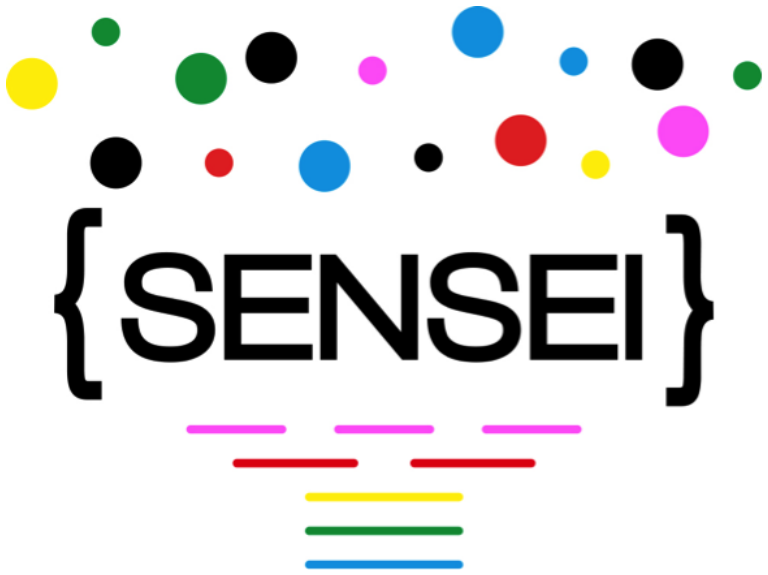


# D3.3 – Report on the para-semantic parsing of conversations (spoken and text)

| | |
|---|---|
| Document Number | D3.3 |
| Document Title | Report on the para-semantic parsing of conversations (spoken and text) |
| Version | 1.0 |
| Status | Final |
| Workpackage | WP3 |
| Deliverable Type | Report |
| Contractual Date of Delivery | 31.10.2016 |
| Actual Date of Delivery | 31.10.2016 |
| Responsible Unit | UNITN |
| Keyword List | Parasemantics, Mood, Empathy |
| Dissemination level | PU |

**Editor**
Fabio Celli (University of Trento, UNITN)

**Contributors**

| | |
|---|---|
| Fabio Celli | (University of Trento, UNITN) |
| Evgeny Stepanov | (University of Trento, UNITN) |
| Firoj Alam | (University of Trento, UNITN) |
| Shammur Absar Chowdhury | (University of Trento, UNITN) |
| Frederic Bechet | (Aix Marseille Université, AMU) |
| A. R. Balamurali | (Aix Marseille Université, AMU) |
| Mickael Rouvier | (Aix Marseille Université, AMU) |
| Benoit Favre | (Aix Marseille Université, AMU) |
| Jeremy Auguste | (Aix Marseille Université, AMU) |

**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 | 2016-07-20 | Draft | Fabio Celli (UNITN) | Initial Outline |
| 0.1 | 2016-08-11 | Draft | Fabio Celli (UNITN) | Executive Summary |
| 0.1 | 2016-08-12 | Draft | Fabio Celli (UNITN) | Introduction |
| 0.2 | 2016-08-25 | Draft | Firoj Alam (UNITN) | Mood section |
| 0.2 | 2016-08-25 | Draft | Fabio Celli (UNITN) | Edited Mood section |
| 0.2 | 2016-08-26 | Draft | Fabio Celli (UNITN) | Competitive overlapping |
| 0.2 | 2016-08-26 | Draft | Fabio Celli (UNITN) | Empathy |
| 0.2 | 2016-08-27 | Draft | Fabio Celli (UNITN) | Introduction and follow-up |
| 0.3 | 2016-08-30 | Draft | Frederic Bechet (AMU) | Overview |
| 0.3 | 2016-08-30 | Draft | A.R. Balamurali (AMU) | Problematic call detection |
| | | | J. Auguste (AMU) | |
| | | | F. Bechet (AMU) | |
| 0.3 | 2016-08-30 | Draft | M. Rouvier (AMU) | Stance Detection |
| 0.3 | 2016-08-30 | Draft | Shammur Absar Chowdhury (UNITN) | Overlapping |
| 0.3 | 2016-08-30 | Draft | Firoj Alam (UNITN) | Empathy |
| 0.3 | 2016-08-30 | Draft | Evgeny Stepanov (UNITN) | Introduction and follow-up |
| 0.3 | 2016-09-10 | Final | Mijail Kabadjov (ESSEX) | Scientific review |
| 0.3 | 2016-09-20 | Final | Elisa Chiarani (UNITN) | Quality check |
| 0.4 | 2016-09-23 | Final | Fabio Celli (UNITN) | Editor's corrections |
| 0.5 | 2016-10-07 | Final | Elisa Chiarani (UNITN) | Final Quality check |
| 0.4 | 2016-10-14 | Final | Giuseppe Riccardi (UNITN) | Approval for submission |

# List of Acronyms and Abbreviations

| Acronym | Meaning |
|---|---|
| ACOF | Agent Conversation Observation Form |
| BART | Beautiful Anaphora Resolution Toolkit |
| CLI | command-line interface |
| CRF | Conditional Random Field (learning algorithm) |
| CSS | Cascading Style Sheets |
| CNN | Convolutional Neural Networks (learning algorithm) |
| DAs | Dialogue Acts |
| FBK | Fondazione Bruno Kessler |
| GATE | General Architecture for Text Engineering |
| GPL | GNU Public License |
| GUI | graphical user interface |
| HTML | HyperText Markup Language |
| HTTP | HyperText Transfer Protocol |
| ISF | Interactive Summarization Framework |
| JAPE | Java Annotation Patterns Engine |
| JSON | JavaScript object notation |
| ML | machine learning |
| NER | named entity recogntion |
| NLP | natural language processing |
| PDTB | Penn Discourse Treebank |
| PHP | PHP Hypertext Preprocessor |
| PNG | Portable Network Graphics |
| POS | part of speech |
| QA | quality assurance |
| REST | Representational State Transfer |
| RMSE | Root Mean Squared Error |
| ROUGE | Recall-Oriented Understudy for Gisting Evaluation |
| SQL | Structured Query Language |
| SMO | Sequential Minimum Optimization (learning algorithm) |
| SVM | Support Vector Machines (learning algorithm) |

# Executive Summary

The objectives of WP3 is to automatically generate a structured semantic and parasemantic representation of human-human conversations in three languages: English, French,and Italian. In this deliverable we report the work done on the representation and extraction of parasemantic features for speech and social media over Period 3. The work done in WP3 has a strong impact on the applications of parasematic analysis tools in summarisation and in the monitoring of the 2016 Brexit referendum conversations, reported in D5.3 and D6.3 respectively.
Following the use case scenarios designed in Period 1 for speech and social media, we identified different parasemantic tasks: empathy detection from speech, mood detection for both speech and social media, identification of problems in telephone conversations, competitive overlap detection in speech, and stance detection in social media. A final sixth task addresses the parasemantic and discourse structure parsing: the Agreement/Disagreement relation extraction, which is reported in D4.3.

# Contents

# 1 Introduction

The main objectives of WP3 is to automatically generate a structured semantic and para-semantic representation of human-human conversations. These structured semantic representations will be obtained for the three languges of the SENSEI project (English, French and Italian) through a parsing process done at three linguistic levels: syntactic, semantic and para-semantic. This WP is made of three tasks corresponding to different levels of parsing: syntactic, semantic, and para-semantic parsing, as well as to multi-domain and multi-modality adaptation. Syntactic parsing is the segmentation of each conversation into the syntactic dependencies existing between the main syntactic elements, or chunks. Semantic parsing is a predicate/argument extraction based on a FrameNet model. Finally the para-semantic level described in this deliverable, concerns pragmatics, addressing tasks such as competitive overlapping in speech and mood extraction from social media textual data.
WP3 is divided into 3 tasks: 1) Task 3.1. Feature extraction from FrameNet Semantic Parsing. The aim is to automatically produce frame-semantic representation of a sentence. 2) Task 3.2. Extraction of features for para-semantic prediction. The goal is the development of extraction methods of non-verbal cues to predict para-semantic features such as mood, emotions and behavioral patterns. Task 3.3. Multi-domain and cross-media model adaptation. This task is about the use of unsupervised or weakly supervised methods for adapting syntactic and semantic parsing models from one application-domain (or modality) to another application-domain (or modality).
During P1 we performed Task 3.1 and 3.3. We used a combination of models from flat entity-based annotation to FrameNet-based models.The frame parser is based either on off-the-shelf tools adapted on the input and output channels to the tasks or corpus-specific parsers developed when some level of annotation is available on a given corpus on which new models can be trained. In D3.1 we reported about 1) the Adaptation of a syntactic parser to process spontaneous speech with limited supervision; 2) the Frame annotation of the French RATP-DECODA corpus; 3) the Evaluation of the state-of-the-art FrameNet parser SEMAFOR on SENSEI data and 4) Cross-language methodology adaptation, in other words when no available parser or annotated corpus is available for a given language, we translated the source document to English, using a generic parser then align the output with the source document. We applied this methodology to the Italian LUNA data and we compared it to the output of corpus-specific parsers, developed during the LUNA project.
During P2 we addressed mainly Task 3.3, using generic rich linguistic resources available in the three SENSEI languages in conjunction with cross-language and cross-domain adaptation methodologies. The three adaptation approaches are addressed: cross-language adaptation via Statistical Machine Translation cross-domain adaptation through re-ranking of n-best lists of generic or in-domain parsers, and cross-domain and cross-language adaptation of word embeddings. The first two approaches are addressed on the FrameNet semantic parsing task, the latter on the tasks of sentiment lexicon translation and also frame-semantic parsing. We have observed that the cross-language adaptation with re-ranking methodology performs significantly worse than the in-domain semantic models. Moreover, the in-domain Italian semantic parser improves significantly with the re-ranking methodology. Therefore, we abandon the cross-language adaptation with re-ranking methodology and use the re-ranking methodology, in case any in-domain data is available in the desired language. From the cross-language adaptation of word embeddings, we have observed that adapting an embedding space is good for sentiment lexicon translation. For the cross-domain word embedding adaptation, the proposed approach

outperforms state-of-the-art Conditional Random Field approach on the frame-semantic parsing tagging task when little in-domain adaptation data is available. This deliverable is structured as follows: in the following sub-section we review the follow-up recommendations from the second year review report, in section 2 we briefly review the approaches to para-semantic analysis in NLP, then the sections 3 and 4 describe the parasemantic tools we developed for speech and social media respectively.

## 1.1 Follow-up to Period 2 Activities

In this deliverable we report the work done for Task 3.2: the extraction of parasemantic features. This task, scheduled to start in P2 and end in P3, involved two partners of the consortium: UNITN and AMU. At the end of Period 1, in the deliverable D3.1, we have presented the semantic models and the parsing methodology developed in WP3 for processing the Human-Human SENSEI conversations for social media and speech data. The different methods for producing semantic representations were introduced from the three SENSEI languages English, French, and Italian either using corpus-specific or generic tools. At the end of Period 2 we have presented cross-languages adaptation methodology with subsequent cross-domain re-ranking for generic tools in the resource-rich languages like English and compared this methodology with the re-ranking of the output of corpus-specific semantic parser. As expected, the outcome of period 2 demonstrated that the use of generic models with or without cross-language methodology produces lower performance. However, re-ranking of domain-specific models further improves the performance. Thus, using generic tools with cross-language methodology is left as the last resort in the case of absence of any in-domain annotated data.
The work done on adaptation with vector-space models, either cross-language or cross-domain, opened new pathways for other SENSEI task and WPs. In particular, we designed cross-language systems for parasemantic feature extraction in WP3 (mood extraction), WP4 (Agreement/Disagreement structures), in summarisation and for monitoring and predicting brexit (see D6.3).

**Follow-up to Recommendations from Previous Review** A recommendation from the second year review related to WP3 among others, addressed the issue of two summarisation techniques, developed respectively by AMU and UNITN, trained on different languages (French and Italian respectively). The recommendation was to apply the techniques to at least one common language, in order to have an evaluation of the system performances on the same data.
Given that the evaluation of social media is in English, we designed language-adaptive systems for the extraction of mood from social media (reported in this deliverable) and the extraction of Agreement/Disagreement structures (reported in D4.3) in order to exploit them in the evaluation of social media summrisation in English, even if the training set was in Italian. The results of the social media evaluation are reported in D6.3.

# 2 Overview of Approaches to Para-semantic Analysis in NLP

Thinking about para-semantics in NLP, the most popular tasks are opinion mining [1] and sentiment analysis [2, 3]. In the last decade these tasks strongly attracted the attention of the media, in particular because they allowed researchers and companies to automatically analyze large datasets from social media and microblogging websites, tracking reputation of public figures and opinions about companies products. Existing tools for NLP, such as dictionaries and lexical resources, are the main method used in literature for the extraction of features from text, and in recent years there has been a great effort towards the creation of lexica designed for specific domains [4], in order to capture domain-specific sentiment clues as the presence of intensifiers, emoticons/emojis, abbreviations and hashtags.
Previous work in the detection of parasemantic dimenstions in conversations can be divided into three areas: 1) the definition of parasemantic tasks, the development and annotation of corpora and lexical resources, 3) the prediction of parasemantic dimensions or the extraction of parasemantic features from synchronous conversations, such as dyadic spoken calls, and asynchronous conversations, such as blog comments in social media.

**Overview of Para-semantic tasks for spoken conversations** Parasemantics concerns the communication levels traditionally classified in domains different from semantics, like pragmatics. This includes several tasks ranging from sentiment analysis and opinion mining to stance detection, mood and Agreement/Disagreement. In general, what we find in conversations can be defined in terms of shared public commitments, that ground the speech acts performed by the participants within the conversation [5]. What differs from one task to another is the problem to be solved and how it is addressed. From an operational point of view, the definitions of parasemantic tasks differ considerably in synchronous and asynchronous conversations. In synchronous conversations, such as call center conversations, numerous studies exist for call center related research related to behavioural and conversational analysis dimension [6, 7, 8, 9]. In this case parasemantics is related to the evaluation of the *quality* of a call, in terms of efficiency or customer experience.

Indeed, evaluating the quality of a spoken conversation can be seen, on the customer/client side, as a particular case of opinion/sentiment analysis, assuming that callers satisfaction can be inferred from the expression of positive or negative opinions and sentiments [10]. On the operator/agent side, the recognition of speaker personality traits can characterize some aspects of a dialogue. Research programs such as the *Interspeech Speaker Trait Challenge* [11] have focused on this task, aiming at recognizing traits that can be relevant to this study such as *Conscientiousness* (Efficient, organized, etc.) or *Agreeableness* (Compassionate, cooperative etc). If most of these studies have been made on broadcast speech, spontaneous conversational speech was also studied within the same framework [12]. It was shown in [12] that if acoustic features were the dominant set of features to predict personality, lexical features played also an important role.

A quality management system to asses the service of call centers was introduced by previous work [6]. The system uses a set of features to classify each call as good or bad based on different aspects of the quality questionnaire. The focus is on procedural aspects of the call. Studies suggest that a customer

service can vastly be improved by studying the customer behaviour [8]. However, this cannot be easily performed as most often data cannot be used for analysis because of the privacy issues.

Recently [13] described a system called $QA^{RT}$ dedicated to evaluate both Quality assurance (QA) and customer satisfaction (C-Sat). This is very similar to some of the SENSEI goals, the main difference being that some analysis of their method relied on supervised learning with specific annotations. In our approaches we only use unsupervised or *free* annotations already existing in call centers.

**Overview of para-semantic tasks for social media** On the other hand there are several parasemantic tasks that address asynchronous social media conversations. For example, apart from the very popular sentiment analysis and opinion mining tasks, there are emerging parasemantic tasks such as stance detection [14] and Agreement/Disagreement structures extraction [15]. For example the extraction of relationships among participants to a multiparty conversation, expressed at message level, a post or turn text unit [16]; the same can be done between pairs of sentences belonging or not to the same thread [17]. Another way to define a stance detection or Agreement/Disagreement task is the extraction of Quote-Response DAs pairs and triplets. These pairs and triplets are linked by the structure of the thread, where each message is about the same topic [18]. Regarding mood and emotions, the affective dimension of text has been mainly analyzed in terms of positive and negative polarity [19, 20, 4] for decades, although more detailed dimensions are proven to be very useful. Mood extraction tasks differ for the dimensions they can extract and predict, such as tension, depression, anger, vigor, fatigue, and confusion in tweets are found to be good predictors of stock market exchanges [21]. It has also been demonstrated that it is possible to predict anger, sadness, and joy from Livejournal blog with performances up to 78% of accuracy [22].
Among the corpora and tools available for parasemantic tasks, there are corpora and lexica for sentiment analysis, such as SenticNet [4] or other multilingual resources [23]. Not so many corpora are available instead for tasks like stance detection in social media, such as the IAC corpus [18], a large collection of political debates in English extracted from the website *4forums.com*. Among the existing resources none was available in Italian and French. We produced a resource for Italian (the CorEA corpus, reported in D2.1 and 2.2) that we used for training systems for mood detection and Agreement/Disagreement argument structures.

# 3 Para-semantic Parsing of Spoken Conversations

## 3.1 Problematic Call Detection

Detecting problematic calls in telephone call centre is critical for companies in order to assess the quality of the service provided. The definition of a *problematic call* can differ according to the point of view considered: for call centre companies, agents' behaviour is the most relevant dimension: did the agent followed the company's guidelines, was he efficient, polite, .... For callers, problematic calls are the ones where the request expressed at is not properly resolved by the agent. From a general point of view, problematic calls are those containing conflicts between participants. These dimensions or perspectives can affect the approach that is followed to identify them. Previous studies have focused on supervised approaches for detecting such calls[6]. However, Supervised approaches for call center conversation quality monitoring, suffers of two main drawbacks: the need for expensive human annotation corresponding to the kind of *problems* targeted; the severe unbalanced distribution between *normal* and *problematic* dialogues.

This study addresses these issues by proposing two strategies for detecting several ranges of problematic calls at no extra cost:

1. Targeting agent behaviours and relying on a supervised model trained on *free* annotations collected directly at the call centre level in the Quality Assurance department (the ACOF forms presented in WP1)
2. Using an unsupervised clustering method based on features produced in WP3 and WP4 in order to group together conversations containing conflicts and users frustration.

We use in this study the RATP-DECODA corpus [24] already presented in several other studies in the SENSEI project. We recall in the following subsection the data we use in order to evaluate agent behaviour (the ACOF forms presented in D1.3), then we introduce the features used for the unsupervised clustering process.

**Understanding Agent Behaviour** It is a routine practice in call centers to assess the quality of the calls based on the agent behaviour. They perform this through the means of quality monitoring questionnaires. It contains various functional and operational related questions from the behaviour of the agent. A quality supervisor goes through each of questions mentioned with respect to call from agent's perspective. He/she marks if the agent PASS(ed) or FAIL(ed) with respect to the questions asked. This is a standard operational procedure followed by most of the call centers. The document thus prepared is later used for evaluating the performance of the agent. The information contained in these quality monitoring questionnaire can be used to detect problematic call arising from agent behaviour. Thus, this can be seen as a collection of *freely* available data that could be used to model problematic calls. Table 1 shows a typical Quality Monitoring (QM) parameters which are tracked by the QM supervisors.

Table 1: Quality Monitoring Parameters Evaluated

| ID | Quality Monitoring Parameters |
|---|---|
| 1 | Agent respects opening procedure |
| 2 | Agent listens actively and asks relevant questions |
| 3 | Agent shows the information in a clear, comprehensive and essential way |
| 4 | Agent manages the objections reassuring the customer and always focusing on client satisfaction |
| 5 | Agent manages the call with safety |
| 6 | Agent uses positive words |
| 7 | Agent follows the closing script |
| 8 | Agent is polite and proactive with the customer |
| 9 | Agent is able to adapt to the style of client's communication always maintaining professionalism |
| 10 | Agent Management: he negotiates the wait always giving reasons |
| 11 | Ability to listen |

Typically, in a call center, an agent is evaluated by one QM supervisor. Readers are informed that not all call are monitored due to near impossibility of manually tracking every call. Even though this data is freely available due to the operational aspects of call centers, it is not created with an objective of developing a training set to build a supervised model to identify problematic calls. Hence, further analysis about the usability of this data is required. We set up an annotation task to understand the quality of the agent behaviour assessment of the agent by these QM supervisors.

We selected 3 annotators for this task. The annotators are professional quality monitoring supervisors in one of largest call center company in Europe (*Teleperformance*), partner of the SENSEI project. The audio and respective transcript was provided to the annotators. Based on the specific questions mentioned in table 1, annotators were asked to mark the conversation as *PASS*, *FAIL* and *NA*. The category PASS reflects the fact that the annotator is satisfied with the specific objective mentioned in the QM questionnaire. If they are unsatisfactory, then they are marked as FAIL. If the annotators do not have sufficient information to make decision they are marked as NA. This includes cases in which the service does not provide any actions of up-selling, or if agents do not collect specific information of the customer like name, surname etc, or if the agent's objective (qualitative or quantitative) is ambiguous.

Fleiss Kappa is used to measure the inter-annotation agreement [25]. The kappa agreement along with data statistics of annotations are provided in Table 2. Call samples are assigned labels using a majority based annotator voting.

We observe, on an average, that there is a moderate inter-annotation agreement. In some cases, for instance question 2, 4 and 6, it is poor or no agreement at all. Given this premise, the QM task is highly subjective. The reason for that subjective nature are many. Human bias is an important factor in such annotation tasks. For instance, Question 6 has no agreement at all. This QM tries to find out whether the *Agent uses positive words* during the conversation with the customer. However, since agents do not use very negative (rude words for example) or very positive words, the association of words with sentiment labels is highly subjective. In addition, agents recorded in the Decoda corpus were not instructed to behave so that they obtain a good score with the QM questions. Another factor

Table 2: Inter-annotation agreement using Fleiss Kappa along with annotation category selected based on majority voting

| Quest. ID | Pass | Fail | NA | Kappa | Agreement |
|---|---|---|---|---|---|
| **1** | 346 | 0 | 2 | 1.0 | Perfect |
| **2** | 303 | 15 | 30 | 0.08 | Slight |
| **3** | 334 | 6 | 8 | 0.44 | Moderate |
| **4** | 240 | 25 | 83 | 0.15 | Slight |
| **5** | 329 | 15 | 4 | 0.51 | Moderate |
| **6** | 182 | 76 | 90 | -0.13 | No |
| **7** | 332 | 3 | 13 | 0.54 | Moderate |
| **8** | 341 | 4 | 3 | 0.65 | Moderate |
| **9** | 330 | 14 | 4 | 0.53 | Moderate |
| **10** | 206 | 3 | 139 | 0.90 | Perfect |
| **11** | 326 | 16 | 6 | 0.25 | Fair |

is the different dimension each QM parameter tries to address. Question 3 has a moderate agreement because it is a multi-faceted. It tries to assess agents behaviour on three dimension - clarity of speech, comprehensive and finally on picking necessary information from the customer. From the perspective of QM supervisors', this question is overloaded and assessing it can be very difficult.

From the annotation statistics in Table 2, only question 1 and 10 have high degree of agreement. These two questions could be evaluated automatically. Conversely, critical quality monitoring parameters like Question 2,3,4, 8 and 11 are more difficult to answer for an automatic system. Moreover, number of FAIL samples are too few to create a sound supervised approach for evaluate each QM parameter.

Lastly, it appears also that although annotators would agree on the fact that a conversation is problematic, they might differ on the *FAIL* parameters given to justify their decision. Therefore, as the agreement at the question level is moderate, it would be difficult to design a high confidence automatic QM evaluation system which answers separately to each of them. By taking all the questions as a whole, it is possible to reduce the effort of QM supervisors. A semi-automatic system for QM monitoring based on Table 1 can be developed in order to flag suspicious call conversations based on any of the parameters mentioned in Table 1 and forward problematic conversations to QM supervisors.

**Flagging Problematic calls based on Agent Behaviour** Ideally, a binary classifier needs to be set up for each type of behaviour that is analysed in the questionairre mentioned in table 1. However that is not feasible, as there is high bias in the judgement of the quality monitoring supervisors. This is evident from the inter annotation agreement results presented in the table 2. To mitigate this bias, we propose to flag problematic calls based on the *overall behaviour* of the agent. Here, we are targeting all the parameters mentioned in Table 1 at once.

Such a flagging of the problematic calls would reduce the effort borne by QM supervisors which in turn can lead to more number of calls sampled for quality monitoring. Flagging is posed as a supervised binary classification problem where the thrust is to detect the negative samples. For this, all samples which are considered to be fail for any of the parameters mentioned in Table 1 are labeled as non-compliant samples (**NON-PASS**). The rest of the samples are considered as positive samples (**PASS**). After this abstraction, there were 247 total samples, with 60 non-pass and 187 pass samples.

We categorized features into three types. They are:

- Meta Features (MF): **Conversation Time Features (TI):** The time spoken by each participant during the call can be a measure of the agent behaviour. The conversation time of the agent, customer and the automatic machine is captured using the audio file. The normalized conversation time for each participant is used a features. **Conversation Length Features (T):** Amount of words spoken by the agents/customer can be a signal of service that is being carried out. The conversation length in terms of words spoken by the agent, customer and the automatic machine is captured using audio file. The normalized conversation length for each participant is used a features. **Wait Time Features (WA):** One of the parameters which is often taken in the case of call center conversation is the waiting time. The waiting time refers to time taken by agent to respond to a query. It can also refer to the agent as well as the customer. Average normalized wait period is extracted for the agent as well as customer. **Turn Features (TU):** Total number of turns, along with turns of customer and agent is captured and used a features.
- Speech Features (SF): Different set of experiments were performed to identify a set of features to be used for this study. The Interspeech Emotion challenge feature set of 2009 was used as base features [26]. However, it was observed not all features were necessary for classification. After pruning, final set of features include fundamental frequency, voicing probability and the loudness contour.
- Text Features (TF): **Sentiment Features (E):** We developed a sentiment lexicon to extract sentiment related features. To do so, canonical words are selected from Tweets. Thereafter their Pointwise Mutual Information (PMI) is calculated with respect to the sentiment label of the tweet. The sentiment label of the tweets is obtained from the hashtags present in the tweets. Tweets with hashtags #positive/#negative are chosen to select the words. Real valued sentiment scores are obtained for the words extracted in this manner. Based on threshold these words are marked as positive or negative. Thereafter they are used for feature engineering. There are 27,820 entries in the lexicon. **Agent Utterance Features (UA):** The utterances of the agent from the call transcript are stemmed and terms with frequency more than 1 and less than 5,000 are used as features. QM supervisors most often focus on the vocabulary of the speakers to check the compliance of the call with the standard.

**Detecting conflict from the customer point of view** Conflict of the agent with customer is an instance of problematic call. One reason for such a conflict is the frustration that either of the parties have. For the smooth functioning of call centers, it is imperative that such conflict ensuing instances are reduced.

A conflict episode is defined as a three stages phenomenon: A's statement; B's counterstatement; A's counterstatement to B [27, 28, 29, 30].

A conflict can be detected through interactional and discursive cues. For example, since Sacks et al.[31], it is well known that turns-at-talk are organized as a real system in which pauses and overlaps are minimized. So, overlaps, interruptions, failed attempt to take a speech turn can be relevant parameters to characterize a potential conflict episode. In the same way, in this kind of interaction (call center), politeness constraints, at least from the agent, are stronger than in casual ones. The presence of ritualized sequences like opening and closing are very expected and their absence, partial or full, could indicate a conflict. At the discursive level, negative comments, interjections, insults, negative

categorization (like “mess” instead of “situation”), the presence of explicit opposite or negative markers such as “but” or “no” may reveal a conflict.

Here it is worth noting that all these quoted cues might be poly functional and require taking into account not only a larger discursive context but also the situational one to discriminate the current function.

Due to the lack of annotated data, and considering the difficulty of collecting such data in a running call center, we decided to develop an unsupervised clustering method in order to select calls likely to contain a conflict between the operator and the caller. We use Gaussian mixture models (GMM) for clustering the problematic calls. Multiple features based on agent-customer interactions and inspired by the descriptions of Section 3.1 are designed to capture frustration and conflicts. These are described below:

- Overlaps

  The proportion of overlaps between turns that occur in a dialogue is used. Only the number of overlaps is taken into account because the duration of each overlap is not reliable enough in our corpus.

- Word polarity

  The polarity of each word is retrieved from a word polarity dictionary. Each word’s Pointwise Mutual Information is calculated by looking at the type of smileys found in tweets containing our word. Two features are created using this polarity. Both of them calculate the mean polarity of the words in a dialogue, but the first feature uses words said by both speakers whereas the second feature only uses words said by the caller.

- Dialogue Acts

  The dialogue acts are predicted using an automatic dialogue act classifier on our dialogues. We use 4 dialogue acts for our features which are: Opening, Closing, Declaration and Interruption. The opening (resp. closing) dialogue acts describes introduction turns (resp. conclusion turns). The declaration dialogue act describes neutral turns where the speaker adds a piece of information that doesn’t necessarily need an interaction with the other speaker. The interruption dialogue acts describe return of comprehension turns in which the speaker simply signals to the other speaker that he understood what he was saying. This dialogue act also describes turns in which a speaker (usually the agent) asks the other speaker to wait. The first features take into account the proportion of openings (resp. closings) in a dialogue. Another feature takes a look at the proportion of declarations said by the caller. Finally, the last two features look at the proportion of interruptions said by both speakers and by the caller.

**Dialog flagging classifier:** Stochastic Gradient based linear predictors are used for classification [32]. Since the dataset is comparatively small, a stratified 5 fold sampling is performed. For speech related features, opensmile is used [33]. For extracting these features FrameSize=0.025 and FrameStep=0.010 are used. Speech features thus extracted were aggregated over the individual turn time of the speakers and used as features. Standard classification metrics of recall, precision and fscore are used for evaluation.

Table 3: Results of Flagging based QM system on various feature sets for manually annotated dataset (top) and on ASR output(below). Best combination consists of features that includes agent utterances, turns and wait time

| Feature Set | PASS | | | Non-PASS | | | Overall | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | Fscore | Precision | Recall | Fscore | Precision | Recall | Fscore |
| **SF** | 0.76 | 0.60 | 0.67 | 0.25 | 0.42 | 0.31 | 0.64 | 0.56 | 0.59 |
| **TF** | 0.75 | 0.82 | 0.78 | 0.23 | 0.17 | 0.19 | 0.63 | 0.66 | 0.64 |
| **MF** | 0.73 | 0.67 | 0.70 | 0.19 | 0.23 | 0.21 | 0.60 | 0.57 | 0.58 |
| **SF+MF** | 0.76 | 1.00 | 0.86 | 0.00 | 0.00 | 0.00 | 0.57 | 0.76 | 0.65 |
| **SF+TF** | 0.77 | 0.61 | 0.68 | 0.27 | 0.43 | 0.33 | 0.65 | 0.57 | 0.60 |
| **MF+TF** | 0.75 | 0.80 | 0.78 | 0.21 | 0.17 | 0.19 | 0.62 | 0.65 | 0.63 |
| **SF+MF+TF** | 0.75 | 0.84 | 0.79 | 0.23 | 0.15 | 0.18 | 0.63 | 0.67 | 0.65 |
| **Best Combination** | 0.79 | 0.85 | 0.82 | 0.39 | 0.30 | 0.34 | 0.69 | 0.72 | 0.70 |
| **On ASR output** | | | | | | | | | |
| **SF** | 0.76 | 0.60 | 0.67 | 0.25 | 0.42 | 0.31 | 0.64 | 0.56 | 0.59 |
| **TF** | 0.76 | 0.87 | 0.81 | 0.28 | 0.15 | 0.20 | 0.64 | 0.69 | 0.66 |
| **MF** | 0.75 | 0.80 | 0.78 | 0.25 | 0.20 | 0.22 | 0.63 | 0.65 | 0.64 |
| **SF+TF** | 0.76 | 0.80 | 0.78 | 0.27 | 0.22 | 0.24 | 0.64 | 0.66 | 0.65 |
| **SF+MF** | 0.75 | 1.00 | 0.86 | 0.00 | 0.00 | 0.00 | 0.57 | 0.75 | 0.65 |
| **MF+TF** | 0.76 | 0.89 | 0.82 | 0.29 | 0.14 | 0.18 | 0.64 | 0.70 | 0.66 |
| **SF+MF+TF** | 0.76 | 0.89 | 0.82 | 0.29 | 0.14 | 0.18 | 0.64 | 0.70 | 0.66 |
| **Best Combination** | 0.79 | 0.84 | 0.81 | 0.39 | 0.31 | 0.34 | 0.69 | 0.71 | 0.70 |

**Unsupervised dialog clustering:** in order to evaluate the different clusterings, new annotation has been done on the decoda corpus. Annotators had to say for each speaker if he's calm (cold), angry (hot) or between the two (medium) at the start, middle and end of the dialogue. Using this "heat" annotation, 158 dialogues have been annotated. For our evaluation, dialogues are splitted into two classes: Hot and Cold. If in the heat annotation a dialogue has at least one Hot or Medium label, then we consider the dialogue to be Hot. Otherwise, the dialogue is Cold. This gives us 50 Hot dialogues and 108 Cold dialogues.

Scikit-learn's [32] implementation of GMM which uses the expectation-maximization algorithm to estimate the parameters is used for clustering experiments. Through empirical testing, we found the best results were obtained by using diagonal covariance matrices and 5 components. The results reported are based on that setting. The metrics used to evaluate the clusterings are the purity of the clusters, and the precision, recall and f-score for the Hot class. The precision and recall are calculated on the cluster that has the most Hot dialogues in it.

**Automatic Speech transcriptions:** in order to have realistic assessment, performance comparisons are done on automatic transcriptions (ASR) as well. The ASR transcriptions used in this study are described in [34]. They are obtained thanks to the LIUM system based on the Kaldi decoder [35] with DNN acoustic models as well as LIUM rescoring tools [36]. The average WER is 34.5%. This high error rate is mainly due to speech disfluencies and noisy acoustic environments.

**Problematic calls based on Agent Behaviour** Table 3 shows the result of the flagging system for different feature sets. The result suggests that features can identify the NON-PASS samples from PASS samples. Importance is given to detect the NON-PASS class over the PASS class. Conventional approaches for classification would fail as the system developed would not be optimized to detect the

Table 4: Results of the clusterings depending on the used feature set

| Feature set | Purity | Precision | Recall | F-score |
|---|---|---|---|---|
| Polarity | 0.728 | 0.561 | 0.637 | 0.597 |
| Overlaps | 0.684 | 0.374 | 0.338 | 0.355 |
| Acts | 0.684 | 0.480 | 0.486 | 0.483 |
| Acts+Overlaps | 0.747 | 0.582 | 0.718 | 0.643 |
| Polarity+Overlaps | 0.746 | 0.595 | 0.618 | 0.606 |
| Polarity+Acts | 0.766 | 0.600 | 0.780 | 0.678 |
| Polarity+Overlaps+Acts | 0.753 | 0.587 | 0.740 | 0.655 |
| **On ASR output** | | | | |
| Polarity | 0.671 | 0.482 | 0.717 | 0.577 |
| Polarity+Overlaps | 0.671 | 0.466 | 0.737 | 0.570 |

under sampled category. To mitigate the effects of class skewness, the optimization performed on the learner is recall based. This ensures that classifiers learn the under-sampled class.

Given that annotation agreement is quite poor, the best f-score of 34% on the NON-PASS class is acceptable. This is obtained for feature set comprising of agent utterances, turns and wait time. The system is able to detect PASS class with acceptable precision and recall. As a result more calls can be sampled in the allotted time. This can lead to an overall improved operational efficiency.

**Problematic calls based on Conflicts and User frustration** Table 4 shows the result of the clusterings depending on the features that were used. We could not evaluate the dialogue acts features on the ASR output because they weren't available.

The results suggest that the word polarity features are the main contributors to the task and when used alone puts into the same cluster 64% of the "Hot" dialogues. But overlaps and dialogue acts are still beneficial since we obtain better scores when used together. When used with the polarity features, dialogue acts and overlaps allow us to even further improve the scores even if overlaps seem to be slightly less beneficial.

The best results are obtained with the word polarity features combined with the dialogues acts features with a purity of 77% and a F-score of 68%.

**Are the problems related?** In order to see if the Agent Behaviour approach and the Conflicts and User frustration approach respond to a similar problem, evaluation of the obtained clusters was also done using the FAIL/PASS annotation. In figure 1, we can see the distribution of the FAIL and PASS dialogues in our clusters generated using the Polarity+Acts feature set. For comparison, the distribution of the Hot and Cold dialogues is shown in figure 2. The instances of failed classes are distributed across different clusters, suggesting different dimensions of problematic calls.

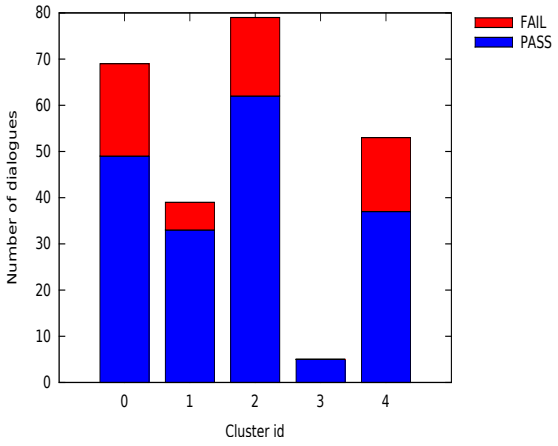


Figure 1: Distribution of the dialogues in the clusters using labels obtained from Section 3.1

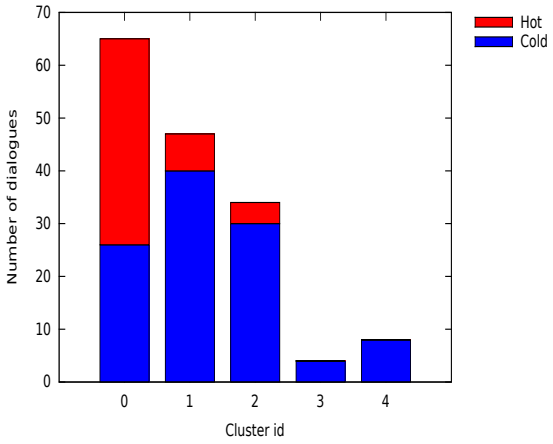


Figure 2: Distribution of the dialogues in the clusters using heat annotation

## 3.2 Empathy and Emotion Detection in Speech

In the context of automatic behavioral analysis, we aim to classify empathy and other basic and complex emotions in human-human spoken conversations. Empathy underlies to the human ability to recognize, understand and to react to emotions, attitudes and beliefs of others. While empathy and its different manifestations (e.g., sympathy, compassion) have been widely studied in psychology, very little has been done in the computational research literature. In this work, we investigated the occurrences of empathy on the agent's channel and other basic and complex emotions, such as anger, frustration, satisfaction and dissatisfaction, on the customer's channel, which were collected from call-centers. We have designed binary classification systems to detect the presence of each emotional state. The automatic classification system has been evaluated using call centers' spoken conversations by exploiting and comparing performances of the lexical and acoustic features.

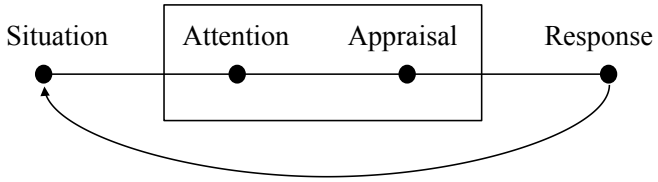


Figure 3: The modal model of emotion [38].

**Corpus** The corpus includes $1,894$ randomly selected customer-agent conversations, which were collected over the course of six-months, amounting to $210$ hours of audio. These conversations were recorded on two separate channels of $16$ bits per sample and $8kHz$ sampling rate. The average length of the conversations was $406$ seconds.

**Annotation Scheme for Spoken Conversations** For the annotation, we followed the psychological definition of Hoffman [37], which states empathy as "*an emotional state triggered by another's emotional state or situation, in which one feels what the other feels or would normally be expected to feel in his situation*". In order to design the operational model of empathy annotation, we adopted the *modal* model of emotion by Gross [38], where the appraisal processes of the unfolding of emotional states are modeled sequentially.

In the psychological literature, it has been shown that temporal unfolding of emotional states can be conceptualized and experimentally tested [39]. Gross has provided evidence that concepts such as *emergence* — derivation from the expectations of relationships — and *unfolding* — sequences that persist over time — may help in explaining emotional states. The modal model of emotions [38, 40] emphasizes the attentional and appraisal acts underlying the emotion-arousing process. In Figure 3, we provide the original schema of Gross model. The individuals' core *Attention-Appraisal* processes (included in the box) are affected by the *Situation* that is defined objectively in terms of physical or virtual spaces and objects. The *Situation* compels the *Attention* of the individual; it triggers an *Appraisal* process and gives rise to coordinated and malleable *Responses*. It is important to note that this model is dynamic and the situation may be modified (directed arc from the *Response* to the *Situation*) by the actual value of the *Response* generated by the *Attention-Appraisal* process. In this model, emotional states are seen as a way of experiencing the world: they are distinct functional states [41], and the appraisal acts describe the content of those functional states within a context.

Therefore, we believe that Gross' model provides a useful framework for describing the dynamics of emotional states within an *affective scene*[1] [42], because not only it does focus on appraisal, but also considers how responses are feeding back to the initial communicative situation.

In order to make it applicable in a real-life domain like the call center conversations, we operationally defined empathy as *"a situation where an agent anticipates or views solutions and clarifications, based*

[1] *"The **affective scene** is an emotional episode where one individual is affected by an emotion-arousing process that (a) generates an emotional state variation, and (b) triggers a behavioral and linguistic response. The affective scene extends from the event triggering the unfolding of emotions on both individuals, throughout the closure event when individuals disengage themselves from their communicative context."* It is defined based on the emotion sequence between interlocuators. For example, the sequence of emotional states between an agent and a customer could be Frustration (C) → Empathy(A) → Satisfaction(C). A - Agent, C-Customer.

*on the understanding of a customer's problem or issue, that may help in relieving or preventing the customer's unpleasant feelings"*. For designing the annotation scheme, we have performed an extensive analysis of one hundred conversations (more than $11$ hours), and selected dialog turns where the speech signal showed the emergence of both basic emotions, such as anger, and complex emotional states such as frustration, and empathy.

Our qualitative analysis supported the hypothesis that the relevant speech segments were often characterized by perceivable variations in the speech signal.

As expected, such variations sometimes co-occurred with emotionally connoted words, but also with functional parts of speech, such as adverbs and interjections, which could play the role of lexical supports for the variations in emotional states. On the basis of the above observations, we have designed an annotation scheme for empathy and other emotional states by taking into account the perception of the variations in the speech signal as well as variations in the linguistic content of the utterances [43].

The annotation scheme includes the following recommendations for the annotators:

1. Annotating the onset of the signal variations that supports the perception of the manifestation of emotions;
2. Identifying the speech segments preceding and following the onset position.
3. Annotating the context (left of the onset) and target (right of the onset) segments with a label of an emotional state (e.g., frustration, empathy etc.).

The context of the onset is defined to be neutral with respect to the target emotional state label. We have introduced *neutrality* as a relative concept to support annotators in their perception process of an emotional state while identifying the support of the situational context.

In the annotation process, given the limited resources, our goal was to maximize the number of annotated conversations. For this reason, we annotated only the first occurrence of a segment pair (e.g., neutral-empathy) within each conversation. Once candidate segment pair was selected, the annotators could listen the speech segments as many times as needed to judge if the selected segment pair could be labeled. After that, the annotator tagged the right of the onset of the segment pair with an emotional label, and left of the onset was labeled as neutral. During the annotation process, annotator also needed to focus on the boundaries of the speech segment.

**Evaluation of the Annotation** For our experiment, the annotation task was performed by two expert annotators who worked on non-transcribed spoken conversations by following the annotation scheme reported above. The annotators used the EXMARaLDA Partitur Editor [44] to perform their task. They annotated *Empathy* on the agent channel, and *Anger, Frustration, Satisfaction, Dissatisfaction* and on the customer channel. The annotators labeled *Neutral* on segment that appeared before any emotional segment to define the context, as mentioned earlier. The inter-annotator agreement of the annotation task is $0.74$.

**Experimental Methodology** The importance of the automatic classification of empathy and other emotional states has been highlighted in [45, 46] where behavioral analysis experiment has been conducted by human experts in workplaces such as the call centers to evaluate the interlocutors' affective

behavior during the phone conversations. For designing the automatic systems, we first prepared the dataset by applying some sampling techniques, then experimented with different set of features.

**Data Preparation** For the classification task, we designed binary classifiers for each emotion category by considering whether particular emotion category exists in a conversant channel or not. For each binary classifier, we prepared dataset using the following approach. If a conversation contains at least one emotional event we labeled that conversation as a positive example, otherwise we labeled it as negative, as also shown in Figure 4. Therefore, for empathy, positive are those that contain empathic emotional marker and negative are those that are neutral. For frustration, satisfaction, dissatisfaction and anger, positive examples are those that contain respective emotional marker and negative are those that include neutral and other emotional markers.

In Table 5, we present the original distribution of the dataset, which we prepared for the classification experiments. However, such skewed distributions result in lower classification performance. Therefore, we down-sampled examples of majority classes by randomly removing them to make a balanced class distribution for each category as shown in Table 6,.

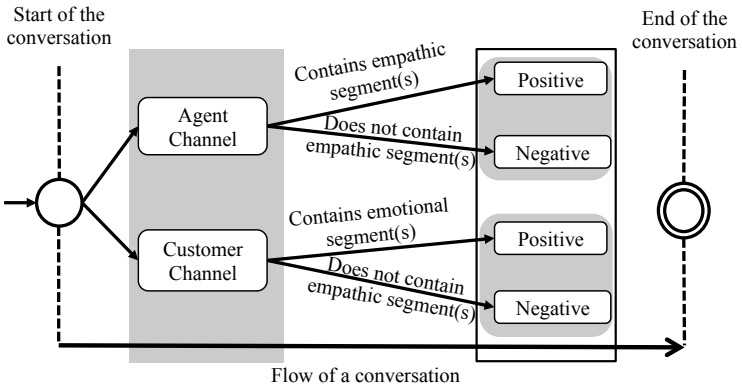


Figure 4: Data preparation for the classification experiments.

Table 5: Original distribution of the class labels.

| Class | Y | N | Total | Y (%) | N (%) |
|---|---|---|---|---|---|
| **Emp** | 525 | 1250 | 1775 | 0.30 | 0.70 |
| **Ang** | 118 | 1776 | 1894 | 0.06 | 0.94 |
| **Fru** | 338 | 1556 | 1894 | 0.18 | 0.82 |
| **Dis** | 382 | 1512 | 1894 | 0.20 | 0.80 |
| **Sat** | 735 | 1159 | 1894 | 0.39 | 0.61 |

Table 6: Class distribution after down-sampling of the majority class.

| Class | Y | N | Total | Y (%) | N (%) |
|---|---|---|---|---|---|
| Emp | 530 | 636 | 1166 | 0.45 | 0.55 |
| Ang | 118 | 141 | 259 | 0.46 | 0.54 |
| Dis | 367 | 403 | 770 | 0.48 | 0.52 |
| Fru | 338 | 405 | 743 | 0.45 | 0.55 |
| Sat | 736 | 883 | 1619 | 0.45 | 0.55 |

**Classification System** In Figure 5, we present the architecture of the automatic classification system, which takes a spoken conversation as input and generates a binary decision regarding the presence *or* absence of an emotional state. The recognition system evaluates the cues present throughout the spoken conversation and then commits to a binary decision. In order to evaluate the relative impact of lexical features, we used transcriptions obtained from an Automatic Speech Recogniser (ASR). We extracted acoustic features directly from the speech signal and designed the classification systems. For both acoustic and lexical features, we applied feature selection algorithms. We also investigated feature combination to investigate the performance of different configurations of the system.

The ASR system that we used to transcribe the conversations was designed using a portion of the data containing approximately 100 hours of conversations. The system has been designed using Mel-frequency cepstral coefficients (MFCCs) based features with a splice of three frames on each side of the current frame. Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) feature-space transformations were used to reduce the feature space. We trained the acoustic model using speaker adaptive training (SAT) and also used Maximum Mutual Information (MMI). Word Error Rate (WER) of the system is $31.78\%$ on the test set split [47].

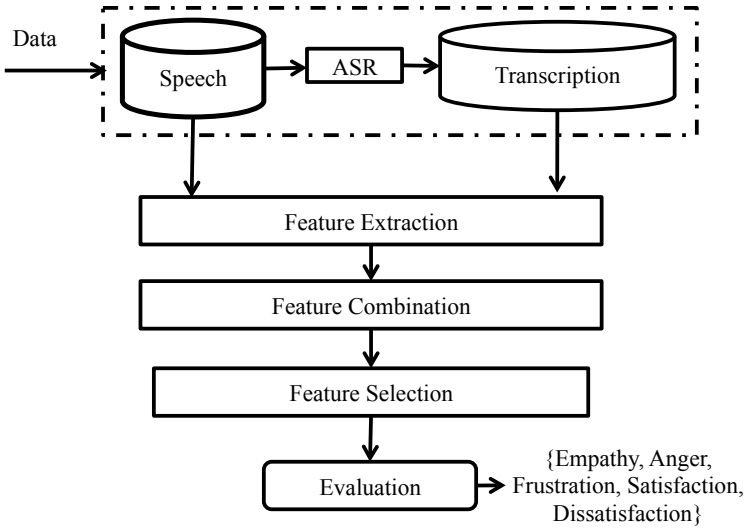


Figure 5: System for classification.

**Feature Extraction** We used features of two types, Acoustic and Lexical. The use of large-scale acoustic features was inspired by previous studies in emotion and personality recognition tasks, in which low-level features were extracted and then projected onto statistical functionals [48, 49]. For this study, we extracted features using openSMILE [33]. Before extracting features, we automatically pre-processed speech signals of the conversations to remove silence at the beginning and end of the recordings. We also removed silences longer than 1 second. The low-level acoustic features were extracted with approximately $100$ frames per second, with $25 - 60$ milliseconds per frame. These low-level descriptors (LLDs) were then projected onto single scalar values by descriptive statistical functionals. The details of the low-level features and statistical functionals are given in Table 7.

Table 7: Low-level acoustic features and statistical functionals

| **Low-level acoustic features** |
|---|
| **Raw-Signal:** Zero crossing rate |
| **Energy:** Root-mean-square signal frame energy |
| **Pitch:** F0 final, Voicing final unclipped, F0 final - nonzero |
| **Voice quality:** jitter-local, jitter-DDP, shimmer-local, log harmonics-to-noise ratio (HNR) |
| **Spectral:** Energy in bands 250-650Hz, 1-4kHz, roll-off-points (0.25, 0.50, 0.75, 0.90), flux, centroid, entropy, variance, skewness, kurtosis, slope band (0-500, 500-1500), harmonicity, psychoacoustic spectral sharpness, alpha-ratio, hammarberg-index |
| **Auditory-spectrum:** band 1-10, auditory spectra and rasta |
| **Cepstral:** Mel-frequency cepstral coefficitnts (mfcc 0-3) |
| **Formant** First 3 formants and first formant bandwidth |
| **Statistical functionals** |
| Relative position of max, min |
| Quartile (1-3) and inter-quartile (1-2, 2-3, 3-1) ranges |
| Percentile 1%, 99% |
| Std. deviation, skewness, kurtosis, centroid, range |
| Mean, max, min and Std. deviation of segment length |
| Uplevel time 25 and rise time |
| Linear predictive coding lpc-gain, lpc0-1 |
| Arithmatic mean, flatness, quadratic mean |
| Mean dist. between peaks, peak dist. Std. deviation, absolute and relative range, mean and min of peaks, arithmatic mean of peaks, mean and Std. of rising and falling slope |

We extracted lexical features from both manual and automatic transcriptions. To utilize the contextual benefits, we extracted trigram features, which eventually results in a very large dictionary. Therefore, we filtered out lower frequency features by preserving 10K most frequent n-grams. We then transformed lexical features into bag-of-ngrams (vector space model) with logarithmic term frequency (tf) multiplied with inverse document frequency (idf) – tf-idf, as presented in the equation 1. Here, we considered the conversation as a document.

$$tf \times idf = log(1 + f_{ij}) \times log\left(\frac{\text{number of conversations}}{\text{number of conversations that include word i}}\right) \quad (1)$$

where $f_{ij}$ is the frequency for word $i$ in conversation $j$. It assigns a weight to each term of the conversation. Its value is highest when the word, $i$, appears many time in a few conversation, which leads to higher discriminating power for the classification. It is lower when the word appears fewer times in a conversation and appears in many conversations, which represents less discriminating power.

**Feature Selection and Combination** We extracted a large number of features for both acoustic and lexical sets. In order to reduce the computational cost and avoid overfitting we have chosen Relief [50] as a feature selection technique. In a previous study [51], we comparatively evaluated this technique against other algorithms such as information gain, and it performed best in terms of classification performance and computational cost. In order to select the best set of features, we ranked the features according to the Relief's score and generated feature learning curve by incrementally adding batches of ranked features. Before applying feature selection, we discretized the feature values into $10$ equal frequency bins, where each bin contains an approximately equal number of values. For the feature fusion, we merged acoustic and lexical features into a single vector to represent each instance.

**Classification and Evaluation** We designed the binary classification models using Support Vector Machines (SVMs) [52]. Linear kernel of the SVMs was chosen in order to alleviate the problem of higher dimensions of lexical and combination of $acoustic + lexical$ features. We optimized the penalty parameter $C$ of the error term by tuning it in the range $C \in [10^{-5}, ..., 10]$ and the gaussian kernel parameter $G$ in the same range as well, using cross-validation.

At the feature fusion level, we applied feature selection on the combined acoustic and lexical features. We then applied the feature selection process to the merged feature vector to obtain an optimal subset of features.

We measured the performance of the system using the Un-weighted Average (UA), which has been widely used in the evaluation of paralinguistic tasks [26]. UA is the average recall of positive and negative classes and is computed as $UA = \frac{1}{2}\left(\frac{tp}{tp+fn} + \frac{tn}{tn+fp}\right)$, where $tp$, $tn$, $fp$, $fn$ are the number of true positives, true negatives, false positives and false negatives, respectively. Due to the limited size of the conversational dataset we used $10$ folds cross-validation method.

**Results and Discussion** In Table 8, we report the performances of the classification system for a single feature type and feature combination. We report them in terms of average UA of the cross-validation and its standard deviation. We computed the baseline by randomly selecting the class labels, such as empathy i.e., positive, and non-empathy i.e., negative, based on the prior class distribution of the training set.

In the classification experiments, we obtained better results using lexical features compared to the acoustic features. The linear combination of acoustic and lexical features did not perform well due to the complexity of the large feature space. The other reason could be that the feature representation of these two sets is different, i.e., dense *vs* sparse, which may increase the complexity of the task.

For each emotional category, performances are statistically significant with $p < 0.05$ compared to the baseline. The significant test has been computed using two-tailed paired sampled t-test. The reason

to compute *oracle* performance is to understand the upper-bound for each classification model, which shows that a relative improvement, ranges from $\sim 15\%$ to $\sim 21\%$, can be achieved for each case.

The results of anger vary a lot in each cross-validation fold, which we see from a high standard deviation. The reason is that we have a very small number of instances for this emotional class. For dissatisfaction performances are comparatively lower than other categories.

Table 8: 10 folds cross-validation results using different feature set. Ac: acoustic features, Lex (A): lexical features from ASR transcription.

| **Experiments** | **UA Avg (Std)** | | | | | |
|---|---|---|---|---|---|---|
| | **Emp** | **Ang** | **Fru** | **Sat** | **Dis** | **Avg** |
| **Random baseline** | 49.0(4.5) | 48.8(10.3) | 50.2(5.7) | 50.4(4.3) | 49.7(5.0) | 49.6(5.9) |
| **Ac** | 58.8(4.8) | 66.3(7.4) | 61.3(4.9) | 53.3(2.7) | 55.2(5.7) | 59.0(5.1) |
| **Lex (A)** | 61.2(5.3) | 76.3(6.5) | 65.9(7.4) | 62.3(3.1) | 60.8(3.1) | 65.3(5.1) |
| **Ac+Lex (A)** | 59.0(3.5) | 67.6(11.8) | 63.0(4.8) | 54.1(3.9) | 57.6(6.3) | 60.2(6.0) |
| **Oracle** | 74.2(3.8) | 84.7(3.4) | 79.9(6.6) | 80.9(2.1) | 71.0(2.6) | 78.1(3.7) |

## 3.3 Competitiveness in Overlapping Speech

Overlapping speech is a common and relevant phenomenon in human conversations, reflecting many aspects of discourse dynamics. Understanding the dynamics of overlapping speech is crucial for conversational analysis and for modeling agent-client behavior. In this research work, we focus on the pragmatic role of overlaps in turn-in-progress, where it can be categorized as competitive or non-competitive [53]. Previous studies on these two categories have mostly relied on controlled scenarios, and handcrafted feature. On the contrary, in our study, we focus on call center data, with customers and operators engaged in problem-solving tasks. We also analyzed characteristics of overlaps in large feature space using unsupervised techniques [54]. We designed a speech overlap annotation scheme in order to annotate the competitiveness, in overlapping speech [53]. Our goal is to detect and classify the non-competitive or competitive overlaps in the speech by using different feature groups and contextual information. While doing so, we analyzed different acoustic feature groups, linguistic feature groups, their combination and an optimal subset by using feature selection [53]. The details of the study can be found in SENSEI deliverable D4.2 [55].

# 4 Para-semantic Parsing of Social Media Conversations

## 4.1 Stance Detection

Social platforms, such as Twitter, Facebook, Instagram etc., have attracted hundreds of millions of people to share and express their opinions, feelings, emotions... These social platforms provide a wonderful sandbox opportunity for mining opinion in natural language text. In recent years, the microblogging service Twitter has emerged as one of the most popular and useful sources of user content, and recent research has begun to develop tools and computational models for tweet-level opinion and sentiment analysis. Recent research have been done in order to extend the detection of a generic sentiment, by detecting stance in tweets.

Stance detection is the task of automatically determining from text whether the author of the text is in favor of, against, or neutral towards a proposition or target. The target may be a person, an organization, a government policy, a movement, a product, etc. This task is distinct from sentiment analysis in that an in favor or against stance can be measured independently of an author's emotional state.

For example, this tweet concerning the Brexit : "*The next James Bond will just be him spending 2 hours in passport control at De Gaulle*". We can infer that the user is in favor to remain Britain in Europe. Similarly, people often express stance towards various target entities through posts on online forums, blogs, Twitter, Youtube, Instagram, etc. Automatically detecting stance has widespread applications in information retrieval, text summarization, and textual entailment.

Stance detection task is often modeled as a classification problem which relies on features extracted from the text in order to feed a classifier. Classical approach is to use neural network like Convolutional Neural Network (CNN) and/or Recurrent Neural Network (RNN).

In machine learning and more precisely in neural network it's common to begin learning of any new task from scratch (ie. by randomly initializing the parameters of a neural network). This disregards any knowledge gained by similar algorithms when solving previous tasks. Pre-trained hidden layer can on the other hand, store the knowledge gained in one context and apply it to different, related problems. This type of approach is particularly appealing when one lacks sufficient quantity of in-domain labeled training data, such as when there are only a few hundred known examples of a target.

One strategy for performing pre-trained is to train the parameters of a neural network on multiple tasks: first on an auxiliary task with plentiful data that allows the network to identify meaningful features present in the corpus, then a second time using actual task data to tune and exploit those features learned in the first pass.

In this study we propose for stance detection task to pre-train the hidden layer of neural network with several sub-tasks such as hashtag prediction or sentiment analysis.

Our experiments on the SemEval 2016 corpus show that pre-train hidden layer obtains 65.93% f-measure, whereas randomly initialize hidden layer obtains 59.60% f-measure (an absolute gain of 6 points).

The same method was applied for the SENSEI WP6 demo on the Brexit usecase.

**Task and Evaluation** In this paper, we use the Stance dataset proposed during the SemEval 2016 [56]. The Stance Dataset was partitioned in two tasks: Task A (supervised framework) and Task B (weakly supervised framework). In Task A, a training data is available and can be used in a standard supervised stance detection task. In Task B, no training data labeled with stance towards was available. However, some very minimal labeling is permitted. For example, labeling a handful of hashtags.

Possible stances were : FAVOR, AGAINST and NONE where the latter served both as an indicator of a tweet not being related to the target, as well as the tweet being neutral towards the target.

In **Task A**, the dataset consisted of 2814 tweets in the train set and 1249 in the test set divided into five targets : Atheism (Atheism), Climate Change is a Real Concern (Climate), Feminist Movement (Feminist), Hillary Clinton (Hillary), and Legalization of Abortion (Abortion).

Class balance varied between topics, with some topics showing significant skew (e.g. Climate Change is a Real Concern with 4% AGAINST and 54% FAVOR) while others were more balanced (e.g. Feminist Movement with 49% AGAINST and 32% FAVOR). Approximately 74% of the provided tweets were judged to be either in favor or against,

In **Task B**, the dataset contained only test data, no training. We worked on the *Donald Trump* target: tweets in favor or against Trump as a candidate for the 2016 US presidential election. Table 9, shows the number and distribution of instances in the Stance Dataset.

Table 9: Stance distribution according to targets and corpora

| | Overview of instances in Train | | | | Overview of instances in Test | | | |
|---|---|---|---|---|---|---|---|---|
| **Targets** | **#Train** | **%Favor** | **%Against** | **%None** | **#Test** | **%Favor** | **%Against** | **%None** |
| Atheism | 513 | 17.9 | 59.3 | 22.8 | 220 | 14.5 | 72.7 | 12.7 |
| Climate | 395 | 53.7 | 3.8 | 42.5 | 169 | 72.8 | 6.5 | 20.7 |
| Feminist | 664 | 31.6 | 49.4 | 19.0 | 285 | 20.4 | 64.2 | 15.4 |
| Clinton | 689 | 17.1 | 57.0 | 25.8 | 295 | 15.3 | 58.3 | 26.4 |
| Abortion | 653 | 18.5 | 54.4 | 27.1 | 280 | 16.4 | 67.5 | 16.1 |
| Trump | - | - | - | - | 707 | 20.93 | 42.29 | 36.78 |

The official evaluation measure was the macro-average of F-score for FAVOR and AGAINST across all targets, meaning that weak F-score performance on an unbalanced label distribution for a target could be compensated for by the overall good performance on other targets. Note that the label NONE was ignored during the evaluation. Consequently, misclassifying FAVOR or AGAINST as NONE (or vice versa) was penalized less than misclassifying FAVOR as AGAINST (or vice versa).

**A CNN-based method** Deep learning models have been shown to produce state-of-the-art performance in various domains (vision, speech, etc...). Convolutional Neural Networks (CNN) represent one of the most used deep learning models in computer vision (LeCun and Bengio, 1995). Recent work has shown that CNNs are also well suited for sentence classification problems and can produce state-of-

the-art results (Tang et al., 2014a; Severyn and Moschitti, 2015). The difference between CNNs applied to computer vision and their equivalent in NLP lies in the input dimensionality and format. In computer vision, inputs are usually single-channel (eg. grayscale) or multi-channel (eg. RGB) 2D or 3D matrices, usually of constant dimension. In sentence classification, each input consists of a sequence of words of variable length. Each word $w$ is represented with a n-dimensional vector (word embedding) of constant size. All the word representations are then concatenated in their respective order and padded with zero-vectors to a fixed length (maximum possible length of the sentence).

The word embeddings we use are based on an approach for distributional semantics which represents words as vectors of real numbers. Such representation has useful clustering properties, since it groups together words that are semantically and syntactically similar (Mikolov et al., 2013). For example, the word coffee and tea will be very close in the created space. The goal is to use these features as input to a CNN classifier.

**Pre-trained Layer** In the baseline system our CNN Stance classifier system is directly trained on the training corpus for each target presented in table 9. Our contribution is to perform pre-training of the CNN layers on related subtasks for which larger quantities of training corpus were available with respect to the stance classification task.

We considered here 3 sub-tasks for pretraining the network, described as follows:

- Sentiment analysis: Sentiment analysis is closely related to stance detection. The ideological stance of a person normally influences his/her sentiment towards different ideological topics. Accordingly we decide to identify the overall sentiment of each sentence in the given tweet and use this sentiment as a possible indicator of a tweeter's stance. Our assumption is that since tweets are inherently short in length, we can assume that the expressed sentiment targets the ideological topic the tweet is discussing. We used the SEMEVAL tweet opinion corpus to pre-train our CNN to predict sentiment labels.
- Hashtag prediction: hashtags can be seen as *concept tags* added to disambiguate a short message. The sub-task considered here consists in training the CNN to predict hashtags for a given text string We first collect tweets related to the target, list all the hashtags occurring in these tweets and keep the $N$ most frequent ones Then we train the CNN on predicting these $N$ labels (and the label `none`) The network obtained after training is used as pre-training for the stance classification task.
- Hashtag-stance prediction: For some targets the hashtags are highly indicative of the stance. For example, in the tweet `Rethink your beach clothes. Bc it may oppress some people!! #thisoppresseswomen` the hashtag is fully indicative of the stance. Similarly, for the *Brexit* target, a hashtags such as: `#strongerin` is not ambiguous relative to the global meaning of the messages containing it. This pretraining is quite similar to the hashtag prediction one, except that we list a small set of non-ambiguous hashtags and group them according to their opinion (favor or against). We perform here a semi-supervised training for the stance classification task.

**Results on Semeval 2016 - TASK A** We tested the system on the data provided by the Semeval 2016 shared task 6: detecting stances in Tweets. The results for task A are reported in Table 10.

Table 10: Results on semeval 2016 Task A and B (Semeval score)

| | Task A | | | | | | Task B |
|---|---|---|---|---|---|---|---|
| Methods | Overall | Atheism | Climate | Feminism | Hillary | Abortion | **Trump** |
| Baseline | 59.60 | 63.04 | 41.27 | 44.94 | 54.69 | 53.68 | **27.64** |
| Sentiment analysis | 59.86 | 60.76 | 36.91 | 49.61 | 53.19 | 52.91 | - |
| Hashtag prediction | 65.45 | 67.20 | 39.75 | 54.20 | 53.96 | 65.66 | - |
| Hashtag-Stance | 65.76 | 65.64 | 40.65 | 54.91 | 53.47 | 61.94 | - |
| All | 65.93 | 62.60 | 42.21 | 53.79 | 54.04 | 64.20 | **37.99** |

As we can see pre-training always brings an improvement over the baseline, however as expected this improvement is more important if the pretraining is closely related to the task. *Sentiment analysis* which consists of a different task on a different corpus that the target ones only brings a very small gain. *Hashtag-Stance* that is very close to the final task on a similar corpus is the best pre-training method, at a very reasonable cost since it only needs supervision for selecting a small set of non-ambiguous tags. The *Hashtag prediction* is a very interesting pre-training method since it does not require any supervision, while achieving improvement over the baseline (over 5% absolute improvement).

For Task B were no training data was available, our pre-training process gave a large improvement of 10% compared to the baseline trained on the corpora of Task A.

Table 11: results on the Brexit usecase

| Class | Recall | Precision | FMeasure |
|---|---|---|---|
| FAVOR | 0.7319 | 0.7799 | 0.7551 |
| AGAINST | 0.7658 | 0.7412 | 0.7532 |
| Accuracy | | 0.7569 | |

**Application to the Brexit usecase** The same method has been applied to the SENSEI *Brexit* usecase studied in year 3. From the *Brexit* corpus collected in WP2 we selected a set of tweets with hashtags. The top 200 hashtags have been manually labelled as *pro* or *against* Brexit. All the tweets containing a *pro* or *against* hashtag were selected to form a training corpus for our CNN stance classifier. We obtained a training corpus of 411K tweets, balanced between the two stances.

For the test corpus we used only 4 non ambiguous hashtags: `#strongerin, #strongerout, #betterin, #betterout`. All tweets containing `#strongerin` and `#betterin` were considered as against Brexit, and all tweets containing `#strongerout` and `betterout` were considered as in favor of the Brexit. We removed these tweets from the training corpus and considered them as the test corpus (1115 tweets). This was a *cheap* method for obtaining a test corpus with very limited supervision. One issue with this method is that it is not possible to have a *none* label in the evaluation corpus.

The results obtained after the same pre-training method as the one used for Semeval are displayed in table 11. As we can see we obtained an accuracy of 75% although very little supervision was needed

to obtain the training corpus.

## 4.2 Mood Extraction in Social Media

The extraction of mood from social media text is becoming more and more important, and can be considered an evolution of sentiment analysis. Data from social media includes news articles or other multimedia contents, user's generated content such as likes, dislikes, emotions and tastes. Exploiting the CorEA dataset collected during periods 1 and 2 of the SENSEI project, we used 5 classes of user-declared mood dimensions: indignation, sadness, amusement, worry and satisfaction. In Figure 6, we present a spider plot with reference mood scores selected from a set of comments.

**Experimental Methodology** We investigated two different approaches for the prediction of mood from social text: 1) we trained a system that can predict the mood score for each mood category from the articles and comments, 2) we trained classification models to classify either positive or negative mood. For the experiments, we utilized a different set of features, which include word-ngrams, character-ngrams, stylometric, psycholinguistic and ngrams of part-of-speech. Our investigated mood categories include amused, satisfied, disappointed, worried and indignant.

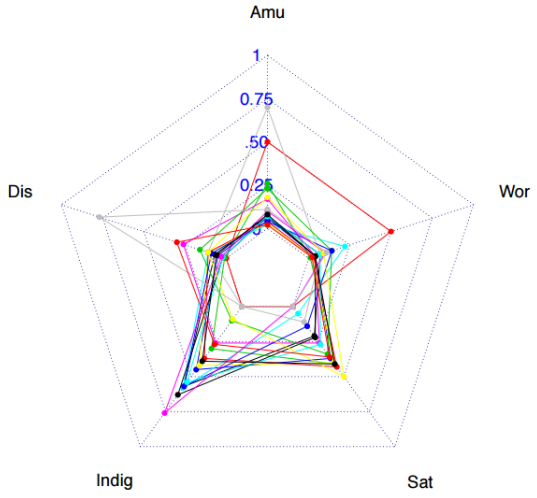


Figure 6: Spider plot of the reference mood scores from the selected comments. Amu- Amusement, Dis-disappointment, Indig-indignation, Sat-satisfaction, Wor-worried.

**Feature Extraction** We experimented with different sets of features for the prediction of the mood in social media. The features we used are reported below:

**Word-ngram** We investigated the bag-of-word-ngrams, with $3 >= n >= 1$, and their logarithmic term frequencies (tf) multiplied with inverse document frequencies (idf) – TF-IDF. Even-though bag-of-words model has many drawbacks such as data sparsity and high dimensionality. However, it is simplest and has been work well in most of the text-based classification task. As the bag-of n-grams approach

results in a large dictionary, which increases computational cost, therefore, we select 5K most frequent n-grams.

**Character-ngram** Similar to the bag-of-word-ngrams, we also extracted and evaluated bag-of-character-ngrams, with $6 >= n >= 2$. The motivation of experimenting with this feature set is due to its success in sentiment classification task as reported in [57].

**Part-of-Speech features (POS)** To extract POS features we used TextPro [58] and designed the feature vector using bag-of-ngram approach, with $3 >= n >= 1$.

**Stylometric Features** The use of stylometric features has its root in the domain of authorship identification [59, 60, 61, 62]. Its use has also been reported for the text categorization and discourse classification problems [63, 64]. In authorship identification task, the stylometric features is defined as different groups such as lexical, syntactic, structural, content specific, idiosyncratic and complexity-based [63, 60, 62]. In this work, we use the term *stylometric* to refer to the complexity-based features reported in [65, 66]. Stylometric features are reported in detail in Figure 7.

**Psycholinguistic Features** To extract the psycholinguistic features from the articles and comments we utilized Linguistic Inquiry Word Count (LIWC) [67], which is a knowledge-based system developed by Pennebaker et al. over the past few decades. The uses of these features have been studied in different research areas of psychology and sociology, and mostly used to study gender, age, personality, honesty, dominance, deception, and health to estimate the correlation between these attributes and word use [68, 69]. The success of these features has also been reported in literature [70, 49, 42]. The types of LIWC features include the following:

- General: word count, average number of words per sentence, a percentage of words found in the dictionary and percentage of words longer than six letters and numerals;
- Linguistic: pronouns and articles.
- Psychological: affect, cognition, and biological phenomena.
- Paralinguistic: accents, fillers, and disfluencies.
- Features about personal concern include work and home.
- Punctuation and spoken categories.

Since it is a knowledge based system, therefore it is packaged with dictionaries for different languages including Italian. For this work, we used the Italian version of the dictionary [71]. It contains a total of 102 features. The LIWC feature processing differs according to types of features, which include counts and relative frequencies, see [69].

**Mood Prediction System** We defined the prediction of moods as a regression task, where we have to predict one score for each of the five mood dimensions. For the mood prediction experiments, we

| Features Details. |
|---|
| **General** |
| • word count = $N$<br>• dictionary size = $V$ |
| **Length-based features:** |
| • Average word length<br>• Short word ratio (length = 1-3) to $N$ |
| **Frequency-based Ratios** |
| • Ratio of Hapax Legomena to $N$<br>• Ratio of Hapac Dislegomena to $N$ |
| **Lexical Richness using transformations of $N$ and $V$:** |
| • Mean Word Frequency = $N/V$<br>• Type-Token Ratio = $V/N$<br>• Guiraud's $R = V/sqrt(N)$<br>• Herdan's $C = log(V)/log(N)$<br>• Rubet's $K = log(V)/log(log(N))$<br>• Maas $A = (log(N) - log(V))/log^2(N) = a^2$<br>• Dugast's $U = log^2(N)/(log(N) - log(V))$<br>• Lukjanenkov and Neistoj's $LN = (1 - V^2)/(V^2 * log(N))$<br>• Brunet's $W = N^(V^( - a)), a = 0.172$ |
| **Lexical Richness using Frequency Spectrum:** |
| • Honore's $H = b(log(N)/a - (V(1, N)/V)), b = 100, a = 1$<br>• Sichel's $S = V(2, N)/V$<br>• Michea's $M = V/V(2, N)$<br>• Herdan's $V = sqrt(sum(V(i, N) * (V(i, N)/N)^2) - 1/V)$<br>• Yule's $K = a(-1/N + sum(V(i, N) * (V(i, N)/N)^2)), a = 1$<br>• Simpson's $D = sum(V(i, N)(V(i, N)/N)(V(i, N) - 1)/(N - 2))$<br>• Entropy = $V(i, N)(-log((V(i, N)/N))^s(V(i, N)/N)^t, s = t = 1$ |
| • Lists 23-52.<br>• Length ratios 30 features |

Figure 7: Stylometric features in detail.

utilized Random Forest as a learning algorithm [72]. It is a decision tree based algorithm where each decision tree is generated by randomly sampling instances and features, then the score of the forest is computed by averaging the scores from the trees. For this experiment, the number of the tree is set to 100.

We used development set to do some preliminary experiments, then, to obtain the results on the test set, we trained the model by combining training and development set. For each task and feature set, we normalized each feature to have zero mean and unit variance.

Since mood dimensions do not require manual annotation and are available as metadata from *corriere.it*, the source of data of the CorEA corpus, we decided to collect much more data than the one provided in the CorEA corpus for training. We collected 2200 articles (CorEA has 26) and the associated 300K comments (CorEA has 2900). As a part of preprocessing, we filtered some data to remove the outliers, for each mood category, for articles and comments, respectively. Outliers are computed based on the mood scores appeared independently in each mood category of the articles and comments. After that data was partitioned into the train, development and test set with 60%, 20%, and 20%

respectively. As a part of preprocessing, we removed URLs from the text even though the URL itself represents some information. We observed that there are similar distributions in the mood categories for both articles and comments. A lexical analysis has been performed on articles and comments to understand the complexity of the task. We observed that for the article the average number of the token is 550 per article with maximum 3,188 and minimum 44. Whereas for comment, the average is 44 with maximum 285 and minimum 1 token.

**Results and Discussion** We measured the performance of the mood prediction system using Root-Mean-Square-Error (RMSE). We computed the baseline results by randomly generating the scores using the gaussian distribution based on the prior mean and standard deviation, as presented in 12 and 13.

Table 12: Baseine results. Randomly selected from gaussian distribution based on prior mean and standard deviation.

| Type | Amu | Dis | Indig | Sat | Worr | Avg |
|---|---|---|---|---|---|---|
| **Articles** | 0.13 | 0.15 | 0.38 | 0.37 | 0.13 | **0.23** |
| **Comments** | 0.17 | 0.18 | 0.35 | 0.23 | 0.17 | **0.22** |

Table 13: Detailed results for the prediction of mood scores. Prediction scores are evaluated using RMSE, the lower is the better.

| Class | Word-ngram | Char-ngram | POS | Style | LIWC |
|---|---|---|---|---|---|
| **Articles** | | | | | |
| **Amused** | **0.100** | 0.100 | 0.102 | 0.120 | 0.102 |
| **Disappointed** | **0.108** | 0.112 | 0.116 | 0.128 | 0.120 |
| **Indignant** | 0.266 | 0.274 | 0.280 | **0.247** | 0.278 |
| **Satisfied** | 0.267 | 0.276 | 0.271 | **0.166** | 0.275 |
| **Worried** | **0.095** | 0.096 | 0.099 | 0.118 | 0.099 |
| **Avg** | 0.167 | 0.172 | 0.174 | **0.156** | 0.175 |
| **Comments** | | | | | |
| **Amused** | 0.118 | 0.119 | 0.119 | 0.119 | 0.120 |
| **Disappointed** | 0.126 | 0.128 | 0.127 | 0.128 | 0.128 |
| **Indignant** | 0.244 | 0.246 | 0.246 | 0.246 | 0.247 |
| **Satisfied** | 0.164 | 0.165 | 0.165 | 0.165 | 0.166 |
| **Worried** | 0.117 | 0.118 | 0.118 | 0.118 | 0.118 |
| **Avg** | 0.154 | 0.155 | 0.155 | 0.155 | 0.156 |

We obtained better results with word-ngram for the mood categories of the articles. For the comments, we also obtained better results with word-ngram and almost similar results across other feature sets. Our results for the articles and comments are statistically significant ($p < 0.05$) compared to the random baseline. The overall results for the articles are better with the stylometric feature set, where word-ngram provides second best results.

# 5 Conclusions

We presented our contributions to the design of computational models for five different para-semantic tasks, including empathy, mood, competitiveness, stance and problem detection for automatic conversation analysis. We selected the tasks in such a way to maximize usefulness in the target SENSEI objectives, as well as novelty for the scientific community. In particular there are few previous works on empathy, stance and Agreement/Disagreement treated as computational tasks for automatic conversation analysis. All the results of the systems reported here, both on call center and social media conversations are very promising compared to the baselines. Many of the systems presented here, as well as Agreement/Disagreement structures classifiction, reported in D4.3, were succesfully exploited for summarisation of human-human conversation (described in D5.3) and provide useful information not covered by the semantic modules. The evaluation of the effectiveness of the para-semantic modules is reported in D1.4.

# References


[1] Akiko Murakami and Rudy Raymond. Support or oppose?: classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 869–875. Association for Computational Linguistics, 2010.

[2] Jose M Chenlo, Alexander Hogenboom, and David E Losada. Sentiment-based ranking of blog posts using rhetorical structure theory. In *Natural Language Processing and Information Systems*, pages 13–24. Springer, 2013.

[3] Valerio Basile and Malvina Nissim. Sentiment analysis on italian tweets. *WASSA 2013*, page 100, 2013.

[4] Erik Cambria, Catherine Havasi, and Amir Hussain. Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *FLAIRS Conference*, pages 202–207, 2012.

[5] Alex Lascarides and Nicholas Asher. Agreement and disputes in dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 29–36, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[6] Youngja Park. Your call may be recorded for automatic quality-control. Technical report, IBM Research Report RC24574 (W0806-018) June 5, 2008.

[7] Frederik Cailliau and Ariane Cavet. Mining automatic speech transcripts for the retrieval of problematic calls. In *Computational Linguistics and Intelligent Text Processing*, pages 83–95. Springer, 2013.

[8] Anat Rafaeli, Lital Ziklik, and Lorna Doucet. The impact of call center employees' customer orientation behaviors on service quality. *Journal of Service Research*, 10(3):239–255, 2008.

[9] L Alan Witt, Martha C Andrews, and Dawn S Carlson. When conscientiousness isnt enough: Emotional exhaustion and performance among call center customer service representatives. *Journal of Management*, 30(1):149–160, 2004.

[10] Nathalie Camelin, Frederic Bechet, Geraldine Damnati, and Renato De Mori. Detection and Interpretation of Opinion Expressions in Spoken Surveys. *IEEE Transactions on Audio, Speech, and Language Processing*, 18:2:369–381, 2010.

[11] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, et al. The interspeech 2012 speaker trait challenge. In *Proc. of INTERSPEECH*, 2012.

[12] Firoj Alam and Giuseppe Riccardi. Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 955–959. IEEE, 2014.

[13] Shourya Roy, Ragunathan Mariappan, Sandipan Dandapat, Saurabh Srivastava, Sainyam Galhotra, and Balaji Peddamuthu. Qa rt: A system for real-time holistic quality assurance for contact center dialogues. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[14] Swapna Somasundaran and Janyce Wiebe. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 226–234. Association for Computational Linguistics, 2009.

[15] Lu Wang and Claire Cardie. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. *ACL 2014*, page 97, 2014.

[16] Emily M Bender, Jonathan T Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In *Proceedings of the Workshop on Languages in Social Media*, pages 48–57. Association for Computational Linguistics, 2011.

[17] Jacob Andreas, Sara Rosenthal, and Kathleen McKeown. Annotating agreement and disagreement in threaded discussion. In *LREC*, pages 818–822. Citeseer, 2012.

[18] Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. A corpus for research on deliberation and debate. In *LREC*, pages 812–817, 2012.

[19] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, pages 1320–1326, 2010.

[20] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! In *ICWSM*, 2011.

[21] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*, pages 1–10, 2011.

[22] Thin Nguyen, Dinh Phung, Brett Adams, Truyen Tran, and Svetha Venkatesh. Classification and pattern discovery of mood in weblogs. In *Advances in knowledge discovery and data mining*, pages 283–290. Springer Berlin Heidelberg, 2010.

[23] Josef Steinberger, Mohamed Ebrahim, Maud Ehrmann, Ali Hurriyetoglu, Mijail Kabadjov, Polina Lenkova, Ralf Steinberger, Hristo Tanev, Silvia Vázquez, and Vanni Zavarella. Creating sentiment dictionaries via triangulation. *Decision Support Systems*, 53(4):689–694, 2012.

[24] Frederic Bechet, Benjamin Maza, Nicolas Bigouroux, Thierry Bazillon, Marc El-Beze, Renato De Mori, and Eric Arbillot. Decoda: a call-centre human-human spoken conversation corpus. In *LREC*, pages 1343–1347, 2012.

[25] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[26] Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. In *Proc. of Interspeech*, pages 312–315, 2009.

[27] Douglas W Maynard. On the functions of social conflict among children. *American Sociological Review*, pages 207–223, 1985.

[28] Helmut Gruber. Disagreeing: Sequential placement and internal structure of disagreements in conflict episodes. *Text-Interdisciplinary Journal for the Study of Discourse*, 18(4):467–504, 1998.

[29] Helmut Gruber. Questions and strategic orientation in verbal conflict sequences. *Journal of Pragmatics*, 33(12):1815–1857, 2001.

[30] Marjorie Harness Goodwin, Charles Goodwin, and Malcah Yaeger-Dror. Multi-modality in girls game disputes. *Journal of pragmatics*, 34(10):1621–1649, 2002.

[31] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *language*, pages 696–735, 1974.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[33] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proc. of the 21st ACM international conference on Multimedia (ACMM)*, pages 835–838. ACM, 2013.

[34] Carole Lailler, Anas Landeau, Frdric Bchet, Yannick Estve, and Paul Delglise. Enhancing the ratp-decoda corpus with linguistic annotations for performing a large range of nlp tasks. In *Proceedings of LREC*, 2016.

[35] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *Proc. of Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–4, 2011.

[36] Paul Deléglise, Yannick Estève, Sylvain Meignier, and Téva Merlin. The lium speech transcription system: a cmu sphinx iii-based system for french broadcast news. In *Interspeech*, pages 1653–1656, 2005.

[37] Martin L Hoffman. Empathy and prosocial behavior. *Handbook of Emotions*, 3:440–455, 2008.

[38] James J Gross. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271, 1998.

[39] David Sander, Didier Grandjean, and Klaus R Scherer. A systems approach to appraisal mechanisms in emotion. *Neural Networks*, 18(4):317–352, 2005.

[40] James J Gross and Ross A Thompson. Emotion regulation: Conceptual foundations. *Handbook of Emotion Regulation*, 3:24, 2007.

[41] James J Gross and Lisa Feldman Barrett. Emotion generation and emotion regulation: One or two depends on your point of view. *Emotion review*, 3(1):8–16, 2011.

[42] Morena Danieli, Giuseppe Riccardi, and Firoj Alam. Emotion unfolding and affective scenes: A case study in spoken conversations. In *Proc. of Emotion Representations and Modelling for Companion Systems (ERM4CT) 2015,*. ICMI, 2015.

[43] Morena Danieli, Giuseppe Riccardi, and Firoj Alam. Annotation of complex emotion in real-life dialogues. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors, *Proc. of 1st Italian Conf. on Computational Linguistics (CLiC-it) 2014*, volume 1, 2014.

[44] Thomas Schmidt. Transcribing and annotating spoken language with EXMARALDA. In *Proc. of LREC 2004 Workshop on XML-based Richly Annotated Corpora*, pages 69–74, 2004.

[45] E Stepanov, B Favre, F Alam, S Chowdhury, K Singla, J Trione, F Béchet, and G Riccardi. Automatic summarization of call-center conversations. In *In Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, 2015.

[46] Danieli Morena, Riccardi Giuseppe, Emma Barker, Jonathan Foster, Adam Funk, Rob Gaizauskas, Mark Hepple, Emina Kurtic, Massimo Poesio, Letizia Molinari, and Vincenzo Giliberti. *Preliminary Version of Use Case Design. SENSEI project deliverable D1.1*. University of Trento, 2014.

[47] Shammur A Chowdhury, Giuseppe Riccardi, and Firoj Alam. Unsupervised recognition and clustering of speech overlaps in spoken conversations. In *Proc. of Workshop on Speech, Language and Audio in Multimedia - SLAM2014*, pages 62–66, 2014.

[48] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. Paralinguistics in speech and language state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39, 2013.

[49] Firoj Alam and Giuseppe Riccardi. Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition. In *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 955–959, May 2014.

[50] Igor Kononenko. Estimating attributes: analysis and extensions of relief. In *Proc. of Machine Learning: European Conference on Machine Learning (ECML)*, pages 171–182. Springer, 1994.

[51] Firoj Alam and Giuseppe Riccardi. Comparative study of speaker personality traits recognition in conversational and broadcast news speech. In *Proc. of Interspeech*, pages 2851–2855. ISCA, 2013.

[52] John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Microsoft Research, 1998.

[53] Shammur Absar Chowdhury, Morena Danieli, and Giuseppe Riccardi. Annotating and categorizing competition in overlap speech. In *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[54] Shammur A Chowdhury, Giuseppe Riccardi, and Firoj Alam. Unsupervised recognition and clustering of speech overlaps in spoken conversations. In *Proc. of Workshop on Speech, Language and Audio in Multimedia - SLAM2014*, pages 62–66, 2014.

[55] Mijail Kabadjov, Evgeny A Stepanov, Fabio Celli, Shammur A Chowdhury, Benoit Favre, Adam Funk, Udo Kruschwitz, and Massimo Poesio. The sensei discourse analysis tools.

[56] Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. Stance and sentiment in tweets. *arXiv preprint arXiv:1605.01655*, 2016.

[57] Ahmed Abbasi, Hsinchun Chen, and Arab Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12, 2008.

[58] Emanuele Pianta, Christian Girardi, and Roberto Zanoli. The textpro tool suite. In *LREC*. Citeseer, 2008.

[59] G Udny Yule. On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3/4):363–390, 1939.

[60] Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):7, 2008.

[61] Shane Bergsma, Matt Post, and David Yarowsky. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337. Association for Computational Linguistics, 2012.

[62] Marco Cristani, Giorgio Roffo, Cristina Segalin, Loris Bazzani, Alessandro Vinciarelli, and Vittorio Murino. Conversationally-inspired stylometric features for authorship attribution in instant messaging. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1121–1124. ACM, 2012.

[63] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

[64] Fabio Celli, Evgeny A. Stepanov, and Giuseppe Riccardi. Tell me who you are, i'll tell whether you agree or disagree: Prediction of agreement/disagreement in news blog.

[65] Kumiko Tanaka-Ishii and Shunsuke Aihara. Computational constancy measures of texts—yule's k and rényi's entropy. *Computational Linguistics*, 2015.

[66] Fiona J Tweedie and R Harald Baayen. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352, 1998.

[67] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71, 2001.

[68] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res.(JAIR)*, 30:457–500, 2007.

[69] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[70] Thin Nguyen, Dinh Phung, Brett Adams, and Svetha Venkatesh. Mood sensing from social media texts and its applications. *Knowledge and information systems*, 39(3):667–702, 2014.

[71] F. Alparone, S. Caso, A. Agosti, and A. Rellini. The italian liwc2001 dictionary. Technical report, LIWC.net, Austin, TX, 2004.

[72] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.