{SENSEI}

D2.4 – Data Collection Report Y3

Document Number	D2.4
Document Title	Data Collection Report Y3
Version	1.0
Status	Draft
Work Package	WP2
Deliverable Type	Report
Contractual Date of Delivery	31.10.2016
Actual Date of Delivery	31.10.2016
Responsible Unit	Websays
Keyword List	Data collection
Dissemination level	PU

Editor

Marc Poch (Websays SL, Websays)

Contributors

Hugo Zaragoza (Websays SL, Websays)

Vincenzo Giliberti (Teleperformance,TP)

Vincenzo Lanzolla (Teleperformance,TP)

Cosima Caramia (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	22/07/2016	Draft	Marc Poch (Websays)	Table of Content
0.2	30/08/2016	Draft	Marc Poch (Websays)	Web Data Collection Section (2.2)
0.3	02/09/2016	Draft	Marc Poch (Websays)	Follow-up and Approach sections (1.1 and 1.2)
0.4	06/09/2016	Draft	Marc Poch, Hugo Zaragoza (Websays)	Public Data Access (partial), Executive Summary (draft), Sharding and dumps sections (2.2.6 and 2.2.7)
0.5	22/09/2016	Draft	Cosima Caramia, Vincenzo Lanzolla (TP)	Annotations (partial), Approach (Section 1.2), Data Access (Section 1.2,1.3) and Collection of Call Center Data (Section 2.1)
0.6	26/09/2016	Draft	Marc Poch (Websays)	conclusions and minor remarks
0.7	27/09/2016	Draft	Vincenzo Giliberti (TP)	Section 2.1.2 Annotations edited
0.8	27/09/2016	Draft	Elisa Chiarani (UNITN)	Quality check completed
0.9	28/09/2016	Draft	Benoit Favre (AMU)	Scientific review
0.10	06/10/2016	Draft	Marc Poch (Websays)	Improvements
0.11	07/10/2016	Draft	Elisa Chiarani (UNITN)	Final check
1.0	07/10/2016	Final	Marc Poch (Websays)	Finalisation of the document. Ready for submission

INDEX

Executive summary

Deliverable D2.4 describes the data collected during Period 3 (P3) of the project, the annotation efforts and developed tools.

The document presents the call center data collection, together with its annotation and indexation tasks.

The deliverable describes the web data collection and the work carried out for data extraction, pre-processing and data indexing.

Finally document reports information about data publication and sharing beyond the consortium, and the methods to obtain copyrighted free materials.

LIST OF ACRONYMS AND NAMES

Acronym	Meaning
ACOF	Agent Conversation Observation Form
API	Application Program Interface
CRF	Conditional Random Field
QA	Quality Assurance
RATP	Régie Autonome des Transports Parisiens
REST	Representational State Transfer
SVN	Apache Subversion (often abbreviated SVN, after its command name svn) is a software versioning and revision control system distributed as free software under the Apache License
SSH	Secure Shell (SSH) is a cryptographic network protocol for operating network services securely over an unsecured network. The best known example application is for remote login to computer systems by users.
TRS	TRanScription

1. Overview

The main objectives of WP2 is to collect, process and annotate the data required for all other WPs and to inter-link such collections for enabling multi-channel and multi-modal analysis. WP3, WP4 and WP5 require the use of data collected with specific characteristics in terms of content, but also in terms of media, size, annotation, time-spans, etc.

For speech data, multi-lingual call-centre dialogues have been collected and pre-processed, both in speech audio and transcription forms, with appropriate multi-layered linguistic, semantic and behavioural annotations.

For textual data, social media online conversations in blogs, forums and news sites, as well as conversations in social media (e.g. Twitter) have been collected and pre-processed. Data sets cover on-going political events (RATP, Sochi Winter Olympics, Ukraine Crisis, Charlie Hebdo terrorist attack, plane accidents, EU-Referendum in the UK, etc.).

1.1. Follow-up to period 2 Activities

During Period 2 (P2) Speech data collections (DECODA and LUNA) were indexed in Elastic Search and can be queried using Kibana. The SENSEI ACOF tool was updated and improved to its version 2 (see D2.3).

Social Media data was crawled constantly during P2 and new case studies were introduced ("Charlie Hebdo terrorist attack", etc.). Parsers were improved and better indexing techniques were introduced.

In P3 Teleperformance has annotated 300 LUNA synopsis to support the abstractive template-based summarization approach described in D5.3.

During P3 Social Media has been continuously crawled and processed. Old case studies have been stopped and new ones introduced. A full new crawling Profile has been configured for the "EU Referendum" in the UK. Some techniques for handling large amounts of posts per day have been introduced (BREXIT reached 1M posts a day). To be able to share this data with SENSEI partners easily some automatic data dumps were developed and configured.

- Speech:
  - DECODA;
  - LUNA.
- Social Media v2
  - EU Referendum;
  - General News Topics;
  - Newspaper Publications;
  - RATP;
  - Orange.

Some tasks necessary to achieve this deliverable are briefly described here:

- Creation of a specially dedicated profile for EU Referendum;
- SENSE-EU website design and management tasks;
- Creation of a Solr shard only for EU Referendum;
- Development of new topics for EU Ref. and their validation;
- Update and bug fixing of crawling and parsing tools (websites change constantly);
- Adaptation of the Websays Election tools to the EU Ref. poll;
- Clustering of the conversations to create groups which look like each other;
- Creation of slot and template annotations;
- Development of tool to support the template based annotation of synopses;

D2.4 data collection is a continuation of the D2.3 collection.

The main work carried out in D2.4 is listed here:

- Speech:
  - Internal (TP) and external (UNITN) Calibration;
  - Review of TP annotation work on the LUNA and DECODA corpus;
  - Analysis of machine generated annotations;
  - Annotation of selected synopsis.
- Social Media:
  - New parsers added when needed;
  - Update and correct already existing parsers;
  - Development of new dashboard Topics (specially for EU Ref.);
  - Manual configuration, evaluation and reviews for all SENSEI data;
  - Data correction when needed;
  - Technical support for all SENSEI profiles and shards (2 shards);
  - Automatic data dump generation for the EU Ref. SENSEI website daily updates;
  - SENSE-EU website project management;

1.2. Approach

The ultimate goal of WP2 is to provide a unified data view of “conversations” and data, both from speech dialogues and online (typed) dialogues. This however requires a high level of abstraction from the raw data, which is not readily available; indeed, building such an abstraction is one of the main objectives of WP2.

WP2 should provide views on the data in a way that the full original data could be reconstructed. Additional annotation on the data should be provided by other WPs in the form of stand-off annotations on these views.

The xml representation used in the SENSEI repository is the result of a complex task of abstraction to find common mappings between such different scenarios.

The designed data schema represents tokens following the next description:

TOKEN:

- Features
    - category: pos tag
    - kind: word
    - length: length of word in characters
    - root: lemma of word
    - string: text of word
    - turn_id: identifier of turn
    - word_id: word identifier in current turn
    - disfluency: disfluency marker
    - named_entity: named entity label
    - dep_label: dependency label
    - dep_gov: word id of governor in turn
    - id_text: marker for synchronization with TRS
    - morpho: morphological features
    - speaker: speaker id from TRS
    - start_time: start time from beginning of conversation
    - end_time: end time from beginning of conversation
    - eos: whether or not this word ends a sentence type
- type: Token
- id: unique hash of all features of word
- start: character offset
- end: character offset

1.3. Data Access

1.3.1. Public Data Access

Data to be shared beyond the consortium has been prepared following D8.4 “Second Ethical Issues Report” conclusions.

For the Social Media collection, the website provides a data bundle for D2.1: a small sample of 1000 social media items from the Social Media collection, together with the entire list of public URLs of the data crawled for D2.4 final collection. The entire collection (as well as individual parts of the collection) can be made available to the public upon e-mail request to sensei-data@list.disi.unitn.it .

For LUNA data we provide a small complete sample; the entire collection is distributed as-is to partners for evaluation and annotation through the data sharing agreement provided in the Ethical Issues Plan (D8.2).

For DECODA data we provide a small complete sample. The entire collection is distributed by SLDR/Ortolang (http://crdo.up.univ-aix.fr, ID: http://sldr.org/sldr000847). Researchers or practitioners may get access to the annotated corpus of human conversations free of charge by accepting the SLDR/ORTOLANG license.

Teleperformance data (limited to the annotations produced by QA Supervisors during the filling of AOFs and synopsis) are available to the partners internally. Similar to the social media data, the Teleperformance anonymized data can be made available to the public upon e-mail request to sensei-data@list.disi.unitn.it .

1.3.2. Partner's Data Access

For partners, a SVN data repository has been setup on one of the SENSEI servers containing all the data for easy access. In the case of the LUNA collection, the data will be distributed as-is to partners for evaluation and annotation through the data sharing agreement provided in the Ethical Issues Plan (D8.2).

The Websays Dashboard has also been made available to all partners in order to provide a rich visual interface to browse the Social Media portion of the data.

All partner have web access, upon authentication, to the SENSEI ACOF Annotation tool developed by Teleperformance, where they can find LUNA and DECODA selected conversations with integrated the relatives machine annotations and human annotations.

2. Period 3 Data Collection

2.1. Collection of Call-Center Data

2.1.1. Previous work

Deliverable D2.1 described the LUNA and DECODA collections as well as the data model and specifications of the data to be acquired for the Call Center Quality Assurance process.

Deliverable D2.2 described the selected set of data collected during the first year of the project, the call center annotation efforts and the developed tools to annotate the conversations in Italian and French language.

Deliverable D2.3 described the annotation activities carried out in P2 and the changes applied to the SENSEI ACOF tool. It also presented statistics and details of annotated data.

2.1.2. Annotations

In P3 the Teleperformance Quality Assurance Team involved in the Sensei Project has annotated 300 LUNA synopsis with templates and slot labeling.

QA supervisors have identified 61 “Topic Categories” that cover most of the situation encountered in the conversations, their occurrences, code and English translation are reported in the following Table 1.

Table 1: Topic Categories and Frequency of the Annotation

Code	Topic Categories	English Translation	Frequency
&WIND	windows	*windows*	2
£WMALL	web mail e allegato	*web mail and template*	1
£WMDG	web mail di gruppo	*group web mail*	1
%PST	web mail	*web mail*	31
$VDPW	visualizzazione documenti personali sui siti web	*display personal documents on websites*	1
"VLCG	visualizzazione cartografica	*cartographic visualization*	2
!UTNA	utente non abilitato	*User not enabled*	2
/TMBR	timbratura	*Stamping*	4
=TELF	telefono	*Phone*	2
£Tabe	tabella	*Table*	1
&STM	stampante	*Printer*	24
@SPAM	spam	*spam*	2
!SOLL	sollecito	*Reminder*	1
$SVER	server	*server*	2
=SCVD	scheda video	*Video Card*	2
%SCGF	scheda grafica	*Graphic card*	3
7SAPG	salvataggio pagina	*page backup*	1

7SADT	salvataggio dati	*data backup*	1
!INT	connessione internet	*internet connection*	1
&PRIR	protocollo e Iriswin	*protocol and Iriswin*	2
^PRTC	protocollo	*protocol*	15
£PRNE	problemi di	*problem of*	1
=PDRE	procedure	*procedures*	6
"PPBL	pin e puk	*pin and puk*	6
pin &!PIN	pin	*pin*	12
(PNDR	pen driver	*pen driver*	3
?PCNF	pc non funzionante	*PC not working*	35
£PSW	password	*password*	24
@SVHM	servizio human	*human sevice*	1
@MAST	masterizzatore	*burner*	2
@LOTU	Lotus	*Lotus*	14
#PRGR	programma	*program*	6
£PRNE	programma nero	*blank program*	1
%PST	posta elettronica	*email*	4
£MNTR	monitor	*monitor*	9
£ACIT	impossibilità di accesso intranet	*impossibile to internet access*	9
$IVPG	impossibile visualizzare pagina	*impossibile to see the web page*	8
£ICIN	icona internet	*internet icon*	2
&GSPR	gestione pratiche	*practices management*	2
£FLTM	file temporaneo	*temporary file*	1
£DLDT	determina e delibera	*determines and resolves*	13
&CROL	corso on line	*course on line*	1
!CNIT	connessione internet	*internet connection*	2
# CRTF	certificazione	*certification*	1
£BTDV	boot device	*boot device*	3
&BLUT	blocco utenza	*user block*	1
&BRRA	barra	*bar*	2
£VRUS	controllo	*control*	1
#AFAP	Adobe Flash	*Adobe Flash*	4
%PST	account posta	*poste account*	1
=TRNT	accesso su tarantella	*tarantella access*	1
@ACPC	accesso pc	*pc access*	2
&ACGR	acceleratore grafico	*graphics accelerator*	1
$ABLT	abilitazione	*enabling*	3
@LIGN	login	*login*	3
&LNPR	link tra preferiti	*favorite link*	1
£LTSM	lettore smart card	*smart card browser*	1

=MOUS	mouse	*mouse*	3
@NAVI	navigazione	*internet browsing*	5
!CPGW	chiusura pagina web	*closure web page*	2
£VRUS	antivirus	*antivirus*	4
		Total of the Annotations	***300***

During the annotations activities QA supervisors have identified 15 Agent Resolution Categories, their codes and English translation are reported in the following Table 2.

Table 2: The Agent Resolution Categories of the Annotation

Code	Agent Resolution Categories	English translation
#SGNL	segnalazione secondo livello	*second level reporting*
&RCNT	ricontatto	*re-contact*
#SGNS	segnalazione ai sistemi	*reporting to systems*
$MOD	solution mode risolto in tempo reale	*solution mode in real time*
&INTT	intervento tecnico in sede	*technical assistance in the field*
$RNTR	problema rientra da solo	*problem returned alone*
%FNAC	assistenza/consulenza on line	*Support / counseling online*
£NTKT	rilascio numero ticket	*ticket number release*
^CDLN	cade linea telefonica	*telephone line falling*
*NRCH	rilascio numero di chiamata	*release call number*
£CHIN	chiamata interrotta	*dropped call*
%SGAU	segnalazione altro ufficio	*reporting another office*
&PGEN	problema di carattere generale	*general problem*
$SLLC	sollecito	*reminder*
$NINV	numero inventario	*inventory number*

The three slots adopted for the annotations are:

- Identified issue;
- Resolution of the agent;
- Reporting to the second level.

Finally, the template used for the annotations are:

1. web mail;
2. stampante *(printer);*
3. protocollo *(protocol);*
4. Accesso *(access);*
5. pc non funzionante *(pc not working);*
6. password;
7. sollecito *(reminder);*
8. Programma *(program);*
9. Internet;
10. Hardware.

These synopses were annotated with templates and slot labeling using the template based synopsis annotation tool described in the following paragraph.

2.1.3. Template based synopsis annotation tool

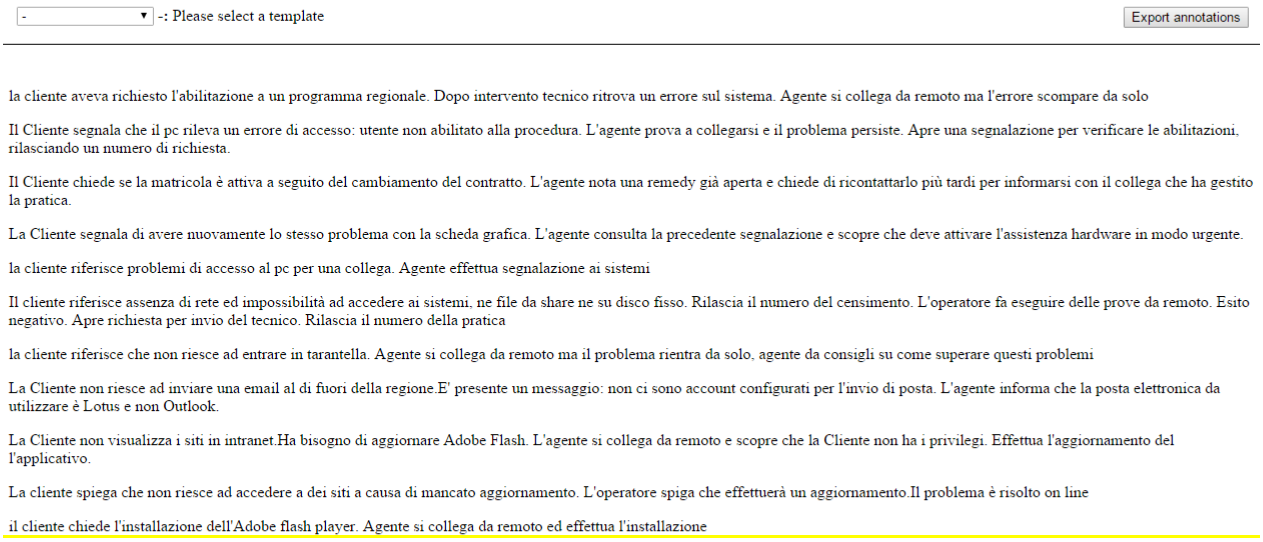
Teleperformance developed a tool in Python to support the QA supervisors in the annotation activity described in the previous paragraph. The tool is based on the one developed for French data by AMU.

The tool allows an administrator to create templates that should be defined following the guidelines below:

- First cluster the conversations to create coherent groups;
- A template shall not cover all aspects of the conversations;
- Multiple templates can be used for the same synopsis;
- Templates should be general enough to represent multiple conversations;
- Slot variables are elements which change from multiple conversations about the same topic;
- Slot variables names are global to all templates, and can be reused in several templates if they have the same role;
- Variable names should be chosen carefully to enable the differentiation of similar entities with different roles.

Once the templates, slots and synopsis have been defined and set up in the tool, QA supervisor can start the synopsis template based annotation activity through a web browser.

At the opening the system shows the list of synopses and a dropdown menu containing the list of templates previously defined



la cliente aveva richiesto l'abilitazione a un programma regionale. Dopo intervento tecnico ritrova un errore sul sistema. Agente si collega da remoto ma l'errore scompare da solo

Il Cliente segnala che il pc rileva un errore di accesso: utente non abilitato alla procedura. L'agente prova a collegarsi e il problema persiste. Apre una segnalazione per verificare le abilitazioni, rilasciando un numero di richiesta.

Il Cliente chiede se la matricola è attiva a seguito del cambiamento del contratto. L'agente nota una remedy già aperta e chiede di ricontattarlo più tardi per informarsi con il collega che ha gestito la pratica.

La Cliente segnala di avere nuovamente lo stesso problema con la scheda grafica. L'agente consulta la precedente segnalazione e scopre che deve attivare l'assistenza hardware in modo urgente.

la cliente riferisce problemi di accesso al pc per una collega. Agente effettua segnalazione ai sistemi

Il cliente riferisce assenza di rete ed impossibilità ad accedere ai sistemi, ne file da share ne su disco fisso. Rilascia il numero del censimento. L'operatore fa eseguire delle prove da remoto. Esito negativo. Apre richiesta per invio del tecnico. Rilascia il numero della pratica

la cliente riferisce che non riesce ad entrare in tarantella. Agente si collega da remoto ma il problema rientra da solo, agente da consigli su come superare questi problemi

La Cliente non riesce ad inviare una email al di fuori della regione.E' presente un messaggio: non ci sono account configurati per l'invio di posta. L'agente informa che la posta elettronica da utilizzare è Lotus e non Outlook.

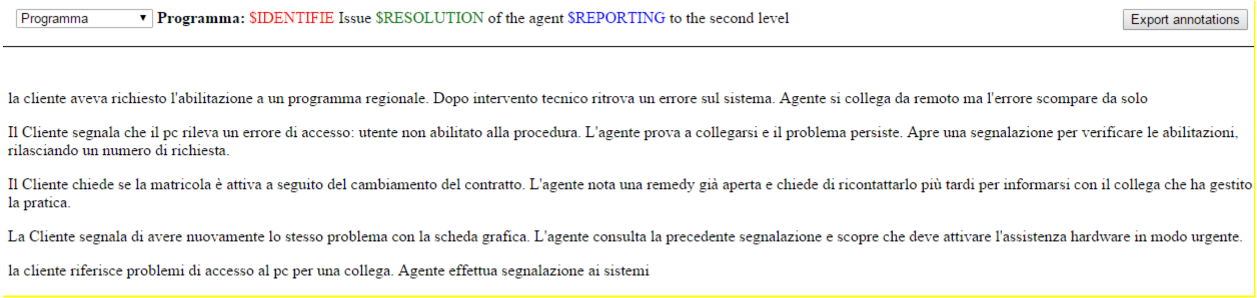
La Cliente non visualizza i siti in intranet.Ha bisogno di aggiornare Adobe Flash. L'agente si collega da remoto e scopre che la Cliente non ha i privilegi. Effettua l'aggiornamento del l'applicativo.

La cliente spiega che non riesce ad accedere a dei siti a causa di mancato aggiornamento. L'operatore spiga che effettuerà un aggiornamento.Il problema è risolto on line

il cliente chiede l'installazione dell'Adobe flash player. Agente si collega da remoto ed effettua l'installazione

Figure 1: Annotation tool - home page

User selects the template he intends to use to annotate the first synopsis and the system shows the slot related to the template selected.



Programma: $IDENTIFIE Issue $RESOLUTION of the agent $REPORTING to the second level

la cliente aveva richiesto l'abilitazione a un programma regionale. Dopo intervento tecnico ritrova un errore sul sistema. Agente si collega da remoto ma l'errore scompare da solo

Il Cliente segnala che il pc rileva un errore di accesso: utente non abilitato alla procedura. L'agente prova a collegarsi e il problema persiste. Apre una segnalazione per verificare le abilitazioni, rilasciando un numero di richiesta.

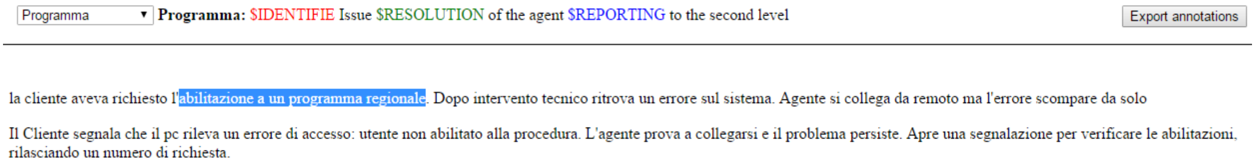
Il Cliente chiede se la matricola è attiva a seguito del cambiamento del contratto. L'agente nota una remedy già aperta e chiede di ricontattarlo più tardi per informarsi con il collega che ha gestito la pratica.

La Cliente segnala di avere nuovamente lo stesso problema con la scheda grafica. L'agente consulta la precedente segnalazione e scopre che deve attivare l'assistenza hardware in modo urgente.

la cliente riferisce problemi di accesso al pc per una collega. Agente effettua segnalazione ai sistemi

Figure 2: Annotation tool – template and slots

User selects unannotated text within a synopsis and clicks on the slot variable which corresponds to the selected text



Programma: $IDENTIFIE Issue $RESOLUTION of the agent $REPORTING to the second level

la cliente aveva richiesto l'abilitazione a un programma regionale. Dopo intervento tecnico ritrova un errore sul sistema. Agente si collega da remoto ma l'errore scompare da solo

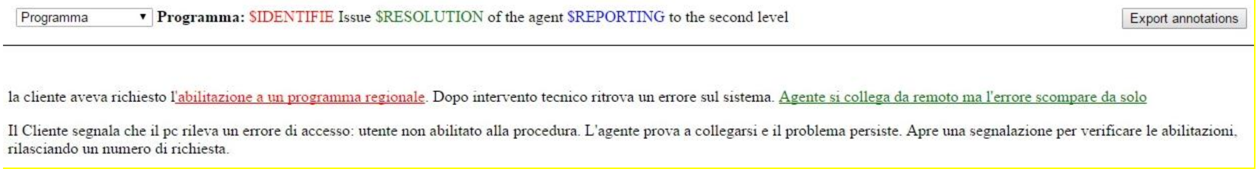
Il Cliente segnala che il pc rileva un errore di accesso: utente non abilitato alla procedura. L'agente prova a collegarsi e il problema persiste. Apre una segnalazione per verificare le abilitazioni, rilasciando un numero di richiesta.

Figure 3: Annotation tool – text and slot selection

The system highlights the annotated text with the color of the slot used.

In case of a mistake, user can click on an annotation to remove it.

User proceeds with the annotation of other part of synopsis using other slots.



Programma: $IDENTIFIE Issue $RESOLUTION of the agent $REPORTING to the second level

la cliente aveva richiesto l'abilitazione a un programma regionale. Dopo intervento tecnico ritrova un errore sul sistema. Agente si collega da remoto ma l'errore scompare da solo

Il Cliente segnala che il pc rileva un errore di accesso: utente non abilitato alla procedura. L'agente prova a collegarsi e il problema persiste. Apre una segnalazione per verificare le abilitazioni, rilasciando un numero di richiesta.

Figure 4: Annotation tool – synopsis annotated with two slots

When all synopses have been annotated, the export function extracts data in the format agreed with UNITN.

2.2. Web Data Collection

2.2.1. Previous work

Deliverable D2.1 presented the definition of a reach data schema for the collection of data and metadata from social media. Many different social media sources were taken into account (blogs, Twitter, Facebook, Youtube, etc.) and newspaper forums were targeted as the main source of data because of their complex dialogue structure.

In D2.2 we presented the first year data collection with all the efforts done in data crawling, manual data curation by Websays analysts, bug fixing, parsers improvements, etc. Meanwhile partners started using the data and reported feedback to fix inconsistencies, make improvements, crawl other data, etc. Period 1 data collection contained over 4 million posts and over a 1.5 million conversations.

In D2.3 Period 2 data collection was presented with more than 10 M posts. It was also presented all work related to data curation, topic definition, infrastructure (sharding) and data navigation (sort by conversation Size).

2.2.2. Data sources

Following previous period work the SENSEI profile was kept alive and data crawling continued during Period 3. For RATP and Orange profiles data crawling was stopped and a new and profile was created. The 23th of June the EU Referendum took place in the UK and it was considered to be a very interesting use case. Therefore a profile was configured and data was crawled for almost 4 months.

Profiles:

- EU Referendum;
- SENSEI (generic):
    - Charlie Hebdo (terrorism);
    - Germanwings (plane crash);
- RATP (Paris public transportation system);
- Orange (Telephone company).

2.2.3. Content Extraction

Content extraction task is composed of three steps. Each step requires a specially designed and developed module adapted to each of the sources aimed by SENSEI.

- Boiler Plate Detection;
- Content Extraction;
- Structure Parsing;

As data sources continually evolve and make changes to their respective web pages the SENSEI specialized parsers have been updated when necessary. When analysts, partners or the system have detected parsing problems for a given source the involved parser has been updated and fixed.

For Period 3, already existing parsers for i.e. the Guardian have been improved and all comments for a given post are parsed and indexed even if they do not match the Profile query. This improvement allows us to get full conversations from especially relevant newspapers for the EU Referendum Profile.

2.2.4. Pre-processing

All documents crawled for the SENSEI data collection is pre-processed using the Websays pipeline.

The main components for pre-processing documents are:

- **Language Detection**: as mentioned in D2.2 it is very challenging for short texts (especially if they contain brands, acronyms, etc.). The method used to detect the language of a post is:
  - Fast look-up for similar texts with language label corrected by a human;
  - Remove terms that can mislead the automatic classifier;
  - Character heuristics for alphabet-specific languages (e.g. Japanese, Russian);
  - Dictionary based frequent expressions;
  - A HMM based on character n-gram is used to detect the most likely languages;
  - A topic-specific error cost-matrix is used to correct biases (or boost specific languages) for each specific topic;
- **Online-Terms Detection**: a set of regular expressions is used to identity URLs, smi-leys, @authors, hashtags, retweet and forward notations, etc.
- **URL normalization**: URLs in text are typically expressed as relative or partially specified paths, and they can use URL shorteners. In this step URLs are normalized and resolved so that they lead to their full unique URL. During 2015 there have been a special effort to improve URL normalization taking into account new trends in parameterization in newspaper content URLs.
- **Named Entity Detection**: a combined approach is used to named entity detection:
  - Dictionary lookup method. Human analysts built the dictionaries;
  - A CRF model trained on a standard generic named entity corpus is used to detect named entities in English, French, Italian, Spanish and Portuguese.
- **Sentiment Detection:** a combined approach is used for sentiment detection:
  - A weighted-dictionary method is used to detect clearly positive and negative expressions for Spanish, Catalan, English, Italian, French and German;
  - A proprietary nearest-neighbour based method is used to detect similar posts.

2.2.5. Data Statistics

In this section we will describe the data collections obtained after all the Web Data Crawling Process.

Data crawling process for the Websays Profiles configured for RATP and Orange were stopped at the beginning of this third period. RATP and Orange collections were described in previous deliverable (RATP with 0.4 million posts and 163k conversations).

The other two main data collection have been the continuation of the SENSEI Websays Profile (**General News and Newspaper Social Media Publications**) and the Profile created specifically to monitor **EU Referendum** (BREXIT).

2.2.5.1. General News and Newspaper Social Media Publications

The total amount of data collected during the SENSEI project using this profile has been **28.3 million posts** with more than 980K conversations (with at least two posts).

In Figure 5 it can be seen the monthly data crawled per month for this profile during the SENSEI project. It shows that more than 750K posts have been crawled and processed per month.

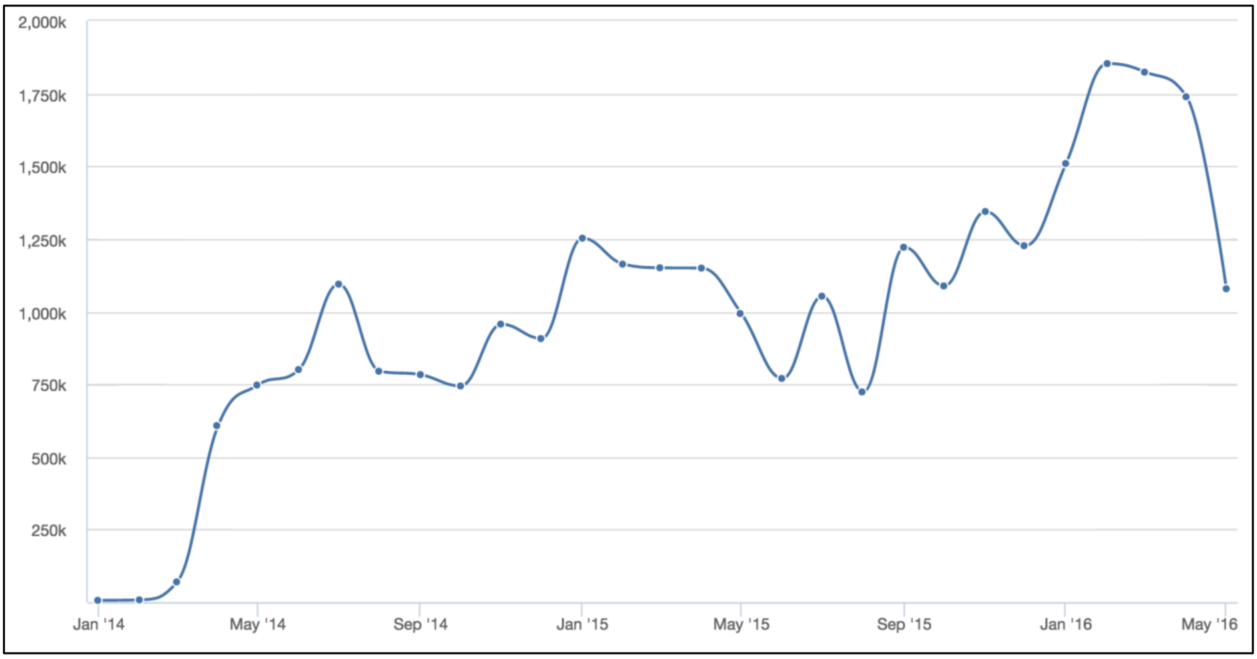


Figure 5: volume evolution

Social Networks represent the 98% of the data as shown in Figure 6.

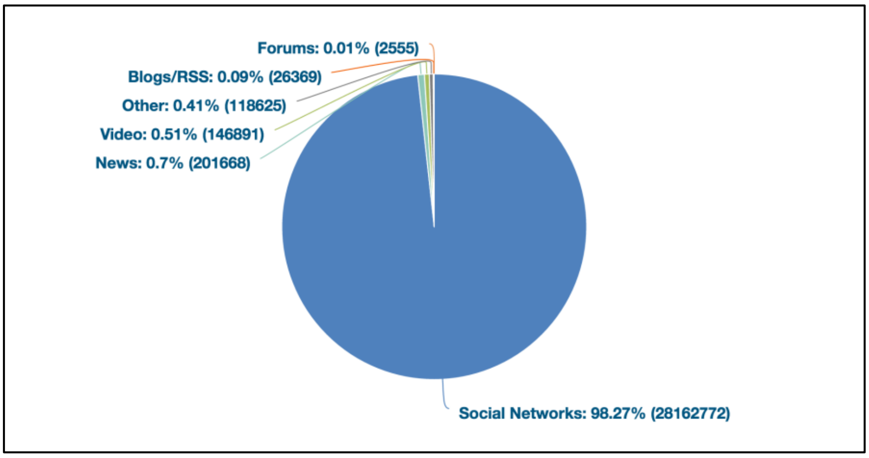


Figure 6: Data sources

Figure 7 shows the data sources in detail. We can see that 41% of the posts crawled are Facebook comments, since the project is focused on conversations this is especially interesting data.

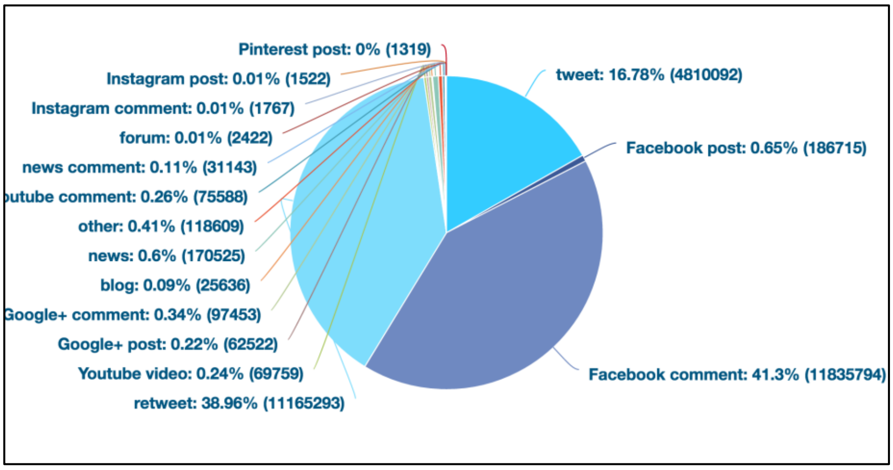


Figure 7: Sources (detailed)

Languages with higher data volumes in this data collection are English (52%), French (19%), and Italian (20%) as it was expected considering the profile configuration.

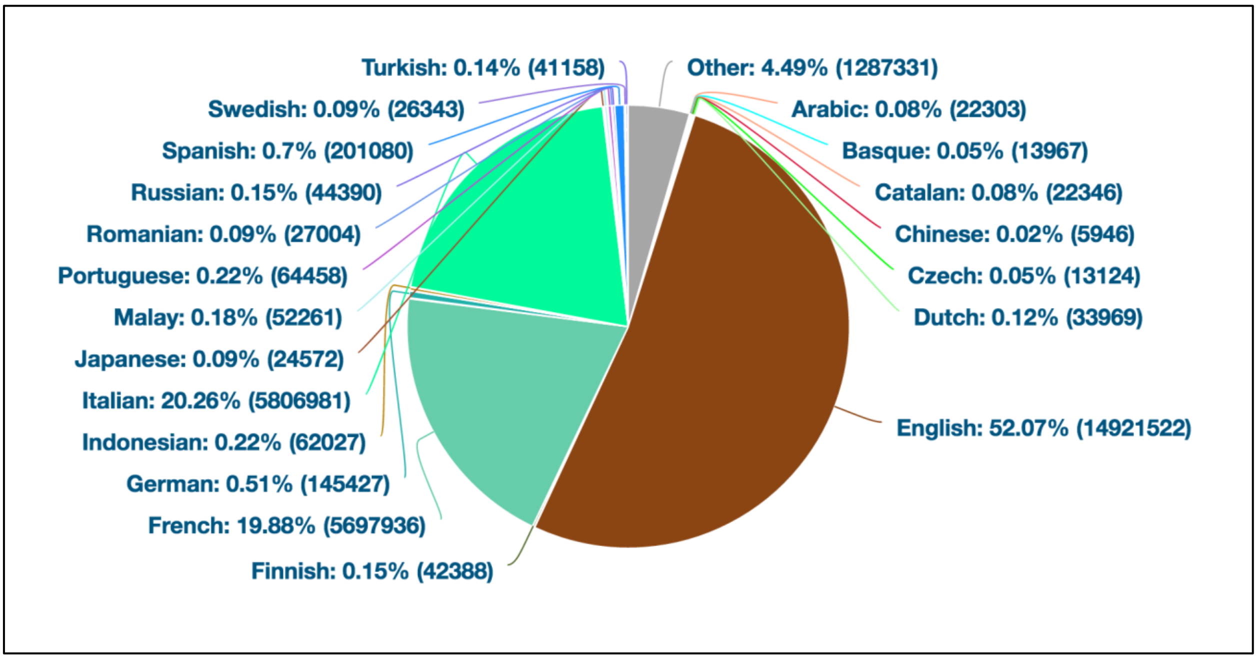
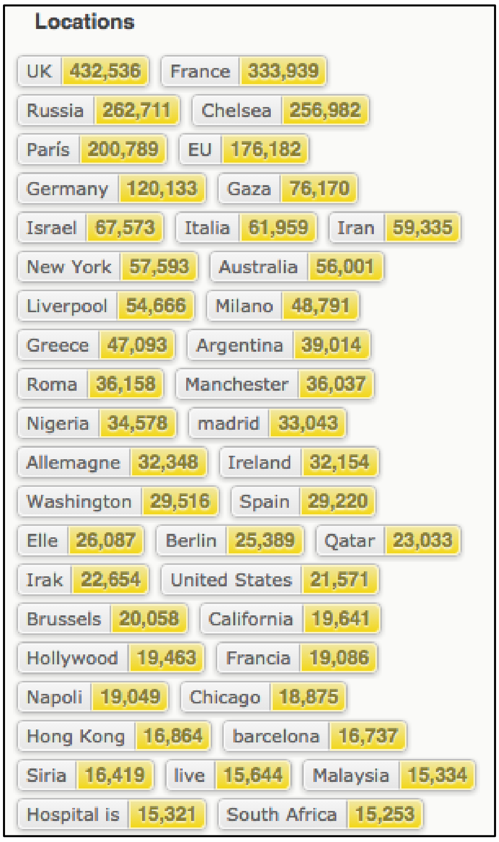


Figure 8: Languages

Next list (Figure 9) shows the names of locations more used in the data set.



Locations

UK	432,536
France	333,939
Russia	262,711
Chelsea	256,982
París	200,789
EU	176,182
Germany	120,133
Gaza	76,170
Israel	67,573
Italia	61,959
Iran	59,335
New York	57,593
Australia	56,001
Liverpool	54,666
Milano	48,791
Greece	47,093
Argentina	39,014
Roma	36,158
Manchester	36,037
Nigeria	34,578
madrid	33,043
Allemagne	32,348
Ireland	32,154
Washington	29,516
Spain	29,220
Elle	26,087
Berlin	25,389
Qatar	23,033
Irak	22,654
United States	21,571
Brussels	20,058
California	19,641
Hollywood	19,463
Francia	19,086
Napoli	19,049
Chicago	18,875
Hong Kong	16,864
barcelona	16,737
Siria	16,419
live	15,644
Malaysia	15,334
Hospital is	15,321
South Africa	15,253

Figure 9: Locations

As it was presented in Deliverable D2.3 during the data crawling process some topics have been configured to help monitor and follow specific events ie. Ebola Crisis, Charlie Hebdo terrorist attack, German wings plane accident, etc. Figure 10 shows the topic volume evolution for some of the topics configured in this account. It can easily be seen that the volume rises during the event dates and drops some days later.

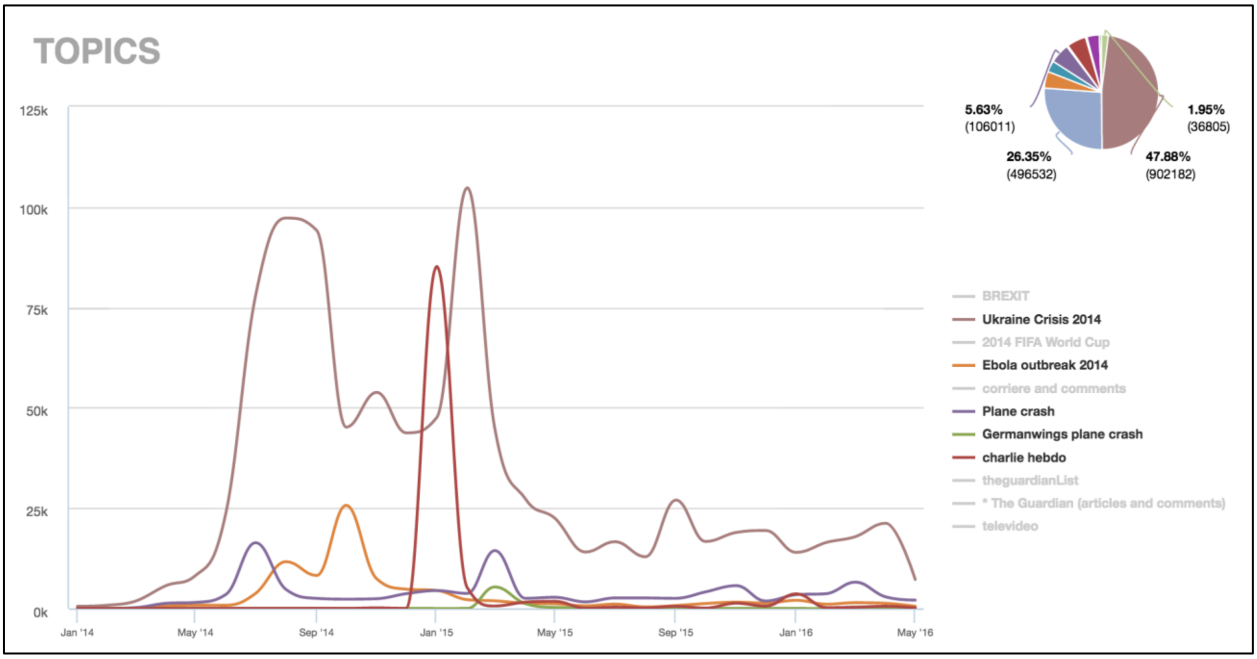


Figure 10: Topic volume evolution

2.2.5.2. *EU Referendum (BREXIT)*

The EU Referendum profile was designed with the aim to obtain a website similar to other elections websites designed and run by Websays. The main idea is to summarize in a single web page the trends and statistics of an Elections or Poll process. Websays has run similar processes for Spanish and Catalan elections (i.e. http://elecciones20d.websays.com/) with rankings, word clouds, topics, etc. with all data obtained from social networks, news, etc. In the SENSEI scenario the goal is to combine this idea with all technology developed in different work packages to get a rich website which is able to provide a general view of what is being said about the referendum and specially to try and predict the result of the poll.

Websays coordinated the SENSE-EU web project and was in charge of the management tasks: SENSEI partners who volunteered to contribute collaborated to obtain the SENSE-EU website[1]. Deliverable D6.3 Section 4 describes de System Infrastructure and Semantic annotations. In this deliverable we will focus on the data crawling and pre-processing tasks only.

Data crawling was started on the 6th of April, the EU Referendum took place on the 23th of June and data crawling was stopped a month after, on the 24th of July. During this interval of almost 4 months **28.6M clippings** have been crawled and processed. **The data set includes 2.1M conversations** with at least two posts.

The profile was configured using the following key words:

EUreferendum, brexit, SayYes2Europe, post-Brexit, brexiteers, #brexit, #no2eu, #notoeu, #betteroffout, #voteout, #eureform, #britainout, #leaveeu, #voteleave, #beleave, #loveeuropeleaveeu, #yes2eu, #yestoeu, #betteroffin, #votein, #ukineu, #bremain, #strongerin, #leadnotleave, #voteremain, #brexitdebate, @LeaveEUOfficial, @nothankseu,

[1] http://www.sense-eu.info

@end_of_europe, @ukleave_eu, référendumeurope, référendumUE, referendumue, referendumuk, #euref

Most of the data crawled comes from social networks (81%) as shown in Figure 11. News data represent 16% of the data.

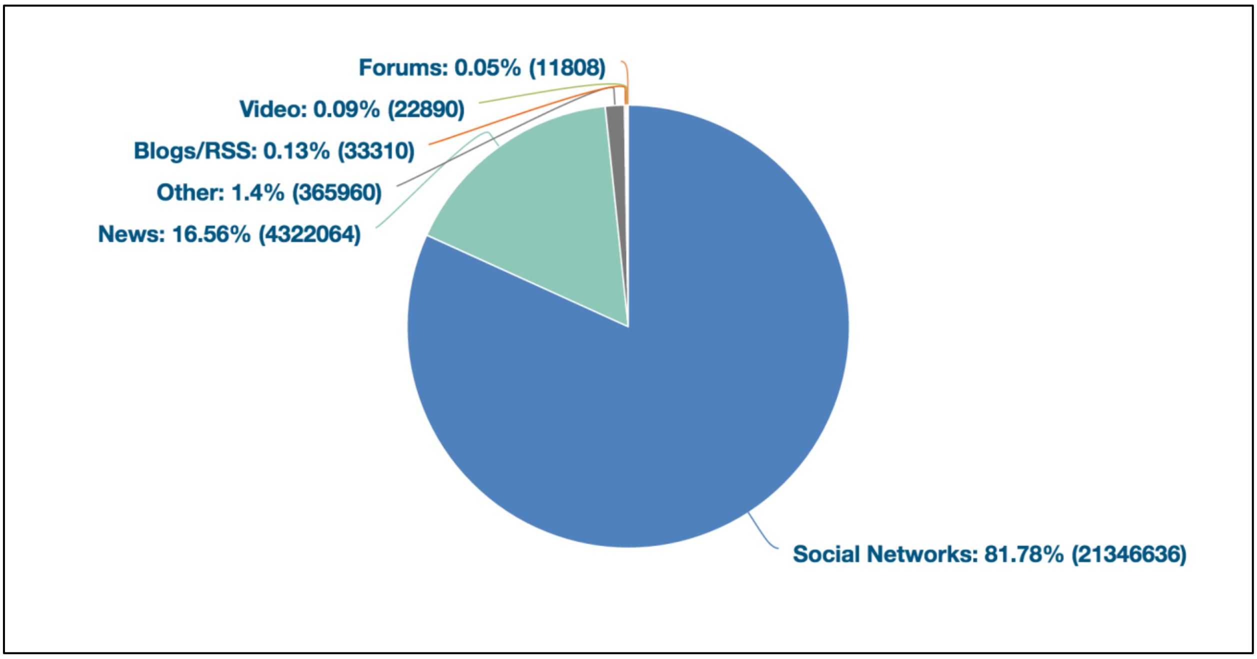


Figure 11: EU Ref. Sources

If we look into Figure 12 we can see that there are more than 3M News comments, which combined, generate rich conversations.

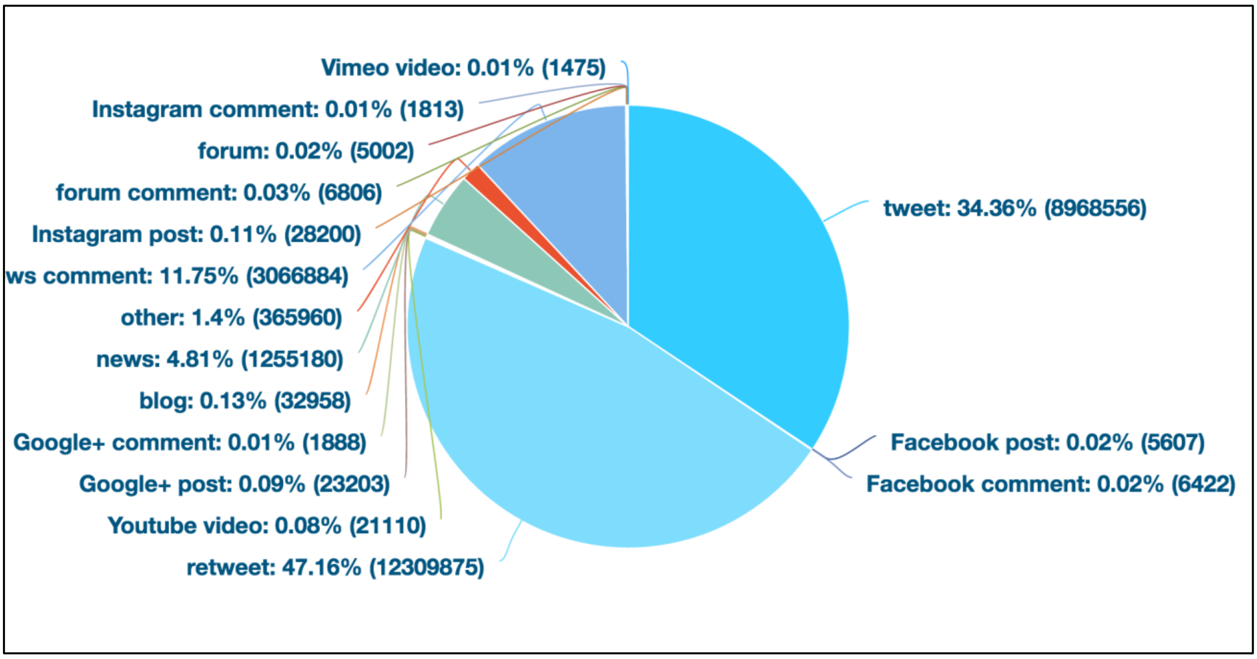
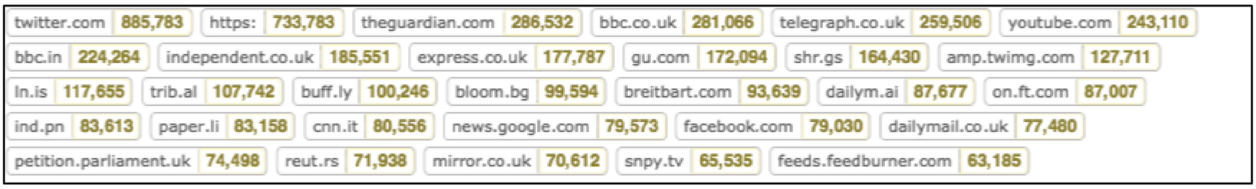


Figure 12: EU Ref. Sources in detail

The following picture (Figure 13) shows the most written domains in the data set.



twitter.com	885,783	https:	733,783	theguardian.com	286,532	bbc.co.uk	281,066	telegraph.co.uk	259,506	youtube.com	243,110
bbc.in	224,264	independent.co.uk	185,551	express.co.uk	177,787	gu.com	172,094	shr.gs	164,430	amp.twimg.com	127,711
ln.is	117,655	trib.al	107,742	buff.ly	100,246	bloom.bg	99,594	breitbart.com	93,639	dailym.ai	87,677
on.ft.com	87,007	ind.pn	83,613	paper.li	83,158	cnn.it	80,556	news.google.com	79,573	facebook.com	79,030
dailymail.co.uk	77,480	petition.parliament.uk	74,498	reut.rs	71,938	mirror.co.uk	70,612	snpy.tv	65,535	feeds.feedburner.com	63,185

Figure 13: EU Ref. Mentioned domains

The dataset includes different languages but most of the posts are in English (84%) as shown in Figure 14.

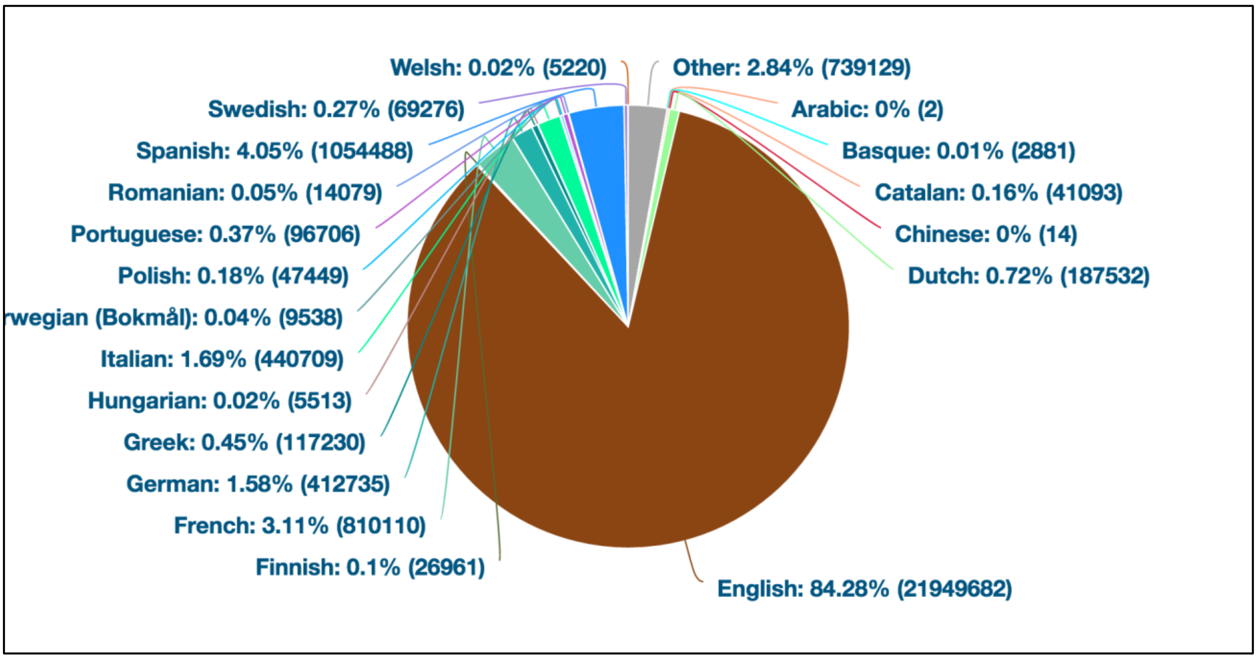
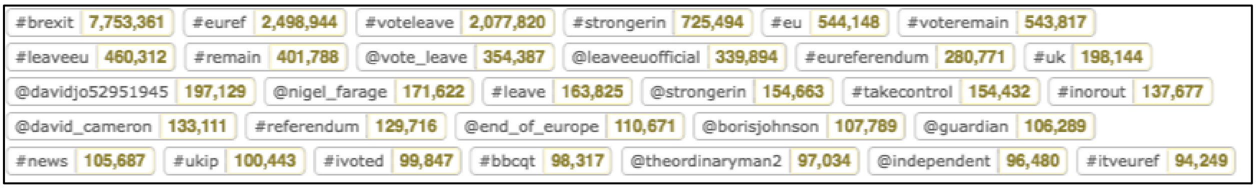


Figure 14: EU Ref. languages

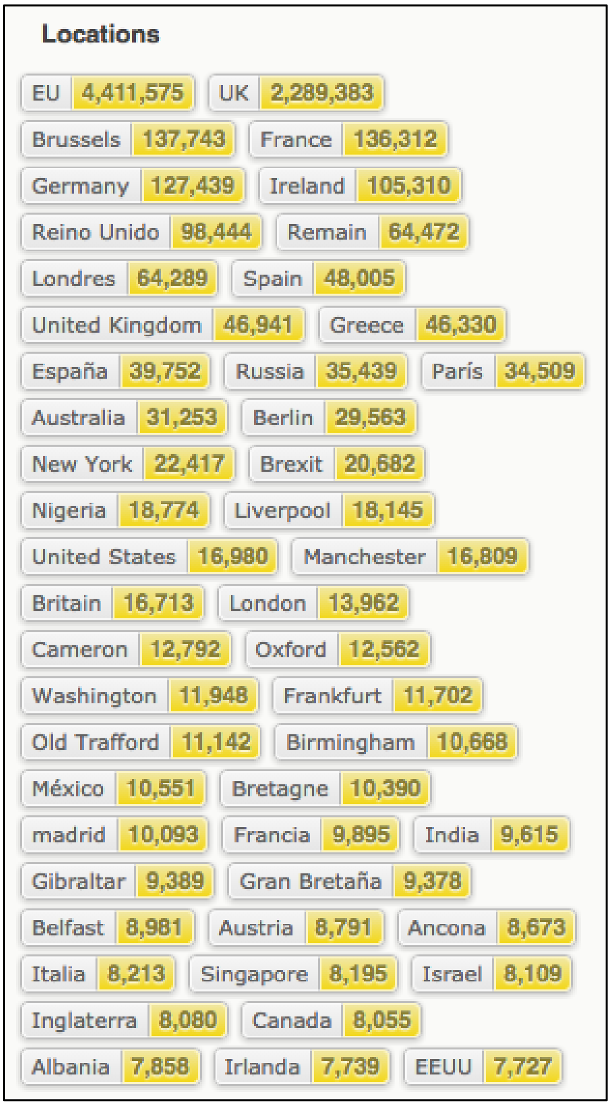
The following list (Figure 15) shows the list of mostly used authors and hash tags in the data set. These hash tags have been widely used during the EU referendum for social network and online newspaper users to easily express their opinions and position (i.e. #voteleave, #strongerin).



#brexit	7,753,361	#euref	2,498,944	#voteleave	2,077,820	#strongerin	725,494	#eu	544,148	#voteremain	543,817		
#leaveeu	460,312	#remain	401,788	@vote_leave	354,387	@leaveeuofficial	339,894	#eureferendum	280,771	#uk	198,144		
@davidjo52951945	197,129	@nigel_farage	171,622	#leave	163,825	@strongerin	154,663	#takecontrol	154,432	#inorout	137,677		
@david_cameron	133,111	#referendum	129,716	@end_of_europe	110,671	@borisjohnson	107,789	@guardian	106,289				
#news	105,687	#ukip	100,443	#ivoted	99,847	#bbcqt	98,317	@theordinaryman2	97,034	@independent	96,480	#itveuref	94,249

Figure 15: EU Ref. mentioned authors and hash tags

EU, UK and *Brussels* are the three most mentioned locations in the data set. Figure 16 shows in detail the top most mentioned locations.



Locations

EU	4,411,575
UK	2,289,383
Brussels	137,743
France	136,312
Germany	127,439
Ireland	105,310
Reino Unido	98,444
Remain	64,472
Londres	64,289
Spain	48,005
United Kingdom	46,941
Greece	46,330
España	39,752
Russia	35,439
París	34,509
Australia	31,253
Berlin	29,563
New York	22,417
Brexit	20,682
Nigeria	18,774
Liverpool	18,145
United States	16,980
Manchester	16,809
Britain	16,713
London	13,962
Cameron	12,792
Oxford	12,562
Washington	11,948
Frankfurt	11,702
Old Trafford	11,142
Birmingham	10,668
México	10,551
Bretagne	10,390
madrid	10,093
Francia	9,895
India	9,615
Gibraltar	9,389
Gran Bretaña	9,378
Belfast	8,981
Austria	8,791
Ancona	8,673
Italia	8,213
Singapore	8,195
Israel	8,109
Inglaterra	8,080
Canada	8,055
Albania	7,858
Irlanda	7,739
EEUU	7,727

Figure 16: EU Ref. mentioned locations

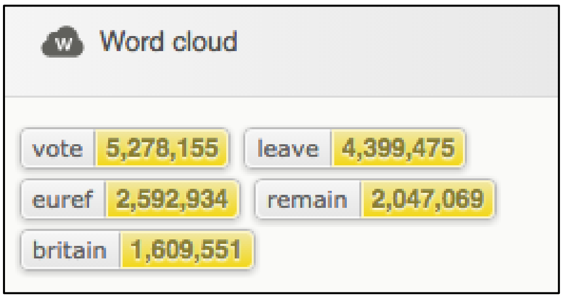
The dataset most mentioned people are:

1. “Obama” (241K);
2. “Boris Johnson” (233K);
3. “David Cameron” (179K);
4. “Angela Merkel” (126K);
5. “Nigel Farage” (108K);
6. “Jeremy Crobyn” (77K);
7. “Michael Gove” (60K);
8. “George Osborne” (56K);
9. “Donald Trump” (39K);
10. “Jonathan Freedland” (36K).

The basic word cloud of the data set is presented in the following pictures (Figure 17 and Figure 18).

leave
britain
remain
vote
euref

Figure 17: EU Ref. Word Cloud


Word cloud

vote 5,278,155
leave 4,399,475
euref 2,592,934
remain 2,047,069
britain 1,609,551

Figure 18: EU Ref. Word Freq. list

2.2.5.3. EU Ref. Topic Definition

For the EU Ref. profile several topics were configured. Websays analysts review data and configure these topics constantly. For each topic a set of keywords are used to define “required terms” and “banned terms”. One post belongs to the topic if it contains any of the required terms and none of the banned.

The main topics for this profile are “Against EU” and “FOR EU”:

AGAINST_EU Topic definition:

- *required*: "#beleave", "#betteroffout", "#Bexit", "#brexitdebate", "#britainout", "#Czexit", "#Dexit", "#FarmersGO", "#FishingGO", "#Frexit", "#go", "#Itexit", "#LeaveEU", "#Lexit", "#Nexit", "#Nexit", "#noTTIP", "#nottip", "#Pexit", "#Spexit", "#StudentsGO", "#TakeControl", "#VoteLeave", "#VoteNO", "#voteout", "@end_of_europe", "@leaveeuofficial", "@NoThanksEU", "@nothankseu", "@ukleave_eu", "@vote_leave", "euroscepticism", "exit", "Fexit", "leave", "leaving", "vote #brexit", "vote brexit", "vote no"
- *banned*: "#betteroffin", "#bremain", "#leadnotleave", "#loveeuropeleaveeu", "#Remain", "#strongerin", "#ukineu", "#votein", "#voteremain", "#VoteYES", "#yes2eu", "#yestoeu", "remain", "SayYes2Europe", "SayYes2Europe", "stay"]

FOR_EU Topic definition:

- *required*: "#betteroffin", "#bremain", "#leadnotleave", "#loveeuropeleaveeu", "#Remain", "#strongerin", "#ukineu", "#votein", "#voteremain", "#VoteYES", "#yes2eu", "#yestoeu", "remain", "SayYes2Europe", "SayYes2Europe", "stay"

- *banned*: "#beleave", "#betteroffout", "#brexitdebate", "#britainout", "#FarmersGO", "#FishingGO", "#go", "#LeaveEU", "#Lexit", "#Nexit", "#noTTIP", "#nottip", "#StudentsGO", "#TakeControl", "#VoteLeave", "#VoteNO", "#voteout", "@end_of_europe", "@leaveeuofficial", "@NoThanksEU", "@nothankseu", "@ukleave_eu", "@vote_leave", "euroscepticism", "exit", "leave", "leaving", "vote no"

2.2.6. Automatic Sharding For Large Data Profiles

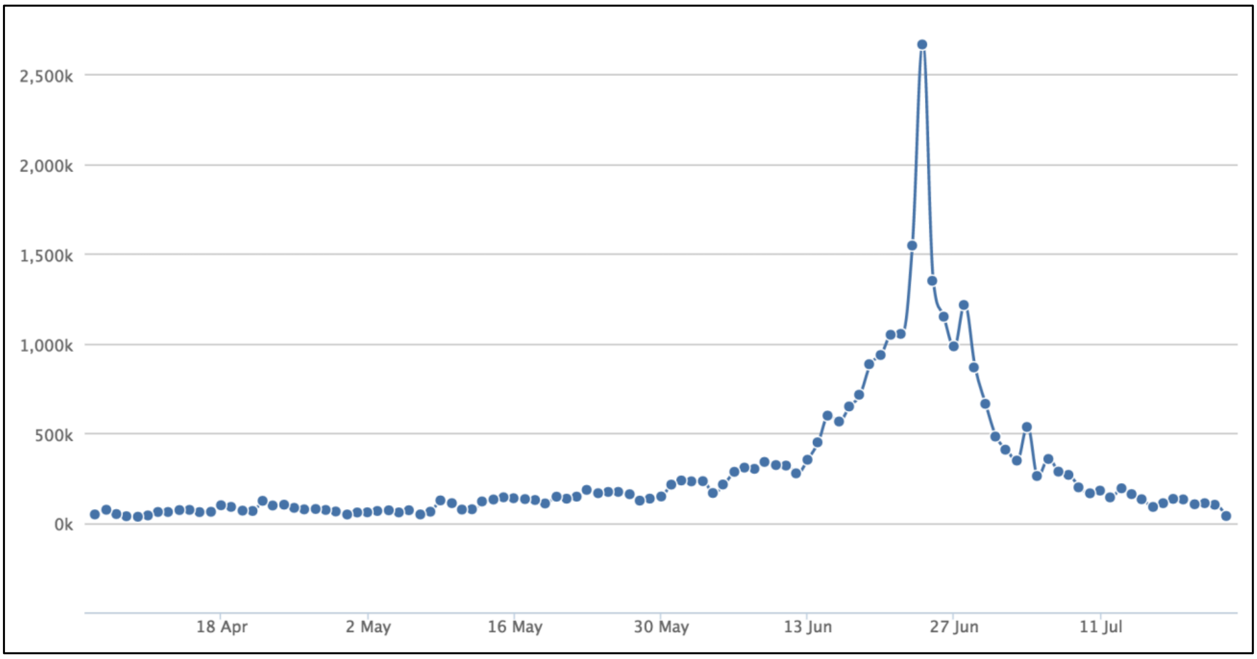


Figure 19: EU Ref. volume evolution

The EU Ref. profile has crawled more than 28M posts in 4 months. This is an average of more than 230K posts per day. In Figure 19 it can be seen the daily volume evolution (posts crawled and processed per day evolution). It can be seen that there are several days with more than 500K posts. During these days more than 5 posts musts be crawled, processed and stored per second. If we focus on the poll date we can see a volume of more than 2.5M posts. This means that during that day more than 28 posts were processed per second.

As presented in D2.3 a technique to split the data collection index in Solr was used to be able to handle a large collection while keeping the index fast (divided into two part or shards). This process was done manually and periodically. When the part where posts were indexed was too full, a part of the posts was moved to the second shard.

For the EU Ref. profile, which has large data volumes in a short period of time, it was necessary to make the sharding process much more often. To this aim, an automatic sharding process was designed and implemented.

The system periodically (daily, during hours with less volume) checked if it was necessary to move data to the second shard. If it was, oldest posts, were moved from one shard to the other.

This way the shard were data is constantly being indexed keeps a reasonable size keeping it fast and able to handle the large amounts of work.

2.2.7. Automatic data dump for data sharing

To be able to update the SENSE-EU website on a daily basis it was necessary to make previous day data crawled and processed for the EU Ref. profile available to other partners. To this aim, it was given access to SENSEI partners to a folder in a Websays server were all data was prepared every day. This data dump included the previous day data:

- Posts dump in the SENSEI xml format;
- Top Authors json (per language and total);
- Top Mentions json (per language and total);

  Word Clouds json (per language and total).

Due to the large amount of data, generating all these dumps could take several hours. The process was started soon after midnight to make sure data was ready for other SENSEI partners early in the morning.

3. Conclusions

With respect to social media data collection and processing, work on Period 2 has been continued. 28.3 million posts have been crawled and processed for the whole project for general news and social media. Updates and improvements have been added to parsers allowing the system to crawl much more user generated content. A new Profile and topics have been specially configured for the EU Referendum in the UK (Brexit). An automatic sharding system has been developed to automatically split data in different Solr shards to allow the system to handle the large amount of input data. Brexit data has been automatically processed and shared with SENSEI partners to build and update the sense-eu.info website on a daily basis. For the Brexit profile 28.6 million posts have been crawled and processed.

With respect to speech, 300 LUNA synopsis have been annotated with templates and slot labeling. LUNA and DECODA corpus annotation have been reviewed and machine generated annotations have been analyzed.

Data to be shared beyond the consortium has been prepared following D8.4 “Second Ethical Issues Report” conclusions. Data has been shared inside the consortium using an SVN repository, sharing server access users (SSH), and data dump automatic processes.

Some of the developments have already been exploited commercially by integrating them into Websays system. The parsers’ updates and improvements benefit all clients by enriching the data quality and diversity. The automatic sharding process makes large profiles able to handle large amounts of data volume per day.

REFERENCES

[Artiles et al., 2007] Artiles, J., Gonzalo, J., and Sekine, S. (2007) "The SemEval-2007 WePS evaluation: Establishing a benchmark for the web people search task", In Proc. of SemEval.

[Artstein and Poesio, 2008] Artstein, R. and Poesio, M. (2008) "Intercoder agreement for Computational Linguistics", Computational Linguistics, 34(4).

[Asher and Lascarides, 2003] Asher, N. and Lascarides, A. (2003) The Logic of Conversation. Cambridge University Press.

[Bagga and Baldwin, 1998] Bagga, A., Baldwin, B. (1998) "Entity-based cross-document coreferencing using the vector space model", In Proc. of COLING/ACL.

[Baker et al., 1998] Baker, F.C., Fillmore, J.C., and Lowe, B.J. (1998) "The Berkeley FrameNet project", In Proc. of COLING/ACL.

[Barzilay and Lee, 2004] Barzilay, R. and Lee, L. (2004) "Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization", In Proc. of NAACL-HLT.

[Bazillon et al., 2012] Bazillon, T., Deplano, M., Bechet, F., Nasr, A., and Favre, B. (2012) "Syntactic annotation of spontaneous speech: application to call-center conversation data", In Proc. of LREC.

[Bechet and Nasr, 2009] Bechet, F. and Nasr, A. (2009) "Robust dependency parsing for Spoken Language Understanding of spontaneous speech", In Proc. of INTERSPEECH.

[Bechet et al., 2012] Bechet, F., Maza, B., Bigouroux, N., Bazillon, T., El-Bèze, M., De Mori, R., and Arbillot, E. (2012) "DECODA: a call-center human-human spoken conversation corpus", In Proc. of LREC.

[Blitzer et al., 2006] Blitzer, J., McDonald, R., and Pereira, F. (2006) "Domain Adaptation with Structural Correspondence Learning", In Proc. of EMNLP.

[Byrd et al., 2008] Byrd, R.J., Neff, M.S., Teiken, W., Park, Y., Cheng, K.S.F, Gates, S.C., and Visweswariah, K. (2008) "Semi-automated logging of contact center telephone calls", In Proc. of CIKM.

[Carlson et al., 2001] Carlson, L., Marcu, D., and Okurowski, M.E. (2003) "Building a discourse-tagged corpus in the framework of rhetorical structure theory". In J. Kuppevelt and R. Smith (eds) Current Directions in Discourse and Dialogue. Kluwer.

[Chambers and Jurafsky, 2011] Chambers, N. and Jurafsky, D. (2011) "Template-based information extraction without the templates", In Proc. of ACL.

[Chen and Martin, 2007] Chen, Y. and Martin, J. (2007) "Towards robust unsupervised personal name disambiguation", In Proc. of EMNLP.

[Coppola et al., 2009] Coppola, B., Moschitti, A., and Riccardi, G. (2009) "Shallow Semantic Parsing for Spoken Language Understanding", In Proc. of NAACL.

[Csomai and Mihalcea, 2008] Csomai, A. and R. Mihalcea (2008) "Linking documents to encyclopedic knowledge", IEEE Intelligent Systems.

[Daumé, 2007] Daumé III H. (2007) "Frustratingly Easy Domain Adaptation", In Proc. of ACL.

[Dinarelli et al., 2009] Dinarelli, M., Quarteroni, S., Tonelli, S., Moschitti, A., and Riccardi, G. (2009) "Annotating Spoken Dialogs: from Speech Segments to Dialog Acts and Frame Semantics", In Proc. of EACL Workshop on Semantic Representation of Spoken Language.