of the forms, the QA supervisors decide if in the observed spoken interaction the agent was able to meet the quality assurance requirements for whom s/he received training. During the first year of SENSEI project we defined specific ACOFs for applying them on the LUNA and RATP-DECODA conversation corpora. During the second year of the project, such ACOFs have been further refined. This activity is described in D2.3, Section 2.1.

The synopses are short summaries of the content of call centre conversations. While ACOFs are focused on the agents' behaviour, the synopses are focused on the semantics of the call, i.e. the reasons why the customers called the contact centre, and the way the agents (hopefully) solved their problems. In SENSEI the synopses of the calls are very short, around 7% of the original length of the transcript of the conversation. While manually filled ACOFs are in place in call centre activities, the availability of synopses for performing agents' supervision is brand new.

In the Social media extrinsic evaluation we investigated how and to what extent SENSEI technologies can assist a user who wants to gain an overview of comments on a news article, in a short period of time (the use case 1 for Social Media, as described in D 1.1 and D1.2, is based on this scenario). The setup involved users carrying out a number of short tasks based on readings of news articles and comment from *The Guardian* newspaper using i) current practice news and comment technology, and ii) current practice technology *and* SENSEI generated summaries of comment, presented via the SENSEI prototype interface. The tasks (which are examples of reading comprehension tasks tailored to the reader comment genre) involved participants answering questions relating to comment content, i.e.: to identify "four main issues" in the discussion and to "characterise opinion" on a given issue. A good summary of comment should help readers to carry out these tasks. We developed novel metrics to quantitatively assess these tasks and we gathered ratings from participants on the usefulness of different systems and system components in the context of completing these tasks. Further feedback was obtained in a post-task group discussion in which we invited participants to comment on their experience during the tasks and using the different systems. So in sum for the Social media evaluation, our approach provided three complementary sets of results, which allowed us to compare how, and to what extent, the different systems helped users in the different task contexts.

As we mentioned above, further activities in WP1 during Period 2 of SENSEI project were about the assessment of the reliability of the annotation of corpora implemented in WP2. This type of assessment was particularly relevant for the speech scenario, because – as we will discuss below, the ACOF annotation by humans is a highly subjective task.

1.2. Motivations for evaluation

SENSEI summary definition is reported in Section 2 of the deliverable D5.2. In this Section we focus our attention on the motivations underlying the evaluation tasks for such different types of summaries.

The summaries generated in SENSEI are reductive transformations of the source data (spoken and written). The data content may be represented into one out of a set of stereotypical reductive transformations. For example, for the speech use case we identified stereotypical reductive transformations that are applicable in contact centre tasks, including the generation of short synopses of the calls (focused on call content), the generation of ACOFs, and the opportunity of navigating the transformed conversations starting from queries specified by the users.

In designing the evaluation tasks for such types of summaries, we could realize that the nature of the SENSEI research places important evaluation requirements to be met. Behind the traditional assessment of the component technologies of the prototype, we had to verify if the summaries produced by the prototype may capture *real behavioural patterns* of individuals as they are shown in both call centre and social media interactions. In addition, we needed to understand if, and in which measure, those automatically generated summaries may be *accepted and used* by potential users. In other words, we needed to perform an intrinsic (technological and quantitative) evaluation, an extrinsic (qualitative and behavioural) evaluation, and an insight-oriented (qualitative) evaluation.

Scientific research in behavioural analytics aims to identify patterns of human behaviour that may help in understanding the intentions, the needs, and the personal traits of individuals. The identified patterns need to be verified for assessing their explanatory potential. In classical behavioural research the needs for verifying the identified traits have been traditionally approached by *grounding* those traits through the submission of different types of surveys and observational protocols. The extrinsic evaluation of the SENSEI summaries relies on qualitative evaluation methods that have been applied both in the speech and in the social media scenarios. For social media the issue of extrinsic evaluation of behavioural patterns is even more challenging than for the speech scenario, because of the lack of well-established protocols for evaluation, while for the dyadic conversations of the speech scenario we may rely on the evaluation protocols of conversational and behavioural research [Zafarani & Liu 2015]. For social media the difficulty is partly due to the huge number of much diversified users of social media, and it is partly due to the novelty of the SENSEI outcomes. We have faced these new research needs by developing an evaluation framework where methods from statistics and behavioural sciences are jointly used.

Another important motivation for the evaluation activities in WP1 is related with the design and development of the SENSEI prototype. Not only the evaluation provides obvious feedback to the technical workpackages about the present performance of the SENSEI technologies, but the adoption of a 'system-in-the-loop' approach, where potential communities of real users have been involved since the beginning, contributes to the refinement of the initial set of user requirements, and it provides more focused and realistic tasks to be developed.

2. Evaluation tasks and scenarios

In WP1 we have involved real users for identifying the requirements of the SENSEI summarisation prototype, and for assessing the quality of the project results. In Section 2 we describe this approach and the related tasks we implemented for the speech and social media scenarios.

2.1. Preliminary evaluation of the speech corpora

2.1.1. The need for task based evaluation in speech use cases

For the speech scenario we identified as potential users the Quality Assurance (QA) supervisors of call centre. In contact centres the QA supervisors measure and evaluate agent adherence to the internal protocols during conversations with customers. Those protocols cover communicative and behavioural features of the agents. QA is critical for companies, and today most of the QA methods require manual evaluation of randomly selected conversations. In a recent number of "The Real-Time Contact Centre Newsletter", Fluss reported that traditional observational methods often target "only 2 to 10 calls (or interactions) per agent per month". She adds that "the traditional QA process is statistically invalid [..]. If agents handle an average of 50 interactions per day, 2 to 10 interactions per month equates to only 0.2% - 1.0% of their monthly interaction volume. Increasing the number of QA analysts or supervisors barely moves the needle to 2% or maybe 3%" [Fluss 2015]. In other terms, while QA is crucial for contact centre and for customers, nowadays it runs the risk of being very labour-intensive and sometimes ineffective.

As we explained in D1.2, in real call centres the live conversations are assessed by QA supervisors and are scored against established contact handling criteria, summarised into a QA questionnaire. In state-of-the-art working conditions the conversations are scored manually. One of the goals of SENSEI is to review and score automatically all the calls, and to summarise the features of agents' behaviour in each call by an automatically generated QA form. For pursuing this goal we have developed two variants of ACOF suited for the conversations of LUNA corpus and of RATP-DECODA corpus respectively. In addition, we have been targeting the goal of producing automatically generated short summaries (synopses) of each call.

Differently from some pioneering current applications, the SENSEI prototype can understand several different aspects of spoken conversations by leveraging different behavioural analytics methods and technologies, as reported in Sections 3.1 and 3.2 of D5.2.

2.1.2. Evaluation of the reliability of the annotated speech corpora

For training the speech analytics algorithms we needed to provide gold standard, human annotated sample calls including DECODA and LUNA conversations. For each conversation in those corpora, the human annotators provided both human generated ACOFs and synopses. Each conversation of the corpus was annotated by two to five different native speakers (or nearly native speakers) QA professionals. The annotators performed their tasks

by following the guidelines described in D2.3, Section 2, where the reader may also find further details about the annotation activity. In this paragraph we report the evaluation of the accuracy and reliability of those annotations.

The estimation of the accuracy and reliability of the corpus annotation has to goals:

1. to assess the reliability of the ACOF, i.e. if the different questions that occur in the ACOF may be understandable and unambiguous for the users;
2. to assess the reliability of the annotated corpora in terms of the reproducibility of the annotation protocol.

The first goal contributes to provide an answer to the general question "*How good is the type of summary produced by SENSEI speech prototype*?" i.e. is it able to capture behavioural patterns that correspond to the real behaviour observed in agents engaged in call centre interactions?

The second goal contributes to the **grounding** of the annotation protocol, i.e. to understand if it may be applied by different observers with the same results.

We have approached the first goal by applying a test-retest paradigm, while the second goal was approached by calculating the inter-annotator agreement. For the latter we selected a subset of the ACOF questions. They were three questions from the questionnaires used by the annotators to evaluate the LUNA conversations, and three from the questionnaire used to evaluate the DECODA conversations. The questions were selected by the QA professionals of the partner TP who deemed them as the most explicative for evaluating the communicative ability of the call centre agents (ACOF questions 3 and 9 for DECODA, and ACOF questions 3 and 10 for LUNA), and the listening competence of the agents (ACOF questions 2 for LUNA, and 10 for DECODA).

2.1.3. Test – retest experiment

The *test-retest* protocol [Weir 2005] is commonly used in experimental psychology as a simple method for evaluating the stability and reliability of a psychological construct over time. The protocol requires that the same test is given to the same subjects in two separate sessions (T1 and T2). The scores on the two occasions are then correlated. This correlation is known as the *test-retest-reliability coefficient*, or the *coefficient of stability*. The closer each respondent's scores are on T1 and T2, the more reliable the test measure is. A coefficient of stability of 1 says that each subject's scores are perfectly correlated. That is, each subject scored the exact same thing on T1 as they did on T2. A coefficient correlation of 0 indicates that the scores at T1 were completely unrelated to the scores at T2; therefore the test is not reliable.

For SENSEI we designed the following test-retest protocol. We recruited two participants who contributed to the annotation of ACOFs of LUNA and DECODA conversations. Each of them received 60 conversations: half of those conversations were extracted from the ones they annotated from LUNA, half from the ones they annotated from DECODA. 34 of the selected conversation had been annotated less than 41 days before the retest, 26 had been annotated more than 41 days and less than 90 days before the retest. The participants were female, Italian native speakers, their knowledge of French was rated C2.

The participants worked independently, and without having access to their previous ratings. They received instructions for re-annotating each item of the ACOFs over the selected data. If necessary, they could take notes during the task. They were also recommended not to worry about the fact that they did not remember the conversations they already annotated. We also asked them *not* to try to remember their previous ratings, as far as that was possible.

We calculated the test-retest correlation by using the *ICC (Intraclass Correlation Coefficient),* whose formula is $ICC = (F - 1)/(F + k - 1)$, where F is the F ratio, and k is the number of tests. We do not have missing observations in this experiment.

Table 1: Test-retest results

T1-T2 (*n* dialogues)	F for subjects (confidence level %)	ICC and Confidence Levels		
		ICC	Lower	Upper
<40 days (34)	29.2 (90)	0.93	0.84	0.87
41-90 days (26)	33.8 (90)	0.80	0.66	0.82

In Table 1 above we report the results of test-retest. In total, 60 conversations have been re-annotated by each participant. The scoring was calculated by counting 1 each time the participant attributed at T2 the same value (Pass, Fail, or NotApplicable) s/he attributed at T1 for each item of the ACOF, 0 in case of difference. The test–retest experiment showed that even when T1 and T2 are in the interval 41-90, reliability is still good. This result supports the hypothesis that the ACOF is a stable evaluation tool over time.

2.1.4. Assessment of reliability of the speech annotations

We calculated the *inter-annotation agreement* for the selected *ACOF* questions that were used for extrinsic evaluation (see the introductory notes to Section 2.2 above). Fleiss Kappa [Fleiss 1981] is used as metric for inter-annotation agreement.

Fleiss's *Kappa* measures reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items. This contrasts with other *kappas* such as Cohen's *kappa* [Fleiss & Cohen, 1973] which only works when assessing the agreement between two raters. The measure calculates the degree of agreement in classification over that which would be expected by chance.[3]

A group of three subjects performed the ACOF annotation for DECODA whereas for LUNA it was performed by five subjects. They all had experience in monitoring of call centre agents, and all were Italian native speakers. The two subjects who worked on the French conversations had C2 and B2 levels of listening comprehension in French. The participants

[3] Wikipedia, url:https://en.wikipedia.org/wiki/Fleiss%27_kappa, accessed: August 31, 2015

worked independently on the same data. Before calculating the agreement, noisy annotations like additional entry or incomplete entry were removed. DECODA and LUNA corpora contained 2 and 5 such entries respectively. In Table 2 below Fleiss Kappa agreement for ACOF questions on DECODA/French corpus is reported:

Table 2: DECODA inter-annotation agreement results

Question	ACOF	Pass	Fail	N/A	Kappa	Agreement
3	Agent shows the information in a clear, comprehensive and essential way	334	6	8	0.47	Moderate
9	Agent is able to adapt to the style of client's communication always maintaining professionalism	330	14	4	0.53	Moderate
10	Agent Management: he negotiates the wait always giving reasons	206	3	139	0.99	Almost perfect

Fleiss Kappa agreement for ACOF questions on LUNA/Italian corpus is reported in Table 3:

Table 3: LUNA inter-annotation agreement

Question	ACOF	Pass	Fail	N/A	Kappa	Agreement
2	Agent listens actively and asks relevant questions	290	7	3	0.16	Slight
3	Agent shows the information in a clear, comprehensive and essential way	288	8	4	0.13	Slight
10	Agent is able to adapt to the style of client's communication always maintaining professionalism	294	2	4	0.38	Fair

Overall it is seen that agreement is *moderate* among the annotators.

Computing the *reliability of annotations on full-text productions*, such as summaries, is not as simple as on categorical tasks such as ACOF due to the natural variability in writing text independently. We chose to use the *ROUGE evaluation metric* for computing such reliability.

The idea is to consider each human annotator as a system and evaluate it against the reference produced by other human annotators. For a given conversation, given *n* human reference summaries, we, in turn remove one annotator from the pool and evaluate its production against the *n-1* other references. This approach has been followed in the DUC/TAC evaluation for computing human topline performance, which was very difficult to

reach for systems [Lloret & Palomar 2012]. This type of reliability measurement has the limitations of the underline measure, ROUGE, which is computed as the ratio between the number of word *n-grams* overlapping between the reference and the hypothesis, and the number of *n-grams* in the reference.

A ratio of 1 means that the reference and the hypothesis have the same words, but that is unlikely unless all human annotators wrote the same text. By looking at words, ROUGE is somewhat limited in modeling semantic equivalence such as synonymy and paraphrasing. However, we can use it at our advantage since it is considered reliable for human annotators to make the same lexical choices. We computed ROUGE-2 recall (word bigram) for each annotator on the subset of the SENSEI-created synopsis for the *Multiling'15 Call-Centre Conversation Summarisation* task.

Compared to participating systems, ROUGE score are much higher (about 0.15 *vs* 0.035) which means that humans are indeed consistent but variance is also greater (0.02 *vs* 0.005) meaning that most of the times annotators agree but when they disagree, all annotators tend to have different lexical usages. This last trend is not unexpected for summarisation where annotators have to make choices towards what is important in a conversation and shall be represented in the limited space of a synopsis, and may make different choices on peripheral aspects. This high variance also comes from the fact that there are a limited number of synopses for each conversation (3-5) and when removing one synopsis from that set, the remaining synopses form an even cruder sample from the true distribution of what 1,000 synopses would look like. It should also be noted that some annotators have much lower scores than others, probably indicating their divergence from average synopses and the need to give them better guidelines in future collections.

Table 4 below reports ROUGE-2 scores for LUNA/Italian synopses:

Table 4: ROUGE-2 scores for Italian synopses

Annotator	ROUGE-2	Sigma
IT1	0.121	0.023
IT2	0.213	0.023
IT3	0.175	0.022
IT4	0.073	0.014
IT5	0.125	0.018

and the following Table 5 reports ROUGE-2 scores for DECODA/French synopses:

Table 5: ROUGE-2 scorse for French synopses

Annotator	ROUGE-2	Sigma

FR1	0.194	0.029
FR2	0.207	0.036
FR3	0.077	0.048
FR4	0.057	0.039
FR5	0.113	0.054

2.2. Speech and social media extrinsic evaluation scenarios

This Section describes the experimental settings of the extrinsic evaluation scenarios for speech and social media. Evaluation results are reported and discussed in Section 3.

2.2.1. Speech extrinsic evaluation scenario 1 – ACOF

The SENSEI prototype classifies the call centre conversations on the basis of the different aspects of agent behaviour as they are represented by the values assigned to the different ACOF questions. We wanted to evaluate its positive predictive value and sensitivity. For reaching this goal we designed an evaluation task where the ratings assigned by the SENSEI prototype are compared with the ones assigned by human evaluators.

The experimental setting was the following. Two participants for RATP-DECODA (French corpus) and three for LUNA (Italian corpus) individually listened to the same set of conversations that were previously analysed by the system. They were asked to assign their ratings (Pass, Fail or Not Applicable) to the three ACOF questions. Then, they had access to the ACOFs that the system generated for each conversation. For each question they had to answer to a single-item questionnaire by choosing a value arranged on a five values Likert scale (0-4). The single question was “How much accurate was the system in filling this ACOF question?”. Assigning 0 means that the human rater completely disagreed with the system evaluation (the system was not accurate at all) and 4 that s/he completely agreed with the evaluation given by the system (the system was perfectly accurate).

The participants to this evaluation task were expert QA supervisors. They were selected among the Teleperformance Italia QA professionals who did not participated in the corpus annotation. The participants were provided with the annotation guidelines previously developed for the corpus annotators. They received training for this evaluation task by two evaluation supervisors. The goal of the training was to familiarize with the annotation guidelines and to understand the evaluation task.

The participants were instructed to focus their attention on the selected subset of ACOF questions. The questions were selected by three expert QA supervisors on the basis of their relevance for assessing the ability of the agent to manage efficiently the time length

constraints, and for evaluating the agent communication attitude. As for the inter-annotation agreement, the selected questions were the following: question 10 (LUNA and DECODA ACOF) for assessing agents' overall ability in call management, while questions 3 (DECODA) and 2 (LUNA), questions 9 (DECODA) and 3 (LUNA) all addresses the agents' communicative style and behaviour.

From the point of view of the participants, the task was a kind of human observation. In designing this task we needed to take into account that the ACOF filling task is inherently biased when it is done by a human annotator. We have approached this by applying the tool of consensus meeting over different evaluation results, and by normalizing the scores obtained by the different evaluator answers to the questionnaire submitted to them.

In Table 6 we report the main features of the experimental setting.

Table 6: Features of speech extrinsic evaluation task - 1

	RATP-DECODA Corpus	LUNA Corpus
Nr of participants and gender	2 (female)	3 (2 female, 1 male)
Native speaker	Italian (2)	Italian (3)
Oral Comprehension Level in French	B2 and C2	---
Nr of calls listened to	30 (randomly selected)	30 (randomly selected)
Questionnaire	How much accurate was the system in filling this ACOF question?	How much accurate was the system in filling this ACOF question?
Further evaluation tool for gaining feedback	Consensus meeting moderated by the evaluation task supervisors	Consensus meeting moderated by the evaluation task supervisors

The results and discussion for this evaluation task are illustrated in Section 3.1.

2.2.2. Speech extrinsic evaluation scenario 2 – Synopses

For evaluating SENSEI prototype with respects to synopses generation, we decided to set up an extrinsic evaluation task aiming at identifying if, and at what extent, the availability of automatically generated summaries may help the QA supervisors in focusing on problematic calls.

Focusing on problematic calls is important because it may potentially reduce the time to completion of tasks related with the supervision of call centre agents. At present for

evaluating the agents a great number of calls need to be listened and assessed in order to identify the potentially problematic ones as soon as they occur in the call centre.

The design of this task is based on a focus group methodology, whose goals are

- the discovery of shared views among the participants,
- the implications behind those views for the SENSEI speech prototype.

We devoted great care to the composition of the focus group. The nature of our task required that the group participants should be representative of the potential population of users of SENSEI speech prototype. In D1.2 we identified quality assurance and human resources professionals as potential users, and participants with relevant professional background were recruited for the focus group. With the exception of one participant, all the others had previous experience with the SENSEI prototype. In particular, two of them performed ACOF filling tasks over the conversations of the call centre corpora both with and without the pre-filled ACOF and system-generated synopses of the calls.

For the focus group the interview was not structured, but the set of topics to be discussed was carefully selected on the basis of the principle of moving from general to more specific issues, and focusing on the most important issues for SENSEI research goals. The group discussion was facilitated by a moderator, but no observer was present. Given the different skills of the involved participants, we needed to recruit subjects who had difficulties in meeting face-to-face in the same town on a given date and hour. So we decided to hold the focus group in remote setting, by conference call, and to record the complete discussion. After we edited the transcript of the conversations, those transcripts were used for data analysis. Table 7 reports the focus group plot that was sent one day in advance to the participants.

Table 7: Focus group plot

The focus group plot
Introduction
We would really appreciate your feedback on some aspects of your experience with the SENSEI speech prototype. We'll be asking you a few questions and would like you to respond as freely as possible. We note that this discussion will be recorded.
[Start recording]
1.ACOF task
Question 1: How did you find using the ACOFs filled by the system?
Sub-questions:
1.1Were the ACOF items useful for highlighting agent's behaviour?
1.2At what extent did you agree with the judges of the automatically filled ACOFs?

1.3Time: could you imagine that automatically filled ACOFs may help you in sparing time in your supervising job?

1.4Do you think that automatically generated ACOFs could be enriched with evidence of the system decisions?

2.Synopsis task

Question 2: We would like to hear more about your experience using the synopses of the call. This is something completely new in your daily activity. Do you have anything to say about this?

Question 3: Do you have a preference for the automatically generated ACOFs or for the summary of the call? Do you find them equally useful, maybe for different tasks? Could you please give us some examples?

3. Queries

Question 4: Could you tell us something about the strategies you used to formulate the queries based on the ACOFs?

4. General

Question 5: Did you fine SENSEI potentially useful or "added value" for your job?

Question 6: Can you imagine different tasks where you would find useful system-generated summaries, for example summaries over millions of conversations, more focused research, retrospective investigations on very large sets of conversations?

5.Any further comment

Question 7: Is there anything else you would like to tell us about your experience with the SENSEI speech prototype ?

[Stop recording]

Final: A big thank you for your collaboration!

As for the evaluation metrics for this task, despite of the widespread use of focus group qualitative method in the social and behavioural research, few explicit guidelines exist on how to analyze data collected with this methodology. Onwuegbuzie and his coauthors [Onwuegbuzie et al. 2009] provide a qualitative framework for collecting and analyzing such data. They term their method *micro-interlocutor analysis*, wherein meticulous information about which participant responds to each question, the order in which each participant responds, response characteristics, the nonverbal communication used, are collected, analyzed, and interpreted. They conceptualization takes profit from conversation analysis techniques, and for analysing the focus group data in SENSEI we had taken profit from this framework. Section 3 will describe the analysed information, their meaning and implications.

Ethics: The experimental settings described above have been designed in accordance with the University of Trento ethics policy. The focus group was moderated by a licensed

psychologist in accordance with the Meta-Code of Ethics of the European Federation of Psychologist Associations, and the Ethics Code of the Italian "Ordine degli Psicologi".[4]

2.2.3. Social media extrinsic evaluation scenario

This Section describes the scenario or "setup" for the Extrinsic Evaluation of the SENSEI social media prototype. We note that this is an interim evaluation and further system development is planned for the final period of the project.

Research Aims: This interim evaluation was guided by two high level research aims:

- Firstly, to assess how well the SENSEI prototype can help users carry out a real world user task. Insights and feedback gathered from the assessment can then inform future technology development.
- Secondly, to test run the evaluation methodology (which involves both novel tasks and metrics) and to consider the effectiveness of this approach as an extrinsic evaluation. Insights obtained will inform the design of a larger scale extrinsic evaluation of the SENSEI Social Media System, due to take place in Period 3 of the project.

We used the following setup for evaluation:

Experiment Design: The experiment design was based on the scenario[5] of a ***participant*** carrying out a user ***task*** -- to read and gain an overview of comment on a news article in a short time period -- using a system to help him carry out that task. The design allowed for a comparative assessment of two systems -- a baseline (S1) and a SENSEI condition (S2). Four participants each carry out two iterations of the task, each time using a different system. The design specified two different topics (each topic T comprising a news article and an associated set of comments), since the participants would acquire knowledge of a topic on the first iteration of the task. Each participant was to use each system exactly once and consider each topic exactly once. We note that the same type of task (which included 3 short sub-tasks)[6] was used for each system–topic iteration (a detail in one of the sub-task questions was topic-dependent). We wanted to control for the possible effects of bias due to the different order in which systems and topics were experienced. Thus the design allowed for the 4 different orderings of the system and topic: two participants were to use systems in the order S1-S2 and two in the order S2-S1 and 2 participants were to do the topics in the order T1-T2 and 2 in the order T2-T1. Thus each of the four possible orderings of 2 systems and 2 topics is considered exactly once.

This is a flexible design, which could be extended to involve greater numbers of participants and topics. For example, in a future evaluation we might want to allow more for the effects of different topics, individual differences among participants and possible biases arising due

[4] The Meta-Code of Ethics, 1995-2005, is available at http://ethics.efpa.eu/meta-code/ (link verified on Oct. 3 2015)

[5] We elaborate on the scenario and tasks, below, see "***Evaluation Scenario***" and "***Experimental Tasks***".

[6] See further details on the sub-tasks in "***Experimental Tasks***" below.

to the allocation of a system-topic combination to different participants. We note that since the system and evaluation method are still in development and the recruitment and management of participants is costly, we concluded that four participants and two topics would be sufficient number to provide useful results and feedback for the interim evaluation.

Finally, the design specifies a single SENSEI system condition. In this experiment the SENSEI system included outputs from multiple SENSEI technologies (e.g. clustering, and summarisation) presented via a multi-faceted, experimental interface -- see Section on "System 2" below for more details.

To fully investigate the respective role of the various outputs/interface features (e.g. summaries vs a graphical pie-chart based on clusters) in the task context, we would need an alternative experiment design involving *multiple* SENSEI conditions (i.e. we factor out the various outputs and interface features in different versions of the interface). However, by introducing further system conditions, we would have to either 1) ask participants to do more tasks (which seemed infeasible) or 2) opt for a different design where participants carry out 1 or 2 system/task combinations, but we don't allow for a full *within* subject comparison of systems. Option 2 would require much greater numbers of participants to account for individual differences between participants.

Given the arguments (see above) against involving greater numbers when the technologies are still under development, we opted for the simple, more coarse-grained design, in which participants assess 1) a current online technology vs 2) a combined SENSEI system. However, we also gathered data on what participants thought about the usefulness of the different system components/interface features in the task context, and which features they mainly used, via the post task questionnaires and group discussion (see Section "Gathering Feedback from Participants below").

Interface Support for the Experiment Design: We developed an interface to support the experiment design described above. This comprised an id-based login for participants, which provided individualised controlled access to the required sequence of system-topic combination for each participant.

Systems: We selected two systems to evaluate: System 1, the Current Guardian News and Comment Facility; System 2, the SENSEI Social Media prototype V1.0, as described in deliverables D5.2 and D6.2. We describe these systems in brief as follows:

- ***System 1:*** makes use of the current Guardian News and Comment online facility. A pre-selected news article and associated set of reader comments is presented via the interface as it appears on the Guardian site. Features for viewing comments include: threads, sort options, expand/collapse threads, etc. In addition users can make use of the web browser "search in screen" option to search for keywords.
- ***System 2:*** has two main components, presented via two "frames" (or windows)[7].
    - **The Guardian component:** The left hand frame presents a pre-selected

[7] For example, the following URL provides access to the SENSEI system displaying the "Heatwave Topic": http://sensei.rcweb.dcs.shef.ac.uk/y2extrinsic/demo.php?data=265726500730208456612606160000
The Heatwave Topic was from the article and comment set: http://www.theguardian.com/uk-news/2014/jul/16/heatwave-alert-england-wales-humid-thunderstorms

news article and associated set of reader comment, via the Guardian News and Comment facility. I.e. this component is identical to System 1.

- **The SENSEI component:** The right hand frame presents features to help people make sense of the comments (the features being the outputs of current clustering and summarisation technology applied to the comments). Features include:
  - a "pie chart" which provides an overview of comment content. This graphical feature is a circle, made up of different sized coloured segments; each segment is based on clusters of comment; the size of the segment reflects the relative proportion of comment in the underlying cluster; descriptive keyword "labels" are assigned to each segment to indicate the character of topical content in the underlying comment cluster.
  - an extractive summary (based on clusters). This summary functions both as an overview of the comment and an index to clusters of related comments, since each comment in the summary is "clickable" and can be navigated to view the underlying cluster of individual comments.

Topics: We wanted to see whether the different systems helped more or less with different topic types, some topics being more complex than others. We selected two contrasting topics, each comprising a news article and a set of associated comments, which varied in terms of overall word length and complexity (see Table 1 for a comparison of simple word and thread counts for the topics). For example, the total word length for the comment on topic 1 "Network Rail" was **4,619** and the average number of words in a comment was **46.2**, while the total word count for comment on topic 2 "Heatwave" was **3,141**, and the average number of words in a comment was **28.8;** The total number of threads was similar across topics: **16** for "Network Rail" and **13** for "Heatwave".[8] Further details of the topics follow:

- Topic 1 (T1), "Network Rail", the more complex topic[9]: the article reported on a recent fine imposed on the UK-owned rail company "Network Rail" due to late running trains. The comments discussed many issues such as the confusing structure of the current mix of private and nationalized companies in the British Rail System; whether Network Rail was responsible for setting ticket prices, the inherent difficulties of fining a publicly owned company; pros and cons of privatisation and nationalization; spending the fine on WIFI, etc.
- Topic 2 (T2), the "Heatwave" topic[10]: included an article about a spell of high temperatures and thunderstorms forecast for parts of the UK. The article included

[8] Total word length calculated based on the set of approximately the first 100 comments (taking the first 100 as ordered by the time of thread posting and rounded up to the nearest complete thread – this resulted in 100 comments for "Network rail"/ 109 comments for "Heatwave.")

[9] Network rail sourced from: http://www.theguardian.com/business/2014/jul/07/network-rail-fined-50m-pounds-late-trains

[10] The Heatwave Topic sourced from : http://www.theguardian.com/uk-news/2014/jul/16/heatwave-alert-england-wales-humid-thunderstorms

details of which regions would be affected; health risks and national records for high temperature. The commenters were more light-hearted in response to the heatwave article and focused on issues such as the pros and cons of air-conditioning versus fans for keeping cool; the definition of a heatwave in the UK relative to other countries where such temperatures were commonplace; how London was a good place to live; media reporting around heatwaves; etc. We judged the comments for the "Heatwave" topic to be less dense and easier to make sense of in a short space of time.

Table 8: Summary statistics for the two Social Media topics[11]

Topic	Article word count	First 100 comments word count	Average number of words per comment	Total number of threads
Network Rail	**730**	**4,619**	**46.2**	**16**
Heatwave	**598**	**3,141**	**28.8**	**13**

Participants: We recruited participants who were native English speakers or who had excellent command of English. Participants were all graduates with background in research in information and language technologies. All participants had experience of using online news and reader comment. To obtain further details on their background we invited participants to answer a few questions relating to their command of English and their background and experience of reader comment at the start of the evaluation session (see Appendix 1).

Scenario: The evaluation is based on the scenario of a generic reader of online news and comment who has a short period of time to read some news and associated comment, as say in a 10 minute coffee break. Ideally he/she would like to gain a comprehensive overview of comment but with only limited time available the reader would be happy to:

I. Identify the main issues addressed in the comments – what were the commenters talking about?
II. Gain a sense of the spread of opinion on a particular issue – i.e.
    - what were the different perspectives and opinion on the issue?
    - areas of consensus and disagreement;
    - the feeling expressed.

We note that this evaluation scenario is a simplified version of the user scenario described in Use Case 1 of the Social Media Use Cases presented in D1.2. Feedback from news and

11 Total word length and thread count calculated based on the set of approximately the first 100 comments (taking the first 100 as ordered by the time of thread posting and rounded up to the nearest complete thread – this resulted in 100 comments for "Network rail"; 16 threads/ 109 comments for "Heatwave";13 threads.).

comment readers (both news professionals and members of the public), in response to our questionnaire on the use cases (also reported in D1.2), confirmed that, given a news article of interest, gaining a comprehensive overview of comment in a short period of time is something that users would like to achieve. However, in current practice, this is an impractical task for most comment readers, due to the complexities of online comment and limitations of current technology. One condition of an extrinsic evaluation is that it assesses how well a technology helps people to carry out a "real world" task or activity. Hence for the purposes of this evaluation, we focused on two activities that readers currently engage in, (with varying degrees of detail and success). These activities are: identifying what people are talking about, i.e. "**identifying issues in the comment**", and "**characterising opinion**" – i.e. obtaining a sense of the spread of opinion on a particular issue. (We note a full overview would identify and characterise opinion on for all the main issues.) For a full definition of the tasks, please see **Appendix 2**.

Experimental Tasks: Based on this scenario, the experiment comprised a sequence of 3 short tasks, each to be carried out within a set time limit:

- Read news article (5 mins); (article only no comments).
- Read associated comments (5mins); (participants view and explore the comments using the specified system condition).
- Answer 2 questions relating to the comments (10 mins); (participants continue to view and explore the comments using the specified system condition):
- Q1: "Identify 4 issues";
- Q2: "Given issue "X", characterise opinion on the issue".

With two different system/topic combinations to explore, participants carried out two iterations of the task sequence described above. We also provided a five-minute break between each iteration.

The short reading tasks are followed by questions relating to the content of the comments: "Identifying Issues" and "Characterise Opinion". The full definition (see Appendix 2), which elaborates on the tasks, was provided in a participant hand-out. While we believe the tasks to be intuitive, we provided this instruction to ensure that participants had a similar level of understanding of what a typical response should include.

Two researchers checked that it was possible to identify at least 5 issues in the comment sets by carefully reading through the comments. We also consulted the SENSEI Gold Standard Summary annotations to check for issues.

We selected the issue for Question 2, "characterise opinion" using a similar procedure. Three researchers selected a candidate issue based on the article, the comments and the SENSEI Gold Standard Summary annotations. The Gold Standard helped us to ensure that candidate issues had been recognized by at least 2 Gold Standard annotators. In addition we required an issue that was discussed by multiple commenters, across different threads and which was substantive enough to make the task feasible. A final discussion of candidate issues resulted in a consensus decision on which issue to select for the task. (See **Appendix 2** for the full questions.)

Gathering Feedback from Participants: We used 2 complementary methods to gather feedback from participants on their experience using the different systems when completing the tasks.

First, we invited participants to complete a short "post-task questionnaire" (see **Appendix 3**); this included:

1. a multi-part question asking participants to rate on a scale of 1-5 how useful they found each of the two systems and then each respective system component, when completing the two experimental tasks ("Identify Issues"; "Characterise Opinion").
2. a more general question asking participants to indicate on a scale of 1-5 the extent they would like to have a SENSEI System available for use in a comment facility, when browsing news and comment.
3. a question inviting participants to provide any further comments or feedback on their experience using the systems to carry out the tasks in the experiment. (This included prompts, such as "was there anything you really liked or disliked?"; "... any possible improvements or things you would like to see included in a system ...").

Second, we invited participants to take part in a group discussion, which was recorded. The discussion was divided into high-level questions about participant experience of i) the experimental tasks; ii) the systems and an open question asking if there was anything else they would like to say about the experiment. For the first 2 high-level questions, we included a list of topics/sub-questions for the discussion to cover, e.g. task difficulty, time to complete tasks, strategies used with different systems for completing tasks, etc.[12] (The Group Discussion Questionnaire is provided in **Appendix 4**).

Further Details of the Experimental Setup: In addition to the tasks and questionnaires described above, the experiment involved: a very brief overview of the session; a demonstration of the different systems and system components (with a demo topic); an opportunity for participants to practice using the different systems (with the demo topic); an introduction to the scenario and instructions on how to answer the content related questions. Before commencing with the "timed tasks" and questions, participants were provided with refreshments. The total time for the evaluation session was not to exceed 2 hours.

We arranged the various tasks in a script, which specified the order and times for the different tasks – sometimes being strict (e.g. the time limited scenario tasks) and others more flexible (e.g. the group discussion). For the full script please see **Appendix 5**.

A series of PowerPoint slides[13] was developed based on the script and the tasks and these slides were used to guide the researchers and participants through the different stages of the evaluation session.

[12] The group discussion is an example of a "semi-structured questionnaire"; it is designed to encourage open discussion and to accommodate conversation as it happens and as it drifts into different topic areas; the researcher leading the discussion may encourage comment using the prompts/sub-questions; he/she need not ask a question if the topic has been sufficiently addressed by participants in response to a different question.

[13] http://sensei.group.shef.ac.uk/extrinsicEvaluation/ExtrinsicEvaluationSlides.pdf

We tested the setup in a run-through with colleagues from the USFD SENSEI team acting as participants and adjusted some of the timings accordingly.

Ethics: The experimental setup described above received ethics approval from the Department of Computer Science Ethics Review Committee, in accordance with the University of Sheffield ethics policy.

3. Evaluation metrics and results

3.1. Speech extrinsic evaluation results

3.1.1. Speech extrinsic evaluation scenario 1- ACOF: Results

The first extrinsic evaluation scenario for speech required the comparison between the human filled ACOF and the automatically generated ACOF (Section 2.2.1). In Table 9 and Table 10 we report the descriptive statistics, i.e. means, followed by standard deviation in brackets, calculated from the scores that the participants assigned to the question "How much accurate was the system in filling this ACOF item?" for the three items of LUNA and RATP-DECODA ACOFs.

Table 9: LUNA results

LUNA				
	Participant A	**Participant B**	**Participant C**	**Total**
ACOF Item 2	2,8 (1,03)	4 (0,01)	4 (0,01)	**3,6**
ACOF Item 3	2,9 (1,97)	4 (0,01)	4 (0,01)	**3,63**
ACOF Item 10	3,8 (0,63)	3,5 (0,7)	3,9 (0,31)	**3,73**
Mean	3,17 (1,14)	3,83 (0,49)	3,9 (0,36)	**3,65**

Table 10: RATP-DECODA results

DECODA			
	Participant D	**Participant E**	**Total**
ACOF Item 3	3,9 (0,3)	3,85 (0,36)	**3,875**
ACOF Item 9	3,85 (0,48)	3,75 (0,55)	**3,800**
ACOF Item 10	3,85 (0,36)	3,95 (0,22)	**3,900**
Mean	3,86 (0,38)	3,85 (0,4)	**3,858**

From this comparison task, it resulted that on average, the participants found that the SENSEI prototype was *accurate* in assigning the Pass/Fail/NA values to the three ACOF items.

From the post-hoc consensus meeting between participants, it resulted what follows:

1. For the DECODA conversations the main issue of uncertainty between human evaluators and the SENSEI prototype was about the ACOF item 10: that item is about the management of the waiting time by the agent. Since in the DECODA corpus is an information delivery service there is no waiting time within the call. In those cases the human evaluators scored the item as NotApplicable, while the system always scored it with Fail.
2. Some discrepancies were also reported both for LUNA and DECODA for the use of the welcoming messages: in some cases, for example, the agents did not say “thank” to the customer after some waiting time. The human annotators were more lenient in those cases, while the system always marked them as Fail.

3.1.2. Speech extrinsic evaluation scenario 2 - Synopses: Results

For the evaluation of the second speech scenario we recorded the one-hour focus group discussion and transcribed the speech. As we explained above, all the participants had familiarity with the speech SENSEI prototype: some of them (2 out of 4, Participant A and B henceforth) personally participated in the first evaluation task, and 2 out of 4 (Participant C and D henceforth) had acquaintance with the system and the research goals, but they were not involved in evaluation task with the role of annotators.

We analysed the data by using metrics from conversational analysis. In particular, we analysed the turn taking structure of the conversation, the discussion of given topics, and the novel topics that emerged from discussion.

As for the turn taking, we could observe that the conversation went on smoothly. In particular there were not instances of overlapping speech. This may be due partly by the remote setting that we adopted: since the only channel of communication was audio, the participants spontaneously adapted to intervene in the discussion in an orderly manner. So, in general, the moderator asked the questions of the focus group plot reported above, and each participant made her remarks and comments in an orderly fashion. However, it is worth mentioning that for each question whose answer required a direct experience with the prototype, Participant A and B took the floor of the conversation first, and C and D commented after listening to their experience.

All the topics given in the plot were explored. In Table 11 we report the content that emerged from the discussion for each question.

Table 11: Comments on focus group questions

Questions	Comments
1 How did you find using the ACOFs filled by the system?	Participant A and B evaluated as positive their experience with using the SENSEI prototype. Their answers were more based

	on the experience with automatically generated ACOF than with synopses, as it will emerge from the comments to the subquestions 1.1 – 1.4
1.1 Were the ACOF items useful for highlighting agent's behaviour?	Participant A and B agreed on judging the ACOF item as useful for focusing on behavioural aspects of call centre agents' performance.
1.2 At what extent did you agree with the judges of the automatically filled ACOFs?	Most of the time participant A and B agreed on the score assigned by the system.
1.3 Time: could you imagine that automatically filled ACOFs may help you in sparing time in your supervising job?	All the participants agreed on the usefulness of automatically generated ACOF for sparing time in the supervisor job, but participant D outlined that the supervision by an automatic system could be accepted by the call centre agents only if the accuracy of the system score can be no less than 80%.
1.4 Do you think that automatically generated ACOFs could be enriched with evidence of the system decisions?	Participant A and B said they could be helped in their supervision tasks if they could access to evidence of the system decisions.
2. We would like to hear more about your experience using the synopses of the call. This is something completely new in your daily activity. Do you have anything to say about this?	Participant A and B said that the synopses of the call might be useful for assessing the issues related with call content. In the evaluation tasks they used to read the synopsis of the call after reading the ACOF.
3. Do you have a preference for the automatically generated ACOFs or for the summary of the call? Do you find them equally useful, maybe for different tasks? Could you please give us some examples?	Participants A, B, and C replied to this question. They agreed that that the synopses of the call might be useful for assessing the issues related with first call resolution, and for identifying the reason of the call for inbound call. A and B said that the synopses should be more informative for being really used, i.e. they should report more on call content.
4. Could you tell us something about the strategies you used to formulate the queries based on the ACOFs?	Participants A, B and C replied to this question; the chosen ACOF items were deemed to be the most informative for focusing on two aspects of agents' behavior, i.e. expertise in call time management and appropriateness of communication attitude.

5. Did you fine SENSEI potentially useful or "added value" for your job? 6. Can you imagine different tasks where you would find useful system-generated summaries, for example summaries over millions of conversations, more focused research, retrospective investigations on very large sets of conversations?	All the participants agreed that the SENSEI generated summaries could provide added value to their job due to the larger number of potentially supervised calls. Participant D added that using the system could help in overcoming the subjectivity issue of listening forms filled by human supervisor, but she also outlined again the issue of acceptance by the observed agents.
7. Is there anything else you would like to tell us about your experience with the SENSEI speech prototype ?	All the participants would like to evaluate tasks where the system runs in quasi – real time and provides the QA supervisors with insights on which agents are having problems with the ongoing calls in order to provide real time interventions to address the reported issues.

Subtasks of speech scenario 2 based on user generated queries

For the second extrinsic evaluation task of the speech scenario, the post hoc discussion in the evaluators' group also brought out an interesting potentiality of the prototype related with *agents' professional profiles.* This comment was based on their experience with the queries defined for selecting the problematic calls. This task was done on the LUNA corpus. In this sub task the problematic calls should report Fail values to any of the ACOF item on which the evaluation was focused. For the LUNA corpus, the results of the extractions are reported in Table 12 below.

Table 12: Types of agents' failure

Patterns of failure			
Cases	**LUNA ACOF Questions**		
	2	**3**	**10**
1	Pass	No Pass	Pass
2	No Pass	Pass	Pass
3	Pass	No Pass	Pass
4	Pass	No Pass	Pass
5	No Pass	Pass	Pass

6	No Pass	Pass	Pass
7	No Pass	Pass	Pass
8	Pass	No Pass	Pass
9	No Pass	No Pass	No Pass
10	Pass	No Pass	No Pass

It can be noticed that there are four possible patterns of failure:

- Pattern 1 is exemplified by cases reported in the rows 1 – 3 – 4 – 8 of the Table above
- Pattern 2 is exemplified by cases reported in the rows 2 – 5 – 6 – 7 of the Table above
- Pattern 3 is exemplified by cases reported in the row 9 of the Table above
- Pattern 4 is exemplified by cases reported in the row 10 of the Table above

The first pattern represents the case of call failure where the system classified the agent as being able to listen to the customer with interest and positive attitude, but at the same time, classify as critical the agents' behaviour with respect to the clarity and exhaustiveness of the information delivered to the customer. *This pattern outlines a professional profile of agent who is capable of empathy, with great ability to listen to the customer, but at the same time with poor communication skills*.

The second pattern outlines an agent who is able to communicate, and to adapt to the speech style of the interlocutor, but with poor listening abilities. According to the evaluators this agent profile is critical because the active listening ability is crucial for focusing on the customer problems and for the efficient management of the call. *This aspect has impact on the time management of the call, because they may increase due to the lack of focus on the real need of the customer.*

The third pattern summarises instances where all the three fundamental requirements of the agent communication have been classified as fail. *This agent profile is obviously very critical*.

In the fourth pattern, despite of actively listening, the agent does not show to be able to clearly communicate and *s/he is not able to adapt to the speech style of the customer*. According to our evaluators, also this agent profile is critical because the quality scores of the service provided run the risk of being low.

This post-hoc analysis allowed the identification of a novel task for the speech SENSEI prototype, i.e. the use of the system classifications for selecting patterns of problematic agents' behaviour on the basis of queries autonomously defined by the prototype users.

3.2. Metrics for the social media extrinsic evaluation

We applied 3 different approaches to interpret the data gathered in the experiment:

1. we summarised data gathered from questionnaire 1 (user background) and questionnaire 3 (user ratings of the different systems/system components, on a 5 point scale) using simple statistics;
2. we assessed and summarised the free text/spoken responses gathered via the feedback questions of questionnaire 3 and the group discussion session using simple qualitative techniques: any patterns in the data are reported and supported with quotes/examples from the participant responses;
3. the written responses obtained from the two questions in questionnaire 23 (the content related questions) were assessed quantitatively using a novel graded scheme, which we describe in detail as follows:

Question 1: "Identify 4 Issues"

In question 1 the task was to "To identify 4 issues". To assess the quality of participant responses to this task we used a four point scale, ranging from 0-3. Judges assigned an individual score to each of 4 issues.

The 4 point scale (given in detail below) takes account of criteria including "evidencing" (i.e. is there evidence for the issue in the comments—is it an accurate description of a "main issue" in the comments?); and "clarity of expression" (how clearly is the issue articulated?).

"Evidencing"

To help judges assess the evidence for an issue, we provided the judge with access to the article and the set of comments (threads expanded) and 2 sets of Gold Standard annotations for the topic, which included: the summaries - an overview of opinion in the comments, and group annotations - these comprised groups of manually assigned comment "labels" (mini summaries of a comment), which point to the original comment they were based on. We advised that judges should first check a candidate issue against the Gold Standard data, but that they should also check for evidence in the original comments. If the candidate issue was not present in the Gold Standard annotations, judges were still to examine the comments by reading through the comments carefully, and by making use of their browsers "search in screen facility".

Evidencing "Main issues"

In accordance with our instructions on "how to identify issues" (see Appendix 2), assessing the degree of evidence in the comments takes into account the amount of comment relating to an issue, relative to the amount of comment referring to other issues. In other words, strong evidencing = the issue was "a main issue", and not something discussed by one or 2 comments.

"Clarity of Expression"

Our instructions on how to identify issues indicated the form of expression an issue should take.

We decided to use a single score as opposed to providing a rating for different criteria respectively, since the criteria of evidencing and clarity are interdependent. For example, if an issue is expressed poorly, i.e. there is a low degree of clarity of expression, it is difficult for a judge to make an assessment of the quality of evidence in the comments, because the ambiguity means he cannot be exactly sure of what he is assessing for evidence.

For example, a candidate issue is given as “ticket prices”. It is unclear from this expression what the “issue” is, i.e. what is being said about ticket prices in the comments. A more clear expression is “ticket prices are too high” while another is “who is responsible for ticket prices”. There is evidence for both these issues in the comments. One is discussed by many comments, the other is discussed by a few, but it is impossible to know which issue the candidate expression is referring to.

We provided the following guidelines for assigning different scores:

Score: 0

• No issue given or issue given but no evidencing apparent; a well-articulated issue with no evidencing in the comment would receive a score of 0.

Score: 1

• The issue is expressed poorly, but some content is indicated and the comments can be seen to address it, for example, for a response “ticket prices” there is evidence of people talking about different things to do with ticket prices in the comments.

• The issue is clearly articulated as a proposition or something that one can believe or not believe but is poorly evidenced, e.g. only 1 or 2 comments discuss the issue.

Score: 2

• The issue is adequately expressed e.g. “fining directors”, but one could imagine it being made more clear and specific, e.g. “fining directors would be a more effective way of ensuring trains run on time”. The issue is of sufficient clarity to assess evidence or strength of support in the comments, which should be good or satisfactory.

• A well-articulated issue but with a low level of evidencing, say 2-3 comments, or when there were relatively many other more significantly discussed issues would get a 2 .

• A clearly articulated issue, with good supporting evidence, but perhaps a better way of characterising the issue could have been chosen – e.g. if the issue represents the minority view “privatisation of railway is needed” – supported by say 1 or 2 comments when the majority focused more on the opposite view “nationalising the railways is the solution to the problems”.

Score: 3

• The issue is clearly articulated/expressed; so it is straightforward to assess evidencing/strength of support, which is good (relative to the overall discussion in the comments).

Question 2: "Characterise Opinion"

In question 2 the task was "To characterise opinion". To assess the quality of participant responses to this task we used a six point scale, ranging from 0-5, with maximum and minimum values given: 0 = no characterisation of opinion present, 5 = excellent characterisation of opinion in the response. An "excellent" rating requires an answer to include: good coverage of opinion on the issue, i.e. details of the different perspectives on the issue/the different sides to the argument; where there was consensus or not; some detail of the respective quantities of opinion; and the characterisation should be accurate, i.e. there should be evidence for the information given in the comments. Judges assigned a single score to the response given for Question 2 for a particular topic.

3.3. Social media extrinsic evaluation results

The results of the social media use case evaluation are reported with respect to the evaluation metrics described above. The evaluations take into account the participants' background in terms of language and experience with online news forums, which we report first.

3.3.1. Participants' background information

Background information about participants was collected using the questionnaire shown in Appendix 1. 3 summarises the background information that we collected for the four participants, who took part in this task:

- All participants are English native speakers
- Two described their roles as "news and comment readers" (i.e. "I read news and/or comments but very rarely provide/post comments"), whilst the other two described themselves as "comment providers" (i.e. "I read news and/or comments and provide/post comments on a regular basis")
- Three participants engaged with online news at least once a day, and one at least once a week

Their engagements with reader comments vary widely: at least once a day (1 participant), at least once a week (1 participant), at least once a month (1 participant), and very rarely (1 participant).

Table 13: The background information about participants in social media evaluation task

		Participant 1	Participant 2	Participant 3	Participant 4
Q1	**English language proficiency**	Native speaker	Native speaker	Native speaker	Native speaker
Q2	**Your current role**	Comment provider	Comment provider	News and comment reader	News and comment reader
Q3	**Engagement with online news**	At least once a day	At least once a day	At least once a week	At least once a day
Q4	**Engagement with reader comments**	At least once a day	Very rarely	At least once a week	At least once a month

3.3.2. Participants' responses to the content questions

A caveat that applies to all the results below is that they are derived from very small numbers of participants and topics and thus we cannot draw strong conclusions at this stage. Nonetheless, some patterns have emerged that provide insights into the current state of the SENSEI technology and the experiment design. The original participant responses to the content questions are provided in **Appendix 6**.

Was one topic easier to answer than another? Was one content question easier than another?

Table 14 and Table 15: show the results for the total scores allocated to participant responses for questions 1 and 2, for the different topics respectively.

Table 14: Participants' responses to T1: the "Network Rail" article

System	Participant ID	Question 1	Question 2	Total Score[14]
System 1	1	10	0	10 (58.8%)
System 1	4	11	3	14 (82.4%)
System 2	2	4	1	5 (29.4%)
System 2	3	5	2	7 (41.2%)
Grand Total		**30**	**6**	**36 (52.9%)**

[14] Total score as a percentage of the maximum possible score. The maximum score per task is 17, which is 12 points for Q1 (4 questions @ 3 points) and 5 points for Q2.

Table 15: Participants' responses to T2: the "Heatwave" article

System	Participant ID	Question 1	Question 2	Total Score
System 1	2	10	0	10 (58.8%)
System 1	3	12	2	14 (82.4%)
System 2	1	9	4	13 (76.5%)
System 2	4	12	4	16 (94.1%)
Grand Total		**43**	**10**	**53 (77.9%)**

The overall score for all responses obtained in T2, the Heatwave topic (53 out of a possible 68 or 77.9%), is clearly higher than that for T1, the Network rail topic (36 out of a possible 68 or 52.9%). This most likely reflects the fact that T1 is a more complex topic with longer comments. T2 is easier to read and interpret in a short space of time and hence people were able to make sense of the comments and answer the questions more effectively (i.e. to identify, check evidence and articulate issues, and to characterise opinion). Indeed, the results show that overall, both content questions (Q1: "Identify 4 issues" and Q2: "Characterise opinion") received better scores for T2 than T1.

Q1 responses in T2 scored a total of 43 out of a maximum of 48 whereas Q1 responses in T1 scored only 30 out of 48. Similarly, Q2 responses in T2 scored a total of 10 out of a maximum of 20 whereas Q2 responses in T1 scored only 6 out of 20. The particularly low scores for the "characterise opinion" question for T1 suggests that such a short period of time (up to 10 minutes plus 5 minutes reading time) is insufficient for most people to characterise opinion for a given issue on a complex set of comments, using either system.

Overall responses (all participants in both system conditions) to Q1 scored higher, proportionally, than responses to Q2 (76% versus 40%, see Table 17). Suggesting that identifying issues is easier than characterising opinion.

Did different systems help with the different content questions?

Interestingly, if we compare the participant scores for the different questions types, in the different system conditions, see Table 16, there is a clear pattern: System 1 scores for Q1 are much higher (89.6% out of maximum possible score) than those for System 2 (62.5%). Whereas for Q2, System 1 scores were lower (a total of just 5 or 25%), whereas the total for System 2 was 11 (or 55%). Two "zero" scores (for P1 and P2) were recorded for the System 1, Q2 condition, and this accounts in part for the low total. However, we note that the zero scores can be explained partly by problems in the participant understanding of the task instructions (as reported in the post task discussion), i.e. P1 did not know to continue to the second question in this first iteration. Meanwhile, P2 was not aware that they could explore beyond the SENSEI component when using System 2, and thus having found 1 relevant comment in the SENSEI component, was unable to do much in terms of characterising opinion. Say if we substituted these scores with an above average score of 3 out of 5; then

the totals for Q2 in System 1 and System 2 would be the same. Suggesting that there is not such a difference between the different systems, in respect of Q2.

Table 16. Total participant scores for different question types in both system conditions

Participant ID	Question 1[15]		Question 2[16]	
	System 1	**System 2**	**System 1**	**System 2**
1	10	9	0	4
2	10	4	0	1
3	12	5	2	2
4	11	12	3	4
Total	**43**	**30**	**5**	**11**
% of total possible scores	**89.6%**	**62.5%**	**25%**	**55%**

The relatively strong performance of System 1 on Q1: "Identify Issues", suggests that current technology provides some support with respect to this task. Results of the post-task questionnaire (see Section 3.3.3 below) in which participants rated the system features, indicated that the Guardian thread was found to be very useful in completing the task, especially in identifying four issues in the comments. This suggests that while threads might not capture all comments relating to an issue, the structure is adequate for indicating topics/issues in the conversation.

Further development in SENSEI clustering technology will focus on providing more useful and coherent clusters of comments. With such advances, we hypothesise that people would do better overall on both questions ("Identifying Issues" and "Characterising Opinion"), since good clusters should represent issues and gather together related comments on issues. We can also hypothesise that an improved SENSEI condition might help people to answer questions in less time. Recording time might thus be something we could include in a revised evaluation.

Table 17: Comparison of participants' response scores for questions 1 and 2

Participant ID	**Question 1[17] (Combined scores)**	**Question 2[18] (Combined scores)**	**Total**
1	19 (79.2%)	4 (40%)	23 (67.6%)
2	14 (58.3%)	1 (10%)	15 (44.1%)
3	17 (70.8%)	4 (40%)	21 (61.8%)
4	23 (95.8%)	7 (70%)	30 (88.2%)
Average (%)	**76%**	**40%**	**65.4%**

[15] Total possible Q1 scores is 48 (4 people @ 12 points)
[16] Total possible Q2 scores is 20 (4 people @ 5 points)
[17] Max scores: 24 (2 tasks @ 4 questions @ 3 points)
[18] Max scores: 10 (2 tasks @ 1 question @ 5 points)

Table 18: Participants' performances in different questions

Rank Position	Question 1	Question 2
1	P4	P4
2	P1	P3, P1
3	P3	
4	P2	P2

Did one system help overall more than another?

If we compare the overall scores for responses within a particular system condition (see Table 19) we see that the scores for a system condition are fairly similar: for System 1 (Guardian comments only) the total was 48, out of a possible total of 68 (or 70.6%); for System 2 (SENSEI + Guardian), the total was lower at 41 out of a total of 68 (or 60.3%)[19].

Table 19: Comparison of overall participants' response scores with different systems

Participant ID	System 1[20]	System 2[21]	Total
1	10 (58.8%)	13 (76.5%)	23 (67.6%)
2	10 (58.8%)	5 (29.4%)	15 (44.1%)
3	14 (82.4%)	7 (41.2%)	21 (61.8%)
4	14 (82.4%)	16 (94.1%)	30 (88.2%)
Total	**48**	**41**	**89**
Average (%)	**70.6%**	**60.3%**	**65.4%**

Table 20: Participants' performances in different systems

Rank Position	System 1	System 2
1	P4, P3	P4
2		P1
3	P1, P2	P3
4		P2

[19] Although we note again that the two zero scores arising from a misinterpretation of task instructions in System 1 Q1, will have deflated the overall System 1 scores, suggesting the overall difference between systems might have been greater than what is reported above.
[20] Max score: 17 (Q1 (4 questions @ 3 points) and Q2 (5 points))
[21] Max score: 17 (Q1 (4 questions @ 3 points) and Q2 (5 points))

In the group discussion (see below) two of the participants P1 and P4 reported that they mostly found answers using the Guardian component of System 2, while P2 and P3 reported that they used the SENSEI component of System 2 (although in the Group discussion P3 said that they had relied on their memory of using the Guardian component in the reading session, when answering questions). As we discuss further below P2 and P3 (see also Table 19 and Table 20) received the lowest scores when using System 2. This suggests that the SENSEI component in its prototype form was found overall to be less helpful than current technology.

While System 2 responses scored lower overall than those for System 1, the responses of P1, which were scored lower relative to the others when using System 1, than for System 2 (see Table 19) appear to be an exception. A possible explanation for this might be in differences in the difficulty of topics: P1 had to answer T1 questions with System 1. However we note again that this participant was not aware of having to answer a second question in the first system topic condition and this meant they received a Q2 score of 0, which notably lowered the total score in the S1 condition, so it is difficult to draw any conclusions from the exception.

Did a particular system help with a particular topic?

The total score for participant responses using System 1, was the same for both topics, while the scores obtained for System 2 were considerably higher in one topic condition than the other. The total score for questions on Topic 1 (T1: "Network rail"), when participants were using System 1, was 24; and the total score for questions on Topic 2 (T2: "Heatwave"), when participants were using System 1, was also 24. By comparison, the total score for questions on T1, when participants were using System 2 was just 12 (half that of the score for responses using System 1); whereas, the total score for responses from participants using System 2 with T2: "Heatwave", was much higher at 29 (i.e. four points higher than the score for System 1 and T2: "Heatwave").

Table 21: Comparison of participants' response scores for topic 1 and 2

Participant ID	Topic 1[22] (Combined scores)	Topic 2[23] (Combined scores)	Total
1	10 (58.8%)	13 (76.5%)	23 (67.6%)
2	5 (29.4%)	10 (58.8%)	15 (44.1%)
3	7 (41.2%)	14 (82.4%)	21 (61.8%)
4	14 (82.4%)	16 (94.1%)	30 (88.2%)
Average (%)	**52%**	**77.94%**	**65.4%**

[22] Max scores: 17 (Q1 (4 questions @ 3 points) + Q2 (5 points))
[23] Max scores: 17 (Q1 (4 questions @ 3 points) + Q2 (5 points))

Table 22: Participants' performances in different topics

Rank Position	Topic 1	Topic 2
1	P4	P4
2	P1	P3
3	P3	P1
4	P2	P2

The results suggest that overall the most difficult combination of System and Topic was using System 2 (SENSEI + Guardian) to answer questions on the more complex topic, T1. While System 2 was perhaps the better-suited system for providing support for answering questions on the simple topic (T2: "Heatwave").

However, if we consider that Topic 1 ("Network Rail") was longer and more complex, and that the experiment design meant that the same two participants (P4 and P1) were assigned System 1 with T1 and System 2 with T2, we can also explain these results due to both topic effects and differences between participants.

Were some participants better at answering the questions than others?

Evidence for differences between participants can be seen if we compare the respective total scores for the 4 participants' responses. Those of one participant (P4) were consistently high (see Table 17 and Table 19); while responses for another participant (P2) were consistently low:

- Overall, P4 responses received a total of 30 out of a maximum of 34, or 88.2%; the total responses for P3 and P1 were fairly similar, with those for P3 given a score of 21 out of 34, or 61.8 % and those for P1 scored as 23 out of 34 or 67.6% respectively; the total P2 responses scored just 15 out of 34 or 44.1% ;
- P4 responses were ranked first for both content questions (see Table 18): for Q1 "identify issues" P4 scored 23 out of a possible 24 (95.8%) and for Q2 "characterise opinion" a total of 7 out of 10 (or 70%); again P2 responses received the lowest scores relative to the other participants' for both Q1 (14 out of 24 or 58.3%) and Q2 (1 out of 10 or 10%).
- P4 responses were also ranked 1st out of all the participants in both system conditions (see Table 20). However, we note that there was less difference between the scores obtained for the different participant responses when using System 1 than when they were using System 2: P4 and P3 responses ranked joint 1st position for System 1, both scoring 14 out of a possible 17 while P2 and P1 both scored 10 out of a possible 17. In comparison when using System 2, P4 responses scored highly at 16 out of 17, with P1 scoring 13 out of 17, while P3 responses were scored much lower at 7 out of 17 and P2 responses at just 5 out of 17.
- Table 22 shows how P4's scores ranked first for both topics. P2's scores ranked last.

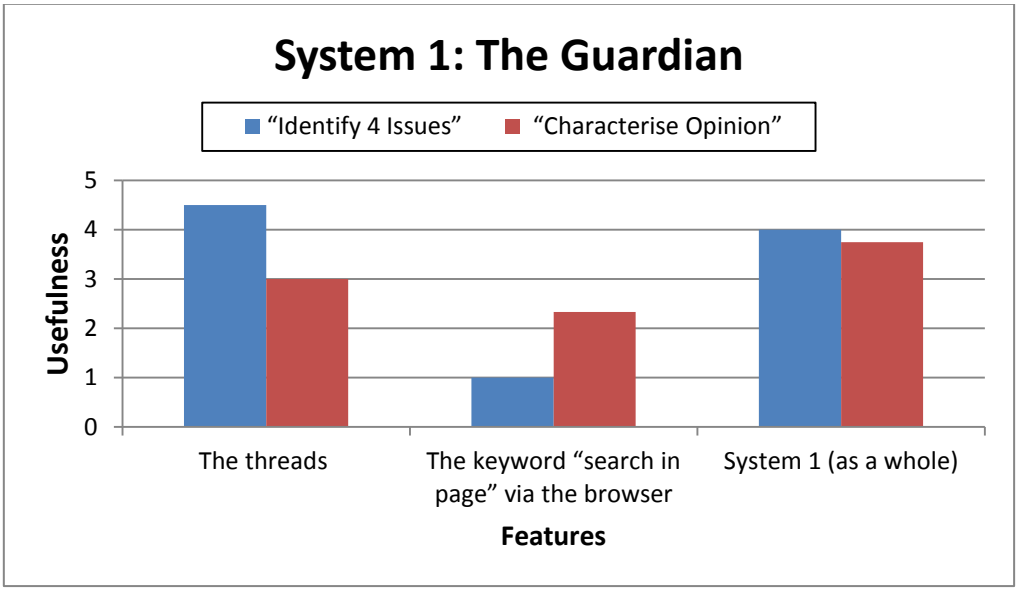
The differences between participants suggests that in future work would benefit from selecting larger numbers of participants in order to ensure against effects of bias of particular systems and topics being assigned to a particular individual.

3.3.3. *The participant experience reports*

The participants' views on usefulness of the different systems for making sense of online reader comments were collected using the questionnaire in Appendix 3.

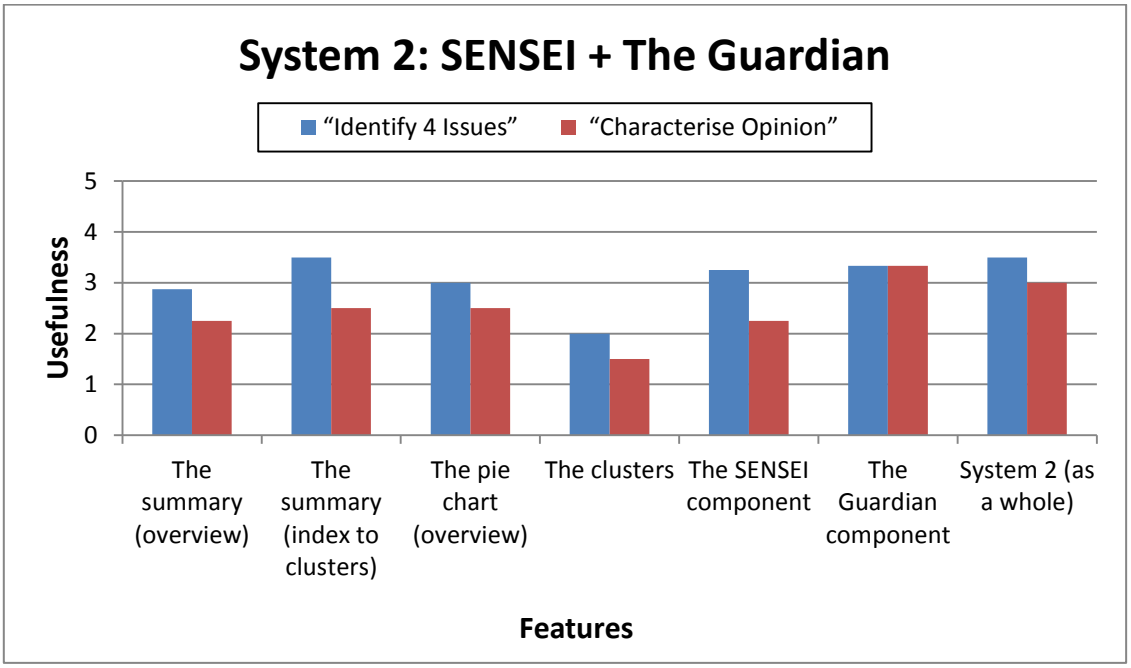
Figure 1 and Figure 2 show a summary of the four participants' ratings of the usefulness of System 1, in comparison to the ratings of the usefulness of System 2, in the context of the two tasks "Identify 4 Issues" and "Characterise Opinion". The participants were asked to indicate on a scale of 1-5 (1=not useful and 5=extremely useful) how useful the different systems/system components were when completing the experimental tasks. The figures show the usefulness scores averaged across the four participants' ratings.

Figure 1: Usefulness rating for The Guardian standard comment view system



Participants indicated that the Guardian thread was very useful in completing the task, especially in identifying four issues in the comments. The keyword search, on the other hand, was not found to be very useful; three participants indicated "1" as its usefulness and one participant did not use the feature at all in the task. In characterising opinion, one found the keyword search to be extremely helpful in finding comments of a specific issue; however the remaining participants did not find this feature to be helpful at all. Overall, they found the Guardian system to be useful in completing both tasks; an average score of 4 and 3.75 for the Guardian's usability in "Identifying 4 Issues" task, and "Characterising Opinion" task, respectively.

Figure 2: Usefulness rating for the SENSEI system



Participants were also asked to indicate the usefulness of each feature found in the SENSEI system. In both tasks, participants found that the summary (as an index to clusters) was the most useful feature in the SENSEI system, followed by the pie chart (as an overview), and the summary (as an overview). The "clusters" feature, on the other hand, was not indicated to be very useful (an average score of 2 and 1.5 for both tasks). Comparing both tasks, they found that all features were more useful for Task 1 (identifying four issues) than for Task 2 (characterising opinion about an issue).

All participants were then asked to assign a usefulness score for each component in the system, i.e. the SENSEI component and the Guardian component, in completing both tasks[24]. In Task 1, the participants identified both SENSEI component and the Guardian component to be as useful in completing the task (average scores of 3.25 and 3.33, respectively). However, in Task 2, they indicated the SENSEI component to be less useful compared to the Guardian component (average scores of 2.25 and 3.33, respectively).

In addition to these ratings the participants were asked to indicate on a scale of 1 (would not like to have) - 5 (would really like to have), the extent they would like to have System 2 (SENSEI + Guardian) available for use in a comment facility, when browsing news and comment. The results are shown in Table 23.

[24] One participant did not provide a usefulness score for the Guardian component as they did not use it when using System 2 (SENSEI system). Therefore, the usefulness score of the Guardian component was averaged between the three participants only.

Table 23: Participants' responses to the question: Would you like to have SENSEI + Guardian system available for use in a commenting facility?

Participant 1	Participant 2	Participant 3	Participant 4	Average
3	1	3	3	2.5

Finally, participants were invited to provide any other comments/feedback about their experience using the systems to carry out the tasks in the experiment. These are the responses we collected.

- Participant 1:
  - "The clusters seem to be pulled from most popular comments rather than relevance to main topic - this takes the edge of usefulness, especially when humour is prevalent."
  - "I didn't notice the 'show more' link."
  - "The summary would be easier to read if the comments were spaced more / or perhaps use shade to differentiate (anyway - too close together to read easily!)"
- Participant 2:
  - "In principle, any summarisation should be useful. In practice, bad summarisation is worse than none and misleading."
- Participant 3:
  - "I had a pretty negative experience in the SENSEI system but mainly because the pie chart and clustering wasn't very good and just made things confusing."
  - "It is also easier to just read the full comments in the Guardian as the short snippets didn't give you a proper understanding of the comments."
  - "If it worked better, I would like it more as I like the concept."
  - "However, I am an exceptionally fast reader and very good at skim reading (I never read such articles properly anyway, I just skim them) so I am probably atypical. A slower reader might find this more useful."
- Participant 4:
  - "5 mins to read the article was a little too long for me."
  - "The colour box and label alignment would be better as top-aligned rather than middle-aligned."
  - "With the small gaps between the labels and the comma-separated list nature it was difficult to see what the label of each colour has."
  - "Adding a link from the summary comments back into the threaded version would make it a lot more useful."

3.3.4. *Group discussion contributions*

To complement data collected in questionnaires, a discussion with all participants that lasted 30 minutes was conducted at the end of the session. This made use of a semi-structured questionnaire (see Appendix 4). Similarly to the user experience questionnaire, the aim of the discussion was to elicit additional user feedback on the usability of the systems and the features they offer, and also to collect suggestions on the evaluation task itself, which will be taken into account for the final evaluations in Period 3. The discussion was broadly divided into 3 high level questions about participant experience of 1) the experimental tasks; 2) the systems and 3) an open question. The responses to the first two questions covered four aspects of the evaluation:

- the evaluation tasks: reading the news article, reading the comment sets using either of the interfaces, content questions (identify 4 issues and characterise opinions);
- the two systems for making sense of comments;
- their strategies in solving the tasks: which features of the systems were used for each of the two question types;
- general usefulness of the SENSEI system.

We summarise the feedback as follows.

User feedback on the task

- The difficulty of the questions

One participant (P3) found the questions very easy, the others concurred that the questions were easy, however, that this was dependent on the topic (article). The discussion on the notion of "issue" (see next Section) implies that the questions are easier to answer if the article attracts more "major issues" in the comments.

- Disagreement on the notion of "issue"

3 participants found that the article "Heatwave" did not have sufficient issues, but a lot of minor "things that were talked about". P3 did not agree with this claiming that there were sufficient issues discussed in the comments.

This suggests that the notion of "issue" was interpreted differently by the participants, despite the definition in the instructions, which was read by all of them and the example based introduction by the researchers leading the evaluation session. We note that the instructions explained that an "issue" could be about something trivial or "everyday", such as "hard floors are better than carpets".

The following conversational contributions by the participants clarify the disagreement between participants:

P3: "It wasn't an issue in the real world, but it was an issue in the discussion."

P2: "It was something that was discussed, it was hardly an issue."

P1: “It felt like the discussion from the 2nd one is quite trivial and not really raising major issues like health or (.) yeah. Cause the article talked about people dying and that didn’t really come out.”

P1: “I wouldn’t kind of want to rigorously investigate that kind of comment thread because it was just too trivial.”

This suggests that according to participants P1, P2 and P4, in order to be defined as an “issue” a statement needs to be of sufficient significance or relevance – P1 and P4 explain “relevance” as in relation to what the news article talked about. P3 on the other hand defines an issue as any statement that comes up in the discussion, independently of the article. Participants were not convinced that a better definition of what constitutes an issue by the researchers conducting the task ahead of doing the task would contribute to them doing the task differently.

We note this suggests that participants had a very strong pre-conception of what an issue is, and that this was different to our given definition of an “issue”. We used the term “issue” to capture what may be equally described as an “argument” or “proposition”. We chose “issue” in the end since it is less formal than these other terms and because we wanted to avoid using the term “topic”, as this has certain connotations, i.e. as something that can be described by a keyword and not a proposition, e.g. “climate change”. One possible modification to the protocol is to choose a new term to avoid the misleading connotations of “issue”.

- Time given for the completion of the tasks

There was a general consensus that the time for reading the article was too long, while that for reading the comments was too short, in particular in the Guardian system.

P1: “We could probably have had 10 minutes to read the comments.”

P1: “To see the comments and to look for the issues in them 5 minutes just wasn’t enough.”

Doing all at the same time resulted in some participants, who are quicker readers getting bored, while waiting for everyone to complete.

The following possible factors that influenced completing the timed tasks were discussed:

- The difficulty of reading and making sense of what is read on the computer screen

P1: “I actually find it difficult to read and comprehend on screen. Particularly on the computer screen, more than on a mobile screen. I don’t know if that affects the task.”

- The general reading speed

P3: “I think one of the issues (…) which will massively affect your results is how fast people typically read.”

- The familiarity with/curiosity about a topic

P1: “I would do the same (read 10 pages of comments in 2 minutes) if it’s a topic I’m interested in. If it’s a topic that is less interesting or I’m less knowledgeable about, then it takes longer to read.”

P3: “I’d do the opposite. If I’m interested in something I’d read it much more slowly and carefully. If I’m not, I’d just flick through it, …”

P4: “… Even if you know about it and read slow, you certainly spot the issues easier, because you know what issues you expect to see.”

- Reading comments before getting the question sheet

There was no general recommendation on this, but pros and cons were discussed:

P2: “I would have rather had more time in total just to read the comments and summarise them as opposed to read them and then summarise them and also read them again.”

P4: “Probably just give us 15 min. to do both (read comments and answer the questions)”

P1: “I suppose what you might get if you people the combined times that some people might uh just write the first few things mentioned rather than…”

We note that this last comment by P1 in part captures the motivation for the task design (in which a “reading time” is separated from question answering time): i.e. that the controlled reading time may encourage participants to spend time gaining an overview; whereas in a combined question answering and reading time participants might be eager to begin answering questions and thus limit their attention more to the first few comments.

User feedback on the system

Standard Guardian threads were found useful as they reflect the real conversation. The SENSEI system was criticised on a number of attributes, in particular the clusters and the pie chart.

Positive comments on the SENSEI interface:

- P4: “I quite like the idea of the kind of the cluster certainly the pie chart.”
- P3: “I did like the concept of it (labelled clusters in the pie chart)”

Problems with the SENSEI interface:

- Alignment of the labels against the colour blocks – it’s difficult to see to which colour block a label belongs
- Link slices to the labels in the legend: There are too many colours to relate the labels to the slices (i.e. too many slices) in the pie chart
- Ordering of the legend and the pie chart
    - Suggestion: P2: “If you ordered the legend by frequency and then ordered the pie chart clock wise also by frequency, then it’s a lot easier for people to match them up”
- Linking between the pie chart and the summary: It’s currently not clear that there is a relationship between the pie chart and summary
    - Suggestion: colour code the summary sentences (e.g. by colouring their bullet points) to match the colour in the pie chart. Spread the summary sentences a bit more.
- Comments in the summary are removed from their original conversational context

and some comments are completely useless without the context.

- The user (commenters') names before comments are links and link to commenter profiles, which is not useful. The top-level comment in the summary hasn't got a user name in front of it.
    - Suggestion: It would be more useful if the commenters' names linked back to the conversation, e.g. in Guardian interface. Remove the commenter completely, include the link from the summary to the original thread for context.
- Automatic clustering/cluster labelling:
    - P3: "I'd say the biggest thing for me about the SENSEI system was that the clusters made just absolutely no sense to me at all… "
    - P3: "If I had to summarise what one of those clusters was about, I couldn't tell you."
    - P2:"These labels are a dump of key words."
    - Suggestion: Use Wikipedia titles instead of key words for better cluster labelling. Link pie chart to the summary and skip the labels completely.
- Linking between the clusters and the article:
    - P1: "…even if those labels are accurate, they just have nothing to do with the story."
- Too many clusters
    - Suggestion: reduce the number of clusters to about 5 to solve the problem of the pie chart, colour matching and reduce the number of labels.
    - Suggestion: Present participants only with the article and pie chart and ask to identify issues to test the usefulness of the pie chart.
- Snippets are not user friendly
    - They don't make sense, just cut off the sentence.
    - Suggestions: Expand should be on the left, before the sentence, so it's aligned in all summary sentences. More space between summary sentences needed. Align the text centrally, not till the end of the column. Do not force the user to expand each single comment. All comments should be visible when expanded. Pop-up as a solution – not good for mobile systems.
- Summary has repeat comments, expanding summary sentences does not show the conversation as expected, but the comments from the cluster, which is confusing.

User feedback on strategies they used to answer the questions

When given the statement "From this, I take that you couldn't use the pie chart to either identify issues or characterise opinions", the participants did not confirm, but claimed that depending on the article and the contents of the pie chart it was useful in identifying issues, although it had the problems of having too many labels as described above.

When using the SENSEI+Guardian interface, P1 and P4 concurred that their main strategy for identifying the issues was to read the comments in the Guardian interface. P2 only used the SENSEI interface, as he/she was not aware that the comments were accessible through

the Guardian interface and therefore did not use it. P3 used the SENSEI interface and relied on her memory using the Guardian component in the reading session.

Expanding snippets in the summary was not helpful for characterising opinions as it was not clear whether all opinions on an issue would be in the cluster because the conversational context is lost.

Key word search was used only by P1 to characterise opinions. P4 didn't find it useful due to the variability of the language in the comments. P3 found there was no need to search.

The summary was more preferred and used than pie chart in particular for characterising opinions.

Is the SENSEI system in the current version adding value over Guardian?

4 out of 4 participants said "No", but if the issues discussed were solved they think it would be.

3.3.5. Concluding discussion

As stated in Section 2.2.3 above, we had two aims in carrying out this interim extrinsic evaluation of the SENSEI social media prototype. The first was to assess how well the SENSEI prototype can help users carry out a real world user task, with a view to gathering insights to inform future technology development. The second was to test run the evaluation methodology – which is novel – with a view to assessing its feasibility and utility and understanding how we might refine it for use in the Year 3 extrinsic evaluation. Both of these aims have been met.

Assessment of System Prototype

The overall assessment of the participants regarding the SENSEI prototype can be perhaps be simply summarised as "Nice idea, but it's not working well enough yet for us to prefer it over the Guardian interface alone". This is not surprising as (a) the system they evaluated was just an initial prototype and (b) several features we had planned to include had not been completed due to technical difficulties.

Participants' criticisms of the prototype can be loosely grouped into two categories: criticisms of the interface and criticisms of the underlying language technologies. The interface problems included:

- number of groups in a pie chart needs to be limited to ensure readability;
- the expand comment button should be moved to before a summary sentence;
- it should be possible to click through from "slices" in the pie chart to the cluster of comments that the slice represents;
- it should be possible to click through from a comment in a SENSEI-generated cluster to that same comment in full context of comment stream in the standard threaded-comment interface.

These problems can themselves be sub-divided into two groups: those like the last two, which identify functionality we were in the process of implementing but had not yet

completed and those like the first two, which identify issues we were not previously aware of. Thus, the evaluation served both to confirm the value some aspects of planned interface development and identify new areas of modification that require our attention. Work addressing both these sorts of interface issues is underway.

Criticisms of the underlying language technologies are harder to be sure about, partly because the language technology outputs (clusters, cluster labels, summary sentences) are always being viewed through the interface, which has its own issues as discussed above, partly because the outputs may not be obviously "wrong" but just not as useful as one might hope they could be and without a clearly better point of comparison it is hard to be certain about these inadequacies. For example, one user commented on the difficulty in interpreting the clusters (P3: "I'd say the biggest thing for me about the SENSEI system was that the clusters made just absolutely no sense to me at all… "). Perhaps if the interface had supported click-through to the original comment stream P3 might have been able to make more sense of the clusters; or perhaps the clusters did simply contain comments that did not really belong together, making interpretation difficult. Overall, from the participant experience reports and the group discussion, it seems clear that participants found the clusters hard to interpret and hence of limited utility and found the summary as an overview of the conversation to be of limited use for either task, but particularly for the "characterise opinion" task. To address these issues we are currently:

- improving the clustering algorithms by incorporating more features;
- replacing the simplistic cluster labeling approach with a more sophisticated approach;
- about to carry out another evaluation (October, 2015) using the same protocol but where the system outputs are replaced by gold standard outputs (summaries and comment clusters) and the system has been improved by addressing many of the interface issues and enhancing the language technologies system as discussed here; this should give us insight into how effective a user finds "perfect" language technology outputs embedded in a fully functioning, refined UI to be for carrying out the evaluation task.

Further discussion of planned developments in the language processing technologies can be found in D5.2.

The evaluation protocol and tasks

The protocol and tasks were carried out without significant difficulties and yielded rich and informative data for comparing systems. The new metrics worked well and the different results, i.e. quantitative assessment of task performance, qualitative assessment of user experience and focused group discussion, complement each other very nicely.

The biggest change required for the final Y3 evaluation is to include more participants and topics. Doing so will us to better factor out individual participant and topic differences. With just four participants and two topics it is hard, arguably impossible, to gain reliable insights into the superiority of one system over another on two tasks, especially as the amount a single participant can be asked to do is limited. Of course larger scale evaluations are harder to organise and analyse and more costly to run. Nonetheless, while the current evaluation

has been invaluable in terms of the system and technology insights it has given us and the assurance it has given us regarding the evaluation protocol and tasks, we plan a larger scale evaluation for the final stages of the project to give us more reliable results.

Aside from this major change there are a number of minor changes we will consider in future too. These include:

- reducing the reading time of the article to 2 minutes to encourage people to skim;
- reducing overall reading time of the comments to 3 minutes;
- further emphasising what we mean by “issue” in the task instructions, perhaps explaining in more detail how the idea is based on the “comment overview” scenario, where the aim is to identify what people are talking about based on numeric strength of comment, as opposed to being based on perceived social significance or newsworthiness (a different use case);
- possibly capturing additional participant-system interaction data, e.g. using screen recording to log interactions – this would allow us to investigate in detail questions such as from where in a comment set answers are obtained.
- possibly recording task completion times;
- possibly factoring the system evaluation into separate sub-evaluations to allow us to independently assess separate system components, such as the pie-chart, the clusters, the summary etc. (this would give better insight into the utility of each of these components individually, but at the cost of much greater effort and with the risk of losing the supporting synergies between the different components).

Just which of these modifications to the evaluation protocol we proceed with in the coming period will be depend upon (a) the results and analysis of the evaluation exercise mentioned above that we are currently carrying out with gold standard data and an enhanced system and (b) striking a balance between how much effort goes into evaluation versus system development in the coming period.

4. Conclusions and further work

In the D1.3 SENSEI deliverable we have been focused on the *extrinsic* evaluation of the SENSEI prototype. We have designed the evaluation scenarios by including tasks based on activities typically carried out by the SENSEI potential users. As far as it was possible we recruited users with real experience in the tasks selected for the evaluation: for the speech use cases the users were professionals of a call centre company, for the social media use case we could not recruit “real” users because of the novelty of the tasks, but all the recruited graduates had experience of using online news and reader comments.

The prototype we have been evaluating generates different types of summaries, including short summaries of call centre calls (synopses), summaries of reader comments, and filled questionnaires used to summarise some aspects of call centre agents’ communication behaviour when they interact with their customers.

Several experiments have been implemented. The evaluation scenario for the speech prototype included the preliminary evaluation of the reliability and accuracy of the automatically generated questionnaires and the collection of insights and feedback based on users’ experience in comparison tasks, with and without the contribution of the SENSEI prototype. The evaluation task designed and implemented for the social media scenario was based on assessing the quality of user outputs and gathering the experience of the users after having performed a set of tasks, with and without the contribution of the SENSEI prototype.

In general the results of the evaluation supported the hypothesis that the evaluation protocol and tasks are realistic, potentially accepted by the users, and feasible on a larger scale. Conversation oriented summaries at very low compression rate (7% of the original conversation) have been judged useful by the users. Those summaries and the system generated ACOFs may potentially reduce the decision-making time in tasks of call centre QA supervising.

Some possible improvements of the experimental settings have been identified. They are mainly about the prototype interface and they will be presented to the technical SENSEI workpackages. As for the prototype system, from the participants’ comments we could appreciate both their preferences for some types of summaries among the ones proposed, and useful insights for improving the underlying technologies, whose critical aspect seems to be related with the accuracy requirement, that was identified in the speech scenario as an essential requirement for the users’ acceptance of this technology.

4.1. Planned activities for WP1 in Period 3

The D1.3 SENSEI deliverable reports about the activities covered in two WP1 tasks: T1.2, focused on baseline parameters, evaluation tasks and metrics, and T1.3, devoted to the incremental evaluation of the SENSEI prototype. While the first task will be completed at the end of the second contractual year of the project, T1.3 will continue until the end of Period 3.

In Period 3, WP1 tasks will be focused on the assessment of the intermediate evaluation results. We have now obtained behavioural data about the design of the use cases, and more focused user requirements. We plan to implement the following main lines of activities:

- to refine the evaluation tasks by setting up subtasks in order to allow the independent assessment of different system components;
- to review the annotation guidelines both for ACOF and synopses writing, with the aims of providing more focus on failure cases (ACOF), and improving the informativeness of synopses;
- to run evaluation tasks on a larger scale in terms of number of recruited participants;
- to refine the quantitative and qualitative predictors of system performance for the speech use cases, and possibly identify ROIs related with the system tasks;
- to improve the interface and data collection features of the experimental prototype, in particular evaluate the opportunity of introducing recording of the time for the tasks completion;
- to validate the final prototype performance.

References

[Fleiss & Cohen, 1973] Fleiss, J. L., & Cohen, J., “The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability”, *Educational and psychological measurement,* 1973

[Fleiss, 1981] Fleiss, J. L. "The measurement of interrater agreement", *Statistical methods for rates and proportions*, Volume 2: 212-236.

[Lloret & Palomar, 2012] Lloret, E. & M. Palomar. "Text summarisation in progress: a literature review", *Artificial Intelligence Review*, 2012 (37.1): 1-41.

[Onwuegbuzie *et al.* 2009] Onwuegbuzie, A.J., Dickinson, W.B., Leech, N.L, and A.G. Zoran, "A qualitative framework for collecting and analyzing data in focus group research." *International Journal of Qualitative Methods*, 2009, 8(3): 1-21.

[Radev, Howy & McKeown, 2002] Radev, D. R., Hovy E., and K. McKeown,. "Introduction to the special issue on summarization.", *Computational Linguistics,* 2002, 28(4): 399-408.

[Weir 2005] Weir, J. P. “Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM", *The Journal of Strength & Conditioning Research*, 2005, 19(1): 231-240.

[Zafarani & Liu 2015] Zafarani, R. and Liu, H. “Evaluation without ground truth in social media”, *Communication of the ACM,* 2015, 58(6): 54-60.

[Zlatintsi *et al*., 2015] Zlatintsi, A., Koutras, P., Efthymiou, N,, Maragos, P., Potamianos, A., and K. Pastra, “Quality evaluation of computational models for movie summarization", 2015, *Proc. QoMEX*.

Appendix 1: Questionnaire 1 (Participant's Background)

1. Please describe your English language proficiency:
   - ☐ Native speaker
   - ☐ Near native / fluent
   - ☐ Very good command / highly proficient
   - ☐ Good command / good working knowledge
   - ☐ Basic communication skills / working knowledge

2. Which of the following best describes your current role:
   - ☐ News media professionals
   - ☐ News and comment reader – I read news and/or comments but very rarely provide/post comments
   - ☐ Comment provider – I read news and/or comments and provide/post comments on a regular basis

3. Please tell us how often you engage with (e.g. read or skim) any on-line news web sites, e.g. The Guardian, BBC, The Independent, Mirror, etc.?
   - ☐ At least once a day
   - ☐ At least once a week
   - ☐ At least once a month
   - ☐ Very rarely (i.e. more than one month intervals between visits)
   - ☐ Never

4. How often do you engage with (i.e. read or post to) the reader comments in on-line news web-sites? *(Please select the option that best describes your experience.)*
   - ☐ At least once a day
   - ☐ At least once a week
   - ☐ At least once a month
   - ☐ Very rarely (i.e. more than one month intervals between visits)
   - ☐ Never

Appendix 2: Questionnaire 2 (Content Questions)

Content Questions for Extrinsic Evaluation of THM-Style Summarization Components

You will be given **two questions** related to the discussion in the comments. These questions will require you to **identify issues** in the comments and to **characterise opinion** on an issue.

Identifying Issues

The first question will ask you to identify **four** of the main issues, other than the one addressed in the second question, that have been discussed in the comment set.

By an **issue** we mean something that one can believe or disbelieve/agree or disagree with. For example “Climate change is directly caused by human activity”, “The US Senate should vote in favour of the Iran nuclear deal”, ”Wheelie bins attract vermin”. Issues do not have to be significant; i.e. they need not be topics attracting front page attention in the national press. They can also be simple statements about everyday issues such as “bus travel is free for the over 60’s”. “Hard floors are better than carpets”.

The key things about our use of the term “issue” are that:
a) In the context of a comment set, issues are things that **multiple commenters discuss**, perhaps assert, deny, clarify, expand upon, qualify, consider the consequences of, etc.

b) While issues may sometimes be expressed telegraphically by nouns or short phrases (e.g. “immigration”, “climate change”) these are to be understood as short forms for a statement that one can believe or disbelieve, i.e. take a position on. Ideally we would like you to identify and describe an issue in the long form (e.g. Climate change is directly caused by human activity”). An issue may also be thought of as something that can be expressed by a “whether or not” phrase, to indicate opposing views in the comments e.g.: “whether or not climate change is directly caused by human activity”.

To **identify** an issue you should supply a statement that you believe best expresses the issue, e.g. “If Jeremy Corbyn becomes leader of the Labour party this will be a disaster for the party”.

Supplying a noun phrase like “Jeremy Corbyn” is not sufficient (as there may well be multiple issues in a comment set in which Jeremy Corbyn figures). A response like “Jeremy Corbyn -- disaster” is better, but still not optimal. Likewise, “Wheelie bins attract vermin” is better than “Wheelie bins and vermin”.
Your responses will be graded on a multi-point scale so that we can distinguish degrees of clarity and correctness in stating an issue.

“Main issues”

Question 1 asks you to "identify **four** of the **main** issues that have been discussed**".** By **main issue** we mean that the issue received a fairly significant amount of discussion relative to other issues in the comment set. We do not expect you to count comments relating to each issue and then accurately rank issues by number of commenters, but we wish you to get an intuitive sense, over the 100 comments you are asked to consider, of what are the issues attracting the most attention and to identify these.

Note that not all comments relating to an issue need mention the issue explicitly. The issue may be explicit in the context or previous comment and a new comment may clearly, but indirectly, address it. For example, in the context of the issue "less frequent bin collection will lead to an increase in vermin" a commenter might assert that "many people use compost bins without attracting vermin". Clearly this comment addresses the issue yet it does not explicitly mention the issue. In assessing the "main-ness" of an issue in a comment set such related comments should be taken as part of the mass of comment "on" the issue.

Note also that threads and issues are not the same thing: comments addressing any one issue may occur in multiple threads and any one thread may contain comments relating to multiple issues. A comment replying to another need not be on the same issue and new issues can emerge as the discussion drifts into new areas. Use your intuition in determining what feels like a new, distinct issue in the overall discussion.

Characterising Opinion

In the second question you will be given a specific issue that has attracted discussion in the comments and we would like you to **characterise opinion** on that issue. Typically characterising opinion involves describing:

- approximately how many people were involved in this discussion in relation to the overall size of the comment set; i.e. was this issue the major focus of discussion across the comments? or did, e.g. only a handful of people address it?
- what views or perspectives did they take with respect to the issue?
- was there consensus? was opinion divided? how was it distributed? (e.g. many in one camp with a few dissenters, versus evenly split, etc.)
- whether the discussion was particularly emotional/heated and if so over what.

Note you need not address every point on the list for every issue nor need you only mention things on the list. **You should be guided by your intuition as to what is appropriate in characterising discussion on the issue** -- how you might sum up the discussion to someone who had not read the comments.

To help you understand what we mean by "characterise opinion" we have provided i) an example question and response and ii) some helpful phrases:

i) **Example question:** *Characterise opinion on what kind of new houses we need to build in the UK.*

Example response:

Around half of the comments discussed what kind of new houses we need to build in the Uk. Opinion was divided. A few believed that new housing in the UK was generally too small in scale and we need to build bigger houses. Many thought that small houses and

apartment style living was a good idea. Many said most people can't afford large houses. There was agreement that new houses need to be affordable. A few noted that apartments are a way of addressing land shortages.

ii) Useful Phrases
In characterising what proportion of the total comment was concerned with the issue you may find it helpful to use phrases such as the following:

"more than half the comments said ..."; "a third of the comments said ..."; "roughly 50% of comments were concerned with ..."; etc.

In characterising opinion on the issue you may find it helpful to use phrases such as the following:

"some commenters said this, others said that"; "many said this; few said that"; '"they were agreed on this"; "they were divided on that"; "the majority said this"; or "an exception said that"; etc.

In addition you may quote directly from the comments if you wish to elaborate or illustrate a point.

For the Question Sheet:

Task 1: Network Rail (article 19)

1. Identify four issues, other than the issue in question 2 that were main topics of discussion in this comment set.
2. Characterise opinion on the issue of whether the proposal to spend the fine imposed on Network Rail on improved wifi services on trains was a good idea.

Task 2: Heatwave (article 3)

1. Identify four issues, other than the issue in question 2 that were main topics of discussion in this comment set.
2. Air conditioning is better than fans to keep cool in a heatwave.

Appendix 3: Questionnaire 3 (Post-task experience)

We would appreciate you feedback on using the different systems and system components today.

1. Please indicate on a scale of 1-5 (1=not useful and 5=extremely useful) how useful the different systems/system components were when completing the experimental tasks.

We invite feedback on the 2 different systems in turn.

Please base your judgement on your experience of each component in the context of the specified system – e.g. in Assessing System 1, "The Threads", you should base your answer on the usefulness of "threads" in the context of "Guardian Comment" only (and not when SENSEI was available).

You may elaborate on your scores in the box "Any further comment?"

1. Assessing System 1: Guardian Comment Facility Only

Feature/System	**"Identify 4 Issues"**	**"Characterise Opinion"**	**Any further comment?**
The threads	1 2 3 4 5	1 2 3 4 5	
The keyword "search in page" via the browser	1 2 3 4 5	1 2 3 4 5	
System 1 (The Guardian system), as a whole	1 2 3 4 5	1 2 3 4 5	

Note: 1 = not useful and 5 = extremely useful

2. Assessing System 2: SENSEI + Guardian Comments Together

Feature/System	"Identify 4 Issues"					"Characterise Opinion"					Any further comment?
The summary – as an overview	1	2	3	4	5	1	2	3	4	5	
The summary – as an index to clusters	1	2	3	4	5	1	2	3	4	5	
The pie chart – as an overview	1	2	3	4	5	1	2	3	4	5	
The clusters	1	2	3	4	5	1	2	3	4	5	
The SENSEI Component (not including the Guardian component)	1	2	3	4	5	1	2	3	4	5	
The Guardian component (not including the SENSEI component)	1	2	3	4	5	1	2	3	4	5	
System 2 (SENSEI + Guardian component), as a whole	1	2	3	4	5	1	2	3	4	5	

Note: 1 = not useful and 5 = extremely useful

2. Please indicate on a scale of 1-5 the extent you would like to have System 2 (SENSEI + Guardian) available for use in a comment facility, when browsing news and comment

Would not like to have – I would never use it 1 2 3 4 5 *Would really like to – I would use it often*

3. Please provide any other comments/feedback you have about your experience using the systems to carry out the tasks in the experiment today. Was there anything you really liked or disliked?

You may also wish to mention any possible improvements or things you would like to see included in a system.

Appendix 4: Questionnaire 4 (Group Discussion Questions)

This is an example of a **"semi-structured" questionnaire** (a common data-gathering tool in qualitative research). There are 3 high level questions and a list of things to cover or "sub-questions" that ideally we would like people to comment on in the discussion.

The researcher leading the discussion goes through the questions inviting responses from the participants. The researcher should allow the discussion to flow freely, allowing the participants to say as much as they want to, and allowing the conversation to drift into new topics if they occur.

The researcher should bring the discussion back to the questions either:

1. when the discussion pauses or seems to reach a natural break; and/or,
2. to keep to the time limit/ensure there is an opportunity for all questions to be addressed within the time available (The time limit we eventually decide on should be sufficient to allow for open discussion on these issues).

Please note:

- The researcher should encourage discussion using the prompts/sub-questions if the issues have not already emerged in the discussion--He/she does not need to ask the question if he/she feels the topic has been sufficiently addressed by participants.
- The order in which these issues are addressed is not important but it is helpful to check they have been addressed.

Introduction: We would really appreciate your feedback on two aspects of the experiments today, **the tasks** and **the systems**. We'll be asking you a few questions and would like you to respond as freely as possible.

We note that this discussion will be recorded.

(start recording).

1. **Tasks**
   Question 1 : How did you find, i.e. get on with, the 2 content-related questions?

   "Finding/Identifying issues" and "Characterise Opinion"

   It might be helpful to invite feedback on "Finding issues" first and then if necessary to invite feedback on "Characterise Opinion".

   Things to cover/prompts/sub-questions:

1.1 Were the questions **easy** or **difficult?**
1.2 **Time:** was there sufficient time for these tasks?
1.3 Did you find it helpful to have time to **read comments /article** beforehand; was the **reading time** too long/too short?
1.4 How did you find the questions in relation to the **different topics**?

(Network rail and Heatwave)

2. **Systems**
Question 2: We'd like to hear more about your experience using the different systems in this experiment. Do you have anything to say about this?

- Guardian only; Sensei +Guardian

Encourage feedback on both systems.

Things to cover/prompts/sub-questions:

2.1 What did you **like** about the different systems?
(try and get feedback on both systems)

2.2 What did you **dislike** about the different systems?
(*try and get feedback on both systems*)

2.3 Could you tell us something about the ***strategies*** you used to **complete the tasks** in the **different system conditions**. (i.e. what functionality you made good use of, and how)

Cover the 2 tasks:

- What did you use/how did you "**Identify issues"?**
- What did you use/how did you **"Characterise opinion"**?
- Were your strategies different for the different question types ("identify issues"/"characterise opinion")?

*--You may find it helpful to illustrate with an **example**:*

"To identify issues in the Guardian only, I scanned comments, thought up possible search terms, and used a key word search to step through comments possibly related comments"

"To identify issues in Sensei I used the pie chart to explore the clusters"

2.4 Did anyone use the key word search in the Guardian only condition: did it help?
2.5 Did you have a **preference** for the **pie chart** or the **summary**—give reasons?
2.6 How easy did you find it to read through comments in a cluster?
2.7 We snippetised comments –did you find this helpful?
2.8 Did you find Sensei useful/or "added value" for completing the tasks?

3. Any further comment

Question 3. Is there anything else you would like to tell us about the experiment today?

Final: A big thank you to all for taking part today ☺

Appendix 5: Researchers' Script

The aim of this document is to give an outline of different parts of the experiment; instructions to participants; a record of times for different parts of the experiment.

Before the session formally begins: welcome participants; check they have read the participant info sheet and complete signing/collect consent forms.

The Session is in two parts: "Preparation" and "Timed tasks/ questions".

Part I -Preparation

1. General Introduction (5-10 mins)

1.1 **Introduction to Sensei**

- High level project aims; the team at Sheffield; researchers present today.
- Why participants are here: to take part in a "user" experiment to evaluate SENSEI summarization technologies in a realistic task setting.
- Outline experiment aims:
  - to see if our technologies can help people make sense of reader comment and whether SENSEI technologies can help people to complete tasks better/more easily, than using current technologies alone.
  - to obtain feedback on SENSEI technologies in a task setting, which may inform future technology development;
  - to inform the design of a large scale SENSEI evaluation with users, due to take place next year.

1.2 Overview of the Evaluation Session

The session today will involve:

Part I-Preparation

1. **General introduction**: session outline, pre-task questions. **(5-10mins)**
2. **Demo of the different system technologies and a short practice session (10 mins).** You will be using two systems today, The Guardian reader comment facility and a Sensei system. We will go through the basics of both so you feel happy using the systems.
3. **Introduction to scenario and related tasks/questions (10-15 mins)**: we will go through the task scenario, and tasks/questions that you will be asked to complete, so you are familiar with the tasks and clear on how to answer the questions.

Part II-Timed tasks and questions

4. **Time limited tasks (45 mins)**:

This is the main part of the experiment. People will be asked to complete a number of time limited tasks as follows:

- reading a news story (5 mins)
- reading associated comments (5mins)
- answering 2 questions relating to the comments (10 mins).

There will be two 2 different news stories and comment sets to explore, each using a different system. So in total there will be 3 timed tasks, a short 5 min break and then a repeat of the 3 tasks but based on a different story/system combination. This is all on the participants' handout, but we will remind you of this again before you commence the timed tasks.

5. **Post –task questions.** We will then ask you to answer some short questions about your experience using the technologies. This will not take more than **5 minutes**.
6. **Group discussion:** Finally we invite you to answer some general questions about the tasks and systems used today in a short (group) discussion.

--Any questions on the overall session today?

1.3 **Inform participants about question sheets and introduce question sheet 1**

We'll be handing out 4 question sheets in total. We will hand these out throughout the session and collect as each is filled in.

Before we begin with the demo we'd like you to complete **question sheet 1.**

Hand out question sheet 1 (pre-task questions on basic experience of news and comment)

Invite people to complete question sheet 1.

Collect question sheet 1

2. System Demo and Practice Session (10 mins)

2.1 Demo. (5mins)

The aim here is to demo the key functionality in each system.

(We do not mention the experimental tasks at this point).

System 1--Guardian Comments: the article, comments, threads, expand/collapse threads, sort options; point out the option of a "search in screen" for finding key words in the comments.

System 2—Sensei System and Guardian comments. Summary; clusters (accessed via summary); pie chart (representing clusters:—size of pie =size of cluster; cluster labels); pie-chart click through to clusters. Etc.

- Note that the underlying clusters are the same as those accessed via the summary.
- Demo how to expand snippetised comments to full comments.
- Demo how to link to comment in Guardian comment stream (if working).

2.2 Practice session. (**5 mins)**. Invite people to spend some time using the different systems on a demo article and comment set. The aim here is to ensure people feel comfortable and confident using the different systems.

(Again—we do not mention the experimental tasks at this point).

--Remember to ask: any questions on using the systems?

3. Introduction to the scenario, and related tasks /questions

(10-15 mins):

The aim here is to ensure people are clear on the scenario and what they have to do in the time limited tasks and to instruct people on how to answer the content related questions before they commence the timed tasks.

3.1 Go through the scenario with participants.

Explain how the tasks people will be doing are based on the following scenario:

> **Scenario:** Imagine you are a general reader of on-line news and comment. You have a short period of time available (e.g. a coffee break) in which to read some news and associated comment. Ideally you wish to get an overview of the commenters' response and opinion. However with limited time available you would be happy to:
>
> i. Identify the main issues addressed in the comments -- what were the commenters talking about?
> ii. Gain a sense of the spread of opinion on a particular issue -- i.e.
>
> - What were the different perspectives and opinion on the issue
> - Areas of consensus and disagreement
> - The feeling expressed

3.2 Introduction and guidance on the tasks

There are 3 time limited tasks.

Go through each task in turn:

- First: **reading the article** (without the comment)
  (5 mins allowed for reading)

- Second: we provide the comment, and the task is to quickly **read and make sense of the comment** as best as possible, **as preparation for the content related questions**. (5mins)
  - Encourage them to skim through the full set of comment and not just the first few comments.
  - Note: we ask them to read and make sense of the comments "in their heads", without writing anything down.

- Third: **answer the content related questions**. (10 mins allowed for content questions)

  We'll give these out on a question sheet.

Content related questions:

Based on the description in the instruction sheet, go through the format of the content related questions and how to provide an answer:

Question 1: "Identify 5 Issues"

Question 2: "Given an issue, Characterise Opinion"

Final Note: you may find these tasks quite hard to complete. Please don't worry if this is the case. The main thing is that they should encourage people to engage with the systems in a focussed way, such that we can gather feedback on the experience in the final questions.

--Any questions before we begin Part II ?

Offer people a Break (5mins);

Part II- Timed tasks and questions

4. Timed tasks.

With the scenario in mind we now invite people to complete the short, timed reading tasks and the questions related to comment content.

This sequence will be carried out twice with a short break in between.

Remind participants:

- You will not receive the questions until you have completed the reading tasks. i.e. you will only be asked to answer the questions once the reading time is complete.
- We will let you know when it is time to complete or begin a task.

Ensure timing of tasks:

Topic 1

Open and read the news article (5 mins max)
Open and read comments (5 mins max)
Hand out question sheet 2
Answer content related questions (10 mins max)
Collect question sheet 2

break and refreshments (5mins)

Topic 2

Open and read the news article (5 mins max)
Open and read the comment (5 mins max)
Hand out question sheet 3
Answer content related questions (10 mins max)
Collect question sheet 3

5. **Experience Questions (5mins)**
Hand out Question sheet 4
Questionnaires complete.
Collect in sheet 4.

6. **Group Discussion. (5-10 mins)**

Remind people this session will be recorded.

Start recording.

Lead/guide discussion via questions but allow people to talk freely if relevant topics emerge.

7. Collecting Questionnaires.

Note: be sure we have collected in and stapled together the 4 question sheets for each participant before they leave the session.

Total 95 mins (based on max times suggested above). Note Ethics application says we will NOT exceed 120 mins (2 hours).

Appendix 6: Results of Questionnaire 2

Article: "Network Rail" (System: The Guardian)

Participant 1

Q1. Identify four issues, other than the issue in question 2, that were main topics of discussion in this comment set.

1. Ticket prices are too high.
2. The disconnect between Network Rail and the train operating companies causes inefficiencies.
3. The rail network should be re-nationalised.
4. Fines are a contentious issue – who will benefit if anyone?

Q2. Characterise opinion on the issue of whether the proposal to spend the fine imposed on Network Rail on improved Wi-Fi services on trains was a good idea.

Participant 4

Q1. Identify four issues, other than the issue in question 2, that were main topics of discussion in this comment set.

1. Fining NR just means less money to fix problems causing the delays
2. NR is not responsible for the price of tickets
3. Should the money to fix the network only come from train tickets not general taxation
4. Proposed rail improvements seem to always be focused around London and the south to the *detriment(?)* of the rest of the country (i.e. we plan to spend more money on improvements to the area already with the best delay numbers)

Q2. Characterise opinion on the issue of whether the proposal to spend the fine imposed on Network Rail on improved Wi-Fi services on trains was a good idea.

The issue attacked some comments, but probably less than 20% of the total. I couldn't find a single, non-sarcastic, comment that thought it was a good idea. Pretty much every comment on the issue suggested using the money to improve current network services either to improve punctuality or to add more trains to overcrowded routes.

Article: "Network Rail" (System: SENSEI + The Guardian)

Participant 2

Q1. Identify four issues, other than the issue in question 2, that were main topics of discussion in this comment set.

1. Why should my taxes pay for rail improvements when I don't even use trains?
2. Naturalisation doesn't improve anything.
3. Maybe the money from the fine would be better spent on something more useful - not wifi that few people use.
4. -

Q2. Characterise opinion on the issue of whether the proposal to spend the fine imposed on Network Rail on improved Wi-Fi services on trains was a good idea.

One comment hardly indicates a consensus. The single commenter thinks it would be better spent improving the service itself.

Participant 3

Q1. Identify four issues, other than the issue in question 2, that were main topics of discussion in this comment set.

1. Whether it's worth spending the money on Wi-Fi since lots of people don't use it on trains anyway or even want to use it
2. Whether punctuality is more important than ticket price or not
3. The issue of whether people who don't use trains (much or at all) should have some of their tax go towards funding these things
4. The issue that (like many other things) the problem is mainly focused on London and the South East

Q2. Characterise opinion on the issue of whether the proposal to spend the fine imposed on Network Rail on improved Wi-Fi services on trains was a good idea.

The fast majority of people seem to think it is a bad idea, for a number of reasons – as given in the 4 issues discussed in question 1. There was some defence against the arguments that it was a bad idea, but the defence was mainly in the form of explaining why the arguments were not valid, rather than providing any real reasons why the fines were a good idea. No one really came out in support of spending money on Wi-Fi.

Article: "Heatwave" (System: The Guardian)

Participant 2

Q1. Identify four issues, other than the issue in question 2, that were main topics of discussion in this comment set

1. The hot weather is uncomfortable, especially for native Brits; others argue it's fine – summer is meant to be hot.
2. This just sensationalism by the media, as usual – nothing to worry about, summer is hot, winter is cold (shock! Horror!)
3. This is mainly London / the South, who cares – what about the rest of the country?
4. Good weather? – Let's go to the pub!

Q2. Characterise opinion on the issue of whether air conditioning is better than fans to keep cool in a heatwave.

-

Participant 3

Q1. Identify four issues, other than the issue in question 2, that were main topics of discussion in this comment set

1. How to use a fan to get maximum benefit from it
2. The fact that there is no real news in this story and it's making sensationalism out of nothing – including the fact that the advice is all pretty obvious
3. People who like hot weather versus people who don't and the one-upmanship of people who live / have visited hot countries and survived, especially without air-con etc.
4. Ways to keep cool in the heat other than fans and AC, e.g. making sheets damp before sleeping.

Q2. Characterise opinion on the issue of whether air conditioning is better than fans to keep cool in a heatwave.

Opinion was pretty much divided on this topic, but the real discussion boiled down to whether people liked or hated air con (also divided) and how best to use a fan (in swing rather than stationary, only when you're in the room).

Article: "Heatwave" (System: SENSEI + The Guardian)

Participant 1

Q1. Identify four issues, other than the issue in question 2, that were main topics of discussion in this comment set.

1. There is a north/south divide at the issue of a heatwave – is it happening or not?
2. Health issues connected to the weather
3. Weather issues are over-estimated or reported
4. People have different preferences for dealing with hot weather

Q2. Characterise opinion on the issue of whether air conditioning is better than fans to keep cool in a heatwave.

Can't find much on *this(?)*

Several people would like air con in heatwave conditions. There is divided opinion over whether air con is needed in the UK, as not hot for extended periods. Some people feel that air con is a waste of money and impacts the environment. One person highlights additional health problems caused by fans. There is no real consensus of whether to use fans / air con / or nothing at all.

Participant 4

Q1. Identify four issues, other than the issue in question 2, that were main topics of discussion in this comment set

1. Is 30 degrees really a heatwave, or is it just a summer temperature
2. Leaving the windows open means more insects, particularly the dreaded Scottish midge, getting into the house
3. Does sleeping under a wet sheet help you get more sleep when it's hot
4. Reporting of whether events seem to only focus on London and the South and ignores the rest of the country

Q2. Characterise opinion on the issue of whether air conditioning is better than fans to keep cool in a heatwave.

About 30% of comments seemed to be about AC versus a fan. Opinion on which was best seemed to be split 50/50. The major opinion seemed to be to go with fans in the UK as

cheaper than AC and easier in rented accommodation. AC seemed to be preferred in countries where temperatures were consistently higher.