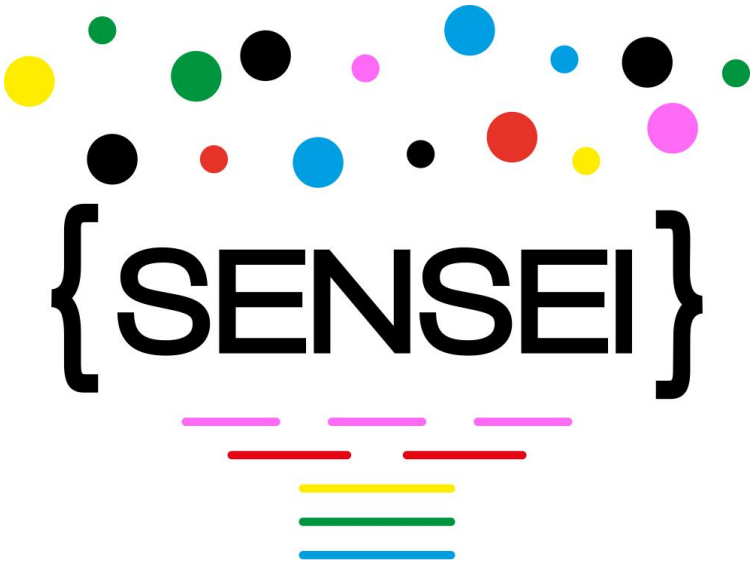


D1.4 – Final Report of Prototype Evaluation

Document Number	D1.4
Document Title	Final Report of Prototype Evaluation
Version	1.0
Status	Final
Work Package	WP1
Deliverable Type	Report
Contractual Date of Delivery	31.10.2016
Actual Date of Delivery	31.10.2016
Responsible Unit	UNITN
Keyword List	speech summary evaluation, social media summary evaluation, evaluation metrics, extrinsic evaluation, insight-oriented evaluation, user acceptance
Dissemination level	PU

Editors

Morena Danieli (University of Trento, UNITN)

Emma Barker (University of Sheffield, USFD)

Contributors

Frederic Bechet (Université Aix-Marseille, AMU)

Cosima Caramia (Teleperformance, TP)

Fabio Celli (University of Trento, UNITN)

Benoit Favre (Université Aix-Marseille, AMU)

Carmelo Ferrante (University of Trento, UNITN)

Robert Gaizauskas (University of Sheffield, USFD)

Letizia Molinari (Teleperformance, TP)

Adele Palumbo (Teleperformance, TP)

Monica Paramita (University of Sheffield, USFD)

Giuseppe Riccardi (University of Trento, UNITN)

Vincenzo Lanzolla (Teleperformance, 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	25/07/2016	Draft	Morena Danieli (UNITN)	Table of Content
0.2	03/08/2016	Draft	Morena Danieli (UNITN) Emma Barker, Rob Gaizauskas (USFD)	Revision and Final Table of Content
0.3	30/08/2016	Draft	Morena Danieli (UNITN) Letizia Molinari (TP)	Description of the evaluation scenarios, methodology – added a subsection for training
0.4	31/08/2016	Draft	Morena Danieli (UNITN) Letizia Molinari (TP)	Added sections 2.4.1 and 2.4.2
0.5	08/09/2016	Draft	Emma Barker (USFD) Monica Paramita (USFD) Morena Danieli (UNITN)	Section 3- 3.4 added- full draft. Appendices added. Full draft of speech evaluation sections (with the exception of questionnaire section)
0.6	09/09/2016	Draft	Morena Danieli (UNITN)	Speech Extrinsic Evaluation - Questionnaire results added
0.7	14/09/2016	Draft	Vincenzo Lanzolla (TP) Morena Danieli (UniTN)	Speech Evaluation Prototype section added
0.8	14/09/2016	Draft	Monica Paramita (USFD) Benoit Favre (AMU)	Section 3.5 added: Quantitative results of the social media extrinsic evaluation. Review and comments on speech results (AMU)
0.8.1	27/09/2016	draft	Emma Barker (USFD) Monica Paramita (USFD)	Section 3.5-social media results: extended and revised interpretation of quantitative results and added qualitative analysis;

				Added results from issues. Updates to automatic table numbering throughout entire draft of v.8.1 Change to section 3.3.3 Participants: figures updated to report the two groups of 32 participants recruited (64 in total).
0.9	28/09/2016	Draft	Morena Danieli (UNITN)	Sections 2 (2.1 – 2.5): added statistical analysis of User success rate; integration of comments and integration of v0.8.1 (USFD) with draft v0.8.0 (comments)
0.10	28/09/2016	Draft	Morena Danieli (UNITN)	Section 3.2.1: added missing information (Fabio Celli's input)
0.11	10/10/2016	Draft	Emma Barker, Monica Paramita (USFD)	Revised text and added figures, tables and results to the section on the Social Media Evaluation
0.12	12/10/2016	Draft	Elisa Chiarani (UNITN)	First Quality Check completed
0.13	12/10/2016	Draft	Morena Danieli (UNITN)	Added 1.1, 1.2, 1.3, and 2.5.1 – first complete draft
0.14	14/10/2016	Draft	Morena Danieli (UNITN)	Completed 2.5.1; edited 1.3 and speech paragraph in 4
0.15	16/10/2016	Draft	Morena Danieli (UNITN)	Completed Reference section
0.16	18/10/2016	Draft	Rob Gaizauskas, Emma Barker (USFD)	Added and revised text in the Executive Summary, Section 1 and Section 3
0.17	20/10/2016	Draft	Morena Danieli (UNITN)	Merged v 0.14, 0.15,0.16 – Complete draft submitted for scientific review

0.18	24/10/2016	Draft	Udo Kruschwitz (UESSEX)	Scientific Review
0.19	26/10/2016	Draft	Emma Barker (USFD), Morena Danieli (UNITN), Rob Gaizauskas (USFD), Monica Lestari Paramita (USFD)	Modifications to sections 2 and 3 based on scientific review by Udo Kruschwitz
0.20	27/10/2016	Draft	Elisa Chiarani (UNITN)	Quality check completed
1.0	27/10/2016	Final	Morena Danieli (UNITN)	Final version
1.0	27/10/2016	Final	Giuseppe Riccardi (UNITN)	Approved for submission

INDEX

Executive summary

D1.4 reports the methods and the results of the extrinsic and insight-oriented evaluation of the SENSEI prototype. The evaluation methodology adopted in SENSEI includes the dichotomy between intrinsic and extrinsic measures. While intrinsic measures have been applied at the technology level to assess and compare performance of different technical approaches, extrinsic measures address questions related to the assessment of usefulness of the project results for potential users. In addition the SENSEI evaluation paradigm includes the concept of "insight-oriented" evaluation to capture qualitative and behavioural variables that may shed light on the features that users find more useful or, on the contrary, problematic for the acceptance of the novel technologies developed in this project.

For SENSEI extrinsic evaluation trials 96 evaluators were recruited, 32 for the speech domain and 64 for social media domain. They were native speakers of English, French and Italian. Each evaluator was asked to perform tasks in her/his native language. The evaluation questions were focused on assessing if and how SENSEI technologies are useful for users in their everyday tasks, and on identifying the variables that could influence their opinion on the new technology. For answering to those questions we applied objective metrics, such as time-to-completion and user success rate, and qualitative analysis of results of the post-task questionnaires and focus group discussion.

The extrinsic evaluation of SENSEI prototype in the speech domain, in French and Italian, required the evaluators to perform the same tasks in different conditions, i.e. by using the SENSEI generated summaries and reports of call center calls, and without those summaries. The results showed that there is a significant difference with respects to user success rate and time-to-completion, and that SENSEI-enabled condition allowed the evaluators to perform their tasks more accurately and efficiently. This result is valid both for the French and for the Italian evaluators. The results collected with the post-task questionnaires for both languages showed that SENSEI prototype provides quite a new technology for the evaluators, and the evaluators expressed difficulties with SENSEI-enabled condition. This discrepancy between objective and subjective results was further investigated by a post-task focus group with a selected subset of evaluators. The analysis of the opinions that emerged from the focus group showed that the evaluators considered synopses and ad-hoc reports insufficient for answering questions about the behavioural attitudes of the call centre agents, while they expressed curiosity and interest for the possible use of synopses and ad-hoc reports for completing tasks that require call classification. A variable that influenced evaluators' opinions was the novelty of the tasks we submitted to them, since they expressed the need to familiarize themselves for a longer period of time with the automatically generated summaries and reports.

In the social media domain two SENSEI prototypes, both of which offered summaries of reader comment conversations in The Guardian newspaper, were assessed in a task-based evaluation in which they were contrasted with a standard reader comment facility like those typically found in on-line news sites today. As with the speech evaluation, each participant was asked to carry out one time-limited task with a SENSEI system and another with the conventional reader comment reading facility. Quantitative measures were used to assess the quality of the participants' task output. Following the tasks, participants were asked to complete a questionnaire about their experience with the systems. The principal research questions addressed by the evaluation were: (1) Were participants able to carry out the task better, i.e. produce higher quality task outputs, with a SENSEI system than with the conventional system? (2) Did participants prefer the SENSEI system to the conventional system for carrying out the

task? Results showed that (1) participants did not score significantly better on the task with one system than another, and (2) on average, participants slightly preferred one of the SENSEI systems to the conventional (baseline) system. On closer investigation the latter results showed that participants were divided between those who strongly preferred one SENSEI system over the baseline and those who preferred the baseline because they either (a) distrusted any automated system to give them a reliable summary or (b) did not like the fact that the SENSEI system made it harder to see key comments in the original conversational context. Insights gained from the evaluation have enabled us to see how to develop a superior interface that should address the concerns of some of the participants.

LIST OF ACRONYMS AND NAMES

ACOF: Agent Conversation Observation Form

1. Overview

In this document we describe and discuss the results of the extrinsic evaluation of the SENSEI summarization prototypes. The evaluation protocol applied in SENSEI has been designed and tested during the first two years of the project. That is a tripartite model where traditional "intrinsic" evaluation of component technologies is coupled with extrinsic evaluation tasks and "insight oriented" evaluation, both aiming at assessing the usefulness of the SENSEI technologies in real life use cases. In this document we describe and discuss the result of the extrinsic evaluation in social media and speech domains. The present section provides an overview of the final prototype evaluation, as well as the follow up of the activities carried on in Period 2 (P2). Moreover, the third paragraph of this section shows how, in the WP1 of the project, we followed Reviewers' recommendations with respects to evaluation metrics for summarization.

Section 2 of this document describes the speech extrinsic evaluation effort in terms of tasks and scenarios, experimental setting, metrics, and results. Section 3 deals with the evaluation activities and results we obtained by experimental evaluation of the SENSEI prototype in the social media domain.

1.1. Overview of the final prototype evaluation

In the human language technology domain the emergence of comparative evaluation has raised the question of how systems that generate linguistic output (in particular automatic translation and summaries) should be compared. In general, intrinsic metrics are more widely adopted in the domain of human language technology, while in the natural language generation domain extrinsic methods and measures are prevalent (see [Daume and Marcu, 2005; Doran et al 2004; Dorr et al. 2005, Mitkov and Rello 2009; Murray et al. 2008] among others). There is debate about the possibility of correlation of the two kind of metrics: for example, an experiment published in [Belz and Gatt, 2008] supported the hypothesis that for natural language generation tasks intrinsic and extrinsic measures do not correlate, suggesting that intrinsic metrics and extrinsic methods can capture different dimensions of how a system performs. The evaluation paradigm adopted in SENSEI applies intrinsic metrics and extrinsic methods to evaluate different aspects of the prototype: intrinsic metrics have been used to assess how the technology performs with respect to baselines, i.e. to evaluate the quality of system output, while extrinsic task based evaluation has been used as an effective methodology for predicting human task performance when using the summarization technology.

The details of the SENSEI evaluation paradigms and their first application in the mid-term evaluation have been reported in D1.2 and D1.3. The evaluation methodology adopted in SENSEI includes the dichotomy between intrinsic and extrinsic measures, without assuming that intrinsic metrics and extrinsic methods necessarily correlate. In particular, the SENSEI extrinsic evaluation methodology addresses questions related to the assessment of usefulness of the project results for potential users. Other subjective dimensions of the evaluation are dealt with by the concept of "insight-oriented" evaluation, where the goal is to capture qualitative and behavioural variables that may shed light on the features that users find more useful or, on the contrary, problematic for the acceptance of the novel technologies they evaluated.

For SENSEI extrinsic evaluation trials 96 evaluators were recruited, 32 for the speech domain and 64 for social media domain. They were native speakers of English, French and Italian. Each evaluator was asked to perform tasks in her/his native language. The evaluation questions were focused on assessing if and how SENSEI technologies are useful for users in their everyday

tasks, and on identifying the variables that could influence their opinions on the new technology. For answering these questions we applied objective metrics, such as time-to-completion and user success rate, and qualitative analysis of answers to the post-task questionnaires and focus group discussion.

1.2. Follow-up of Period 2 activities

During Period 3 (P3) SENSEI WP1 implemented the final prototype evaluation. The design of the evaluation paradigm was set up during Period 1 (D1.1 and D1.2), and a mid-term evaluation was performed in the second part of Period 2 (D1.3) for both the speech and social media domains. Building on the results of the P2 evaluation, during the first part of P3 the activities of WP1 included the refinement of the evaluation scenarios, the definition of the evaluation settings in greater detail, the choice of the quantitative and qualitative metrics to be used, the recruitment of the evaluators, the preparation of the evaluation interfaces for the prototype for the two domains, and the design and editing of the evaluation documents to be used in the training of the evaluators.

As for the speech experimental setup, the main differences with P2 evaluation concern the addition of new types of summaries to the evaluation tasks. While P2 evaluation tasks were focused on assessing the call center agent performance on the basis of automatically generated reports of the agents' behaviour, in P3 trial users were asked to perform both call center agent evaluation (with and without SENSEI generated reports), and to classify call features with and without automatically generated call summaries and ad-hoc reports. In order to accommodate the evaluation of speech summaries (synopses) the interface used for P2 evaluation has been updated.

The analysis of the P2 mid-term evaluation of the speech prototype allowed a setup of the evaluation metrics. From the beginning of the project we focused our attention on the possibility of combining quantitative and qualitative evaluation metrics within the context of extrinsic evaluation. Quantitative measures like time-to-completion and use success rate were used, and this allowed us to test statistically the significance of possible differences in the two settings of the extrinsic speech evaluation. However, given the partially subjective nature of extrinsic evaluation methodology, we also submitted a qualitative questionnaire to the evaluators.

With respect to the P3 social media evaluation, while the experimental task and overall design did not change greatly from that used in the P2 evaluation, there was a key difference in the evaluation setup: in M24 the evaluation session was carried out in the lab, but the P3 evaluation was carried out remotely via an evaluation task interface. So, while in the P2 setup humans supervised the experiments providing training and instructions, imposing time constraints and gathering data from participants (questions and responses were provided on paper forms). In the P3 evaluation we made significant changes to the evaluation interface to include background and training for the prototypes and tasks, to make it simpler and clearer, to allow time constraints to be imposed automatically by the system, and to gather data from the participants via the interface. In addition, the experimental protocol was simplified to include only one of the two tasks run in the P2 evaluation: in P2 the two were tasks "identify 4 issues" being discussed in a set of reader comments and "characterize opinion" on one specified issue. In the P3 evaluation only the first was retained. This simplification was adopted to reduce the task burden on human participants, thus increasing the likelihood of them completing the experiment and yielding more reliable results. Furthermore, with this more efficient means for presenting the tasks and gathering data, in the P3 evaluation we recruited a far greater number of participants (62 in P3 as opposed to just 4 in P2). Finally, two SENSEI systems were

evaluated, each in comparison with a conventional reader comment reading facility, in contrast to only one in the P2 evaluation. Both SENSEI systems, shared common interface components but differed in some of the underlying language processing components. In both cases the language technology components were refined versions of those available at the end of Y2 and in addition included components integrated from WP3 and WP4 that were not available for the P2 evaluation.

1.3. Recommendations from Y2 Project Review

The results of P2 project review reported one recommendation that applies to work done in D1.4, i.e. the suggestion of using other metrics in addition to ROUGE for evaluating the summaries generated by the SENSEI prototype. This recommendation has been taken into serious account.

In the literature all the efforts to improve over ROUGE have been using some kind of semantic similarity. The problem with that approach is that it evaluates jointly summarization and semantic matching task. An alternative could be using ROUGE over a much larger set of reference summaries to approximate the actual semantic population. However the critical issue we had in SENSEI prototype evaluation is the need of evaluating the usefulness of summarization. The approach we chose is based on the methodology of extrinsic evaluation.

While automatic intrinsic measures such as ROUGE use n-gram scoring to produce rankings of summarization methods, the extrinsic evaluation methods like the one we describe in this deliverable concentrate on the use of summaries in specific tasks, e.g. information retrieval, question answering and relevance assessment, with the goal of showing the usefulness of summaries for performing those specific tasks. As we mentioned above, there is a debate in the research community about the correlation of intrinsic and extrinsic task based measures. While some results support the hypothesis the specific extrinsic metrics, for example Relevance-Prediction in [Dorr et al. 2005], correlates with intrinsic ROUGE scores, some other studies suggest that intrinsic and extrinsic measures do not correlate (see [Belz & Gatt 2008]). In our view, despite of the possible correlation between extrinsic and intrinsic metrics, it is worth pursuing the goal of integrating in a comprehensive paradigm different kinds of qualitative and quantitative metrics that may be applied for addressing different evaluation questions. In particular, ROUGE-like metrics have been applied at the technology level in SENSEI for ranking the summarization methods with respects to gold standard, while extrinsic evaluation tasks have been run for determining how well automatic summaries help users to complete their tasks, and to investigate which type of summary perform better than other in a specific tasks.

2. Speech Extrinsic Evaluation

During the first year of SENSEI we designed the evaluation protocol for SENSEI. After the first six months of the project, a preliminary version of use case design was proposed in D1.1 [Danieli 2014] for the speech and social media domains by identifying target end users in both domains, and an initial set of use cases to be considered by the technical SENSEI WPs. For the speech domain the categories of potential users we identified were call centre professionals that needed to listen to a great amount of call centre conversations each day in order to find key indicators of quality of the service provided, i.e. Quality Assurance call centre supervisors. At the end of the first year of the project, D1.2 [Danieli & Gaizauskas 2014] provided a complete scope of the scenarios that were considered in the SENSEI project.

During the second year of the project, several experiments based on a selection of the proposed use cases were run. That intermediate SENSEI evaluation effort had the aim of setting up the different components of the evaluation protocol, including the joint use of qualitative and quantitative evaluation methods in the assessment of the speech tasks.

In the SENSEI deliverable D1.3 [Danieli & Barker 2015] we discussed the motivation of such assessment of the evaluation protocol due to the need of setting baselines for automatically generated summaries, and of fixing possible issues of the evaluation protocol in view of the multilingual trials planned for the third year of the project.

One of the results of Period 2 intermediate speech domain evaluation was the selection and the assessment of different types of extractive summaries that users could exploit to navigate large collections of call centre conversations. In particular, the speech summaries generated in SENSEI are reductive transformations of the source data that 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 survey forms (focused on the behaviour of the call centre agents), and ad-hoc reports that provides the opportunity of navigating the transformed conversations starting from queries specified by the users. On the basis of the intermediate evaluation results (reported in [Danieli & Barker 2015], [Danieli *et al*. 2016]), and by taking into account the users' feedback we got in specific focus group, we could select and refine a set of evaluation tasks to be used in Y3 for the final speech prototype evaluation.

In the following of this section we describe the goals of the speech extrinsic evaluation (paragraph 2.1), the evaluation scenarios (paragraph 2.2), the methodology (paragraph 2.3), the evaluation metrics (paragraph 2.4), the evaluation results (paragraph 2.5), and the discussion (paragraph 2.6).

2.1. Speech extrinsic evaluation goals

We are interested in whether SENSEI speech summarization prototype is effective in assisting the QA supervisors ("the evaluators" henceforth) while they need to process large volumes of call centre conversations. In particular, the evaluation of the speech prototype aims at answering the following questions:

1. Does SENSEI speech technology help the users find information needed to perform their tasks?
2. Does SENSEI speech technology have impact on time-to-completion of the tasks?

3. Does SENSEI speech technology increase user satisfaction?

In order to answer to questions 1-3, we designed a set of evaluation scenarios that we describe in details in Section 2.2.

SENSEI speech prototype generates speech summaries (synopses) and ad-hoc reports in two natural languages, French and Italian. French and Italian call center corpora, DECODA and LUNA respectively, were used for testing. So while the evaluation scenarios have the same structure in terms of the tasks submitted to the evaluators, the specific questions of each task were adapted to meet the different semantic domains of the two corpora[1].

2.2. Evaluation scenarios

The evaluation scenarios that we present in this section were defined on the basis of the analysis of the results of the mid-term extrinsic evaluation [Danieli & Barker 2015]. As mentioned above, the intermediate SENSEI evaluation involved a limited set of SENSEI potential users selected from TP Quality Assurance supervisors. From those results we could get evidence about the potential usefulness of ad-hoc reports and synopses of the call centre calls for the final users. 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, and for reducing the time needed to navigate them. In addition, the users suggested that using the system could help in overcoming the subjectivity issue of their usual listening tasks. For the final speech prototype extrinsic evaluation we have designed a set of evaluation scenarios based on those P2 evaluation results.

Each evaluator is asked to perform two fact gathering scenarios by browsing a set of recorded calls, under two different conditions, C1 and C2. In condition C1 the evaluator may listen to the calls and look at the call transcriptions, while in condition C2 s/he may also read SENSEI generated synopses of the source phone calls, and reports of predicted features (ad-hoc reports). Both spoken call synopses and ad-hoc reports showed in the evaluation were automatically generated by the SENSEI prototype.

Each scenario includes two tasks (T1 and T2 henceforth). For T1 the solution of the task is at the conversation level, while for T2 the solution is at the conversation collection level.

Each task involves answering two evaluation questions about issues in the calls. The questions were designed to capture from the calls facts that may be relevant for the QA supervisors and for their clients. The questions are presented to the subjects as part of a prompt, after their listening to the conversations, or after their inquiries based on call synopses and ad-hoc reports. Given the different types of DECODA and LUNA domains (information seeking, and technical assistance), we identified scenarios and tasks that refer to the specificity of the domains.

Scenarios, tasks, and questions were submitted to the evaluators in their native languages (French and Italian). For ease of reading, in the tables below we report the English translations of the tasks.

[1] In the SENSEI deliverable D6.3, “Report on the Rated Questionnaire and Ad-hoc Report Views” the reader may find the details of the summarization technologies (baselines and experimental systems) that we evaluate in real-world conditions in WP1.

Table 1: French Scenario 1 - Complaints

DECODA Scenario 1: Complaints			
Description: The scenario is focused on how call center agents handle complaints by the caller. In the DECODA corpus complaints may be related to the behaviour of a particular agent, to the service, or to a request of refund.			
Task 1 (conversation level)		**Task 2 (collection level)**	
Characterize the behaviour of the agent in stressing conditions (caller is worried, angry, needs a quick and clear answer)		Estimate the proportion of complaints and the proportion of complaint sources.	
Questions		**Questions**	
Q1	Politeness of the agent	Q1	Proportion of caller complaints
Q2	Efficiency of the agent	Q2	Source of caller complaints

The scenario described in Table 1 was meant to capture elements useful to judge the professional attitude of the agents towards possible complaints of the callers, in particular when they may be angry for reasons related to inefficiency of the service (Task 1). The Task 2 of the scenario was meant to give a quick estimate of the reasons of the calls served by the customer call center.

Table 2: French Scenario 2: Lost Items

DECODA Scenario 2: Lost items			
Description: The scenario is focused on how operators handle the lost items calls, in particular when the callers ask for a lost item that is particularly valuable for them.			
Task 1 (conversation level)		**Task 2 (collection level)**	
Characterize the behaviour of the agents in this scenario when they found or not the item		Estimate the proportion of items that can be effectively retrieved	
Questions		**Questions**	
Q1	Politeness of the agent	Q1	Proportion of retrieved items
Q2	Efficiency of the agent	Q2	Types of lost items

The tasks of the French Scenario 2 are reported in Table 2. As in the previous scenario, the first task was designed to capture behavioural aspects of the call center agents, while in this case the second task was meant to provide a quick qualitative and quantitative description of the content of the calls.

The same task structure, i.e. a first task focused on agent behaviour and a second task focused on the call content, was reflected by the two scenarios designed for the Italian LUNA corpus.

Table 3: Italian Scenario 1: Handling of technical issues

LUNA Scenario 1: Handling of technical requests			
Description: The focus of the scenario is how operators handle the specific technical requests by the caller, if they understand the request, and handle it appropriately and efficiently.			
Task 1 (conversation level)		**Task 2 (collection level)**	
Characterize the ability of the call center agent in understanding the caller technical request.		Estimate the proportion of calls that may be solved directly by the operators	
Questions		**Questions**	
Q1	Politeness of the agent	Q1	Proportion of calls solved by the operators
Q2	Ability to understand the nature of the problem	Q2	Source of technical complaints

As we may see in Table 3 above, in this case the evaluators were asked to judge the agent attitude towards the caller by also taking into account her/his ability to understand the technical nature of the problems (Task 1), and the quantitative task asked to estimate the nature of the technical issues and the proportion of calls that could be solved directly by the call center agents.

The last scenario is reported in Table 4 below. This scenario was meant to capture what happened when it was not possible to solve the technical issue that originated the call to the technical customer care service. Again in this case the behavioural attitude of the agent is evaluated through questions of Task 1, and the call content aspects are taken into account by Task 2 questions.

Table 4: Italian Scenario 2: Unresolved Technical Issues

LUNA Scenario 2: Unresolved technical issues
Description: The focus of the problem is on how call center agents identify the problem of the

call, and what they do if they cannot solve the problem within the call time.			
Task 1 (conversation level)		**Task 2 (collection level)**	
Characterize the behaviour of the agents when they cannot solve the caller problems immediately		Estimate the proportion of calls that are reported to a second level.	
Questions		**Questions**	
Q1	Politeness of the operator	Q1	Proportion of problems sent to a second level
Q2	Proactivity of the operator	Q2	Source of unresolved problems

2.2.1. User Experience Questionnaire

In order to capture the user experience with SENSEI speech prototype, we designed a user experience questionnaire that included seven questions, with some responses arranged on a 5 points Likert scale, other responses with choices among three options, while two final open questions. The questions were common for DECODA and LUNA scenarios. They were submitted to evaluators in their native languages. Questions are reported in Table 5[2].

Table 5: SENSEI user experience questionnaire

User Experience Questionnaire					
	1 Not at all	2	3 Some what	4	5 Completely
1. To what extent did you understand the nature of the tasks you have completed?					
2. To what extent did you find those tasks similar to other tasks that you typically perform?					
	C1	C2	No difference		
3. Which of the two conditions, C1 and C2, did you find easier to learn to use?					
4. Which of the two conditions, C1 and C2, did you find easier to use?					

[2] The user experience questionnaire was designed on the basis of the research presented in [Kelly & Teevan 2033], [Kelly et al. 2007], and [Dang et al. 2007].

5. Which condition was the most useful for completing your tasks?			
6. What did you find useful about working under each of the two conditions?	<free text C1> <free text C2>		
7. What did you not appreciate about working under each of the two conditions?	<free text C1> <free text C2>		

The questionnaire investigates multiple dimensions. The first dimension is captured by questions 1 and 2. It is meant to measure the levels of comprehension of the tasks and the familiarity that the evaluators had with the activities required during the trial.

The second dimension is related with the comparison between working under each experimental condition, i.e. C1 and C2. This dimension is measured by analysis of responses to questions 3 to 5.

The third dimension is insight-oriented: by replying to questions 6 and 7, evaluators may provide suggestions about what they judged useful for their work in each experimental conditions.

2.3. Methodology

2.3.1. Evaluator sample and training

For each language we recruited 16 evaluators, balanced by gender (8 female subjects and 8 male subjects). All the evaluators were Quality Assurance supervisors. They were recruited by SENSEI partner TP in call centers in Italy (Taranto) and France (Bordeaux, Montpellier, Toulouse, Villeneuve d'Ascq). The French evaluators performed the trial during the weeks July 19-29, 2016, and the Italian evaluators during the weeks July 5-15, 2016.

Table 6: sociolinguistic variables of the evaluators

Nr	Native Speaker	Gender	Age	Education	Professional level
8	French	F	37	13	6
8	French	M	34	14	4
8	Italian	F	45	14	7
8	Italian	M	39	14	7

In the Table above the third column report the mean of the age of the participants, the Education column reports the average years of education of the subjects, while the last column report the average years of experience in their professional role.

Both evaluator teams received training before the trial. They were trained in small groups by TP expert QA supervisors. Training sessions were face-to-face for the Italian participants and, due to resource limitation, it was held from remote (webinar) for the French participants. The same material, including information sheets and slide shows, was translated from English into Italian and French. The training sessions were submitted to the evaluators in their native languages.

2.3.2. User interface of the SENSEI speech extrinsic evaluation prototype

We developed an Extrinsic Evaluation software module to implement both the extrinsic evaluation activities and the user experience questionnaire. The module was integrated in the ACOF (Agent Conversation Observation Form) tool that was developed during the first two years of the SENSEI project [Danieli & Barker 205]; it inherits from the host system the look and feel and the database.

2.3.2.1. Evaluation user interface

In the first page of the evaluation user interface, evaluators are asked to choose the corpus (LUNA/DECODA), the scenario, the condition and the task. According to their choice the system submits an evaluation task to the user and starts recording the time to completion.

The structure of the evaluation page is the same for both corpora and scenarios, while it changes depending from the selection of the task and condition.

It is composed by three panels. The top panel is general: it contains the evaluation instructions and the selected condition for the current evaluation. The right panel always contains the audio player and the transcription visualizer of the conversation being evaluated. The left panel presents the evaluation questions. For Scenario 2 this panel presents the search engine and, in C2, it provides the synopses of the conversations.

For example, Figure 1 shows the user interface that the system presents when user selects DECODA service, C1, Scenario 1, task 1.

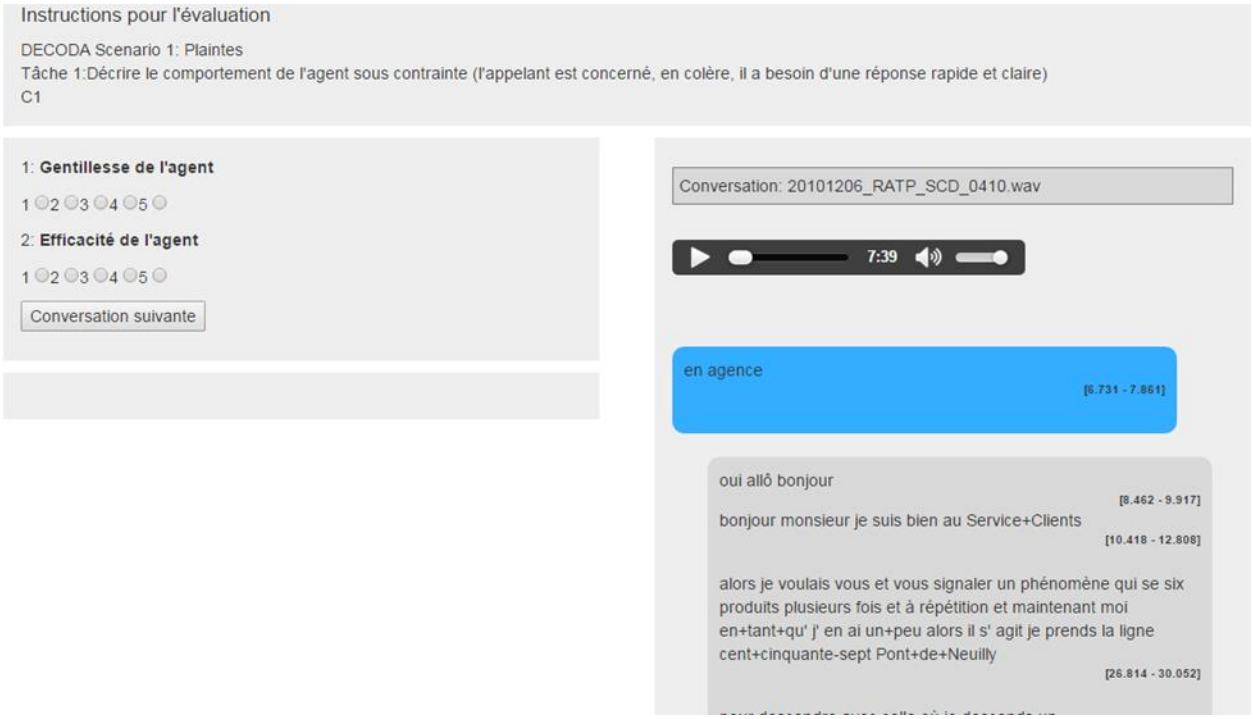


Figure 1: User interface for evaluation – C1

Figure 2 shows the user interface that the system presents when user selects DECODA corpus, C2, Scenario 2, and Task1: as C2 was selected, the SENSEI results are enabled, and the evaluator can read the synopsis and the ad-hoc report of the predicted features of the conversation that were automatically generated by the SENSEI prototype.

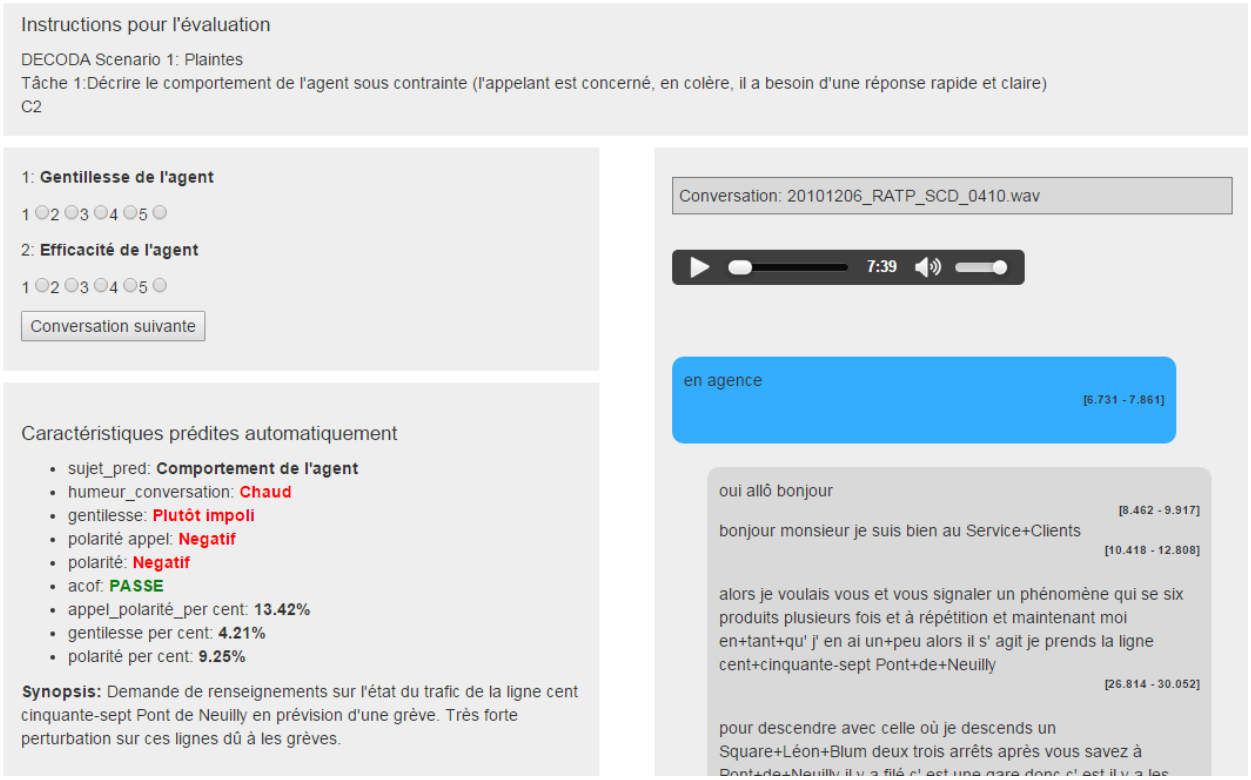


Figure 2: User interface for the evaluation – C2

Figure 3 shows the user interface that the system presents when the evaluators work with LUNA corpus. In this case, the selection included C1, Scenario 1, and Task 2: the search engine is enabled, the evaluator can run full text search and consult the result list. Clicking on the conversation title, the right panel loads the transcription and the audio player.

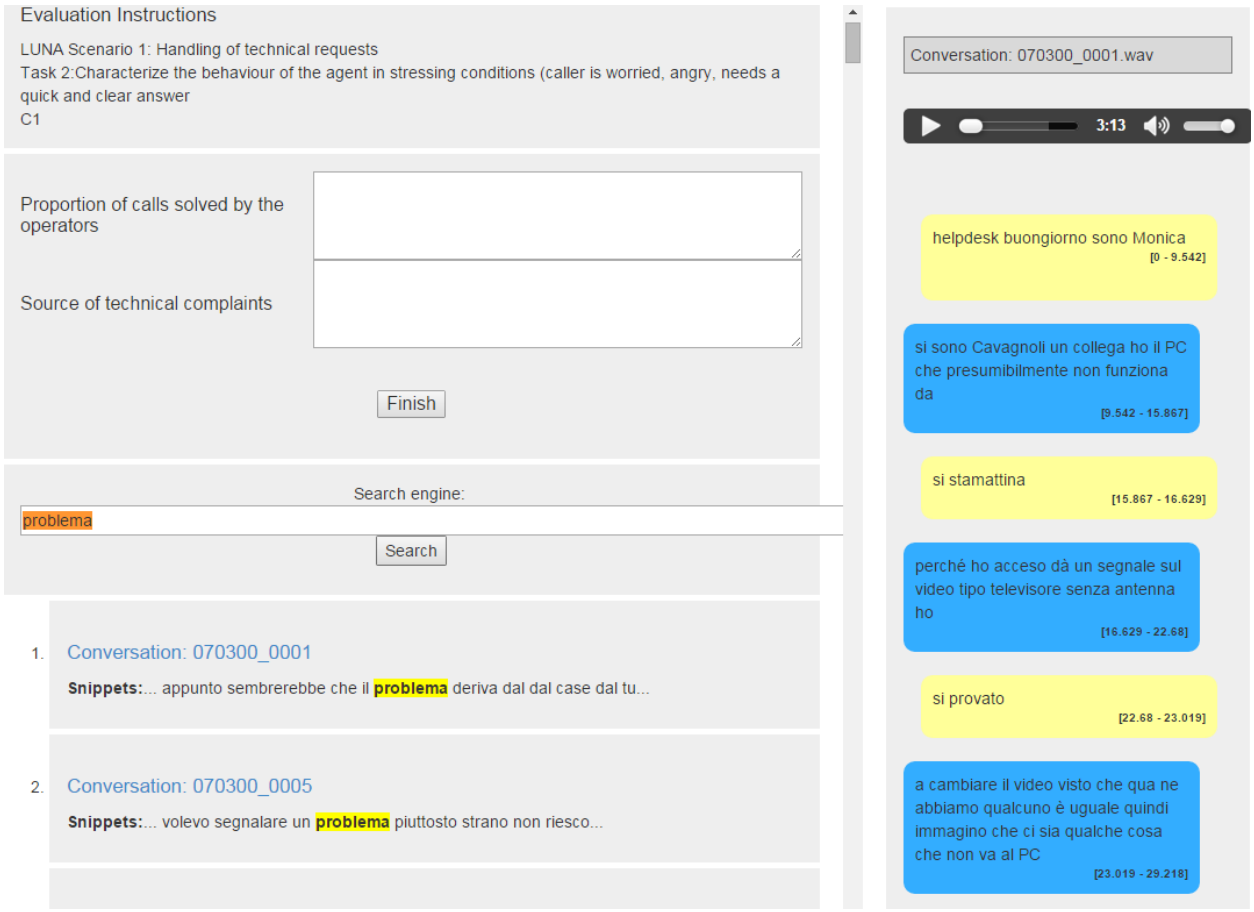


Figure 3: User interface for the evaluation, LUNA - C1

Figure 4 shows the user interface that the system presents when evaluators work on LUNA corpus at a conversation collection level, i.e. Task 2 in condition C2. The search engine is enabled and evaluators can run full text search and read the result list. In C2 other SENSEI automatically generated results are provided to the users, including filtering by polarity of the call, caller polarity and agent polarity. Clicking on the conversation title, the right panel loads the transcription and the audio player.

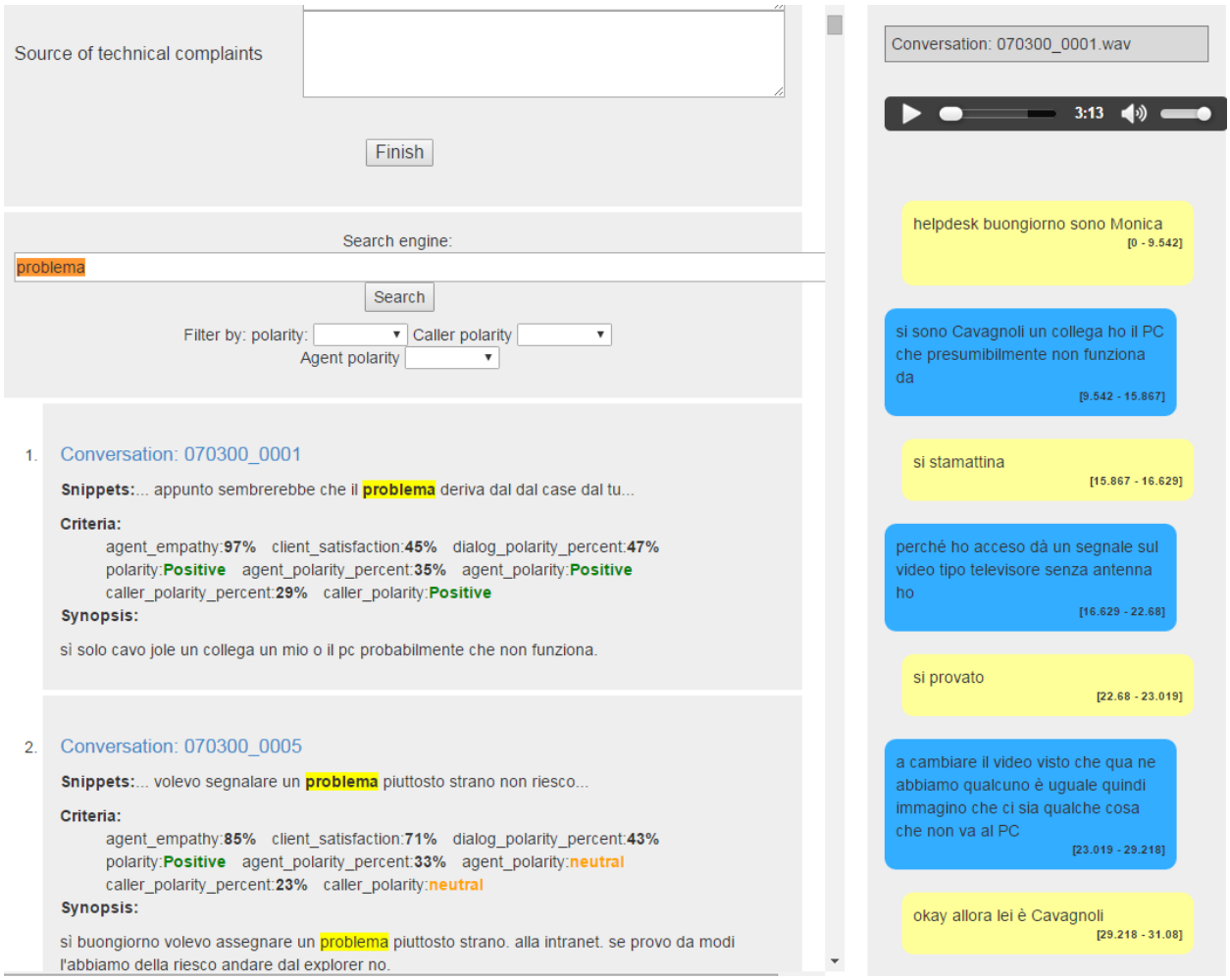


Figure 4: Evaluation interface LUNA – C2

2.3.2.2. Interface of the user experience questionnaire

User experience questionnaire is a simple page that does not vary across corpora, but that may be presented in the evaluators' natural language. Figure 5 shows the English implementation:

User Experience Questionnaire

	1	2	3	4	5
To what extent did you understand the nature of the tasks you have completed?	○	○	○	○	○
To what extent did you find those tasks similar to other tasks that you typically perform?	○	○	○	○	○

	C1	C2	No difference
Which of the two conditions, C1 and C2, did you find easier to learn to use?	○	○	○
Which of the two conditions, C1 and C2, did you find easier to use?	○	○	○
Which condition was the most useful for completing your tasks?	○	○	○
What did you find useful about working under each of the two conditions?			
hat did you not appreciate about working under each of the two conditions?			

Save

Figure 5: User Experience Questionnaire – English version

2.3.3. Experimental methodology

We tested the completion of the supervising tasks under two levels, without SENSEI speech summaries (C1) and SENSEI-enabled (C2), i.e. in this condition the evaluators could use the speech summaries to select relevant calls and to navigate in the call collections. We applied a mixed "within-subjects", repeated-measure design. We tested *within-subjects* one independent variable with two levels, i.e. each subject had been tested under the two conditions.

We know that the learning effect may be related with the order of presentation of the conditions. For example, if subjects are tested under condition C1 first, then under condition C2, they may exhibit better performance under condition C2 simply due to prior practice under condition C1. We compensated for this by placing evaluators in groups and presenting conditions to each group in a different order. To set the order we referred to a balanced Latin square.

The Latin square in Table 7 is 8 x 8, reflecting the condition/scenario/task combinations of SENSEI extrinsic evaluation of the speech prototype (see above the description of scenarios). In particular, the subjects had to complete (for each language, French and Italian) two scenarios that included two tasks each under C1 and C2[3]. Task-1 and task-2 in each scenario were

different for being done at the conversation and at the conversation collection level respectively. In the table below we introduce the labels corresponding to the combination of Condition/Scenario/Task as follows:

Table 7: Labels of the condition – scenario – task combinations

Value Label	Condition	Scenario	Task
A	C1	S1	T1
B	C1	S1	T2
C	C2	S1	T1
D	C2	S1	T2
E	C1	S2	T1
F	C1	S2	T2
G	C2	S2	T1
H	C2	S2	T2

The rows of the Latin square design of Table 8 reports the order in which each evaluator of the G1 group performed the tasks, for example Evaluator 1 (male) will run first "Condition C2, Scenario 1, Task 2", then "Condition C1, Scenario 2, Task 1", and so on[4].

Table 8: Latin square design (male evaluators)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	D	G	A	F	B	C	H	E
2	A	B	E	H	G	F	D	C
3	B	C	H	A	F	D	E	G
4	F	H	C	D	E	A	G	B
5	C	A	G	B	D	E	F	H
6	G	E	F	C	H	B	A	D
7	H	D	B	E	A	G	C	F
8	E	F	D	G	C	H	B	A

The following Latin square shows the order of tasks for G2 evaluators:

Table 9: Latin square design (female evaluators)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	B	F	E	G	A	D	C	H
2	G	E	F	B	D	H	A	C

3	C	A	B	D	E	F	H	G
4	E	C	G	H	F	B	D	A
5	F	B	C	A	H	E	G	D
6	A	H	D	F	C	G	E	B
7	D	G	H	C	B	A	F	E
8	H	D	A	E	G	C	B	F

Since we had 32 evaluators available, 16 for each language, the design described above may be replicated with a second group of 8 male and 8 female evaluators. Evaluators were randomly assigned to each group.

2.4. Evaluation Metrics

In this paragraph we introduce the quantitative and qualitative metrics used in the SENSEI extrinsic evaluation.

2.4.1. Measures

The speech extrinsic evaluation of SENSEI (French and Italian) speech prototype addressed the evaluation questions illustrated in section 2.1. For ease of reading we report here the three evaluation questions:

1. Does SENSEI speech technology help the users find information needed to perform their tasks?
2. Does SENSEI speech technology have impact on time-to-completion of the tasks?
3. Does SENSEI speech technology increase user satisfaction?

To answer these questions we designed the evaluation scenarios and tasks described in § 2.2, and the user experience questionnaire described in § 2.2.1. The evaluation scenarios were focused on getting data to allow comparisons between user performances in two conditions, C1 (without SENSEI) and C2 (SENSEI-enabled). Each of the above evaluation questions refers to specific metrics: user success rate is used to provide responses to question 1, time-to-completion is used to collect data for statistical evaluation of differences between C1 and C2, while answers to questionnaire are evaluated by qualitative description and non-parametric control of the experimental hypothesis.

To evaluate **Question 1** the metric used is user success rate that measures user ability to complete their experimental tasks. We measure the percentage of tasks that were completed by replying to all the evaluation questions, and in the given time constraints. Given the subjectivity that may characterize this extrinsic evaluation tasks, we could not define 'correctness of the task' in terms of the content of the participants' judgements. We assign the tag Success (S) to participants, who completed their tasks completely, Failure to evaluators who failed to complete their tasks, and Partial Success to evaluators who completed successfully only some part(s) of the task, for example that provided answers only to one of the two questions of each task.

To evaluate **Question 2** the metric used is time-to-completion. For each one of the tasks completed by the users within the time assigned (1200"), the time to complete the task was

recorded by the extrinsic speech evaluation prototype. We statistically verify the hypothesis that time-to-completion under C1 is greater than time-to-completion under C2 (t-test).

To evaluate user experience (**Question 3**) the questionnaire included two questions whose replies were arranged on a 5 values Likert scale, three questions where 3 replies were allowed, and two open questions. The latter are evaluated qualitatively, the previous two sets of questions are evaluated by comparing observed and expected frequencies (Chi-square test).

2.4.2. Valid data

We define "valid data" the numerical and qualitative data that we may extract from the evaluation tasks that were both completed, and completed within the time assigned for the evaluation tasks (1200").

Both for French and Italian we submitted the users with 8 tasks to be completed. The tasks were labelled with letters from A to H.

We expected to have for each language, 16 valid tasks for each combination Condition-Scenario-Task (8 completed by female subjects, and 8 completed by male subjects). By applying the exclusion criteria introduced above, we have the following valid data (Table 10):

Table 10: Italian and French valid tasks

Task	Italian valid tasks	French valid tasks
A	7	7
B	14	13
C	10	7
D	13	13
E	11	6
F	16	13
G	12	8
H	16	14
TOTAL	99	81

From the data above we may observe that users completed a greater proportion of tasks with the SENSEI enabled system both in Italian and, although at a lesser extent, in French. In Italian we have 48 tasks completed in Italian without SENSEI, and 51 SENSEI enabled. In French we have 39 tasks completed without SENSEI and 42 SENSEI-enabled. In total we collected in the two languages 93 C2 valid tasks, and 87 C1 valid tasks.

2.5. Evaluation Results

2.5.1. Evaluation Question 1

While the Table 10 in the last section reports the number of valid tasks, in order to calculate user success rate we need to distinguish between complete success (S in Table 11) and partial success (P in Table 11).

Table 11: Success, Partial Success and Failure

	Italian			French		
	S	**P**	**F**	**S**	**P**	**F**
A	5	2	9	6	1	9
B	10	4	2	12	1	3
C	8	2	6	6	1	9
D	11	2	3	13	0	3
E	9	2	5	4	2	10
F	12	4	0	12	1	3
G	10	2	4	8	0	8
H	13	3	0	13	1	2
Total	78	21	29	74	7	47

In total, we observed 128 attempts to perform the tasks for each language. For Italian, of those attempts, 78 were successful and 21 were partially successful. For French, of those 128 attempts 68 were successful and 13 were partially successful.

We gave each success a point, each partial success half a point, while no point is assigned to failure.

For Italian the success rate was 0.6914. For French the success rate was 0.6054. The success rate under C1 was 0.6525 for Italian, and 0.5703 for French, while under C2 the success rate was 0.7265 for Italian, and 0.6406 for French.

2.5.2. Evaluation Question 2

To reply to this evaluation question we statically tested the experimental hypothesis according to which the use of speech summaries reduced the time-to-completion of evaluation tasks. In the following we report the results for French and Italian extrinsic evaluation.

French

Table 12 reports the timing of French valid data for each task: for each task (A-H) we may find the combination (Condition/Scenario/Task), the cardinality of valid tasks, the time-to-completion observed for each task and the mean.

Table 12: French valid time data

Value Label	Condition	Scenario	Task	N valid tasks (F and M subjects)	Total (Female; Male)(*Mean*)
A	C1	S1	T1	7 (4 F; 3 M)	4724 (2738 F; 1986 M) (*M*: 674.85)
B	C1	S1	T2	13 (7 F;5 M)	8258 (5092 F; 3166 M) (*M*: 635.30)

C	C2	S1	T1	7 (3 F; 4 M)	5511 (2215 F; 3296 M) (*M*: 787.28)
D	C2	S1	T2	13 (6 F; 7 M)	5750 (2367 F; 3383 M) (*M*:442.31)
E	C1	S2	T1	6 (3F; 3M)	4190 (1944 F; 2246 M) (*M*: 698.33)
F	C1	S2	T2	13 (6 F; 7 M)	7042 (2759 F; 4283 M) (*M*: 541.69)
G	C2	S2	T1	8 (4 F; 4M)	4679 (2023 F; 2656 M) (*M*: 584.87)
H	C2	S2	T2	14 (8 F; 6 M)	6557 (4772 F; 1785 M) (*M*: 468.36)

From the data above we got the following statistics:

Table 13: French time-to-completion statistics

	Group 1	Group 2
Mean	6053.5000	5624.2500
SD	921.5600	772.9100
SEM	147.5677	119.2626
N	39	42

The two-tailed P value equals 0.0255. By conventional criteria, this difference is considered to be statistically significant. As for the confidence interval, the mean of Group 1 minus Group 2 equals 429.2500. 95% confidence interval of this difference is from 54.0483 to 804.4517. The intermediate values used in calculations are t equals to 2.2772, df equals to 79, standard error of difference equals to 188.501.

Italian

Table 14 reports the timing of Italian valid data: for each task (A-H) we may find the combination (Condition/Scenario/Task), the cardinality of valid tasks, the time-to-completion observed for each task and the mean.

Table 14: Italian valid time data

Label	Female subjects		Male subjects		Total Time	
	N Female valid tasks	Time Fem (*m*)	N Male valid tasks	Time Male (*m*)	N Total valid tasks	Time (*m*)
A	4	4439 (1109,7)	3	2284 (961,3)	7	6723 (*872,9*)
C	7	2353 (336,1)	7	7513 (1073,3)	14	9866 (704.7; *426.47*)

B	3	2899 (966,3)	7	5571 (795,8)	10	8470 (847)
D	7	4830 (690)	5	4241 (848,2)	13	9071 (697,8)
E	4	4146 (1036,5)	7	3159 (451,3)	11	7305 (664,1)
G	8	2619 (327,4)	8	2745 (343,1)	16	5364 (335,2)
F	5	5198 (1039,6)	7	5184 (740,6)	12	10282 (865,2)
H	8	6825 (8853,1)	8	5500 (687,5)	16	12325 (770,3)

From the data above we got the following statistics:

Table 15: Italian time-to-completion statistics

	Group 1	Group 2
Mean	8195.0000	9284.0000
SD	156.5100	2670.0000
SEM	24.7464	347.6044
N	40	59

The two-tailed P value equals 0.0116. By conventional criteria, this difference is considered to be statistically significant. As for the confidence interval, the mean of Group 1 minus Group 2 equals -1089.0000. 95% confidence interval of this difference is from -1929.2373 to -248.76.27. The intermediate values used in calculations are: *t* equals to 2.5723, *df* equals to 97, standard error of difference equals to 423.352.

2.5.3. Evaluation question 3

Question 1 and **Question 2** of the user experience questionnaire aimed to understand the degree of similarity of the required tasks with tasks that the evaluators may perform in their ordinary work (2) and the degree of understanding of the tasks they were asked to perform in this evaluation (1). The question statements are (non-continuous) Likert items. We calculated central tendency (summarized by median and mode), variability, and analyzed the data with non-parametric analysis. Central tendency for each language are reported in Table 16, where we may observe similar results for French and Italian evaluators about Question 1, i.e. the evaluators reported they understood the tasks they were asked to perform. For Question 2 both Italian and French evaluators report the novelty of the tasks with respects to their ordinary work. And actually in ordinary conditions they listen to the ongoing calls and do not have transcripts of the conversations.

Table 16: Central tendency statistics – Questions 1 - 2

	Italian median	Italian mode	Coefficient of variation	French Median	French mode	Coefficient of variation

To what extent did you understand the nature of the tasks you have completed?	4	4	O,1664	3	4	0.4386
To what extent did you find those tasks similar to other tasks that you typically perform?	3	3	0,1481	2	2	0,5484

Questions from **3** to **5** investigated three dimensions: ease of learning (**Question 3**), ease of use (**Question 4**), and usefulness for task completion (**Question 5**). Possible responses were categorical. For all the three the evaluators could express their preference for C1, C2 or No Difference. Results are reported in Table 17.

Table 17: Evaluators' responses to questions 3 - 5

	Italian			French		
Questions	C1	C2	No Diff	C1	C2	No Diff
Which of the two conditions, C1 and C2, did you find easier to learn to use?	13		3	14		2
Which of the two conditions, C1 and C2, did you find easier to use?	14	1	1	15		1
Which condition was the most useful for completing your tasks?	16			14	1	1

As we may observe the great majority of the participants found that Condition C1 was easier to learn (Question 3), and easier to use (Question 4). While for the totality of the Italian participants C1 was also the most useful configuration for completing the evaluation tasks (Question 5), one French evaluator expressed a preference for the use of call synopses, and another one said that there was no difference with respects to usefulness.

As for the last two questions of the survey, we collected qualitative reports from the evaluators. **Question 6** and **Question 7** asked the evaluators to provide free text responses about what they found useful for completing their tasks in both conditions, and what they did not like. The majority of Italian and French evaluators' responses outlined the usefulness of listening to the conversation with the aid of the automatic transcripts. For example, they wrote that

> *"The call transcripts are useful for speeding up the process of recorded call evaluations",*
>
> *"By reading the call transcriptions I was able to focus attention on the speech of the agent".*

Only a few (2/32), replied that the transcripts did not add value to the task execution, and only a few (5/32) find useful to be guided by the call synopses for navigating the conversation collections. In general, call synopses were judged more useful for executing tasks of Scenario 2 in both languages, i.e. the tasks focused on call content. For example, Italian evaluators wrote

> *"Using combination of concepts could be useful for assessing the reasons of inbound calls"*
>
> *"If we were to assess why people call, for instance, customer care, call summaries would be ok"*

According to evaluators, the weak aspects of the automatic generated synopses are related with their degree of informativeness, but again that is more reported for tasks of Scenario 1 – that is more focused on behavioural attitudes of the agent, than for tasks of Scenario 2, where the French and Italian evaluators that were able to complete the tasks successfully recognized that synopses could improve the search of the call collections.

2.6. Discussion

In the extrinsic evaluation of the SENSEI speech technologies we have collected both quantitative and qualitative results. The quantitative results have been measured with objective metrics, i.e. user success rate and time-to-completion of the evaluation tasks. As we could observe from the results and statistical analysis reported in the previous paragraph, both metrics allowed to assess for both language groups of evaluators a superiority of SENSEI enabled condition. In particular, both Italian and French evaluators reported better results of task success and a reduction in time to completion when they executed their tasks in SENSEI-enabled condition. We may hypothesize that in this condition the availability of speech summaries played a role for increasing the accuracy and the efficiency of the evaluation tasks. So the objective evaluation of quantitative results allows us to respond positively to the first and to the second evaluation questions.

To reply to the third evaluation question we examined different dimensions. First of all, we wanted to examine the degree of novelty of the evaluation tasks in comparison with evaluators' usual working conditions, and the degree of their understanding with respect of the submitted tasks. For both dimensions we got central tendency measures that support the view that the SENSEI prototype provides a quite new technology for the execution of tasks that are normally completed based only on the online listening of ongoing conversations. In addition, the questionnaire explored ease of learn and ease of use of the two experimental conditions provided by the SENSEI prototype. As we described in the first conditions, evaluators could access to automatic transcript of the speech conversations, while in the second one they could also use the call synopses to orient the selection process and to navigate the collection of conversation. Despite of the objective evaluation results, that show superiority in terms of time-to-completion and accuracy of the second condition, the evaluators express a preference for condition 1. We hypothesize that this result could be biased by the fact that working in condition C1 they felt a closer similarity with their usual working condition, and/or that the novelty of the technology required a longer period of acquaintance with the system. To explore these hypotheses we organized a focus group with a subset of the Italian evaluators.

2.6.1. Post-task focus group on user experience

A focus group with a subset of Italian evaluators[5] has been organized to investigate the reasons of the discrepancy between user performance and user preferences described above. The goal

[5] For organizational reasons we could not come back in touch with the group of French evaluators.

of the focus group was to understand the attitudes of the evaluators with respects to the introduction of innovative technologies in their ordinary working conditions. Since the focus group was held two months after the prototype evaluation, we could not rely on reliable memories of the difficulties they possibly had during the evaluation sessions, however the questions asked during the focus group could facilitate the recovery of impressions and personal judgments on the technologies they used.

The focus group was conducted by a moderator assisted by two recorders. The group discussion was held in video conference: the moderator was in Trento, while the two recorders and the evaluators were in Taranto. Six subjects were recruited to take part in the focus group. We had 4 female and 2 male participants[6]. They were randomly selected from the subset of Italian evaluators who fully completed the evaluation tasks. The evaluators sat around a large table in a silent room, arranged in a semicircle, and the recorders were sitting in front of them. The task of the recorders was taking structured notes silently, while the group discussion was guided by the moderator.

The set of questions used during the discussion are reported in the box below:

(1) What do you remember about the evaluation experience?

(2) At what extent the training had been effective for task completion?

(3) Which degree of familiarity did you have with the domain of the calls (customer care) prior to the evaluation sessions?

(4) [The moderator briefly reports the objective evaluation results] Do you find that those results reflect your evaluation experience?

(5) In the evaluation you were asked to work by listening to conversation, and by using ad-hoc reports and conversation summaries. Do you believe that the use of summaries could help your ordinary work for analyzing great amount of calls? And for evaluating agents' behaviour?

(6) Are there any tasks in your ordinary work that could be done more easily or efficiently by using the technologies developed in the SENSEI project?

All the evaluators said that they were able to remember their experience with the system (Question 1). All of them expressed appreciation for the training they received before the experiment (Question 2), however all of them considered the tasks difficult to solve, in particular the ones that required them to use actively synopses and ad-hoc reports. The moderator asked them to quantify the degree of difficulty by scoring it with a number from 1 (not difficult at all) to 10 (extremely difficult). The average of their voting was 8.

The participants said that they were very familiar with the listening tasks, since they usually listen to and evaluate the call centre conversations in real time. On the contrary they had less familiarity with navigating collection of recorded calls in order to reply to specific questions or for classifying them (Question 3). In other terms, some of the tasks that were submitted to them were quite novel for them. When the moderator reported the results of the evaluation (Question 4), the answers of the participants stressed again the difference between the tasks based on complete listening of the calls, which they usually carry out with the guide of a behavioural

[6] Initially, seven subjects were selected but for personal reasons a male subject could not be present in the focus group discussion.

evaluation grid, and the tasks based on navigation of call collections (Question 5). However, the participants expressed interest and curiosity for the possible use of SENSEI results (synopses and ad-hoc reports) for completing tasks related with collecting data for reporting the call centre operation to their clients. Finally all the participants said that they would be available and interested in using the system again and for a longer time.

In summary, from the discussion of this focus group, we could identify some concepts that played a critical role in the evaluators' appreciation of the SENSEI speech prototype. The first, and perhaps more important, one is the novelty of the task and the time we left to them to familiarize with the system. They understood the tasks that we submitted to them, but they found that at least some of them were new and difficult. Moreover, the participants expressed interest for using SENSEI to facilitate tasks that require quick and efficient classification, like reports about the argument of inbound calls, the proportion of first call resolution, and so on.

2.7. Summary of findings and future work

The results of this evaluation exercise contributed to shed light on the several variables that may affect the extrinsic evaluation of new technologies. Those variables include the objective performance of technologies and the subjective attitude of participants with respect to novelty of the tools they are asked to familiarize with. That subjective attitude is influenced by factors such as prior experience, ease of use, culture and time left to become acquainted with the new technologies. While the intrinsically subjective nature of such factors may make them difficult to reach, nevertheless future research for identifying moderating factors of acceptance is necessary, as showed also by recent general studies on user acceptance, like [Sun & Zhang, 2006], [Holden & Karsh 2010] among others. The results of the SENSEI speech extrinsic evaluation support the view that the key constructs implied in the acceptance of speech summarization technologies go beyond good levels of technology performance, but include perceived usefulness and ease of use, previous working habits and facilitating conditions.

The results of the SENSEI extrinsic evaluation had to deal with several of such key constructs. Those results allowed answering to the evaluation questions reported at the top of this section. They objectively showed that the different kinds of abstractive speech summaries that the SENSEI technology can generate are useful for improving both the quality of users' tasks, and for reducing the time needed to complete such tasks. In general, the analysis of the different results of this extrinsic evaluation showed that potential users of SENSEI would be interested in using the prototype systems both for tasks that they usually perform in real conditions, and for new possible tasks like listening tasks aiming to assess large sets of recorded calls.

3. Social Media Extrinsic Evaluation

In the past fifteen years there has been a tremendous growth in on-line news and, associated with it, the new social media phenomenon of on-line reader comments. Virtually all major newspapers and news broadcasters now support a reader comment facility, which allows readers to participate in multi-party conversations in which they exchange views and opinion on issues in the news. One problem with such conversations is that they can rapidly grow to hundreds or even thousands of comments. Few readers have the patience to wade through this much content. One potential solution is to develop methods to summarize comment automatically, allowing readers to gain an overview of the conversation. In this project, we have developed two different systems to summarise reader comments. This section describes the extrinsic evaluation of these systems.

3.1. Evaluation Scenarios

Consider the scenario of a reader of on-line news and comment who has limited time (e.g. a 10 minute coffee break) to read some news and associated comment. One possible objective of a reader in this scenario is to obtain an overview of the ensuing debate, i.e. to identify the main issues discussed in the comments and get a sense of the spread of opinion on them. For the purposes of this extrinsic evaluation we have developed a task based on this scenario, in which participants are given a news article to read, and then are asked to use a particular reader comment system to answer a question about the comments:

> Evaluation Task
>
> First, participants are given a news article to read. They then use a reader comment system to access a set of associated reader comments. The task, within a 10 minute time limit, is to use the system to identify and report 4 main issues in the comments.

By issue we mean a *question* or *controversy* that people take a position on. An in-depth analysis of reader comment has shown that for the most part readers exchange viewpoints on issues and they may argue their points. Different comment posters may express similar views and often there are contending views. The task for participants is to make sense of the comments and to identify what it is they are arguing about (i.e. *what are the issues*?).

The task of identifying issues from comments is essentially a “reading comprehension” task. It encourages participants to use a reader comment system in a focused manner. All participants will have the same objective (to identify issues), but they may use the system to help complete the task in whatever way they want to. Participants can provide feedback on their experiences using the system via a post task questionnaire. In addition, by completing the task, participants provide outputs (the reported “*issues*”) which we can assess for “quality” (see Section 3.4 on Evaluation Metrics below). By giving participants a different system in two iterations of the task we can compare their experiences of the systems and compare how effective the different systems are for helping users carry out the task. More details on the evaluation methodology are provided below in Section 3.3.

3.2. Evaluation Prototypes

We developed three systems to evaluate: Systems A, B and C. Each provides a user with access to a comment set via an interface. System A, the baseline system, presents comments as they would appear in a typical reader comment facility. System B and System C, the SENSEI evaluation prototypes, present outputs from SENSEI technologies, (e.g. a set of clustered comments), via User Interface (UI) features such as a clickable pie chart. These features are linked to the original set of threaded comments and provide a very different mode of access to that which the baseline provides. We describe the three systems in more detail as follows:

3.2.1. Three systems for evaluation

System A is based on the features used in *The Guardian* reader comment system. In this system, the source comments are shown in their original threads and are displayed in a chronological order. The user may choose whether they want to see comments ordered by the 'oldest' or 'newest' comments. This system allows the user to view a set of comments with the threads: *expanded*, i.e. all the comments in the thread are displayed, or *collapsed*, only the first few comments in a thread are shown but more replies can be seen if required, or *unthreaded*, i.e. comments are simply listed in order of time of posting, without thread structure. We also display the comment username, and the username of the comment they replied to. Figure 6 shows a screenshot of the Baseline System A interface.

The Guardian (and many other reader comment providers) allow readers to "recommend" comments. Furthermore, The Guardian system allows readers to sort comments based on the number of "recommends". We did not include either the "comment recommends" data or the "sort by recommendation" feature in System A (nor did we include this data or sort feature in System B or System C) in order to reduce the number of variables in the evaluation.

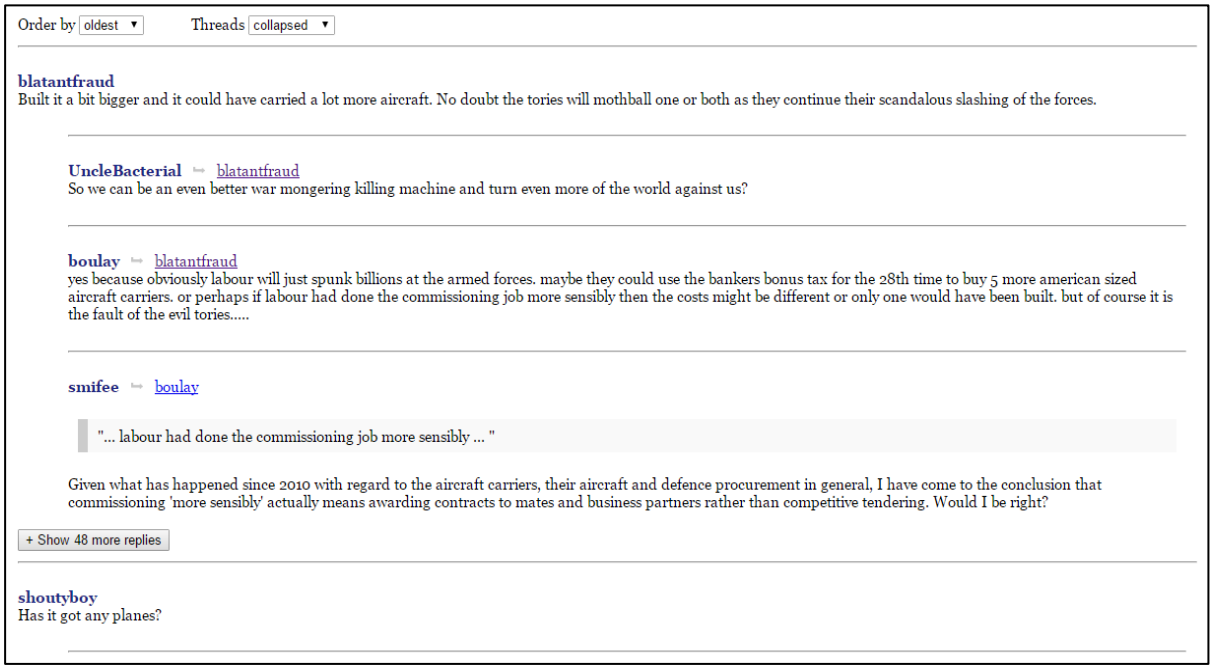


Figure 6: System A interface showing threads collapsed

System B is based on the SENSEI Social Media prototype v1.0 that has been described in D5.2 and evaluated in the M24 evaluation, as reported in D1.3. We have carried out significant changes to improve the system since the M24 evaluation. We report on the current evaluation prototype in D6.3. The system takes the following SENSEI outputs: i) a set of comment clusters generated from the source article and comment set (for details of the source texts see Section on 'Topics' below; the clustering technique is described in D5.3 Section 3.3); and ii) a set of cluster labels, generated for the set of comment clusters (the cluster labelling technique is described in D5.3 Section 3.5), and for each cluster, the most representative quote from the comments in that cluster. Our general approaches to comment clustering and cluster labelling have been reported in [Aker et al., 2016a] and [Aker et al., 2016b], respectively.

This information is presented via an interface, which is shown below in Figure 7. The system has three panels that summarise "***What readers say***". The first panel on the left shows a list of "***topics***" discussed in the comments (i.e. the labels for the clusters). These topics (which represent the clusters) are also depicted graphically in *a* ***pie chart***; the size of each pie segment represents the ***proportion*** of the total number of source comments that are in that cluster. If a mouse is hovered over a pie segment, a pop up window displays the cluster label, e.g. "aircraft carriers" and the ***total number of comments*** in that cluster e.g. "38".

The middle panel shows ***selected quotes*** from each comment cluster. (Colour is used to indicate the different clusters to which a quote belongs.) All ***comments in a cluster*** may be viewed by either clicking on a segment in the pie chart or by clicking on a corresponding selected quote. Following this action the right panel then displays all comments in that cluster. The "***topic***" (i.e. the cluster label) is shown at the top above the comments. If a selected quote is clicked, the ***source comment*** from which this quote has been selected appears in ***pink*** to distinguish it from other comments in the cluster.

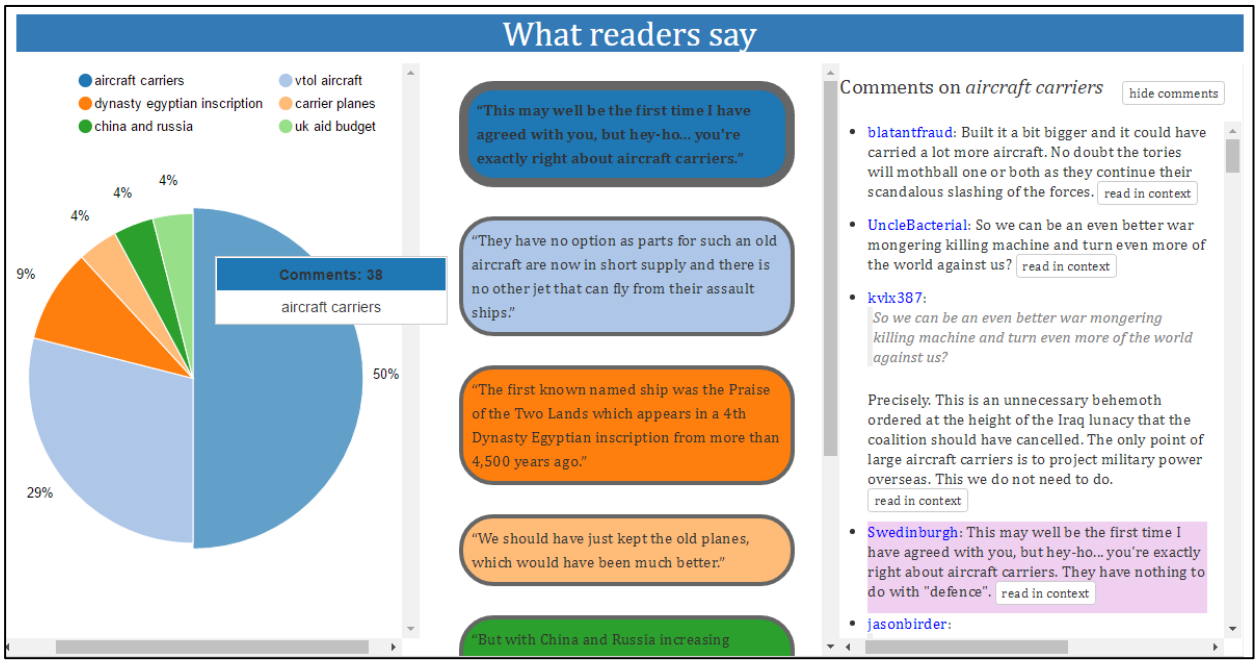


Figure 7: System B interface

Comments in a cluster, as shown in the right hand column, come from various points in the original conversation. Users may read a comment in the context of the original thread by clicking on the “read in context button”. This activates a pop up window, which displays the comment in the full threaded set of comments. The user may scroll up and down to see the full set of comments. The comment that a comment replies to is also indicated here.

In the third system, **System C**, additional inputs such as the template based summary, mood information and agreement/disagreement information have been integrated into the System B prototype (comprising the pie chart, selected quotes and options for viewing clusters of comments). A screenshot of the System C interface is shown in Figure 8, below.

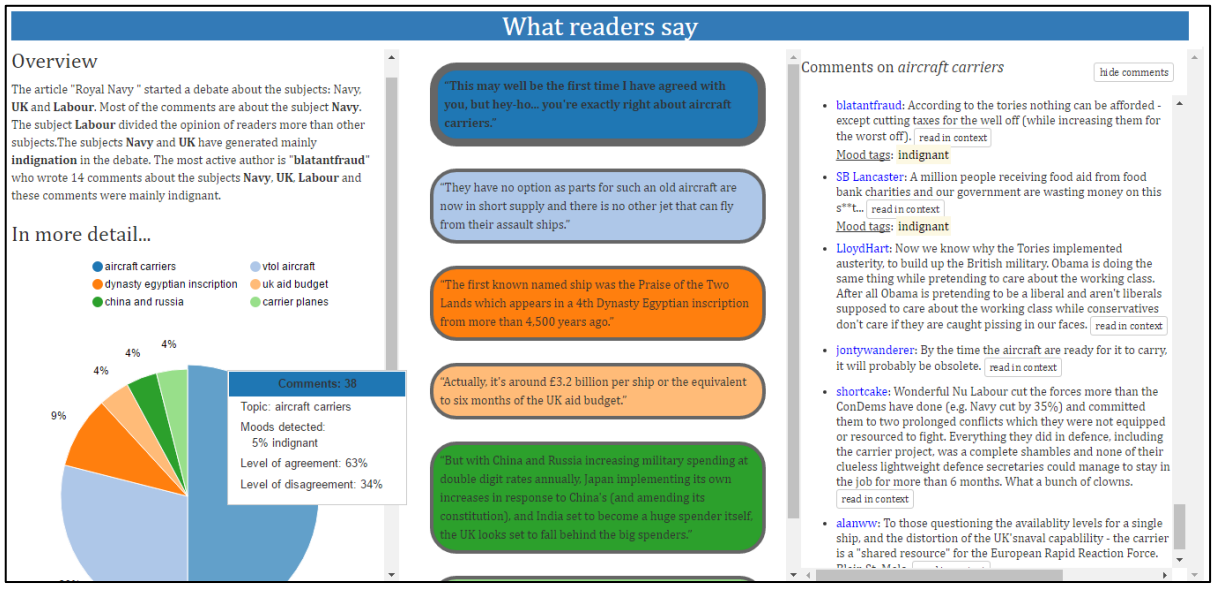


Figure 8: System C interface

The template based summary is generated as described in D5.3 Section 3.8. In the interface we refer to this summary as the “Overview”, and it is presented in the top left hand panel, above the list of topics and the pie chart. It provides: a high level description of the various subjects addressed by the comments. For example, “Navy, UK and Labour”; details such as which subject attracted the greatest number of comments, what subject divided opinion, the moods expressed, and who posted the most comments.

The pie chart in system C is as described for System B, with the clusters generated from the USFD clustering methodology. The labels for each segment in the pie chart, however, also display additional information about detected levels of agreement or disagreement for the comments in each cluster, and also information for moods detected in each cluster. Given the set of source comment texts, the agreement/disagreement information is computed on each comment to the comment it replied to (see D4.3 section 5.1 and D5.3 section 3.2.1). Please note that this feature is different to the “recommendation” information in The Guardian, which instead represents the number of readers that recommended each comment. Taking a set of USFD comment clusters, the agreement/disagreement information has also been aggregated to show this information at the cluster level. Similarly, mood information is detected in each comment in the source texts (see D3.3). This information is further aggregated to show the proportion of moods detected in each USFD cluster.

Finally, we note that the mood information detected for a comment is also displayed for single comments in the clusters shown in the right hand panel.

3.3. Methodology and set up

3.3.1. Overview of experimental set up

The experimental setup used in this final evaluation is similar to that used in the M24 interim evaluation, reported in D1.3 and [Barker et al., 2016], but with some key differences. In this final evaluation the experiments were carried out not in the lab (as they were in the M24 evaluation) but remotely via a purpose built evaluation task interface. This includes a set of training pages for the systems and the tasks. We tested the interface extensively prior to the final evaluation to ensure that the training was communicated as simply and as clearly as possible.

The full task interface can be seen at http:/sensei.group.shef.ac.uk/senseiEvaluation/. We embedded the different systems in the evaluation task interface together with forms for gathering responses from participants. Each participant was to complete two tasks (a task being "to identify and report 4 issues", as described above), using a different system, article and comment set in each of the respective tasks. Before starting the evaluation exercise we asked participants to answer some questions about their background and prior experience of reading online news and comment. We also invited participants who had completed the tasks to provide feedback on the systems based on their experience in the task. More detail on the methodology follows below.

3.3.2. Experiment design

We used an experimental design which was based on a 2x2 "Latin Square", as in the M24 interim evaluation (reported in D1.3). This design allowed for a comparative assessment of two systems -- a baseline (S1) and a SENSEI condition (S2). A participant carries out two iterations of the task, each time using a different system, with a different topic. There are 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. To control for the possible effects of bias due to the different order in which systems and topics were experienced, the design allowed for 4 different orderings of the system and topic. For example, in a 2x2 Latin Square with 4 participants, two participants would use systems in the order S1-S2 and two in the order S2-S1; two participants would experience the topics in the order T1-T2 and two in the order T2-T1. Thus each of the four possible orderings of the 2 systems and 2 topics is considered exactly once.

In this evaluation we took this basic design and increased the number of participants in each topic/system condition to allow us to see variation across individuals (8 participants completed a task in each of the 4 topic/system conditions, a total of 32 participants – see Section 3.3.3: "Participants" below). Ideally we would have also increased the number of topics to explore variation across topics. However given that each participant could only feasibly carry out two tasks, this would have required a further significant increase in the overall number of participants which is not feasible given the resources available for the current evaluation (e.g. while 4 participants are required for each iteration of a task controlling for system and topic order with 2 topics and 2 systems, 12 participants would be required for the analogous setup with 3 topics).

In this final evaluation we used this basic 2x2 Latin Square in **two experiments**, each using the same task and topics, but involving a different pair of systems (see Table 18 below, which shows the resulting composite Latin Square for the two experiments):

- Experiment 1: Baseline (System A) vs SENSEI USFD evaluation prototype (System B)
- Experiment 2: Baseline (System A) vs SENSEI USFD/UNITN evaluation prototype (System C)

Table 18: Composite Latin Square Design Used for Two Experiments, Each Comparing 2 Systems and Using Two Topics

			Task I		Task II	
Combination	No. of participants	Experiment	System	Topic	System	Topic
1	5	1	A	1	B	2
2	5	1	A	2	B	1
3	5	1	B	1	A	2
4	5	1	B	2	A	1
5	5	2	A	1	C	2
6	5	2	A	2	C	1
7	5	2	C	1	A	2
8	5	2	C	2	A	1

We note that a participant could take part only once in either experiment. Also, this design does not allow participants to obtain a direct comparison of the two SENSEI prototype systems. However, since the task, topics, baseline system and general setup were the same in each experiment we can compare how the results from people using the two different prototypes fare against the baseline.

3.3.3. Participants

In this final evaluation we recruited an overall total of 64 participants, with 32 participants completing each of the two experiments (i.e. each system pairing) and 8 participants in each of the topic/system combinations. This represents a significant scaling up of the interim M24 evaluation (in which 4 participants completed the tasks).

We invited people with experience working as media professionals and also members of the public with an interest in online news and/or reading comments. The majority of participants were native English speakers, with others reporting a good to excellent command of English.

3.3.4. Source texts

For the purposes of this evaluation we selected two “topics” (T1 and T2), each comprising a news article from the Guardian and a set of the first 100 comments on this article (as ordered by the time of thread posting and rounded up to the nearest complete thread). We selected these topics because the article and comment sets were of comparable length and did not require specialist background knowledge. The topic T1 included an article reporting on a vote to reduce the frequency of bin collection by a local council and T2 included an article reporting on a

Government fine imposed on the UK rail company "Network Rail" for late running trains. The topics were taken from original article and comment sets, available online at:

(T1) https://www.theguardian.com/uk-news/the-northerner/2014/jul/17/rubbish-bury-council-votes-to-collect-wheelie-bins-just-once-every-three-weeks

(T2) https://www.theguardian.com/business/2014/jul/07/network-rail-fined-50m-pounds-late-trains

Table 19 below shows the summary statistics (i.e. total word counts for articles and comment sets, average number of words per comment, number of threads) for these topics. As the comment average word lengths indicate these were fairly substantial topics with rich contributions from comment posters on a variety of issues.

Table 19: Summary statistics for the two Social Media topics

Topic	Article word count	First 100 comments word count	Average number of words per comment	Total number of threads
Bury Bin Collection	935	5290	52.9	9
Network Rail	730	4,619	46.2	16

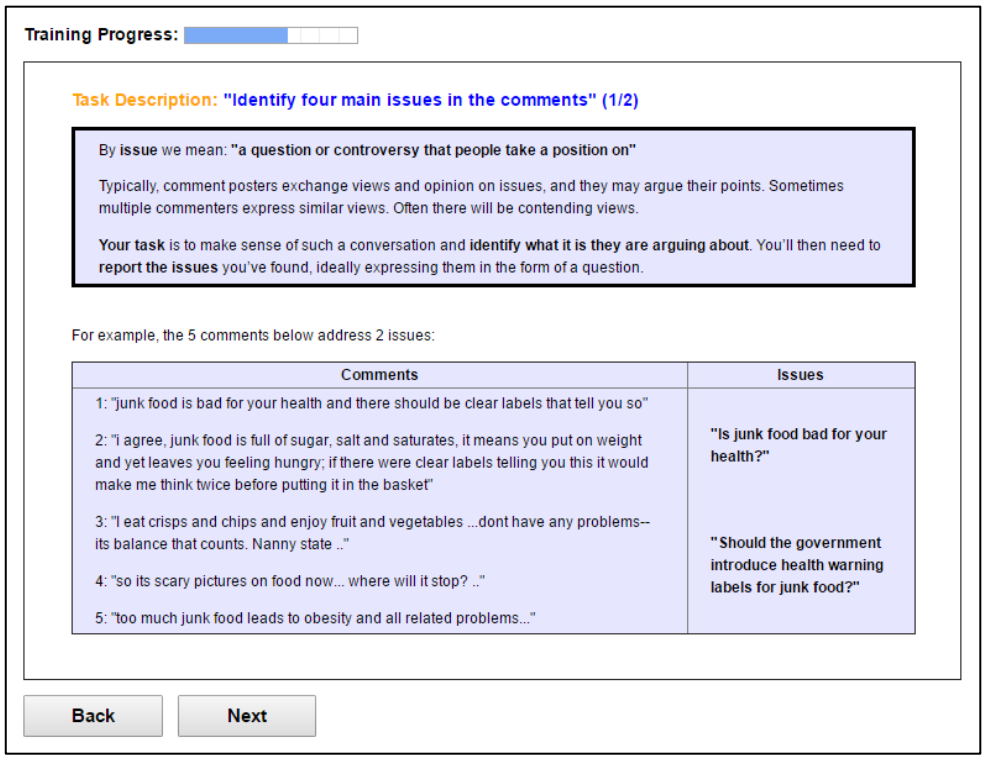
Total word length and thread count calculated based on the set of approximately the first 100 comments (taking the first 100 as ordered by the time of thread posting and rounded up to the nearest complete thread – this resulted in and 100 comments for T1 "Bury Bin Collection"; 9 threads and 100 comments for T2 "Network rail"; 16 threads).

3.3.5. Training

The training pages included a short video demo about how to use the systems to access reader comment:

- System A: https://www.youtube.com/watch?v=kuisdlDSA6E
- System B: https://www.youtube.com/watch?v=gKJ5QYDgsMY
- System C: https://www.youtube.com/watch?v=H3GRjjJVU_I

Participants were also provided with an overview of the evaluation scenario, what's involved in the tasks and instructions on the task question "identify 4 main issues". We show screenshots of the training pages for "identifying issues" in Figure 9 and Figure 10.



Training Progress:

Task Description: "Identify four main issues in the comments" (1/2)

By **issue** we mean: **"a question or controversy that people take a position on"**

Typically, comment posters exchange views and opinion on issues, and they may argue their points. Sometimes multiple commenters express similar views. Often there will be contending views.

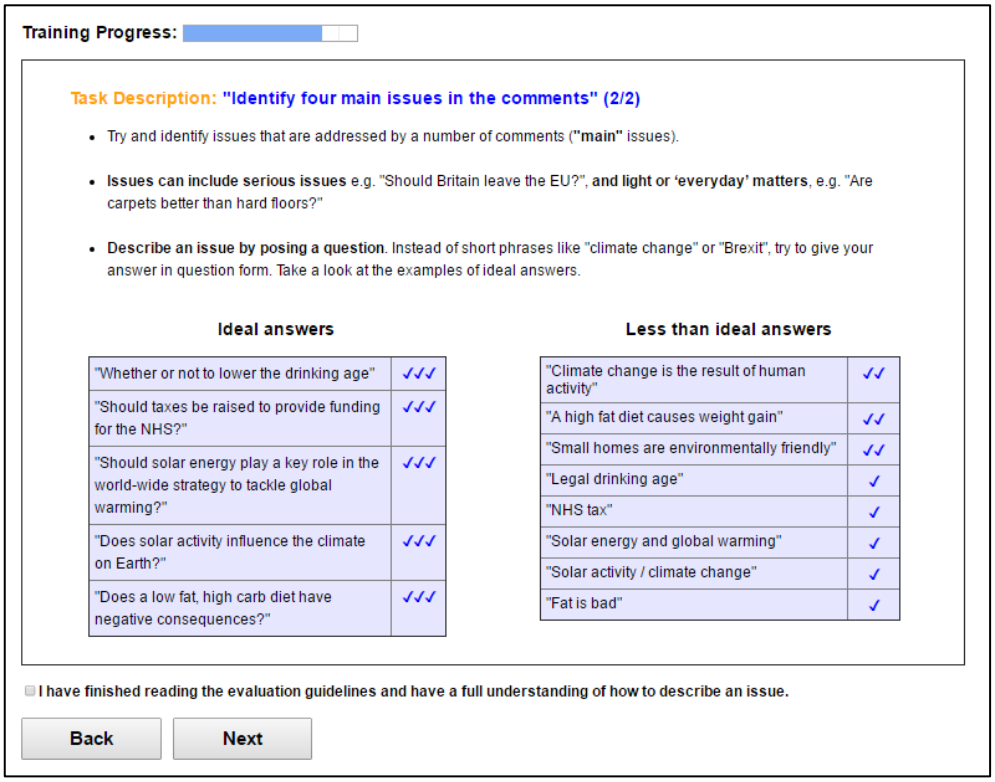
Your task is to make sense of such a conversation and **identify what it is they are arguing about**. You'll then need to **report the issues** you've found, ideally expressing them in the form of a question.

For example, the 5 comments below address 2 issues:

Comments	Issues
1: "junk food is bad for your health and there should be clear labels that tell you so" 2: "i agree, junk food is full of sugar, salt and saturates, it means you put on weight and yet leaves you feeling hungry; if there were clear labels telling you this it would make me think twice before putting it in the basket"	**"Is junk food bad for your health?"**
3: "I eat crisps and chips and enjoy fruit and vegetables ...dont have any problems-- its balance that counts. Nanny state .." 4: "so its scary pictures on food now... where will it stop? .." 5: "too much junk food leads to obesity and all related problems..."	**"Should the government introduce health warning labels for junk food?"**

Back Next

Figure 9: Excerpt of the task training material (1/2)



Training Progress:

Task Description: "Identify four main issues in the comments" (2/2)

- Try and identify issues that are addressed by a number of comments (**"main"** issues).
- **Issues can include serious issues** e.g. "Should Britain leave the EU?", **and light or 'everyday' matters**, e.g. "Are carpets better than hard floors?"
- **Describe an issue by posing a question**. Instead of short phrases like "climate change" or "Brexit", try to give your answer in question form. Take a look at the examples of ideal answers.

Ideal answers

"Whether or not to lower the drinking age"	✓✓✓
"Should taxes be raised to provide funding for the NHS?"	✓✓✓
"Should solar energy play a key role in the world-wide strategy to tackle global warming?"	✓✓✓
"Does solar activity influence the climate on Earth?"	✓✓✓
"Does a low fat, high carb diet have negative consequences?"	✓✓✓

Less than ideal answers

"Climate change is the result of human activity"	✓✓
"A high fat diet causes weight gain"	✓✓
"Small homes are environmentally friendly"	✓✓
"Legal drinking age"	✓
"NHS tax"	✓
"Solar energy and global warming"	✓
"Solar activity / climate change"	✓
"Fat is bad"	✓

I have finished reading the evaluation guidelines and have a full understanding of how to describe an issue.

Back Next

Figure 10: Excerpt of the task training material (2/2)

3.3.6. *Data gathering*

We collected data from participants in three stages:

Pre-task questionnaire

We collected information about professional background, English proficiency and experience using reader comments before training, via a "pre-task questionnaire". This is shown in full in Appendix B.

Reporting Issues

Figure 11 shows the interface for collecting responses from the evaluation question "Identify 4 main issues". Having started the task via a Start button provided by the interface, participants were given 10 minutes to identify the issues and report them in the form shown at the top of the page. Participants were allowed to continue to the next page before the time was up. If the time ran out before participants completed four issues, their answers were automatically saved and they were referred to the next page. We recorded the time taken to complete each task.

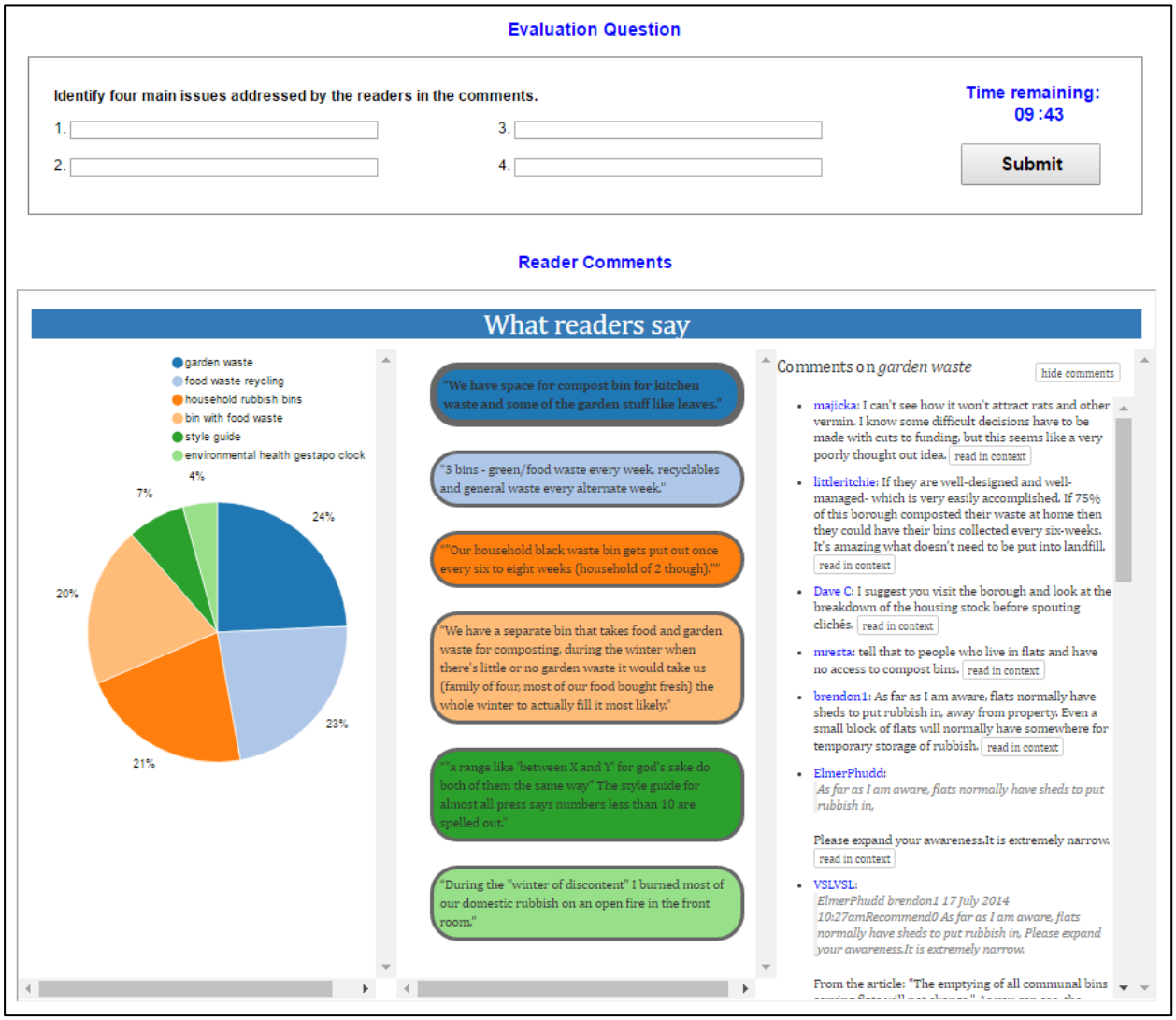


Figure 11: Evaluation question interface

Post-task questionnaire

We invited participants to provide feedback on the systems and their features via a short "post-task questionnaire". This is based on the questionnaire we used in the interim evaluation, but with a few modifications, e.g. to the list of specified system features, to reflect the updated system prototypes. We also added a new question (Question 2, the overview question) and a comment box so that participants could provide a reason for their rating in question 3. The questions can be summarised as follows (the full questionnaire is in Appendix C):

- A multi-part question asking participants to rate on a scale of 1-5 how useful they found each of the respective system components (e.g. the pie chart, the selected quotes etc.), and each system as a whole, when completing the experimental tasks. In addition a comment box is provided for additional feedback on each of the features/systems.
- Participants can indicate which, if any, of a set of listed system features provided an "overview" of the comments.
- A more general question asked participants to indicate on a scale of 1-5 the extent they would like to have a SENSEI type system available for future use in a reader comment facility. This question included a box for participants to explain why they gave the rating they did.
- In addition we invited them to provide any other comment they had about the systems and their experience in the task. (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 ...").

3.4. Evaluation Metrics

We scored participant responses to the issue task using our graded scheme (which we first applied successfully in the interim evaluation, D1.3). One assessor carried out the assessment of the issues. (A second assessment by a different assessor is currently underway and will be reported in a journal paper after the end of the project.) First, we gave the assessor the topic texts (i.e. the news article and the comment set) and asked them to familiarise themselves with the topic prior to scoring the issues. We also provided the assessor with the "issue definition" and examples, as given to participants via the task interface (see Figure 9 and Figure 10 above).

We scored each issue on a four point scale (ranging from 0-3). The 4 point scale 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?). Guidelines for assessing issues are shown in Table 20. We added the scores for the four issue responses together, giving a total for each task. We also calculated a participant average for each task and a total average across both tasks (a task involving a different topic/system pair).

Table 20: The guidelines for assessing issues

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, (i.e, vaguely) but some content is indicated and the comments can be seen to address it, for example, a response such as "ticket prices" where there is evidence of people talking about different things to do with ticket prices in the comments receives a score of 1. The same score is given if the issue is

	more clearly articulated e.g. as a proposition, but is poorly evidenced, e.g. only 1 or 2 comments discuss the issue.
Score 2	The issue is adequately expressed e.g. “we should fine the directors”, but one could imagine the space of possible positions being more clearly indicated e.g. a “would fining the directors be an effective way of ensuring trains run on time?”. The issue should be of sufficient clarity to assess evidence or strength of support in the comments, which should be good or satisfactory. A score of 2 is also given to a well-articulated issue but with a low level of evidencing, say 1-2 comments, or when there were many other candidate issues to choose from, which were much more significantly discussed.
Score 3	The issue is clearly articulated/expressed in the form of a question; so it is straightforward to assess evidencing/strength of support, which is good (relative to the overall discussion in the comments).

Time to complete the issue task

We have recorded the time taken by participants to complete the task and this information can be aggregated and averaged for the different system conditions, and for the participants. In addition we can use the time to complete data in conjunction with the overall score for issues to compare system effectiveness. E.g. A comparatively high issue score with a low time to complete would suggest a system helps participants to carry out a task.

Post-task questionnaire

The text responses gathered in the post task questionnaire are analysed using simple qualitative techniques. Data from the user ratings of the different systems/system components are summarised using simple statistics.

3.5. Evaluation Results (Experiment 1: System A and System B)

In this section, we discuss the results of Experiment 1. First, we describe the background of participants. We then report on participant performance in the reading comprehension task. Lastly, we report on our analysis of participant feedback provided in response to the post-task questionnaire.

3.5.1. Participant background

Almost all participants in Experiment 1 were native speakers or fluent in English, 1 had a good command of English. Over two-thirds (69%) had experience working as a media professional.

Participants varied in terms of how often they read or post comments. Almost 1/3rd were regular visitors to comment, engaging on a daily basis; a larger group, 41%, reported engagement at a rate of at least once a week, (so, 72% engage at least once a week with reader comments, nearly half of this group engaging every day). Only 25% of participants were occasional users of comment (reporting engagement as “at least once a month” or “very rarely”), and one annotator had never engaged with reader comments. We show this information in Table 21.

Table 21: Participants background (Experiment 1)

1. English language proficiency	Native speaker	30	94%
	Near native / fluent	1	3%
	Very good command / highly proficient		0%
	Good command / good working knowledge	1	3%
	Basic communication skills / working knowledge		0%
2. Media professional experience	Yes	22	69%
	No	10	31%
3. Engagement with reader comments	At least once a day	10	31%
	At least once a week	13	41%
	At least once a month	2	6%
	Very rarely	6	19%
	Never	1	3%

3.5.2. The reading comprehension task: Identify 4 Issues

For each topic, each participant was asked to identify four issues raised in the reader comments in a fixed time period (10 minutes).

On average, participants took slightly less time to complete the task when using System A: a total average task time of 6:39 minutes was found for System A and 7:17 minutes on average for System B (a difference of 38 seconds). We found a strong correlation between the times spent on System A and System B (Pearson correlation coefficient = 0.71, $p<0.01$).

Taking the results for each system in turn, there was very little difference in the time taken to complete a task in the two topic conditions (Table 22). There was a slight difference of 16 seconds between topics for System A tasks, but just 2 seconds difference on average between topics for System B tasks. The slight difference in the average task times for the two Systems can possibly be explained by the fact that there were more levels of information to explore in System B (which included the pie chart and the pop-up windows; the selected quotes and comment clusters; in addition to the option for viewing comments in the original threaded comment stream). But also possibly because participants were more familiar with the type of system we present in System A, they were able to get going on the task more quickly, while in System B they perhaps required a little extra time to become familiar with the various features and functionality.

Table 22: Time spent to complete the task

System Name	Topic Id	Average Time Spent
A	1	6:47
A	2	6:31
B	1	7:16
B	2	7:18

In using System A, 31 of 32 participants answered all the four issues, whilst one participant only identified three issues. When using System B, 30 participants identified four issues in the given time, whilst the remaining two submitted three issues. The issues were scored by an annotator using the metrics described above (unanswered issues were scored 0). We show the average scores for the 32 participants when using the different systems in Table 23.

Table 23: Average scores

System	Overall	Topic 1	Topic 2
A	2.36	2.28	2.44
B	2.30	2.44	2.17

We show that participants were able to achieve similar scores when using both System A and System B, with the average scores only differing by 0.06. The two-tailed P value equals 0.6678, which indicates that the results are not statistically significant. 95% confidence interval of this difference is from -0.198861 to 0.308236. The intermediate values used in calculations are t equals to 0.431 and df equals to 61.95.

When analyzing the differences across topics, we found that annotators were able to identify higher quality issues in Topic 1 (Bury bin collection) when using System B. However, when identifying issues in Topic 2 (Network Rail), annotators were able to do that better when using System A.

One possible reason for this is that the threads in Topic 2 are very topic-focused, i.e. each thread discusses a particular topic. Annotators, therefore, could identify the topics easily by simply following the threads. Meanwhile, some of the threads in Topic 1 discussed many topics at once, making it harder for annotators to identify the issues by relying on the threads (System 1). The features provided in System B were able to present the issues better to help participants identify and list the issues in the reader comments.

3.5.3. Post-task questionnaire

We asked participants to provide answers to four questions: (Q1) to rate the level of usefulness of each of the systems and the different system features in the task, (Q2) to identify features that helped to provide an overview of reader comments, (Q3) to indicate how much they would like to have System B available more generally for future use. And a final question, (Q4), asked them to provide general comments on the Systems and their experience in the tasks. We also provided participants with the option to comment on the systems and/or system features listed in question 1, and in question 3, the “future use” question, we asked them to explain their score. We now report on the feedback they provided in response to the questionnaire.

Q1. How useful were the different systems/system features when completing the task "identify four issues"?

Participants assigned a score, using a 5-point Likert scale, to each System and to each feature listed for System A and System B. We show the average scores for the respective Systems and features in Figure 12.

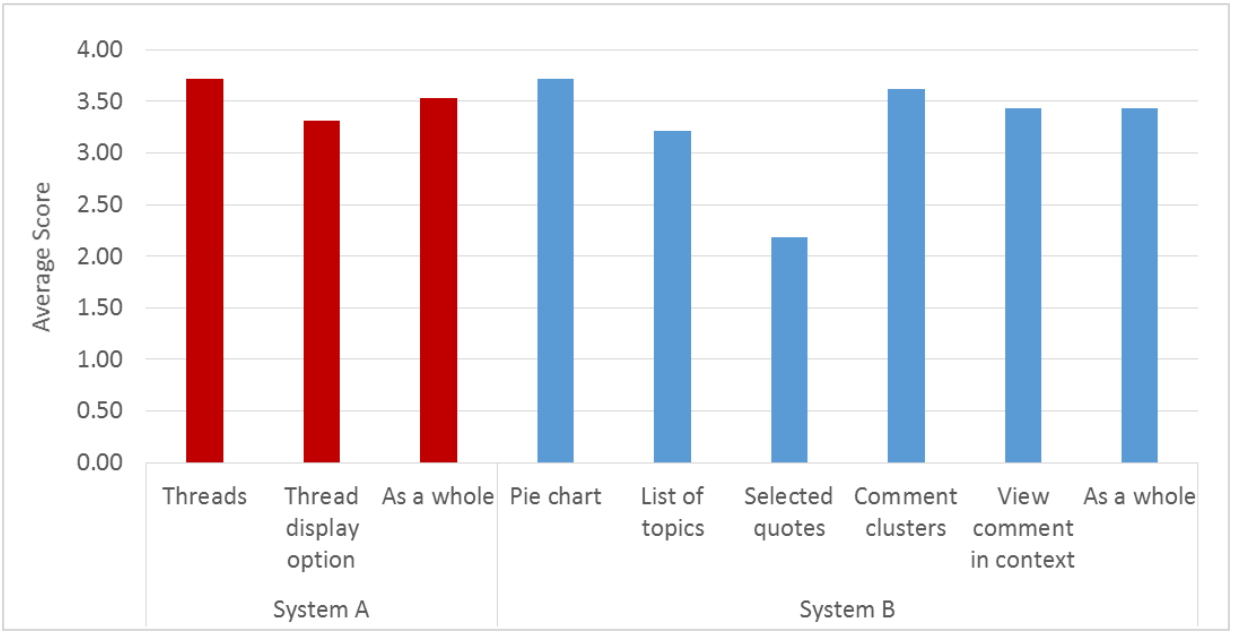


Figure 12: Usefulness of Features in System A and System B

The total average ratings for Question One show that System A and System B were assessed to be of a similar level of usefulness, with both systems and nearly all system features receiving positive scores, i.e. a score greater than the scale mid-point of 3.

The System A "threads" were rated fairly highly with an average score of 3.72 out of 5. The thread display option, i.e. the option to collapse or expand a thread, was given a slightly lower score of 3.31. The average rating for System A, "as a whole" was 3.53, which suggests that the participants found it fairly useful for the "issues" task.

The various features in System B, with the exception of the selected quotes, received similar ratings to those of System A, with most features in System B receiving an average score of above 3.2. A key finding was that on average, participants found the "pie chart" to be as helpful as the "threads" for the task of identifying issues: both the pie chart and threads received an average score of 3.72. **However, this average for the pie chart can be adjusted to 3.81, i.e. a higher rating than that given to the threads**, if we exclude the rating of one participant, an outlier who reported that they were unable to load the pie chart when viewing the system7. The "comment clusters" (i.e. the groups of topically related comments) were also rated fairly highly, with an average score of 3.63. Participants felt less strongly about the usefulness of the "selected quotes" feature, which scored an average of 2.19 out of 5. We elaborate on why this was so below in the summary of the qualitative responses.

The overall scores for each system (i.e. the ratings for the system "as a whole") were also very similar: System B having an average of 3.44, just slightly lower than System A's average of 3.53.

- Do Participants Prefer One System or Another?

[7] Their comments on the respective features in System B suggest that it was just the pie chart that did not load and therefore they were able to carry out the task using the other features and provided scores accordingly. We have included their ratings for the other features and System B in this results report.

We examined the global average ratings for System A and for System B "as a whole" more closely by arranging the individual participant ratings for each system into one of three categories: low = a score of 1 or 2; medium = a score of 3 and high a score of 4-5. There is a pair of scores (one score for each system) from each participant and each respective score is placed in a category (low to high). We wanted to see **whether or not people favoured one system or another**, so we arranged each pair of scores in a table (see Table 24), which shows the respective ratings for each system distributed in the 3 categories.

Table 24: Participant ratings pairs for systems A and B in 3 categories (low, medium, high)

System B Score	System A Score (from scale 1-5)			Total
	Low (1, 2)	Med (3)	High (4, 5)	
Low (1, 2)	0	0	6	**6 (18.8%)**
Med (3)	1	0	8	**9 (28.1%)**
High (4, 5)	3	10	4	**17 (53.1%)**
Total	**4 (12.5%)**	**10 (31.2%)**	**18 (56.3%)**	**32 (100%)**

The table shows that both systems received a relatively small number of low ratings: just 12.5% of ratings for System A and 18.8% of ratings for System B were less than 3. We can also see clearly that the majority of ratings for both system A and B respectively were in the high category (i.e. in the case of both systems, the majority of ratings were scores of 4 or 5).

Interestingly, there were no cases of a participant giving both systems a low score of less than 3. In fact the majority of people who rated one system low gave a correspondingly high rating to the other system: 9 of the 10 who gave a low rating to one system gave a high score of 4 or 5 to the other system. (There was just one exception who gave system A a low score of 2 and system B a score of 3). Furthermore just 4 of the 32 participants were to give a high score to both systems. There were no cases where people gave both systems a medium rating of 3 and all except 1 of the 19 participants who gave a medium rating of 3 to one system gave a high level rating to the other. **These results suggests a pattern of polarization**: while most participants found one system very helpful, they were unlikely to have found the other system as helpful and, if they really didn't find a system to be helpful then they were most likely to have found the other system very helpful.

Question 1: Qualitative Responses

Around half of all participants (15 out of 32 for System A and 18 out of 32 for System B) provided additional "free text" comments on the usefulness of the various Systems and System features. (The full set of responses to the question 1 comment option can be viewed at http://sensei.group.shef.ac.uk/extrinsicEvaluation2016-results/sheffield.php). This qualitative data provided some insights into what people liked or disliked about the two systems. We summarise the findings for key System features as follows.

- Pie Chart

Many participants liked the pie chart because it was useful for quickly seeing what readers are talking about and what was most talked about. For example, there were remarks such as:

> *"V helpful seeing the **main themes**"; "Useful to see **broad areas** of comments";*

*"Useful to get an understanding of the **types** of comments";*

*"Super useful. I can tune into **broad issues** quickly and easily";*

*"It was useful to see the **major areas** of conversation".*

*"This is a great fast way to see what is **concerning people most**."*

*"Useful to see ... what is **popular**."*

*"Useful for identifying "**most talked about**" issues".*

A few comments elaborated on why the indicative qualities of the pie chart were useful. One noted how it could help to decide if a comment discussion was worth looking into at all. Another (media worker) suggested that by identifying broad areas and what is popular in the comments, the pie chart "could help to lead to future stories on the areas".

Not all comments were entirely positive however. There were a few who while liking the idea of the pie chart remarked on problems with the segment labels, one suggesting that they didn't "fit" the comments and two others noting that there was overlap between the segment labels:

"Some of the segment topics overlapped almost to the point of being synonymous."

"A good idea but I didn't find the words used to separate the pie slices to be that useful. In fact, they seemed quite similar - especially when working at speed".

Nonetheless, many liked how they could click through from a pie chart segment to explore the relevant comments. And there were several comments that recognized how the pie chart and comment clusters worked together to provide an overview.

- Comment Clusters

The majority of comments on the "comment clusters", (8 out of a total of 10 comments), were very positive and suggest that people found the comment clusters easy to use, and made good use of this feature in the task. For example, the feedback included the following remarks:

"Read these a lot and were logical"; "Spent majority of my time reading these";

"Allowed you to quickly skim read comments";

"More friendly to the user than the threaded comments";

"Easy to use and then expand";

"A good way, with the pie chart, to get an overview and then a tradition thread system if needed";

"This is more useful - it's similar to the collapsing threads feature and allows a broad brush overview without too much clicking".

The latter two comments suggest that the comment clusters helped some people to gain an overview. And a couple of the remarks above suggest that people were able to follow up the comments in the full comment stream, by using the "comment in context" function.

Again however, there were some diverging views. Some other participants clearly found it difficult to work with the comment clusters: one found it "the most confusing" and another reported that:

"Clicking through from a selected quote, I found it difficult to follow the conversation, where the comment had originated, what other comments it referred to. I did a lot of scrolling up and down, which felt like wasted time".

In System B, the original conversation structure is typically lost in a comment cluster, (the clustering algorithm can select individual comments from different points (and threads) in the comment stream and sequences of comment replies are not necessarily found together in a cluster). This means there is often a lack of rhetorical coherence or "flow" across the comments in a cluster and this may create difficulties for a person reading through a set of individual comments. Also, some comments may be particularly difficult to comprehend when viewed out of conversational context due to anaphor and ellipsis. We did provide the functionality view "comment in context" to address these problems, but some users evidently found this difficult or inconvenient to use. We elaborate on this issue further in the report on the question 3 results, below.

- Comment in Context

The majority of participants who left a comment approved of the view "comment in context" feature. For example, comments included:

> *"Probably the feature I used most in this system!"*
>
> *"Helpful to see the full conversation".*
>
> *"Brings further information and related comments which was helpful".*
>
> *"Really useful".*
>
> *"Good.".*

A couple of participants found it difficult to navigate: finding where the original comment was in the full conversation stream was one difficulty. One suggested this was such an important feature that it should not be in a separate page. (We presented the comment stream in a pop up window). But two other comments stated they did not use this feature.

- Selected Quotes

With a couple of exceptions the majority of comments on the "selected quotes" were critical of the feature, which fits with the low average feature rating (2.19 out of 5) we reported above.

The main concern was that the quote was just a quote and not a full comment – i.e. that it was presented out of context, and this lead to two problems. Firstly, as one said it had the "potential to misrepresent posters' views". And related to this, one speculated that comment posters could exploit the selection process to promote their own interests:

> *"I'd worry this could become a target for spam or trolls. They could end up having their banal tosh featured here if they were persistent or prominent enough. Though a lot of this depends on the security of the comments section, of course."*

Secondly, it could be difficult to *understand* a quote on its own. To understand it meant clicking through to see further context, and some did not find this to be a straightforward process. (To view the source comment for a selected quote a participant had to click on the quote, the source comment then appeared within the group of comments in that topic, in the right hand window. The source comment for the selected quote was distinguished by a pink background. A user needed to click again to "view in context".)

A couple of participants did not see that it added anything to the pie chart functionality. Two others observed that the quotes did not fit the parent topic. There were also several participants who voiced concerns about how the comment was selected:

"All the same concerns as above - who chooses these quotes?".

Not all comments were negative however, one clearly said they "enjoyed having highlighted comments" and another that it was a "nice feature" and a couple of others said that it gave a "snapshot" of "opinions on key topics" or "a comment"; but this was qualified by the fact you had to click down to understand it, which was not always straightforward.

-Threads

Opinion on the threads was divided. Some commenters liked the presentation of comment in threads because it reflected the natural organization of the conversation and allowed them to follow the flow of the discussion across comments. E.g. one commenter reported that it was:

"Very easy to follow a conversation developing between one or more people and follow it like a normal conversation - i.e. reading people's posts and then the replies in chronological order".

Other commenters however reported that it was *difficult* to follow arguments in threaded comment:

"Readers of comments get lost in regular commenters discussing things. There's a lot of arguing that is difficult to wade through".

"They [the threads] were helpful but topics jumped around a lot so chain of thought / responses were difficult to follow".

Finally, while one acknowledged that the threads were easy to read, they said they were not well suited to identifying issues.

This pattern of divided opinion continued across the comments on the two systems, when each was viewed as a whole.

- System A as a whole

The thirteen comments on the usefulness of System A were very clearly divided. On the one hand were those who liked System A: they liked the simplicity of presentation, its familiarity; found it easy to read comment and to navigate and liked that it allowed the user to follow the natural flow of the conversation. For example:

"Easy to use in general - am very used to reading in this way.

"Seeing the 'first' comment then subsequent replies in one place allowed me to follow 'conversations', which made sense in context. Then following on with further comments from oldest to newest allowed me to make sense of what was being said chronologically - I followed the comments in the same order as the commenters. "

"Perhaps it's because I'm used to this system but I actually found it easier to identify key topics by skimming through the comments and picking my own key words".

"Overall a good system, intuitive and easy to use and especially like being able to easily read 'conversations' between two or more people posting repeated replies to each other".

By contrast, around half of the comments were highly critical of this system, citing reasons such as it was not user friendly and was difficult to follow related content:

"Unfriendly and difficult to navigate";

"I find thread systems like this a little bit of a drag, especially if you've expanded the threads, found one that's a bit dull, and then have to scroll all the way past it to find another decent thread. The expansion of them all is irritating".

"It is clunky way of reading comments and can be time-consuming when you're having to skim read past random off-topic rows but because it is what I'm used to I find it OK to navigate".

One comment while finding System A easier to use acknowledged that it doesn't provide an overview. And another suggested a compromise to address this, which involved integrating elements of System B.

- System B as a whole

Of the ten comments on System B, most were encouraging and some of these were very positive about its usefulness, indicating preference for System B over System A. Among those who liked it there was some agreement that it helped to provide a quick digest or overview of the main content:

"Much better. Easier to digest and get a feel of the general vibe/argument being made. Easier to compare and contrast different views".;

".. But it does seem much more user friendly. Especially if there was a particular angle you were interested in or you didn't want to keep scrolling down to get a quick handle on the range of opinion".

The more moderate feedback was positive about the idea of having some of the functionality, but there was a common feeling that it was too complex for their requirements, and possibly too time consuming to use:

"I think in theory it's good but to work at speed with it there was too much going on. I think the pie chart + comments on the right would be enough for me.";

"A more engaging and useful system but almost provided too much info. I'm not sure how often I would click onto the pie chart and read the right hand column as it's quite time consuming. I'd probably only go beyond the pie chart on issues that interested me a lot.

A few participants found that it took a while to get used to the presentation of information in the new system:

"It took me a while to figure out the presentation and some of it seemed to be quite repetitive (i.e., comments on same thing in different categories)".

"I liked it, but not sure how helpful the middle column was. I still find the older system easier but I think that is because I am used to it."

There was also a small but definite minority who stated that they found System A (the thread system) much easier to use. They echoed the concerns above about the complexity of presentation in System B:

"Not easy on the eye - much preferred the text only system for ease of use."

"I found it very confusing, but this may be because I am not used to it".

To sum up the results from the Question One qualitative remarks, many people liked System B and found it useful, but there was a sub-group who while liking the idea of the System B topical overview, had concerns about the overall complexity of System B interface, which was not suited to everyday, time-limited reading activities.

We further analysed how the usefulness scores of these features vary between participants in the different system-topic combination (as described in Table 18). These data are shown in Table 25 (System A) and Table 26 (System B). Not much variation was observed when analysing the feature usefulness across the different topics in System A and System B. This suggested that these features were perceived as helping in a similar way across the different topics.

Table 25: Usefulness of System A Features (across topics)

Topic	System A		
	Threads	Thread display option	As a whole
All	**3.72**	**3.31**	**3.53**
1	3.88	3.56	3.63
2	3.56	3.06	3.44

Table 26: Usefulness of System B Features (across topics)

Topic	System B					
	Pie chart	List of topics	Selected quotes	Comment clusters	View comment in context	As a whole
All	**3.72**	**3.22**	**2.19**	**3.63**	**3.44**	**3.44**
1	3.81	3.06	2.31	3.44	3.38	3.38
2	3.63	3.38	2.06	3.81	3.50	3.50

Q2. Please tick the features or systems which in your opinion helped to provide an overview of the reader comment discussion. (1="yes" and 0="no")

In Q2, participants were asked to select one or more features that they thought were useful in providing an overview of the discussion. We show these results in Figure 13.

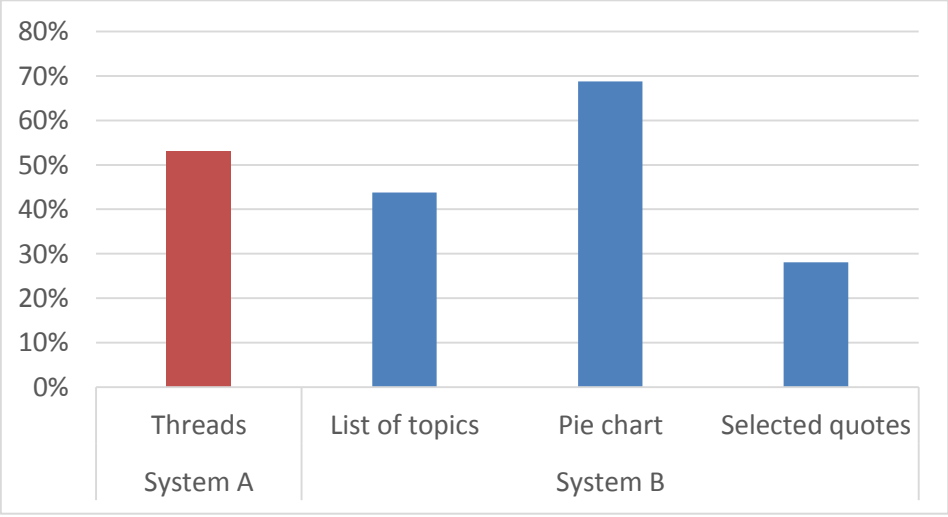


Figure 13: Does a feature give an "overview"? (Results show proportion of participants responding "yes".)

Over two-thirds of participants (69%) specified that the "pie chart" helped provide a discussion overview. Again, we can adjust this total: if we exclude the rating of the participant who reported

that they were unable to load the pie chart when viewing the system, the total percentage rating the pie chart as an overview is 71%. Around half of the participants selected the “threads” feature, and slightly fewer of those thought the list of topics were also good in providing an overview. Only 28% of participants judged the “selected quotes” to provide an overview.

When analyzing them across the different topics (shown in Table 27), we found that a much higher percentage of participants selected the “pie chart” as providing an overview when they had used System B to carry out the task in Topic 1 (Bury Bin Collection), compared to those who used System B for Topic 2 (Network Rail). Not much difference, however, was observed across the different system-topic combination for the rest of the features.

Table 27: Proportion of participants preferring each feature (across topics)

Topic	System A	System B		
	Threads	List of topics	Pie chart	Selected quotes
All	**53%**	**44%**	**69%**	**28%**
1	56%	44%	81%	25%
2	50%	44%	56%	31%

Q3a. How much would you like to have System B (Text and Graphics Reader Comment System) available in a news and comment browsing facility?

Participants were asked to indicate on a 5-point Likert scale how much they would like System B available in a news and comment browsing facility. The average score across the 32 participants is 3.19, showing that the participants do like System B and that it has a potential for further use in a news and comment browsing facility.

When looking at the distribution of participants across the different scores (shown in Figure 14), over 45% participants provided a high score (of 4 or above) on System B, compared to around 30% participants who gave a lower score (of 2 or lower). We further investigated this pattern by comparing their scores with the question 1 system/system feature ratings data.

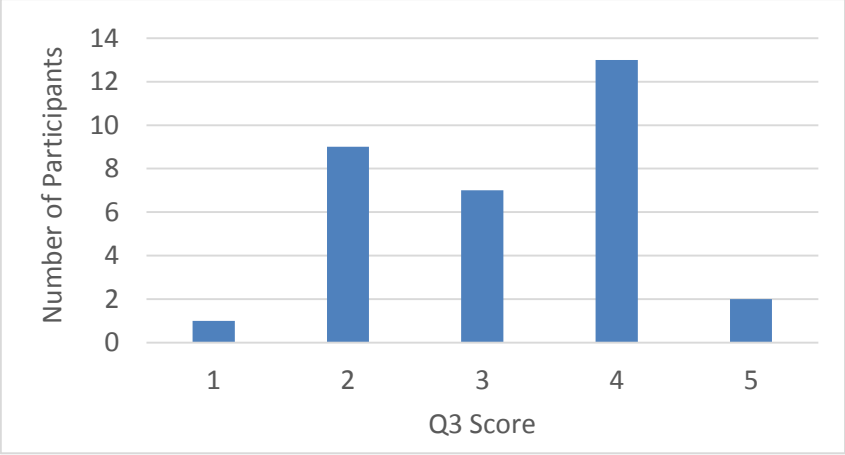


Figure 14: Distribution of Q3 Scores

- Is there a correlation between how the participants rated the systems/system features and the rating they gave to the future use question?

We investigated how well the ratings given by participants in response to Q1 for specific system features correlated with the scores they gave to the future use question, Q3. We calculated the Pearson correlation coefficient r of the participant scores[8] for each of the Q1 features and systems as a whole as compared with the participant scores for Q3. The resulting coefficients are shown in Table 28. Pearson correlation coefficients of .40 to .59 are generally interpreted as "moderate", those between .60 and .79 as "strong" and those of .80 to 1.0 as "very strong".

Table 28: Correlation between participants' ratings of system features and their score for how much they would like the system available in a future news and comment browsing facility

System	System feature(s)	Pearson coefficient
A	Threads	-0.44
	Thread display option	-0.16
	As a whole	-0.70
B	Pie chart	0.68
	List of topics	0.45
	Selected quotes	0.19
	Comment clusters	0.43
	View in context	0.38
	As a whole	0.80

Most notable in these results are the strong correlation (r=0.68) between the ratings for the pie chart and future use and the very strong correlation, (r=0.80) between the ratings for System B overall and future use. There is also a strong negative correlation (r = -.70) between the rating for System A overall and the rating for future use of System B and a moderate negative correlation for System A threads (r=-.44) with future use of System B.

These results suggest that participants who liked the pie chart were quite likely to say they would like system B in a future reader comment system or conversely that for participants who would like system B in future it was the pie chart feature more than any other feature that they rated most highly about system B in helping them to complete the task. Also, perhaps unsurprisingly, those who more highly rated system B overall were more likely to want it for future use. Finally those who more highly rated system A overall were less likely to want system B in future. This result fits with the pattern of polarization we reported based on the respective system ratings for Q1.

Question 3b: Qualitative Responses

Participants provided a comment to explain their score for having System B (the SENSEI system) available for future use in a reader comment facility. This was a mandatory field and we

[8] When calculating the correlations between the two sets of results for questions 1 and question 3 results, we excluded the scores for q1 and 3a from the outlier participant who reported that they could not load the System B pie chart.

obtained 32 comments to accompany the 32 scores for “future use”. In addition participants provided 17 comments in question 4 (further feedback). We carried out a lightweight thematic analysis of this data, one researcher adding a short interpretative comment to each participant response, and where possible using common descriptors to identify recurring themes. This analysis suggested a number of **main themes**, and the results strengthen the initial findings we reported for the more partial question 1 comments.

- A topical overview and filter for reader comment

The majority of the Q3b responses (18 out of 32) remarked on the “topical overview”, that System B provides, their comments indicating that this kind of feature was helpful to have and/or that they liked it. Many of these 18 singled out the pie chart as a useful feature and several stated that the pie chart or overview helped them to follow a topic or issues of interest, indicating that they liked the idea of filtering comments by topic. These results correspond to the strong correlation we found between the Q1 pie chart ratings and the Q3a ratings. People who liked the pie chart were likely to want System B for future use. Comments on the System B style overview or pie chart include, for example:

> *“A good visual aid, helps group comments by topic, thus allowing the reader to remain on a topic at a time rather than context-switch based on comment chronology”.*
>
> *“It provides both a clear overview of a topic and allows you to delve deeper into issues of interest”.*
>
> *“The pie chart gives a quick, easy to understand overview of the main points of discussion”.*
>
> *“I really liked the pie chart because it summarised the type of comments made, which helped me assess the comments quickly and see how people had responded to the article.”*
>
> *“It would really help reflect the audiences' viewpoint quickly and comprehensively. We're always second guessing what our audience thinks, but this tool makes it much easier to understand.”*
>
> *“Very good visual presentation and the pie chart would let me know if I really want to read all of the comments or not”.*

Figure 15 shows the respective scores for Question 3a: “would like to have System B”, for the 18 participants who commented on System B providing a topical overview. They clearly gave a positive rating for future use of System B, an average of 3.7 out of 5 (recall the average for all participants was lower, at 3.19).

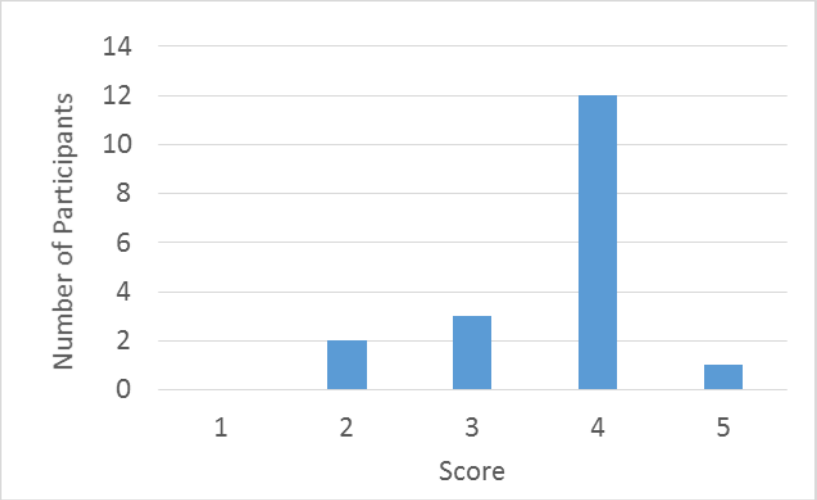


Figure 15: Q3a scores for participants who remarked on the System B overview in Q3b

A small minority of comments (4 out of 32 responses, question 3b) suggest that some people did not require an overview when reading comment as it wasn't something they were interested in. One said they preferred to read just one or two "quality" comments:

> *"… Not sure how interesting the pie chart is really - in my opinion 1 'quality' well-argued and well- written comment is worth reading more than 100+ rubbish comments, so merely presenting the number of comments/proportion of comments on a topic isn't particularly useful, the guardian "pick" comments that are highlighted at the top of the page is a better system in my opinion".*

Others stated that they preferred simply to read some of the conversation:

> *"Just think this system is way too complex, when I'm reading news, I don't really care about the percentage of the comment topic, I just want to read some conversation, that's enough".*

> *"I am not really sure how interested I am in evaluating what everyone thinks about a particular article. I often skim the comments just to get a snapshot of some different views, but think a list view does that quite well (particularly if you collapse the threads where people tend to go off in a tangent. I thought system B was certainly better when it comes to completing this specific task, but I am not sure how much more useful I would find it as someone who just wants to have a quick browse of some of the most recent comments left on any given article".*

Another participant objected to the topical organization of comment because it "could make it [comment] feel more like fact, which could be pretty scary".

- Quality of outputs

Adding to the concerns we identified in the question 1 responses, 6 out of the 32 responses in question 3b expressed concern with the quality of the various System B outputs. These 6 participants gave an average score for question 3a of 2.8. (This is lower than the total participant average of 3.19). There were concerns with the topic labels (described as "repetitious", and "unrelated"); whether the labels represented the issues in the conversation was also raised as a concern: "not sure the words on this… pie chart that helpful"; "not useful for identifying topics"; "issues spanned a number of different pie-chart categories"; "I have

concerns about how the list of topics is generated". The selected quotes were also criticized – to one they seemed like an "arbitrary" selection. 3 out of the 6 indicated that they liked the idea of having the topical overview, the implication being if these issues of quality were addressed.

- System B: A user friendly Interface?

A number of participants remarked on whether or not System B was "user friendly" or easy to use or not. Opinion was divided. Some praised the clarity of presentation, liked the use of graphics and colour and described it as "user friendly":

> *"The system is much more user-friendly, as it is very visual and easy to understand. The pie chart and colour coding made it particularly accessible".*

By contrast, others described it as "too complex", or "cumbersome", difficult to use and/or to navigate. Although a couple acknowledged that with practice they may find it easier to use and one was aware that this might be due to individual preference. We elaborate further on the issue of navigation below.

- Criticisms of System A

Echoing those who reported difficulties using the Threads in the question 1 feedback, there were a few participants (see responses in Q3b and Q4) who reported the difficulties of following the main conversational topics using System A. As the example above suggested, reading comments in chronological order involves "context switching"; others found it can be "tedious" or distracting. For example:

> *".. I did find it [B] easier to get an overview of the concerns people had rather than having to keep scrolling down, which gets tedious quite quickly".*

> *"The text only system is the familiar one but it is easy to get bogged down in irrelevant issues and arguments between commenters."*

These comments were a small minority (we identified just 3) and as the report of the next theme shows, a larger group felt quite differently about reading comments in threads.

- "Following the Flow": reading comment conversations in chronological/reply to order

Around one third of participant comments (10 out of the 32 in question 3b and 5 of the 17 in question 4) referred to being able to read through the comments in order, as they reply to one another, i.e. to follow the flow of a conversation. For example:

> *"I didn't like having to click on 'view in context' to see the threads/replies. In my opinion one of the best things about the good comment systems – like on Facebook – is the threads as it is much easier to see conversations developing, which is the best bit of reading comments (i.e reading two people arguing against each other and disputing each other's points) -and to fully appreciate this the 'context' is key, you need to read the replies in the order they were given. A big debate like this has the potential to cover a number of 'topics' and so grouping individual comments into topics might break up these conversations which would be detrimental in my opinion."*

> *"Comment threads really do follow their own flows, and can be fascinating. Comment is often just that, comment…"*

The majority of participants who expressed this concern suggested that following the flow of the conversation was difficult to do in System B and/or indicated that the threaded system best allows participants to do this:

> *"*I think system B would not be beneficial for a reader with a short time span. With system A the reader can just scroll through the comments and skim read them with ease."

> "I preferred System A. This may be because I'm more used to it, so navigated it more easily. But I didn't like struggling to follow who was replying to whom and comments being out of chronological order*."*

The average score for the future use question (Q3a), for the 10 participants who referred to being able to follow the comments in conversational context in question 3b was only 2.6 (see Figure 16), i.e. a much lower rating than the total participant average of 3.19. Just 1 of the participants to give a 4 or 5 future use score mentioned seeing the comments in chronological order. These results suggests that the difficulty of seeing comments in full context in System B was a limiting factor for some when deciding if they would like to have System B available for future use.

(We also observed that just 1 of the 10 who mentioned reading the comments in order/context, also raised a concern with the quality of any information in System B - e.g. the pie chart labels. This suggests that quality of outputs was possibly not perceived as the major problem in System B for this group of participants.)

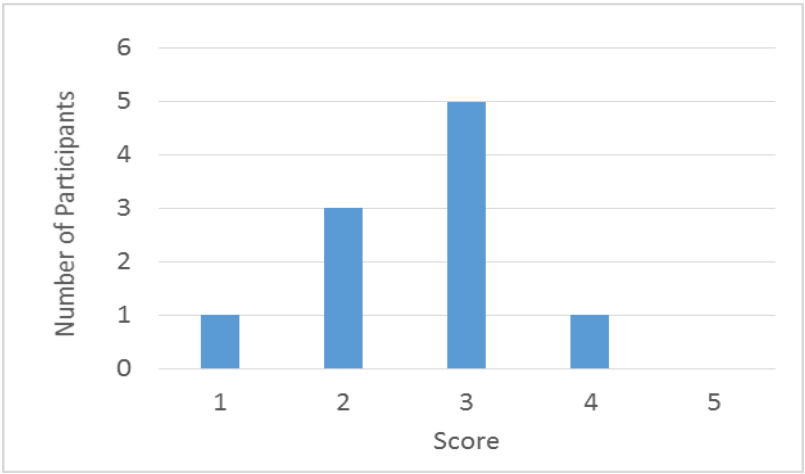


Figure 16: Q3a scores for participants who remarked on seeing the comments in the order they appear in Q3b

The results above do not preclude people liking any aspect of the System B interface. We observed that 5 of the 10 participants who liked seeing comments in conversational context, were also positive about the System B overview feature (i.e. these 5 participants were also part of the larger group of 18 who liked having an overview, with an average future use score of 3.7). The average future use score for the 5 was 3. For example:

> *"I found that while the pie chart and list of topics were helpful in giving an overview of the topics discussed by readers, it was less easy to find main issues using the selected quotes and comment clusters than using the threads."*

- *Embedding System B Features into System A?*

Several comments suggested having a combination of the two systems, the idea being to have a System A style facility which included a pie chart or overview embedded in it, for example:

> *"The pie chart is a very nice idea but would be best incorporated into System A".*
>
> *"In the future I think it would be nice if there was a facility which combined the text layout of System A with the graphics of System B (possibly as an overview section at the top of the page)".*

So, why might people like seeing comments in their threads, with the reply-to structure? The qualitative feedback to questions 3b) and 4) provided the following insights:

- *Enjoyment/Fun*

As we have seen above, for some dipping into the conversation, i.e. following the interchange between comment posters and reading the comment conversation in the order it was generated was something they simply enjoyed or liked to do when engaging with comment, for example (see also the discussion of the theme "topical overview", especially the final 3 examples, which say they didn't need an overview):

> *"I think if I felt it represented the conversation fairly I'd find it a good overview system, but I'd still want to get lost in the conversations and rants and I'm not sure the words on this particular pie chart were that helpful. Sometimes watching the comments provoke surprise responses is half of the fun - as a reader, anyway…"*

- *Reading Comprehension*

As we found in the question 1 qualitative data, some people had difficulty reading or fully comprehending the comments when presented out of chronological context. The data from questions 3 and 4 provide further evidence for this. Some said they were able to identify issues most easily with threaded comments in their original reply to sequence, and other examples illustrate the point that reading comment out of context was difficult:

> *"I found that while the pie chart and list of topics were helpful in giving an overview of the topics discussed by readers, it was less easy to find main issues using the selected quotes and comment clusters than using the threads. For example, the threads appear more like a conversation so it is easier to identify an issue which people are debating and opinions regarding the issue".*
>
> *"The selected quotes on system B were helpful in quickly identifying themes, but scrolling through ALL the threads in order in system A gave me a deeper grasp of the arguments and how they had emerged"*[9]*.*
>
> *"… The right-hand column was confusing and much worse than the current system for reading comments on news articles".*

- *Seeing the Complete Picture*

[9] Note: this participant appeared to have got mixed up when describing the respective systems in their feedback, but the comment is sufficiently detailed for us to be sure that they were referring to A when they said B etc. And so to assist readability we have changed their labels accordingly in this example.

A few responses suggested that when reading through the threads in System A they were more confident they weren't missing out on comments, for example:

> *"I am not sure I enjoy it [System B] as much as scrolling through. I think you miss out on some of the other comments".*

3.6. Evaluation Results (Experiment 2: System A and System C)

This section discusses the results of Experiment 2: System A (baseline) and System C (USFD/UNITN SENSEI system). Similar to Section 3.5, we first start with the background of annotators participating in the experiment. Their performances using the different systems in the reading comprehension task were analysed. Lastly, we report the feedback given by the annotators in the post-questionnaire. As previously discussed in Section 3.2.1, System B and System C share many of the same features, i.e. System C contains all of the features in System B, and a few additional features (i.e. overview summary, mood and agreement information). Since our finding of the original features (i.e. System C features that also appear in System B) was very similar to Experiment 1, in this section, we focus our qualitative analysis on the features that are only available in System C.

3.6.1. Participant background

Most of the annotators participating in Experiment 2 were native speakers or fluent in English, whilst two reported a very good command of English.

A key difference between Experiments 1 and 2 is that in 1, 69% of participants had media experience but in 2 none of the participants had media experience. However, Experiment 2 participants did report a very similar level of engagement with reader comments to those in Experiment 1: in both experiments 72% either read or write reader comments at least once a week, (with 3% more – a total of 34% – engaging on a daily basis in Experiment 2), whilst 29% engaged with reader comments very occasionally (about once a month or less), as compared with 25% in Experiment 1. We show these details in Table 29.

Table 29: Participant background (Experiment 2)

1. English language proficiency	Native speaker	26	81%
	Near native / fluent	4	13%
	Very good command / highly proficient	2	6%
	Good command / good working knowledge		0%
	Basic communication skills / working knowledge		0%
2. Media professional experience	Yes		0%
	No	32	100%
3. Engagement with reader comments	At least once a day	11	34%
	At least once a week	12	38%
	At least once a month	4	13%
	Very rarely	5	16%

	Never		0%

3.6.2. The reading comprehension task: Identify 4 Issues

In general, participants spent 6:19 minutes to do the task using System A, and 6:46 minutes to use System C. When using System A, participants spent less time on Topic 1 compared to Topic 2, 6:13 and 6:26, respectively. Participants spent significantly longer when using System C to assess reader comments in Topic 1 (7:12). However, System C seems to help participants do the task slightly quicker when working on Topic 2, compared to System A, i.e. 6:20 and 6:26, respectively. Overall participants worked faster to complete the tasks than they did in Experiment 1, but the times are still fairly similar. The times to complete for System A are comparable with those reported in experiment 1: topic 1 was completed 34 seconds faster with System A in experiment 2, and topic 2 was just 5 seconds faster in experiment 2. The slowest overall task was Experiment 1 System B, Topic 2 (7:18) which was almost a minute slower than the time to complete for Topic 2 using System C (6:20).

Table 30: Time spent to complete the task

System Name	Topic Id	Average Time Spent
A	1	6:13
A	2	6:26
C	1	7:12
C	2	6:20

At the reading comprehension task, each participant was asked to identify four issues raised in the reader comments. In using System A, all of the 32 participants answered the four issues. When using System B, 27 participants identified four issues in the given time, four participants submitted three issues, whilst one only submitted two issues. Similar to the previous results, each of these issues was scored by two annotators (unanswered issues were scored 0). We show the average scores for the 32 annotators when using the different systems in Table 31.

Table 31: Average scores

System	Overall	Topic 1	Topic 2
A	2.19	2.12	2.25
C	2.01	2.08	1.94

We show that annotators were able to achieve an average score of 2.19 when using System A. Their average scores when using System C is slightly lower (2.01). When analysing the differences across topics, we found that the qualities of issues provided by annotators in Topic 1 were very similar, 2.12 and 2.08 for System A and System C, respectively. However, similar to the findings in Experiment 1, annotators were able to identify issues in Topic 2 better when using System A.

3.6.3. *Post-task questionnaire*

Similar to the post-questionnaire task in Experiment 1, we asked participants to provide some feedback about the systems and their features by answering four questions: (Q1) to rate the level of usefulness of each of the systems and the different system features in the task, (Q2) to identify features that helped to provide an overview of reader comments, (Q3) to indicate how much they would like to have System C available more generally for future use, and a final question, (Q4) to provide general comments on the systems and their experience in the tasks. We report the feedback they provided for each question.

Q1. How useful were the different systems/system features when completing the task "identify four issues"?

Similar to the finding in Experiment 1, features of System A were given relatively high scores (around 3.5) in this experiment, as shown in Figure 17. Features in System B, meanwhile, tend to have slightly lower scores. The exception is the highest scoring feature in System B is the "view comment in context", which is the comments displayed in their original thread, similar to the "threads" feature in System A. Most of the original features used in System B, such as the pie chart, list of topics and the comment clusters again scored between 3 and 3.5, whilst the new features added in System C, such as overview, mood, and agreement scored between 2.5 and 3.

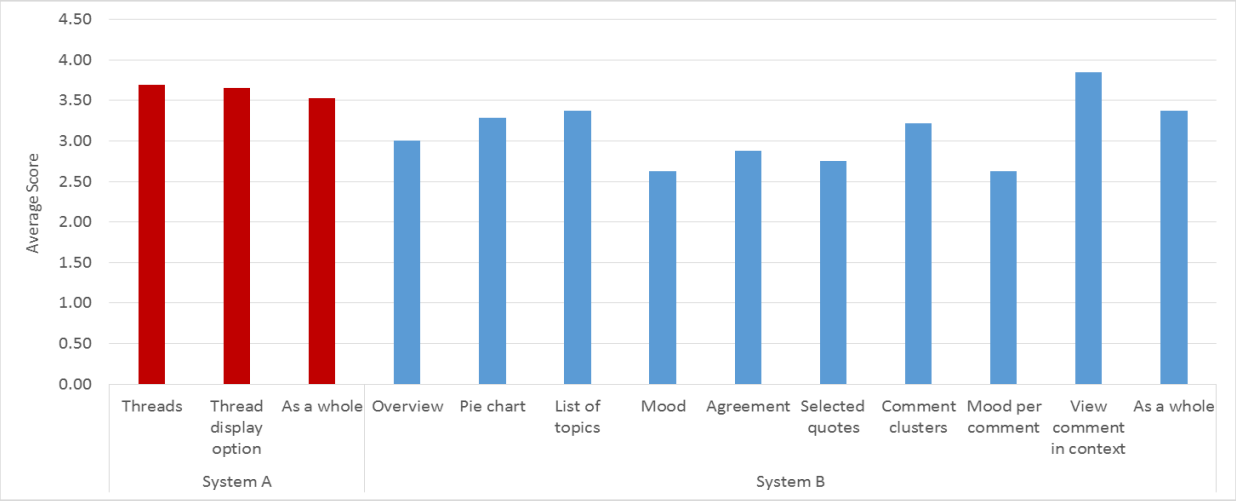


Figure 17: Usefulness of Features in System A and System C

We show the usefulness scores of features across different topics in System A in Table 32 and System C in Table 33. Overall, the usefulness scores given for the features in both systems are similar across the different topics. The scores, however, in general are lower compared to features in Experiment 1. We further investigate these results below.

Table 32: Usefulness of System A Features

Topic	System A		
	Threads	Thread display option	As a whole
All	3.69	3.66	3.53

1	3.44	3.81	3.25
2	3.94	3.50	3.81

Table 33: Usefulness of System C Features

Topic	System C									
	Over-view	**Pie chart**	**List of topics**	**Mood**	**Agree-ment**	**Selec-ted quote s**	**Com-ment clus-ters**	**Mood per com-ment**	**View com-ment in con-text**	**As a whole**
All	**3.00**	**3.28**	**3.38**	**2.63**	**2.88**	**2.75**	**3.22**	**2.63**	**3.84**	**3.38**
1	3.00	2.94	3.13	2.44	2.94	2.44	3.00	2.81	3.69	3.06
2	3.00	3.63	3.63	2.81	2.81	3.06	3.44	2.44	4.00	3.69

System C includes some para-semantic parts that are not present in System B, namely the template-based summary (called "overview"), mood, agreement and mood per comment. As for system B, evaluators entered notes and comments in a free text form that can we used for a qualitative evaluation of these components.

The feedback provided about the Overview depicts it as a useful summary that "delivered a general idea of the topic in discussion and reader emotions", but it is generally considered as "wooden and clearly not user written".

The general impression that evaluator had about mood can be summarised as "not easy to understand and not very accurate"; the agreement part is considered "very useful to identify controversial topics", although some evaluators claim that it is not very intuitive and "may be subjective".

The qualitative evaluation of the mood per comment part can be summarised as "Interesting to see what the computer thought but people can work out the mood of a single comment without it being presented to them" and therefore, did not find it to be very useful.

Lastly, although annotators found System C to be “much more convenient if you want to quickly search through the comment and better showed the general mood and ideas on the discussed topic”, the interface is judged to be too complex with respect to the baseline system. This can be partially due to an effect of rejection towards innovation, summarised by the statement of one annotator, "I like things as they are ;)".

Q2. Please tick the features or systems which in your opinion helped to provide an overview of the reader comment discussion. (1="yes" and 0="no")

In Figure 18, we show the proportion of participants who indicated whether each feature was useful in providing an overview of the discussion. Almost 60% participants specified that the “threads” feature is useful to give an overview, whilst half of the participants thought that the “pie chart” was a good feature as an overview. These numbers dropped to between 25%-34% for the remaining 3 features, i.e. overview, list of topics and selected quotes.

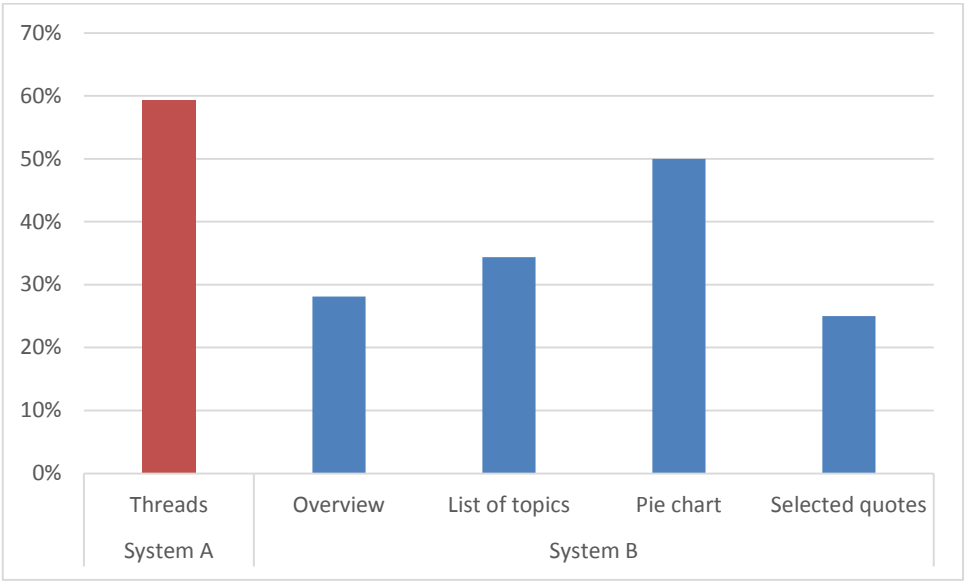


Figure 18: Proportion of participants preferring each feature

When analysing these scores across different topics (as shown in Table 34), we could see some differences occurring between the two topics. Participants who used System C to assess Topic 2 (Network Rail) tend to think these features (of both system) to be more useful in providing an overview compared to those who used System C to assess Topic 1. This might indicate that the quality of the features shown in System C across the different topics are different.

Table 34: Proportion of participants preferring each feature (across topic)

Topic	System A	System C			
	Threads	Overview	List of topics	Pie chart	Selected quotes
All	**59%**	**28%**	**34%**	**50%**	**25%**
1	44%	19%	25%	31%	25%
2	75%	38%	44%	69%	25%

Q3. How much would you like to have System C (Text and Graphics Reader Comment System) available in a news and comment browsing facility?

When asked to provide a score (between 1-5) on how much the participants would like System C, an average score of 2.69 was achieved. We show the distribution of participants choosing each score in Figure 19.

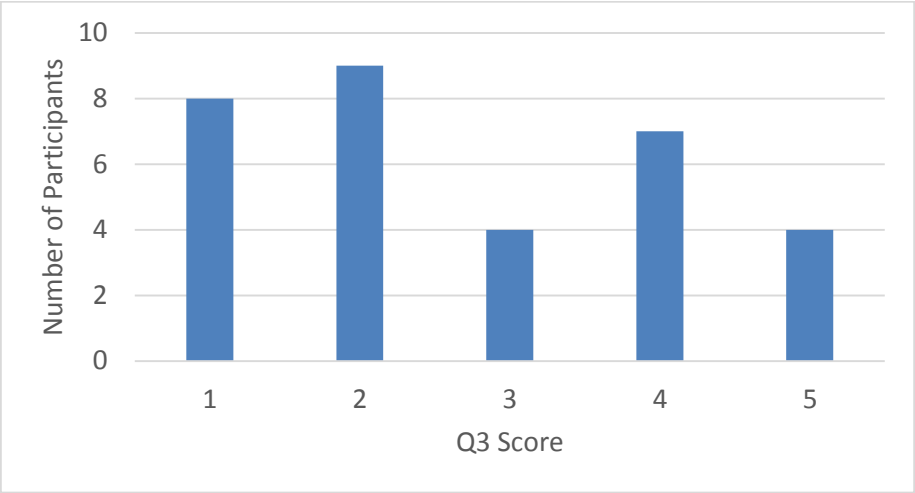


Figure 19: Distribution of Q3 Scores (System C)

Four participants gave a score of 5 for System C, compared to two for System B. Overall, 11 people (34%) scored 4 or higher, mentioning that they could see the potential of System C and that the mood and topics would help to see a particular topic. Others mentioned that they could see System C being very useful for research/professional purposes, such as researching a topic written by readers, or analyzing the mood in a discussion. Others found System C to be interesting, although would not be very useful in their daily tasks.

As also shown in Figure 19, a significant number of people provided a score of 2 or lower for System C. Many of those gave the reason that System C was very complicated and therefore was not very easy to use and understand. The high number of features in System C also meant that the participants needed much more time to read and understand all the contents. Others mentioned that they did not like the user interface of System C and preferred a simpler interface.

We also analysed the preference scores of the System across the two topics (Table 35). Participants who used System C for Topic 2 provided a higher score on average compared to those using System C for Topic 1. This finding is similar to one in Q2, which suggests that the quality of data in Topic 2 may be higher compared to Topic 1.

Table 35: Preference score of System C (across topics)

Topic	Would like System C?
All	2.69
1	2.44
2	2.94

3.7. Summary of Findings and Future Work

Our analysis of the task outputs (assessed issue responses; times to complete the task) and their feedback (questionnaire responses) resulted in a number of findings.

Experiment 1:

On average there was very little or no difference between the results for System A and System B, when we consider both participant performance (as indicated by the issue scores and time to complete) and the perceived utility of a system overall, when carrying out the “find issues” task.

- Most people preferred one system over the other when assessing a system’s usefulness in the context of the “find issues” task. Those who rated system B more highly were more likely to want it for future use.
- What people liked most about System B was the pie chart. On average participants found the pie chart more useful than the threads for the finding issues task. People liked that it helped them to find topically related comments and that it gave them an overview an overview of the comment discussion.
- System A style threads were seen by many to be better for activities such as reading some conversation and following interchanges between commenters. Some found it easier to comprehend what was being discussed when they read the full threaded comment sequence. Some participants found it easier to find issues using the threads.
- Criticisms of System B included:
    - Many found it difficult to understand comments when presented out of context (in comment clusters or as selected quotes) but others said they preferred skimming topically filtered comment. Some people criticised system B for the difficulties of navigation, when wanting to view comments in context.
    - The labels for the pie chart could be improved to better represent the clusters and to be less similar.
- Reasons why people would not want to use System B in the future included:
    - A minority of people aren’t interested in finding issues when reading comment
    - A minority don’t like the idea of automatic summarisation or selection of comment because they don’t trust how its done, or fear it could be misleading.
- Some suggested having a combination of the two systems, the idea being to have a System A style facility which included a pie chart or overview embedded in it.

Experiment 2:

- Overall System C did not compare as well with System A as System B did. Qualitative feedback suggested that the addition of a textual summary overcomplicated the presentation of information, negatively affecting the rating of all system features, including those shared with System B.
- The feedback on System C features was not entirely negative: in particular people liked the agreement feature; mood was liked by some.

In conclusion the SENSEI Social Media Prototype System B was received very positively by users, many of whom were confident that they would like to see such a system available for use when reading comment.

Future Work

Further work could address the suggestion that we look to combine the most favoured features of Systems A and B, to reach out to the different participant types the evaluation has revealed.

What might such a system look like? As the participants suggest, a straightforward threaded comment view is perhaps the simplest and easiest way to view comment. This could be enhanced by including an option to pop up an overview of the topics in a System B style pie chart at the top of a page. In this new version of the system the pie chart if clicked would not display a cluster of comments, out of context. Instead, the pie chart could be linked directly to the comment threads so that if you clicked on a segment you would return to the threads, where there could be SENSEI style features, e.g. a background colour coding, to identify comments on the selected topic, and a "jump to next comment on topic" button, to assist a user viewing further comments in the topic, if the user required. Such a combined system would address both the concerns of i) those who like the idea of an overview but really feel strongly about being able to read follow the comments as they were posted and ii) those participants who experienced difficulties when using System A as it stands.

4. Conclusions

SENSEI D1.4 deliverable reports the results of extrinsic evaluation of SENSEI prototype in speech and social media domains.

The extrinsic evaluation of SENSEI speech prototype had the goal of answering to three questions. We wanted to estimate if SENSEI speech technology may help the users in finding information they need to solve their tasks, to measure if SENSEI speech technology has impact on efficiency and accuracy of task resolution, and to evaluate the user appreciation of such technology.

The extrinsic evaluation of SENSEI speech prototype showed that the evaluators could complete more efficiently and accurately the experimental tasks when working in SENSEI-enabled condition that in 'without SENSEI' condition. The differences estimated from the results of the two sets of tasks are statistically significant. The evaluators could use the results of SENSEI speech technology to gather information useful for completing the tasks that were submitted to them. From the analysis of the post-task questionnaires and from the discussion in the post-task focus group, we could observe that the novelty related with the use of call centre summaries is possibly related with a cognitive load that influenced the participants' attitude towards the SENSEI-enabled condition. However, the feelings reported by the evaluators support the view that a more prolonged training period may overcome such difficulty.

We also carried out the extrinsic evaluation of the SENSEI social media prototypes, to investigate the following research questions:

1. Do the SENSEI social media prototypes (Systems B and C) help users carrying out a reading comprehension task – "find 4 issues in the comments" – better than when using current practice technology (System A)?
2. Do users perceive that having SENSEI for the finding issues task is more helpful than having current practice technologies alone?

Recruiting a total of 64 participants from the media and the general public we completed two similar experiments. In each, we asked users to find issues in comments, using a SENSEI prototype and a baseline system based on current practice. Participants provided feedback on their experiences via a post task questionnaire.

The rich set of data that was generated by the evaluation is evidence that the issue task and evaluation setup is both effective and robust. Regarding the first research question, the evaluation showed that there was essentially no significant difference between using the SENSEI prototypes and the current practice baseline system for the finding issues task. Regarding the second research question, while on average participants' preference ratings for the SENSEI prototypes and the current practice system were quite similar (though System C was ranked somewhat lower, apparently due to the interface being perceived as overcomplicated), the participants divided into two camps: those who preferred SENSEI System B and those who preferred the baseline system. Our analysis, presented above, provides insights into some of the reasons for this. Based on this analysis we have identified some straightforward changes to be made to the SENSEI B interface that we are confident will satisfy the concerns of many of the participants who dispreferred it. Gaining such understanding demonstrates the value of the extrinsic evaluation exercise.

REFERENCES

[Aker et al., 2016a] Aker, A., Kurtic, E., Balamurali, A.R., Paramita, M., Barker, E., Hepple, M., and Gaizauskas, R. 2016. A Graph-based Approach to Topic Clustering for Online Comments to News. In Proceedings of the 38th European Conference on Information Retrieval (ECIR 2016).

[Aker et al., 2016b] Aker, A., Paramita, M., Kurtic, E., Funk, A., Barker, E., Hepple, M., and Gaizauskas, R. 2016. Automatic label generation for news comment clusters. In Proceedings of the 9th International Natural Language Generation Conference (INLG16).

[Barker et al., 2016] Barker, E., Paramita, M., Funk, A., Kurtic, E., Aker, A., Foster, J., Hepple, M., and Gaizauskas, R. 2016. What's the Issue Here?: Task-based Evaluation of Reader Comment Summarization Systems. In Proceedings of LREC 2016.

[Belz & Gatt, 2008] Belz, A., & Gatt, A. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers* (pp. 197-200). Association for Computational Linguistics, June 2008

[Danieli 2014] Danieli, M. (Ed.). *Preliminary version of use case design*, SENSEI Project deliverable D1.1, April 30, 2014.

[Danieli & Gaizauskas 2014] Danieli, M. & Gaizauskas, R. (Eds). Report on use case design and user requirements, SENSEI project deliverable D1.2, October, 31 2014.

[Danieli & Barker 2015] Danieli, M. & Barker, E. (Eds), Report on intermediate evaluation, SENSEI project deliverable D1.3, October 30, 2015.

[Danieli et al. 2016] Danieli, M., Balamurali,A.R., Stepanov, E.A., Bechet, F. and Riccardi, G. “Summarizing Behaviours: An Experiment on the Annotation of Call-Centre Conversations” In *Proceedings of LREC*, 2016.

[Dang et al. 2007] Dang, H. T., Kelly, D., & Lin, J. J. (2007, November). Overview of the TREC 2007 Question Answering Track. In TREC (Vol. 7, p. 63).

[Daume & Marcu 2005] Daume III, H., & Marcu, D. (2005). Bayesian summarization at duc and a suggestion for extrinsic evaluation. In *Proceedings of the Document Understanding Conference*, DUC-2005, Vancouver, USA.

[Doran et al. 2004] Doran, W., Stokes, N., Carthy, J., & Dunnion, J. (2004). Comparing lexical chain-based summarisation approaches using an extrinsic evaluation. *GWC 2004*, 112.

[Dorr et al. 2005] Dorr, B. J., Monz, C., President, S., Schwartz, R., & Zajic, D. (2005, June). A methodology for extrinsic evaluation of text summarization: does ROUGE correlate?. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 1-8).

[Hanushek & Jackson, 2013] Hanushek, E. A., & Jackson, J. E. (2013). *Statistical methods for social scientists*. Academic Press.

[Holden & Karsh, 2010] Holden, R. J., & Karsh, B. T. (2010). The technology acceptance model: its past and its future in health care. Journal of biomedical informatics, 43(1), 159-172.

[Kelly & Teevan, 2003] Kelly, D., & Teevan, J. (2003, September). “Implicit feedback for inferring user preference: a bibliography”. In ACM SIGIR Forum (Vol. 37, No. 2, pp. 18-28). ACM.

[Kelly et al. 2007] Kelly, D., Wacholder, N., Rittman, R., Sun, Y., Kantor, P., Small, S., & Strzalkowski, T. (2007). Using interview data to identify evaluation criteria for interactive, analytical question-answering systems. *Journal of the American Society for Information Science and Technology*, 58(7), 1032-1043.

[Krueger 1997] Krueger, R. A. (1997). *Analyzing and reporting focus group results* (Vol. 6). Sage publications.

[Mitkov & Rello, 2009] Mitkov, R., Ha, L. A., Varga, A., & Rello, L. (2009, March). Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics* (pp. 49-56). Association for Computational Linguistics.

[Murray et al. 2008] Murray, G., Kleinbauer, T., Poller, P., Renals, S., Kilgour, J., & Becker, T. (2008, September). Extrinsic summarization evaluation: A decision audit task. In *International Workshop on Machine Learning for Multimodal Interaction* (pp. 349-361). Springer Berlin Heidelberg.

[Saville & Wood, 1991] Saville, D. J., & Wood, G. R. (1991). Latin square design. In *Statistical Methods: The Geometric Approach* (pp. 340-353). Springer New York.

[Sun & Zhang, 2006] Sun, H., & Zhang, P. (2006). The role of moderating factors in user technology acceptance. International journal of human-computer studies, 64(2), 53-78.

[Venkatesh & Davis 2000] Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, 46(2), 186-204.

[Vredenburg et al.2002] Vredenburg, K., Mao, J. Y., Smith, P. W., & Carey, T. (2002, April). A survey of user-centered design practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 471-478). ACM.

Appendix A: Participant Information Sheet

July 2016

Researcher

Prof Rob Gaizauskas

Department of Computer Science

University of Sheffield

r.gaizauskas@sheffield.ac.uk

Research project title

SENSEI: Making Sense of Human - Human Conversation

Research project aims

The aim of the EU-funded FP7 SENSEI project is to develop summarisation/analytics technology to help users make sense of human conversation streams from diverse media channels. SENSEI will also design and evaluate its summarisation technology in real-word environments, aiming to improve task performance and productivity of end-users. SENSEI is investigating conversations in two settings: 1) spoken language conversations between call centre staff and customers and 2) written language conversations in social media, specifically in reader comments from on-line news sites.

Purpose of the research

You are invited to take part in the experiments for the evaluation of the project's summarisation technologies. Within the SENSEI project, our team has developed technologies that perform automatic summarisation of readers' comments about online news articles. The aim is to assess the effectiveness of these technologies in a task setting that approximates real world use. This will inform future development of the reader comment summarisation technologies. It will also allow us to develop an evaluation protocol that can be used in future evaluations of online forum summarization systems.

Who will be participating?

We will invite adults with excellent English reading and writing skills, who are either:

1. *Media professionals/ news producers*. These include: colleagues in the Department of Journalism Studies; journalists and editors at local and national newspapers, such as The Independent and The Guardian; press officers; USFD Journalism students. Or,

2. News readers and comment providers. These include anyone who has experience and interest in reading news and/or providing on-line reader comments.

What will you be asked to do?

The evaluation tasks will be presented **online,** via a series of web pages, which we have developed. Participants will be asked to carry out the tasks online and provide their responses via answer forms in the web pages.

You will need to have access to a computer and web browser to participate in the experiment. You may carry out the experiment at any time/place that is convenient to you, before the date for completion.

The evaluation tasks:

We will invite you to complete two short tasks which are based on the scenario of a reader of on-line news and comment who has limited time (e.g. a 10 minute coffee break) to read some news and comment. Each task proceeds as follows:

First you will be given a news article to read.

We will then ask you to read some reader comments posted in response to the news article and to answer a question. The question is a "reading comprehension" style question, i.e. the aim is to answer the question based on a reading and understanding of the reader comments. The time available to read the comments and answer the question will be limited to a maximum of **10 minutes**. But you can submit your answers before that time if you wish.

In each task there will be a different article and comment set to read.

Also, in each task we will provide a different system interface for reading and browsing the reader comments.

Before the two tasks we will ask participants to answer some questions about their background and prior experience of reading online news and comment.

After the two tasks have been completed we will invite participants to complete a short feedback questionnaire, based on their experience of using the two systems to complete the task.

The web pages also include some **training** to help you prepare for the evaluation tasks. The training pages include a video demo about how to use the systems to access reader comment;

an overview of what's involved in the tasks; instructions on the task question and some examples of possible answers.

The time required to participate in the tasks should be between around 40 minutes, up to an hour (this includes time to: read the online training; to read the two news articles; to complete the two task questions, and to complete the two questionnaires). If unforeseen technical problems arise during the experiment participants can contact researchers via email.

You will be able to take breaks during the session at any point, except it is less advisable to do so during the 2 time-limited questions (each is a maximum of 10 minutes).

You may withdraw at any point without need to give any reason.

What are the potential risks of participating?

The risks of participating are the same as those experienced in everyday life.

Each of the 2 evaluation tasks will take a maximum of 10 minutes to complete; a little extra time is required for reading 2 news articles and answering 2 short questionnaires. The total time involved (including the evaluation tasks, completing the online consent form and reading the training information) should be from around 40 minutes to an hour.

As the focus of the task will be on your use of current and novel reader comment technologies in an activity related to your daily experience, there should be no potential for physical and/or psychological harm/distress. Participants will be able to take breaks at various points throughout the session and may withdraw at any point.

All tasks you are asked to carry out and questions you are asked will focus on the use of technologies for reading and making sense of online news and associated reader comment. It is possible that in the course of the exercise, you may recall previous frustrations or difficulties in reading on-line news comments, and there is a very small possibility that our questions may trigger unpleasant memories, e.g. of reading unpleasant or offensive comments, or that the news or comments you are asked to read could cause offense or trigger unpleasant memories. However, since reader comment sites are currently moderated carefully by the news providers we do not anticipate that these will be too distressing, certainly no more so than experiences in daily life, and should you experience discomfort during the exercise you will be able withdraw at any time.

What data will we collect?

A small amount of personal data will be collected: do you have experience working as a news media professional?; how often do you engage (e.g. read or post) online news and reader comments? We will also ask you to indicate your level of English language proficiency. Aside from these questions we will only gather data relating to the task described above, including

your views about the technology you have used in the evaluation, how you have used it and possible future developments of it.

Your responses will be recorded electronically, via the evaluation interface; we will gather and store the following data:

1) Answers to the *pre- and post-task questionnaires*;
2) Answers provided to the two task questions
3) A log file of participant interactions with the reader comment system during the 2 five minute timed tasks (e.g. we will record the time taken to complete a task).

We will also ask you to provide an email address when you begin the evaluation session. You can use this email address to log back into the session should your browser session expire at any point.

For all participants who complete the experiment, we will enter their email addresses into a prize draw. The prize draw will take place soon after the experiment is complete. After the prize draw is complete we will use the email address to notify the winners. Then we will delete **all** email addresses from our records. The participant email addresses we collect will not be stored or linked to any of the responses/data we collect after the experiment.

Finally, participants' names will be not stored with the data gathered and none of the information we gather will allow participants to be identified. All information that we collect during the course of this research will be kept strictly confidential and the participants will not be identifiable in any reports or publications.

What will we do with the data?

The data will be used to gain insights into the utility of reader comment summarisation technologies and how they can be improved to help comment readers; the data will also be used to gather insights into the design of our evaluation methodology. The data will be stored on University of Sheffield computers and only the research team will have access to the data. The results of this study will be included in SENSEI project deliverables and may be published in journal or conference papers, but the raw data will not be redistributed and will not be available to anyone other than the researchers directly involved in this project. Responses to the task questions and questionnaire may be quoted in future publications but will not be attributed to any named individual.

Will my participation be confidential?

All the information that we collect during the course of this research will be kept strictly confidential. You will not be identified in any written reports or publications.

What will happen to the results of the research project?

The results of this study will be included in the internal SENSEI deliverables and may be published in journal or conference papers.

Who should you contact for further information?

If you wish to obtain further information about the project or the task, please contact Prof Rob Gaizauskas (r.gaizauskas@sheffield.ac.uk).

We would like to thank you for taking part in this project.

Appendix B: Pre-Questionnaire: Participant's Background

Before starting with the task, we would appreciate it if you could answer the following questionnaire about your 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. Have you prior experience working as a news/media professional?*

○ Yes ○ No

3. 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 C: Post-Questionnaire

Please complete the questionnaire below.

(Note: questions marked with a * must be answered.)

We refer to the 2 systems as follows:

- **System A: *Text Only Reader Comment System***
  (i.e. the system where the comments were presented in sequence, with no graphics.)
- **System B: *Text and Graphics Reader Comment System***
  (i.e. the system where the comments were presented via text and graphical features, such as a pie chart.)

1. **How useful were the different systems/system features when completing the task "identify four issues"?***
   Please indicate by selecting a score on a scale of 1-5 (1=not useful and 5=extremely useful). You may tell us more (e.g. why something was useful; why something was not useful etc.) in the box "Any further comment?"

System A: Text Only Reader Comment System

Feature/System	**Usefulness*** **(Note: 1=not useful and 5=extremely useful)**	**Any further comment?**
The threads	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	
Thread display option (i.e. expand/collapse threads, unthreaded)	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	
System A (Text Only Reader Comment System), as a whole	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	

System B: Text and Graphics Reader Comment System

Feature/System	Usefulness* (Note: 1=not useful and 5=extremely useful)	Any further comment?
The pie chart	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	
The list of topics	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	
The selected quotes (middle column)	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	
The comment clusters (right hand column)	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	
The "view comment in context" feature	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	
System B (Text and Graphics Reader Comment System), as a whole	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	

2. **Please tick the features or systems which in your opinion helped to provide an overview of the reader comment discussion.** You may tick more than one box.

System A (Text Only Reader Comment System):

- Thread display option (i.e. expand/collapse threads, unthreaded) ☐

System B (Text and Graphics Reader Comment System):

- List of topics (left column) ☐
- Pie chart (left column) ☐
- Selected quotes (middle column) ☐

3. **a. How much would you like to have System B (Text and Graphics Reader Comment System) available in a news and comment browsing facility?***
   Please indicate on a scale of 1-5.

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

b. Please tell us why you gave this score.*

4. **Please provide any other comments/feedback about your experience using the different systems to carry out the tasks.**
   - Was there anything you really liked or disliked?
   - You may also wish to mention any possible improvements to either system, or things you would like to see included in a future reader comment facility.