Semantic Analysis of Conversations and Fuzzy Logic for the Identification of Behavioral Profiles on Facebook Social Network

https://doi.org/10.3991/ijet.v14i07.8832

Youness Chaabi (✉),
CESIC, Royal Institute of the Amazigh Culture (IRCAM), Rabat, Morocco
Chaabi@ircam.ma

Lekdioui Khadija
LASTID, Ibn Tofail University, Kenitra, Morocco

Boumediane Mounia
CESIC, Royal Institute of the Amazigh Culture (IRCAM), Rabat, Morocco
Mohammed V University, Rabat, Morocco

Abstract—In this article we describe a new multi-agent approach for the accompaniment and follow-up of learners (tutoring) in collaborative social networks via network technologies. To assist learners in their collaborative learning process, the system we propose offers the possibility to identify the sociological behavioral' profile of each learner on the basis of the automatic analysis of the asynchronous textual conversations exchanged between learners.

To achieve our aims, we first describe the sociological profiles that we use in our model. Then, we propose the approach for the semantic analysis of the messages exchanged (full text), as well as the proposed indicators for the determination of these profiles. After, we present the results of the implementation of the system developed as part of an experiment that we conducted with the students of the Master Program "Software Quality" in the Ibn Tofail University of Kenitra, Morocco. We did indeed obtain very good performances during tests on corpora of messages.

Keywords—Multi-agent system, Collective Learning, Semantic analysis, social behavior profiles, fuzzy logic, Social Networks.

1 Introduction

Our work is in the field of Computer Environments for Human Learning (CEHL). In this article, we are interested in Computer-Assisted Collective Learning, better known by the name: Computer-Supported Collaborative Learning (CSCL) [1].

In the context of collaborative distance learning, textual communication is of paramount importance [2].

Asynchronous text communication tools such as e-mail and the forum avoid face-to-face constraints [3]. These tools remain to this day the best compromise between flexibility and interactivity for the realization of an online collaborative work. Faced with the large number of messages posted in forums, tutors of a practice community often feel incapable to construct synthetic representation of the activity of individuals and groups.

Hence, the tutor may lack objectivity when he uses it to evaluate the involvement and place of learners in exchanges and to identify their social behaviors [4, 5, 6]. As part of the automatic analysis of collaborative activities of learners, we propose a semantic analysis approach of asynchronous textual conversations between learners to determine their social behaviors. In the context of distance learning where there is no interaction between the tutor and the learner, the data collected tend to be more imperfect than those obtained by the face-to-face interaction. The presence of imperfect information is an important factor that leads to errors in the determination of the learner's social behaviors [7]. These imperfections are the consequence of the approximations involved in the data collection due to the nature of human knowledge. It can also be the consequence of loss of information during the previous steps.

The theory of fuzzy logic is presented as a privileged tool for modeling situations with inaccuracies [8]. One of the main motivations for using fuzzy logic is the improved handling of information imperfections. Indeed, the reasoning of a fuzzy logic system is considered "easy", from the point of view of understanding and / or modification by designers and users. One of the factors that enhance this consideration is human similarity. Fuzzy logic can provide descriptions of knowledge as a human and imitate its pattern of reasoning about vague concepts. This is of particular interest in the design of a system modeling the interpretable knowledge of the learner that is based on the reasoning and conceptualization of the teacher-expert.

2 Human Behavior Profiles

In his work in ethology Robert Pléty has studied the behavior of students working in groups; in particular, he analyzed interactions between learners working in groups of four to solve algebra problems [3, 9]. Based on this work, we studied social behavior patterns in online collaborative work. Thus from the experiments, we managed to find the same patterns of behavior among students working in groups on social networks.

In order to determine these behavioral profiles, four kinds of observations are made for each student: the volume (number) of interventions, the different types of interventions, the communication gesture types (look and movement) and the reactions of other participants (consequences of behaviors). These behavioral profiles generalize behavior patterns and are called profiles in the rest of the paper. Pléty identified four different profiles: Animator, Checker, Seeker and Independent. The characteristics of these four profiles are summarized in table 1.

Table 1. Behavioral profiles of students working in groups [3, 9].

Name	Volume of Intervention	Type of Interventions	Entrained Reaction
Animator	Important	Question or proposal	Followed by positive reactions
Checker	Enough important	Reaction, response and evaluation	No monitoring reactions
Seeker	Little important	(Very doubtful (question))	Questions are well accepted
Independent	Low	Little or no proposal or evaluation.	Interventions remain unresolved

3 System Architecture

The proposed approach is to automatically analyze the content of the messages. Following this analysis, a profile for each learner is determined. The challenge lies in identifying behavioral patterns of learners through the automatic analysis of the content of the messages exchanged. Each of these messages undergoes a sequence of treatments. In this article, we present four different profiles that we have identified and characterized through different criteria. In order to determine a profile, four treatments are performed. The first is to simplify messages by removing unnecessary information. The second treatment consists of a semantic message analysis. Using all the calculated indicators as well as a model of fuzzy logic. The fourth treatment makes it possible to determine a behavioral profile for each human actor in the system. We mention that the interactivity between tutor-learners or learner-learners is essentially through textual exchange. Below is the architecture that describes how the system works (Figure 1)

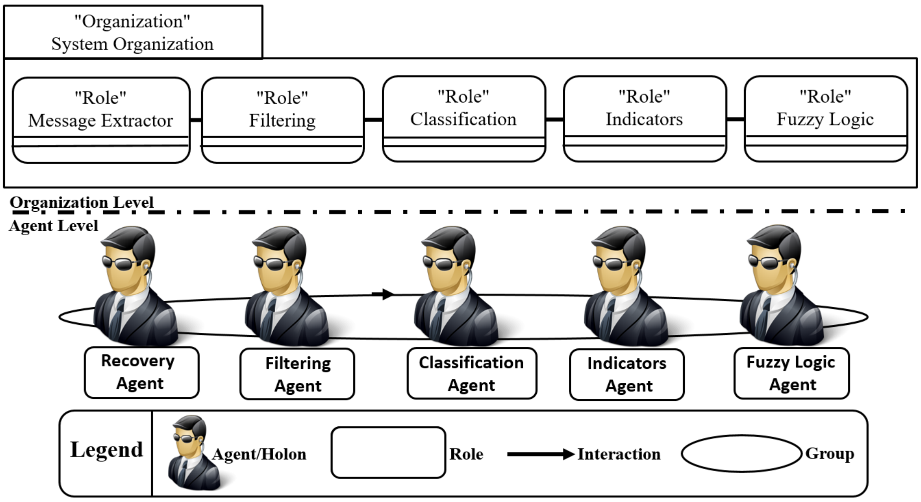


Fig. 1. The organization and group of a Behavior Analysis community in Janus

3.1 Recovery agent

First, a recovery module allows the extraction of interaction messages on social network using open graph protocol for Facebook, as well as their preparation for different subsequent treatments.

According to our experiment, a first treatment of the corpus resides in the correction of misspelling and grammar. Spelling and grammar errors can occur in text analysis for humans as well as for software. A misspelled word (or phrase) can completely change the analysis.

Spelling and grammar correction is obtained using a dictionary word (corpus) associated with an algorithm that takes into account the variation of the language (verbal conjugation, agreement nouns and adjectives). It consists in comparing the words of the text with the corpus, taking into account the context of the sentences. Nonetheless, the usefulness of the spelling and grammar checker, it cannot replace a careful personal check.

3.2 Filtering agent

After message extraction, the pre-filtering treatment automatically deletes the words that do not contain information. Indeed, in text messages, many words provide little information about the message concerned. These words are automatically deleted using "empty words" for each language.

The words that appear most often in a corpus are usually empty grammatical words (empty words): articles, prepositions, linking words, determiners, adverbs, undefined adjectives, conjunctions, pronouns and verbs auxiliaries, etc. These words constitute a large part of the words of a text, but unfortunately are weakly informative on the meaning of a text since they are present on the set of texts. According to Zipf law [10], their removal during message preprocessing allows to save time during the modeling and the analysis of the message.

3.3 Classification agent

The classification agent measures the semantic similarity that a new message belongs to one of the four categories (Animator, checker, seeker, and independent) from the proportion of training messages belonging to that category.

To begin, we want to clarify the context of extraction of training messages. We have worked regularly with a number of tutors on exchanges between learners; we have come to address a set of messages specific to each profile category.

Based on tutors' suggestions, the intuitive analysis of messages shows that messages can be classified as follows: messages that aim to initiate an interaction and to initiate a discussion topic proposition, messages asking for information or expecting a response from others, messages in which an answer to the requests of others is provided, finally previous messages that clarify or deepen a current topic of discussion.

Semantic similarity measurement: In many areas of research such as psychology, linguistics, cognitive science and artificial intelligence, the calculation of semantic similarity between words is an important issue [11]. Semantic similarity (or semantic proximity) is a metric defined on a set of messages or terms, where the idea of the distance between them is based on the similarity of their meanings or semantic contents [12]. On the other hand, as opposed to semantic similarity, we find the type of similarity that can be estimated based on syntactic representations of terms. Mathematical tools are used to estimate the strength of the semantic relation between units of language, concepts or instances, through a numerical description. The latter is obtained by comparing information in support of their meaning or description of their nature.

Semantic similarity can be estimated by defining a topological similarity, using ontologies to define the distance between terms/concepts [13]. For example, a naive metric for the comparison of ordered concepts in a partially ordered set and represented as nodes of an acyclic oriented graph (eg, a taxonomy), would be the shortest path connecting the two concept nodes. Semantic proximity between language units (eg words and sentences) can also be estimated using statistical means such as a vector space to correlate words and textual contexts from an appropriate body of text.

Taxonomy: The concept of semantic similarity is more specific than kinship or semantic relation, since the latter includes concepts such as antonymy and meronymy, while similarity does not. However, much of the literature uses these terms interchangeably with terms like semantic distance [14, 15]. Essentially, the notions of semantic similarity, semantic distance and semantic proximity, provide an answer to the following question: "What is the degree of resemblance between the term A and term B?". The answer to this question is usually a number between -1 and 1, or between 0 and 1, where 1 means extremely high similarity.

Topological similarity measurement: There are essentially two types of approaches that compute the topological similarity between ontological concepts:

- Edge-based approach: uses edges and their types as a data source.
- Content-based approach: the main sources of data are nodes and their properties.

Semantic similarity: Or semantic relation is a concept of measuring the proximity of terms or documents in the context of their meaning. We have two different methods for calculating semantic similarity. One is to define a topological similarity, using ontology to define a distance between words. The other is based on the use of statistical means such as the vector space model to correlate words and textual contexts from an appropriate body of text. We choose the first approach using the WordNet ontology for semantic similarity calculation. The similarity calculation in this approach is based on the fact that the similarity depends on the common and distinct characteristics of the objects.

WordNet: Is a lexical ontology for the English language [16]. It is a semantic network developed by Princeton University that models lexical knowledge in a taxonomic hierarchy. WordNet contains three databases: one for nouns, one for verbs and one for adverbs and adjectives. Terms and concepts are organized in Synsets (List of terms or synonymous concepts). The basic part of WordNet is the Synset which brings together the synonyms of a concept. Synsets are linked in some models by relations such as: hypernymy (type of), meronymy (part of) and antonymy (opposite word) [17, 18]. The semantic similarity in WordNet can be calculated by two methods: the path length and the information content. The first method calculates the number of nodes or relationships between nodes in the taxonomy. The advantage of this method is that it is not dependent on either the static distribution of the corpus or the distribution of words. In our context, we considered only two concepts (relationship and name) in the WordNet hierarchy. We use WordNet 2.1, which contains nine distinct name hierarchies where sometimes the path between two concepts may not exist (see Figure 2). Therefore, we create a root node ("Entity" see Figure 2) that includes all the nine hierarchies given in WordNet.

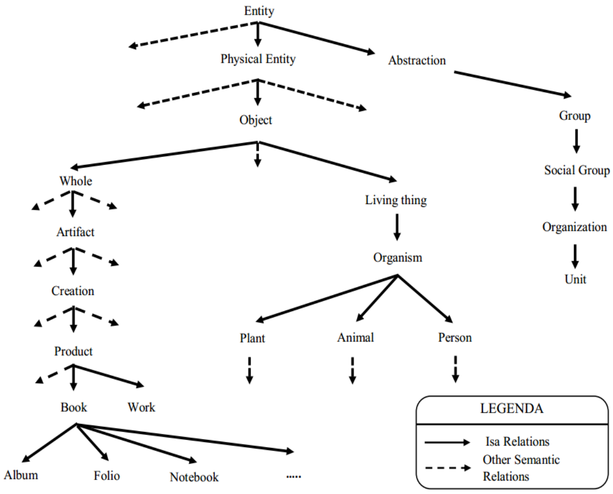


Fig. 2. Janus Extract from the nominal WordNet hierarchy

Semantic similarity measurement process: The classification agent makes it possible to carry out a complete sequence of treatment. The semantic similarity calculation process is illustrated in Figure 3. This process consists of three phases:

- Phase 1: Term construction module
- Phase 2: Calculate semantic module
- Phase 3: Semantic Similarity Measures

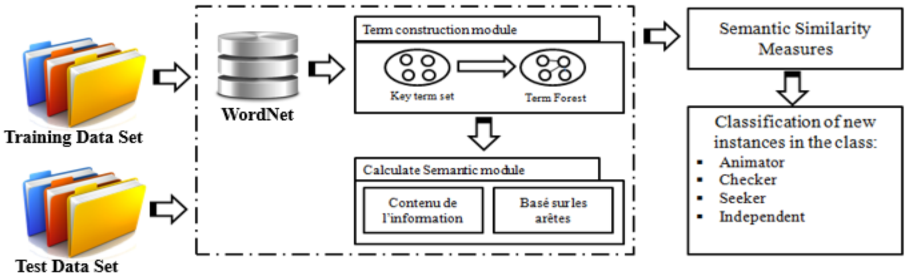


Fig. 3. Semantic Similarity Calculation Diagram

Phase 1: Term construction module: The objective of the module is to select all the words of the text that exist on WordNet and to obtain the relation between these words. We use WordNet to generate a richer text representation. In this module, we used the hyperlinks provided by WordNet as useful features for text analysis.

Phase 2: Calculate semantic module: We use the different algorithms that use the semantic similarity measures to find the appropriate meanings of the words according to the context at the level of the sentence or the text. We quote some algorithms that calculate the semantic similarity:

- Length of the path
- Similarity of Resnik
- Lin's similarity
- Distance from Jiang-Conrath
- Measurement of Wu and Palmer

In what follows, we will explain each of these algorithms.

The path length algorithm: When concepts are organized in a hierarchy, it is appropriate to measure similarity based on structural measures that find path lengths between concepts. In fact, there are a variety of such approaches proposed in English. [19] Rada, et al. (1989), developed a measure based on the length of paths between concepts in the WordNet hierarchy. The shortest path measurement emphasizes the proximity of two concepts in the hierarchy. In a thesaurus hierarchy graph, the shorter the path between two words, the more similar these words are:

- The words are quite similar to the parents;
- Words are less similar to words that are far from them in the network
- Pathlen (c1, c2) = number of edges of the shortest path
- Path-based similarity often involves a logarithmic transformation

The similarity based on the length of the path is (1):

$$simpath(c1, c2) = -\log pathlen(c1, c2) \quad (1)$$

Resnik similarity algorithm: According to (Philip Resnik) 1995 [20], Sun Microsystems laboratories offer an alternative to finding paths through the notion of informative content. It is a measure of specificity attributed to each concept in a hierarchy based on evidence found in a corpus. A concept with high informational content is very specific, while concepts with low informational content are associated with more general concepts. The information content of a concept is estimated by counting the frequency of occurrence of this concept in a large corpus, as well as the frequency of all concepts subordinate to it in the hierarchy. The probability of a concept is determined by a maximum likelihood estimate, and the information content is the negative log of this probability. Resnik defines a measure of similarity according to which these two concepts are semantically related in proportion to the amount of information they share. The amount of information shared is determined by the information content of the lowest concept in the hierarchy that covers the two given concepts. The similarity of the words based on the informative content:

- Still depends on the structure of the thesaurus
- Improves the path-based approach by using normalizations based on the depth of the hierarchy
- Represents the distance associated with each edge
- Adds probabilistic information derived from a corpus

The probability that the random word is an instance of the concept is (2):

$$p(c) = \frac{\sum w \in w_{(c)} \text{count(w)}}{\text{N}} \tag{2}$$

Where: words (c) is the set of words subsumed by the concept c
N is the number of words in the corpus and the thesaurus
P (root) = 1 since all words are subsumed by the root concept
More the concept is low in the hierarchy, the most probability gets weak
We need two other definitions:
The informative content of a concept (3):

$$IC(c) = -\log P(c) \tag{3}$$

Basic information theory
The lowest common subsume: LCS (c1, c2)
This is the lowest node in the hierarchy that is a hyperonym of c1 & c2
The similarity measure of Resnik is (4):

$$simResnik(c1, c2) = -log\ P(LCS(c1, c2)) \tag{4}$$

It estimates the common amount of information between words using the information content of the lowest common subsumer.

Lin Similarity: Lin's similarity is based on that of Resnik. Dekang Lin (University of Manitoba - Canada), 1998 [21] considers the information content of the lowest common subsumer (lcs) and the two concepts compared. For example, Animal and Mammal are subsumes of Cat and Dog, but Mammal is the lowest subsum. Similarity is more than common information. The similarity between A and B decreases if there are several differences between them (5):
Common point: IC (common (A, B))
Difference:

$$IC\left(description(A, B)\right) - IC(commun(A, B)) \tag{5}$$

Where description (A, B) describes A and B.
Theorem of similarity:
The similarity between A and B is measured by the ratio of the amount of information needed to state that there are common points between A and B, on the information necessary to fully describe A and B (6):

$$SimLin(A, B) = \frac{\text{commun(A,B)}}{\text{description(A,B)}} \tag{6}$$

The common information between two concepts is the double of the information in the lowest common subsumer. The final similarity function of Lin for the concepts in the thesaurus is (7):

$$SimLin(c1, c2) = 2 * \frac{\log\ \mathrm{P(LCS(c1,c2))}}{\log P(c1) + \log P(c2)} \tag{7}$$

Distance from Jiang-Conrath: This measure is related to SimLin expressed as distance instead of similarity. Jay J. Jiang (University of Waterloo - Canada) (1997) [14] considers the information content of the lowest common subsumer (lcs) and the two concepts compared to calculate the distance between them (8). The distance is then used in the calculation of the similarity measure.

$$DistJC(c1, c2) = 2 * \log P(LCS\,(c1, c2)) - (\log P(c1) + \log P(c2)) \tag{8}$$

This distance is transformed into a measure of similarity by taking the reciprocal (9):

$$DistJC(c1, c2) = 1/2 * \log P(LCS\,(c1, c2)) - (\log P(c1) + \log P(c2)) \tag{9}$$

Resnik's measure may not be able to make fine distinctions, as many concepts may share the same lowest common subsum and thus have identical similarity values.

Wu and Palmer Measure: Wu and Palmer (1994) [22] present a similarity measure for general English that is based on the search for the most general concept that subsumes the two measured concepts. The length of the path from this shared concept to the root of the ontology is scaled by the sum of the distances from the concepts to the concept that subsumes them.

The similarity measure of Wu and Palmer calculates the most specific common ancestor of the two concepts, with a minimal number of "is-a-bond" in the common subsumer's path (10).

$$Sim = \frac{2*\mathrm{h}}{\mathrm{h1+h2+h}} \tag{10}$$

h: is the depth of the subsume from the root of the hierarchy.

h1 and h2: the minimum number of "is-a-link" from concept c1 and c2 to the most specific common subsum (11).

$$Depth(x) = shortest\ is - a\ path(root, x) \tag{11}$$

The measure of the shortest way:

The measurement of the shortest path emphasizes the proximity of two concepts in the hierarchy (12).

$$Sim = 2 * MAX - \ 1 \tag{12}$$

Where MAX is the maximum path length between two concepts in the taxonomy and L is the minimum number of "is-a-link" between the concepts c1 and c2.

Phase 3: Semantic similarity measures: The semantic vectors for T1 and T2 can be formed from T and corpus statistics. The process of derivation of semantic vectors for T1 (13):

Word w, define

$$Sim(W_1, W_2) = max_{c1,c2}[smin(c1, c2)]$$

$$Sim(T1, T2) = \sum_{i=1}^{n} \left(\frac{sim(Wi,Wi+1}{n}\right) \tag{13}$$

We obtain semantic similarity measurement values for each of the above five algorithms between message 1 and message 2 (14):

- Sim Path (T1, T2) = value1
- Sim Resnik (T1, T2) = value2
- Sim Lin (T 1, T2) = value3
- Sim JC (T1, T2) = value4
- Sim Wu (T1, T2) = value5

$$sim(T1, T2) = Max(valeur1, valeur2, valeur3, valeur4, valeur5) \tag{14}$$

Messages are composed of words, so it is reasonable to represent a message using the words it contains.

Unlike traditional methods that use a precompiled word list containing hundreds of thousands of words, our method dynamically shapes semantic vectors only on the basis of the compared messages. Recent research in semantic analysis is usually adapted to automatically extract a semantic vector of words for a sentence [23]. With two messages T1 and T2, a set of words is formed with (15):

$$T = T1 \cup T2$$

$$= \{W_1, W_2, \ldots\ldots, W_n\} \tag{15}$$

The set of words T contains all the distinct words of T1 and T2. Inflectional morphology can cause a word to appear in a message with different forms that have a special meaning for a specific context. For this reason, we use the word form as it appears in the message.

3.4 Indicators agent

We present the formulas used by indicator agent to analyze the discussions in collaborative works. These heuristics formulas were determined from the work of Pléty and were refined in experiments.

Volume of interventions: The following formula calculates the ratio of participation of a learner by dividing nbMsgLearner(p) which is the number of messages sent by learner P, by NbrTotalMessagesGroup(x) that is the number of messages sent by students of the same group.

This ratio refers to the volume of intervention "VI" for a learner (p) belonging to a group x (16):

$$VI = \frac{\text{nbMsgLearner(P)}}{\text{NbrTotalMessagesGroup(x)}} * 100 \tag{16}$$

Type of interventions: Four expressions are used to calculate the Type of Interventions for each learner. The ratios of interventions is calculated as follows (17):

$$RatiosAnimator = \frac{\text{AnimatorMessage(P)}}{\text{Message(A,C,S,I)}} * 100 \tag{17}$$

In this formula, Animator Message (p) is the number of messages of category "animator" (for Example propose, encourage etc.) sent by learner (p). Message(A,C,S,I) is the total number of messages (respectively animator, checker, seeker and independent) sent by the learner. The Calculations of other ratios types (checker, seeker and independent) are obtained similarly.

Entrained reaction: According to the characteristics of the defined profiles (Table 1), the volume of reactions triggered by a message allows to characterize a behavioral profile. For example, an animator profile requires a very large monitoring of reactions compared to that of a checker. We calculate, for each message, direct reactions (first reaction to a message) and indirect reactions (number of interventions after the creation of the message). According to the tree structure defined for messages, the nodes represent the identifiers of messages sent by the learners and the size of this tree is equivalent to the number of interventions made after the creation of the topic. Two expressions, using the n-ary tree structure of the messages, are used to calculate subsequent reactions of each message:

The direct reaction is the number of direct responses to the messages of the learner divided by the total number of direct answers on posted messages by learners in the group.

$$Reaction_{Direct} = \frac{\sum_{i=1}^{m} ReplyToMessage_i(Learner)}{\text{TotalOfReponse(Group)}} * 100 \tag{18}$$

Indirect reaction is the depth of discussion minus the number of direct reaction divided by the sum of the depths of the subjects send by learners.

$$Reaction_{Indirect} = \frac{Depth - Reponses}{\text{TotalDepth}} * 100 \tag{19}$$

3.5 Fuzzy logic agent

Most of the problems encountered can be modeled mathematically. But these models require assumptions that are sometimes too restrictive, making them difficult to apply to the real world. Real world problems must take into account inaccurate and uncertain information. The knowledge that humans have about the world is almost never perfect. They are almost always tainted with a number of uncertainties and inaccuracies. We are not talking here about scientific reasoning, the purpose of which is precisely to get rid of all imperfections, but of all the other reasoning’s that we make every day, unceasingly, about things, people and thoughts surrounding us. Fuzzy logic therefore seems to reproduce the flexibility of human reasoning in taking into account the imperfections of accessible data. It would therefore be interesting to use it at the heart of expert systems, systems whose purpose is to reproduce the cognitive mechanisms of an expert in a particular field. Fuzzy logic can also be used for a decision-making system during the data analysis phase, for example. It can be useful for decision-making, either to discover rules or fuzzy inferences allowing to better understand the data and thus to enlighten the decisions, or to make requests said vague based on the knowledge of the experts.

Indeed, the fuzzy algorithm takes place in 3 steps:

- Transformation of quantitative variables into fuzzy logical variables;
- Use logical rules to evaluate new fuzzy variables at the output;
- Transformation of these fuzzy variables into qualitative variables.

First step: Fuzzification, or definition of the membership functions of the input and output variables, consists in determining for each variable the linguistic values as well as the form of the membership functions and the degree of belonging to different states that one must define. A fuzzy set is characterized by a membership function f: E → [0, 1], which positions the members of the speech universe E in the unit interval [0, 1]. The value 0 means that the member is not included in the given set and the value 1 describes a fully included member. Values between 0 and 1 characterize fuzzy members. The discourse universe of a variable will cover all the values taken by this variable. In our case, the universe of the speech E corresponds to the following percentages: percentage of intervention, percentage of type of intervention and percentage of direct and indirect reactions. The universe of speech E is discredited into 11 elements {0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100}. For an element x of E, the value f (x) represents the degree of membership of x in a fuzzy subset.

Input variables: The input variables are: (E1) Percentage of intervention, (E2) Percentage of intervention, (E3) Percentage of direct reactions and (E4) Percentage of indirect reactions. Three linguistic variables {Low, Medium, High} qualify our input variables.

Output variable: The output variable is the "Behavioral Level" which is a qualitative characterization of the social behavior of the learner "(Animator, Checker, Seeker and Independent)". (S) Behavioral Levels: {Insufficient, Medium, Good, Excellent}. The discourse universe of each input variable is divided into three fuzzy subsets {Low, Medium, High}.

To represent the linguistic variables of the inputs, we defined in collaboration with the expert teacher the membership function of trapezoidal form. The teacher-expert specifies the degrees of belonging of the learner's behavioral levels to each of the fuzzy subsets obtained. The fuzzy subsets associated with the output variable, "Behavioral Levels" are {Low, Medium, Good, Excellent} defined with line-shaped membership functions (Figure 8). The generation of the output variable is done by the system using the center of gravity method [5], in which the system calculates the output variable rounded to the nearest whole number. The ranges of the output variable have been defined from the intuitive analysis statistics made by the tutors. We estimated the ranges of results according to the statistics of the intuitive analysis by the tutors. If the percentage is between:

- 0% and 20%, then the result is Insufficient
- 20% and 50%, then the result is Medium
- 50% and 70%, then the result is Good
- 70% and 100%, then the result Excellent

Second step: Inference engine Now that we have linguistic variables, we will be able to use them in the inference engine. Each rule of the inference engine is written by the designer of the fuzzy system based on the knowledge he has. Designing a fuzzy rule base is an iterative process. The bulk of the work is in the collection of expert knowledge. Thus, using data corresponding to the different inputs and outputs, the expert teacher provides a series of combinations based on the conditions (Table 1) that characterize each behavioral profile "animator, checker, seeker and independent". One of the interests of fuzzy logic in formalizing human reasoning is that the rules are stated in natural language. For example, here is a rule for determining a learner's social behavior:

If
(Volume of intervention IS High) AND
(Type of intervention as Animator IS High) AND
(Direct Reaction IS High) AND
(Indirect Reaction IS High)
THEN
(Level Behavior as Animator IS Excellent)
(Facilitator animator level is excellent)

Third step: Defuzzification: The last step to having an operational blur is called defuzzification. Once the inference is complete, the fuzzy output set is determined but it is not directly usable to give accurate information. It is necessary to move from the "fuzzy world" to the "real world". To do this, there are several methods and the most used is the calculation of the "center of gravity" of the fuzzy set. Once the value of the "Behavioral Levels" output (animator, checker, seeker and independent) is evaluated using the rule base and then "defuzzified", it gives an estimate of the learner's profile based on indicators.

Finally, our system will have a qualitative assessment of the learner's social behaviors (Figure 4), allowing him to identify his shortcomings and weaknesses and to balance the groups according to their social behaviors.

4 General Context of Experimentation

The purpose of this work is to automate some (laborious) tasks usually performed by a human tutor. In this sense, we carried out a comparative study between the human evaluation and the one produced: result of our model. We conducted intuitive analysis experiments on learner conversations. We are interested here in qualitative and quantitative analysis of 4 tutors. A corpus of messages was elaborated from a sample submitted by 9 groups of 4 learners, over a period of 4 months (from March 2nd to June 2nd),

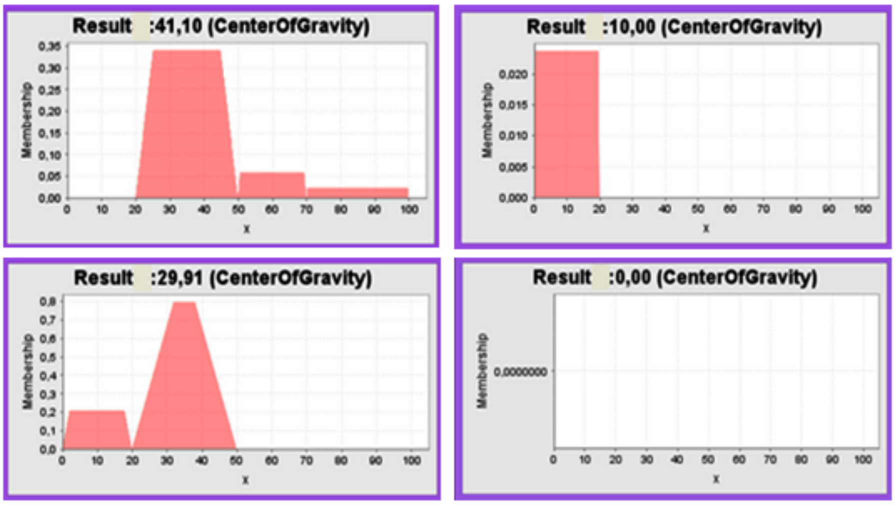


Fig. 4. Centers of gravity Profile animator, checker, seeker and independent

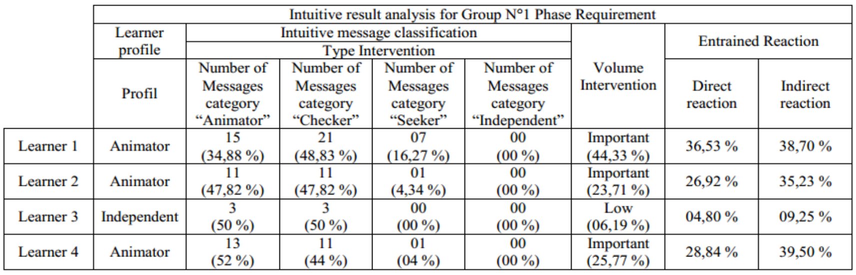
organized in 6 phases, corresponding to different tasks. Our corpus consists of 100 to 120 messages exchanged by the students within their same group for each phase of the project. This analysis of textual conversations is based on the characteristics of the behavioral profiles defined above (Table 1). Intuitive message analysis involves:

- Associating a profile with each message
- Identifying the language acts that determine the profile
- Associating a profile with each student

In the context of project-based pedagogy, for example, these observations provide the supervising teacher with indications to understand, react and intervene with the group. In the same way, for the learners, the perception of the behaviors of the individuals of the group makes it possible to better regulate the collective work.

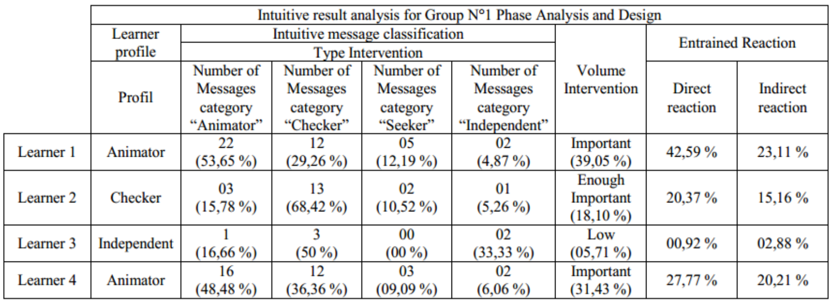
Figures 5 and 6 illustrate the results of the intuitive analysis done by the tutors during the first two phases of the project, whose objective is to associate a behavioral profile to each learner. For greater clarity, in this analysis, each tutor will associate a profile to each learner according to his social behavior. Indeed, after having identified the learning profiles, we asked the tutors to do (by analysis of the contents) a classification of the messages (type: animator, checker, seeker, independent) by identifying the acts of language which characterize them: proposition, message of organization and/or encouragement, intervention to calm a conflict, reaction to a proposal, expression of doubts about an approach or proposal, etc. (see Table 1). From the tutors analysis results for group 1 during periods 1 and 2 (Figure 5 and 6), the profiles of the learners, during each project period, are confirmed by the data collected by the student system through the analysis of message contents (Figure 7 and 8). For example, for group 1 during the requirement period, the learner 1 mainly held animator role that corresponds to the analysis of the data recorded by the tutor: high intervention volume (44.32%) and high level of intervention.

When we submit the same interaction data between learners to the automatic analysis system that we propose, we obtain the results shown in Figures 7 and 8, for the same group and for the same periods.



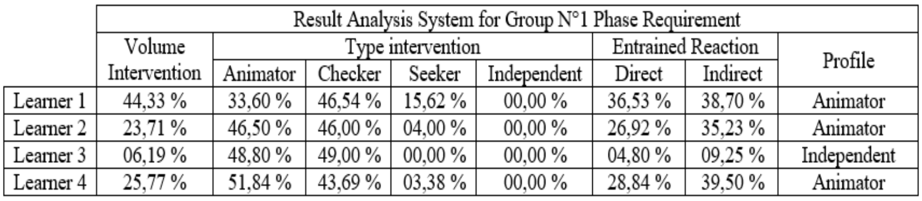
	Intuitive result analysis for Group N°1 Phase Requirement							
	Learner profile	Intuitive message classification				Volume Intervention	Entrained Reaction	
		Type Intervention						
	Profil	Number of Messages category “Animator”	Number of Messages category “Checker”	Number of Messages category “Seeker”	Number of Messages category “Independent”		Direct reaction	Indirect reaction
Learner 1	Animator	15 (34,88 %)	21 (48,83 %)	07 (16,27 %)	00 (00 %)	Important (44,33 %)	36,53 %	38,70 %
Learner 2	Animator	11 (47,82 %)	11 (47,82 %)	01 (4,34 %)	00 (00 %)	Important (23,71 %)	26,92 %	35,23 %
Learner 3	Independent	3 (50 %)	3 (50 %)	00 (00 %)	00 (00 %)	Low (06,19 %)	04,80 %	09,25 %
Learner 4	Animator	13 (52 %)	11 (44 %)	01 (04 %)	00 (00 %)	Important (25,77 %)	28,84 %	39,50 %

Fig. 5. Results of the intuitive analysis for group 1, phase 1 Requirement



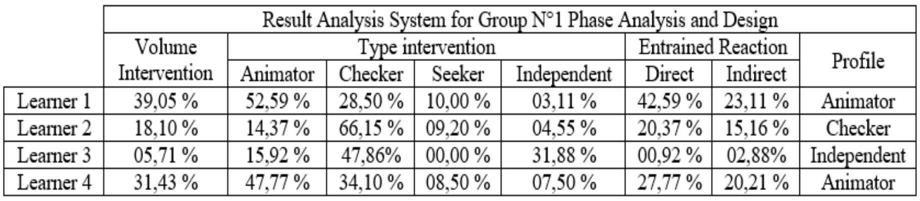
	Intuitive result analysis for Group N°1 Phase Analysis and Design							
	Learner profile	Intuitive message classification				Volume Intervention	Entrained Reaction	
		Type Intervention						
	Profil	Number of Messages category “Animator”	Number of Messages category “Checker”	Number of Messages category “Seeker”	Number of Messages category “Independent”		Direct reaction	Indirect reaction
Learner 1	Animator	22 (53,65 %)	12 (29,26 %)	05 (12,19 %)	02 (4,87 %)	Important (39,05 %)	42,59 %	23,11 %
Learner 2	Checker	03 (15,78 %)	13 (68,42 %)	02 (10,52 %)	01 (5,26 %)	Enough Important (18,10 %)	20,37 %	15,16 %
Learner 3	Independent	1 (16,66 %)	3 (50 %)	00 (00 %)	02 (33,33 %)	Low (05,71 %)	00,92 %	02,88 %
Learner 4	Animator	16 (48,48 %)	12 (36,36 %)	03 (09,09 %)	02 (6,06 %)	Important (31,43 %)	27,77 %	20,21 %

Fig. 6. Results of the intuitive analysis for group 1, Phase 2 Analysis and Design



	Result Analysis System for Group N°1 Phase Requirement							
	Volume Intervention	Type intervention				Entrained Reaction		Profile
		Animator	Checker	Seeker	Independent	Direct	Indirect	
Learner 1	44,33 %	33,60 %	46,54 %	15,62 %	00,00 %	36,53 %	38,70 %	Animator
Learner 2	23,71 %	46,50 %	46,00 %	04,00 %	00,00 %	26,92 %	35,23 %	Animator
Learner 3	06,19 %	48,80 %	49,00 %	00,00 %	00,00 %	04,80 %	09,25 %	Independent
Learner 4	25,77 %	51,84 %	43,69 %	03,38 %	00,00 %	28,84 %	39,50 %	Animator

Fig. 7. System-Calculated Indicators for Group 1 Phase 1 of the Project



	Result Analysis System for Group N°1 Phase Analysis and Design							
	Volume Intervention	Type intervention				Entrained Reaction		Profile
		Animator	Checker	Seeker	Independent	Direct	Indirect	
Learner 1	39,05 %	52,59 %	28,50 %	10,00 %	03,11 %	42,59 %	23,11 %	Animator
Learner 2	18,10 %	14,37 %	66,15 %	09,20 %	04,55 %	20,37 %	15,16 %	Checker
Learner 3	05,71 %	15,92 %	47,86%	00,00 %	31,88 %	00,92 %	02,88%	Independent
Learner 4	31,43 %	47,77 %	34,10 %	08,50 %	07,50 %	27,77 %	20,21 %	Animator

Fig. 8. System-Calculated Indicators for Group 1 Phase 2 of the Project

The analysis of these results in the light of the characteristics of the defined learner profiles (Table 1) allows associating a sociological profile to each learner. Seen the results of the semantic analysis to calculate the type of intervention, two profiles emerge: animator and checker. However, by analyzing the resulting volumes of interventions and associated reactions, we find that they are important and characterize the Animator profile. Thus, for the example considered and for period 1, the learner (1) will be qualified as "Animator". The same approach was used to define the learning profiles in periods 1 and 2 for students in group 1 (Figures 5 and 6).

The idea is to compare the types of intervention which are qualitative. The results are presented in brief in table 2. They illustrate the results of the error margin calculation between the intuitive analysis (by tutor) and the analysis performed by our system.

We have considered the result of intuitive analysis tutors as a reference. All these results are averaged for every learner and profile to produce a single indicator called “Total results”. This later allowed us to verify that the system has an error rate of 2.95 % compared with the intuitive analysis.

Table 2. Comparison between analysis system and intuitive analysis.

	Comparison between analysis system and intuitive analysis.				**Result**
	Animator	***Checker***	***Seeker***	***Independent***	
Learner 1	96,33 %	95,31%	96,00 %	100 %	96,91 %
Learner 2	97,32 %	96,19 %	92,37 %	100 %	96,47 %
Learner 3	97,60 %	98,00 %	100 %	100 %	98,90 %
Learner 4	99,96 %	99,29 %	84,50 %	100 %	95,93 %
Total results :					**97,05 %**

Through this study, we have been able to appreciate the usefulness of the notion of semantic analysis of conversations and fuzzy logic in the evaluation of the learner's behavioral levels.

Fuzzy subset theory provides an appropriate method for incorporating the knowledge of an expert teacher by using qualitative terms that are close to human reasoning. It allows to manipulate inaccurate information and to model subjective knowledge. In addition, the use of fuzzy rules in the system inference algorithm provides the user with greater flexibility and ease of judgment. In addition, we have shown the evolution over time of the profiles (Animator, checker, seeker and independent) of a learner (1) present on the social media discussion groups (Figure 9).

For example, the graph indicates that the learner (1) played a leading role during the first phase of the project. On the other hand, we can notice that this learner became checker at the second phase.

This view makes it possible to identify the role played by the learners in their group through the different phases of the project. This variation is the result of the learners' preference for tasks related to each phase of the project.

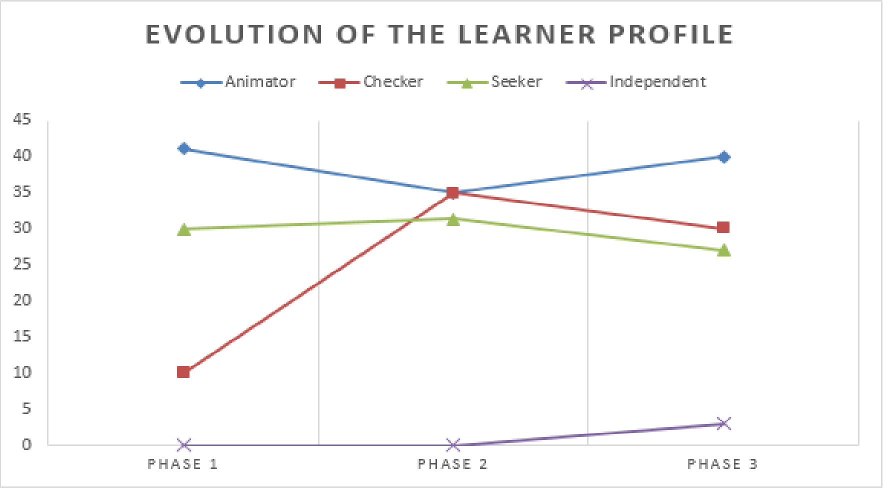


Fig. 9. Evolution of the learner profile (1) over the entire duration of the project

In view of these results, we find that the observations made by the tutors are confirmed by the automatic analysis made by our system, which leads us to affirm that our approach makes it possible to find the behavioral profiles in the groups of learners working remote.

5 Conclusion

The present work is part of the development of automatic interaction analysis systems, widely used to meet the constraints faced by remote tutoring via network technologies.

We propose a complete procedure of "full text" analysis of textual exchanges for the determination of sociological profiles of learners within the framework of collaborative distance learning processes. This analysis consists of 2 steps (Recovery and Filtering) at the end of which, we perform a semantic analysis of conversations that will subsequently contribute to the classification of messages (type animator, checker, seeker and independent). Based on the profiles defined and adapted from the work of Pléty [2], we compute indicators, which, coupled with the message classifications described above, make it possible to assign a sociological profile to each learner on the basis of fuzzy logic. The approach was tested on a real situation, which showed a great concordance between the results observed by human tutors and those automatically determined by our system.

As a development perspective for this project, we plan to integrate a recommendation system. System that generates automatically some recommendations that are suitable for every learner.

6 Acknowledgement

The authors also wish to express their gratitude to the students of the master "software quality" of Ibn Tofail University for helping in carrying out the evaluation part of this work in real-life situation.

7 References

[1] Koschmann, T. (2012). *CSCL:* Theory and practice of an emerging paradigm. Routledge. https://doi.org/10.4324/9780203052747

[2] Sujana, J., Claire, M., & John, K. (2012). A visualization tool to aid exploration of students' interactions in asynchronous online communication. Computers & Education, Volume 58, Issue 1, January 2012, (pp. 30-42).

[3] Mbala A., Reffay C., Chanier T., «SIGFAD : un système multi-agents pour soutenir les utilisateurs en formation à distance». In Actes de la conférence Environnements Informatiques pour l'Apprentissage Humain (EIAH'2003), Strasbourg, France, 2003.

[4] Pléty, R. (1998). Comment apprendre et se former en groupe (Retz, 1998).

[5] Chaabi.Y, al. (2014), «An automatic system for the determination of learner's sociological behavior from textual asynchronous conversations analysis in online collaborative learning". International Conference on Intelligent Systems: Theories and Applications (SITA-14) 7-8 May 2014, IEEE.

[6] Chaabi, Y., Khadija, L., Jebbor, F., & Messoussi, R. (2017). Determination of Distant Learner's Sociological Profile Based on Fuzzy Logic and Naïve Bayes Techniques. International Journal of Emerging Technologies in Learning (iJET), 12(10), 56-75. https://doi.org/10.3991/ijet.v12i10.6727

[7] Zadeh, L. A. (1996). Fuzzy logic= computing with words. IEEE transactions on fuzzy systems, 4(2), 103-111. https://doi.org/10.1109/91.493904

[8] Klir, G., & Yuan, B. (1995). Fuzzy sets and fuzzy logic (Vol. 4). New Jersey: Prentice hall.

[9] Self, J. A. (1988). Bypassing the Intractable Problem of Student Modelling. Proceedings of the 1st International Conference on Intelligent Tutoring System, Montréal, Canada, (june 1-3), 18-24.

[10] BERRI, J., Cartier, E., Desclés, J. P., JACKIEWICZ, A., & Minel, J. L. (1996). A linguistic method for text filtering. *Revista de Procesamiento del lenguaje natural*, *19*, 159-165.

[11] Wu, Z., & Palmer, M. (1994, June). Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics (pp. 133-138). Association for Computational Linguistics. https://doi.org/10.3115/981732.981751

[12] Zhu, G., & Iglesias, C. A. (2017). Computing Semantic Similarity of Concepts in Knowledge Graphs. IEEE Transactions on Knowledge and Data Engineering, 29(1), 72-85. https://doi.org/10.1109/TKDE.2016.2610428

[13] Hua, W., Wang, Z., Wang, H., Zheng, K., & Zhou, X. (2017). Understand Short Texts by Harvesting and Analyzing Semantic Knowledge. IEEE transactions on Knowledge and data Engineering, 29(3), 499-512. https://doi.org/10.1109/TKDE.2016.2571687

[14] Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008.

[15] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, pages 448–453,1995.

[16] Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41. https://doi.org/10.1145/219717.219748

[17] Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41. https://doi.org/10.1145/219717.219748

[18] Miller, G. (1998). *WordNet: An electronic lexical database*. MIT press.

[19] Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. IEEE transactions on systems, man, and cybernetics, 19(1), 17-30. https://doi.org/10.1109/21.24528

[20] Meštrović, A., & Calì, A. (2016, September). An ontology-based approach to information retrieval. In *Semanitic Keyword-based Search on Structured Data Sources* (pp. 150-156). Springer, Cham.

[21] Lin, D. (1998, August). Automatic retrieval and clustering of similar words. In Proceedings of the 17th international conference on Computational linguistics-Volume 2 (pp. 768-774). Association for Computational Linguistics.

[22] Z.Wu and M.Palmer.Verb semantic and Lexical Selection .In Proceddings of 32 Annual Meeting of the association of computer Linguistics (ACL 994), Las Cruces,New Mexico,1994. https://doi.org/10.3115/981732.981751

[23] Navarro-Almanza, R., Licea, G., Juárez-Ramírez, R., & Mendoza, O. (2017, April). Semantic Capture Analysis in Word Embedding Vectors Using Convolutional Neural Network. In World Conference on Information Systems and Technologies (pp. 106-114). Springer, Cham.

8 Authors

Dr. Youness Chaabi is an AR in Center for Informatics Studies, Information Systems and Communication, Royal Institute of the Amazigh Culture (IRCAM), Rabat, Morocco. Main skills being Machine Learning, E-Learning & Coloborative Learning, Software Development more in Java Programming, Classifications, SQL and many more but lastly but not the least Web development. Total achievements are 13 research items and 15 Citations. Chaabi@ircam.ma

Lekdioui Khadija Is with Telecommunications Systems & Decision Engineering Laboratories (LASTID), Ibn Tofail University, Kenitra, Morocco. Lekdioui_khadija@hotmail.com

Boumediane Mounia Professor**,** CEISIC, Royal Institute of the Amazigh Culture (IRCAM), Rabat, Morocco & Faculty of Letters and Human Sciences Mohammed V University, Rabat, Morocco. Boumediane.mounia@gmail.com boumediane@ircam.ma

Article submitted 2018-05-10. Resubmitted 2018-09-13. Final acceptance 2019-01-14. Final version published as submitted by the authors.