Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

An Interval Type-2 Fuzzy Ontological Similarity Measure

Naeemeh Adel[1], Student Member, IEEE, Keeley Crockett[1], Senior Member, IEEE, Daria Livesey[1], Joao Paulo Carvalho[2], Member, IEEE

[1]Department of Computing and Mathematics, Manchester Metropolitan University, Chester Street, Manchester, M1 5GD, UK
[2]INESC-ID, Instituto Superior Tecnico, Universidade de Lisboa, Portugal

Corresponding author: Naeemeh Adel (e-mail: N.Adel@mmu.ac.uk).

ABSTRACT Human language is naturally fuzzy by nature, with words meaning different things to different people, depending on the context. Fuzzy words are words with a subjective meaning, which are typically used in everyday human natural language dialogue and are often ambiguous and vague in meaning and are based on an individual's perception. Fuzzy Sentence Similarity Measures (FSSM) are algorithms that can compare two or more short texts which contain human-perception-based words and return a numeric measure of similarity of meaning between them. This paper proposes a new FSSM called FUSE (FUzzy Similarity mEasure), to assess an individual's perception within a FSSM. FUSE is an ontology-based similarity measure that uses Interval Type-2 fuzzy sets to model relationships between categories of human perception-based words. The FUSE algorithm has been developed over four versions and evaluated on several published and newly created datasets. Typically, results have shown that calculating the semantic similarity of two short texts using FUSE, gives a higher correlation with the average human ratings (AHR) compared to traditional sentence similarity measures that do not consider the presence of fuzzy words. This paper focuses on the second version of the FUSE algorithm, referred to as FUSE_2.0 which has been compared to several state-of-the-art, semantic similarity measures (SSM), including the only published FSSM, FAST (Fuzzy Algorithm for Similarity Testing), which has a limited dictionary of fuzzy words and uses Type-1 to model relationships between categories of human perception-based words. Results have shown that FUSE _2.0 achieves a higher correlation with the average human ratings (AHR) compared to traditional SSM's and FAST. The key contributions of this work can be summarised as follows: The development of a new methodology to model fuzzy words using Interval Type-2 fuzzy sets. This has led to the creation of a fuzzy dictionary for nine fuzzy categories, a useful resource which can be used by other researchers in the field of natural language processing and Computing with Words (CWW) with other fuzzy applications such as semantic clustering.

INDEX TERMS computing with words, natural language processing, FUSE, semantic similarity, fuzzy sets, machine learning, computational intelligence, fuzzy logic

I. INTRODUCTION

When one thinks of Artificial Intelligence (AI), most think about automating tasks and routines. But advances in technology mean AI is now more than just the automation of tasks. With the introduction of Natural Language Processing (NLP) [1] it is now possible to generate text and create interaction between humans and machines. However, there are significant challenges associated with the automatic interpretation and understanding of the human language by machines as they lack contextual awareness. This makes it difficult for machines to understand and interpret human language easily. The motivation behind this work is to investigate human perceptions of subjective words (fuzzy words) used in everyday language that may have different meanings when used in different contexts (for example, in the phrase, *I feel hot*, how do we define the measure for the word *hot*, as it is subjective to each individual). Whilst devices such as Alexa [2] have a good natural language coverage of basic commands, they are limited to sets of instructions identified by sequences of keywords. Such devices are not currently capable of dealing with emotions [3], or of understanding the impact of subjective words within the instruction. For example, consider the following two instructions: (1) "*Alexa - turn the heating up*"; (2) "*Alexa - I'm very cold – turn the* *heating up a little*". In (1), this instruction could lead to the heating being turned up a pre-programmed amount. In (2), understanding of the words *very cold* in the context of the current temperature in the room, may invoke a higher temperature increase. This simple example is focusing on the "devices" understanding the similarity of human utterances and sentences in the English language. According to the Oxford English Dictionary, a sentence is "*A set of words that is complete in itself, typically containing a subject and predicate, conveying a statement, question, exclamation, or command, and consisting of a main clause and sometimes one or more subordinate clauses*"[4]. When a human formulates a sentence, the sentence tends to be made of several verbs, adverbs, adjectives, and nouns, etc. Some of these words will have clearly defined meanings, for example, in English, words such as (*I, Tree, Cat, Sit*); however, words such as (*Hot, Cold, Fast, Young*) do not have a fixed meaning and can vary from human to human depending on the perspective and perception of that person and the context in which they are used. They are subjective and essentially *fuzzy*. In this work, a fuzzy word is defined as "*A word that has a subjective meaning and is characteristically used in everyday human natural language dialogue. Fuzzy words are often ambiguous in meaning since they are based on an individual's perception*" [5]. The challenge and motivation of this work is to have machines (i.e. devices such as Alexa) understand the meaning behind these fuzzy words in a given situation. A human formulates context using other sources such as sight and sound and together they formulate the context of the spoken word. A machine, however, only has the letters and words that are spoken or typed in a specific sequence, with which to infer deeper meaning.

One way for machines to understand context is through the application of semantic similarity measures. Such measures allow comparisons between natural language short texts. Semantic similarity refers to similarity between two concepts in taxonomies such as WordNet [6] or CYC upper ontology [7]. Such measures have been used in many applications from plagiarism detection to information retrieval [8], word sense disambiguation [9], image retrieval [10], multimodal document retrieval [11] and automatic hypertext linking [12]. Traditionally, semantic similarity measures were defined using either ontological, knowledge-based approaches [8], corpus-based methods [8] and more recently deep learning based [13]. One established sentence similarity measure, known as STASIS [14] was first published by Li et. al. in 2006; STASIS is used as the fundamental basis for the work proposed in this paper. STASIS uses a semantic-vector approach [15] which combines word similarity by using WordNet (a lexical database of English [6]). WordNet is used to compute the path lengths between each word. This, combined with the formulation of short word vectors and joint word sets is used to compute the semantic similarity between two short texts. A weakness of STASIS is that it cannot calculate the similarity between fuzzy words in a short text. To address this problem, fuzzy semantic similarity measures (FSSM) were first investigated by Chandran et. al. [16] who proposed a Fuzzy Algorithm for Similarity Testing (FAST). FAST is an ontology-based similarity measure that uses concepts of fuzzy theory [17] to allow for a truer representation of fuzzy based words. Through human experimentation, fuzzy sets were created for six categories of words based on their levels of association with concepts using Type-1 fuzzy sets. These fuzzy sets were then defuzzified and the results used to create new ontological relations between the words. The disadvantage of FAST was its use of Type-1 fuzzy sets, as Type-1 was found not to be able to represent human uncertainty [18]. Additionally, FAST had a very limited collection of fuzzy words [16] resulting in poor coverage of the English language.

This paper is built on work presented in [5], which first introduced the concept algorithm known as FUzzy Similarity mEasure (FUSE), which is an ontological similarity measure that uses Interval Type-2 fuzzy sets to model human perception-based words [5]. FUSE uses its own ontological fuzzy dictionary which has been created following several human experiments and using inputs from English language experts and makes use of the Hao-Mendel Approach (HMA) [19] Interval Type-2 defuzzification method. The original FUSE_1.0 algorithm [5], consisted of six fuzzy categories (Size/Distance, Temperature, Age, Frequency, Worth, Level of Membership). In the work presented in [5], several experiments were conducted on three datasets, two of which were gold standard datasets and the third was approved by English language experts; The two gold standard datasets contained non-fuzzy words and the third dataset contained one or more fuzzy words. The experiments compared the correlation of each dataset with human ratings of FUSE_1.0, its predecessor FAST and STASIS. Results showed that FUSE_1.0 gave a better correlation compared to human ratings than FAST or STASIS on human utterances. The improvement FUSE_1.0 had over STASIS and FAST for the three datasets tested was due to the increased coverage of fuzzy words and the use of the new fuzzy ontology used in FUSE_1.0. Furthermore, using Interval Type-2, as opposed to Type-1 has been shown to contribute towards a higher correlation [5]. However, one of the weaknesses of FUSE_1.0 was the limitations of the fuzzy words in the six categories. To overcome this weakness FUSE_1.0 was expanded to include three new fuzzy categories.

The novel contribution presented in this paper describes the expansion of FUSE_2.0 and the addition of three more fuzzy categories (Strength, Brightness, Speed). The

FUSE_2.0 algorithm comprises of nine fuzzy categories which have been used to formulate a Fuzzy Dictionary that contains a total of 386 fuzzy words that can be found in Appendix A of this paper. Each fuzzy word has a defuzzified value ranging from [-1, 1] that have been derived following extensive human experiments designed with assistance from English language experts. In this work, the new FUSE_2.0 algorithm was evaluated using five datasets against human ratings, two of which were gold standard, and compared with the results generated from two research and one commercially available similarity algorithm. Each of the SSM's selected for comparison do not cater for the presence of fuzzy words in sentences or utterances. Results presented in Section VIII have shown that FUSE_2.0 gives a higher similarity rating in comparison with human ratings across the five datasets.

II. RESEARCH QUESTIONS & CONTRIBUTIONS

This paper builds on initial work in [5] by producing an enhanced version of FUSE_1.0 which can compare fuzzy utterances with an increased fuzzy dictionary of nine categories to cover a larger scale of fuzzy words. The research presented in this paper aims to answer the following research question:

Can Type-2 fuzzy sets be used to represent an individual's perception within a fuzzy semantic similarity-based measure?

The main novel contributions of this paper are:

- The FUSE_2.0 algorithm for determining the semantic and syntactic similarity of short texts. FUSE_2.0 is generalizable and can be used successfully (high correlation with human ratings) with short texts that contain fuzzy and non-fuzzy words;
- A methodology for human similarity modelling of fuzzy words using Interval Type-2 fuzzy sets;
- Evaluation of the FUSE_2.0 measure on two gold standard datasets [20] and three fuzzy datasets;
- A fuzzy dictionary containing defuzzified numerical values derived from the use of Interval Type-2 fuzzy sets to model human perception-based words, which can help researchers in the field of computing with words, numericalise fuzzy words in the context of natural language.

III. PAPER OVERVIEW

This paper is organised as follows:

Section IV starts with an overview of the different semantic similarity measures such as STASIS [14], Dandelion [21] and SEMILAR [22] and discusses the uncertainty of natural language. Section V describes the design of an Interval Type-2 fuzzy ontological similarity measure known as FUSE before moving onto Section VI which describes the evolution of FUSE, from FUSE_1.0 which had six fuzzy categories to FUSE_2.0 which has nine fuzzy categories. Section VI goes on to describe a series of experiments relating to capturing human-based perception words in the suggested nine categories for FUSE_2.0, followed by modelling of the fuzzy words using Interval Type-2 (IT2) Fuzzy Sets (FS) to produce a fuzzy dictionary for each of the mentioned nine categories. The methodology of the experiment is described in Section VII. Section VIII evaluates the results of FUSE_2.0 in comparison with four other SSM's conducted on five datasets to measure the correlation with human ratings. Finally, the conclusions and further work of this paper are presented in Section IX.

This work has received full ethical approval from Manchester Metropolitan Universities Science and Engineering Research Ethics and Governance Committee (EthOS Reference Number: 11759).

IV. SEMANTIC SIMILARITY & UNCERTAINTY OF NATURAL LANGUAGE

A. OVERVIEW

Semantic similarity is an important and fundamental concept in AI and many other fields and refers to the similarity of two concepts in a taxonomy. Examples include word sense disambiguation [23], image retrieval [24], multimodal document retrieval [25] and automatic hypertext linking [26]. The concept of word similarity has been a part of natural language processing for many years. Similarity between words is usually influenced by the context in which those words appear in. An example of this could be the context "*the outside covering of living objects*", this would mean that the words *skin* and *bark* would be more similar in meaning, than the words *skin* and *hair* [27]. However, the larger the number of words in a sentence, the more complex this will become. For example, given the two sentences S1 and S2 below:

S1: A small fish in a big pond

S2: A big fish in a small pond

The two sentences above contain the same words in each sentence with the only difference being the order in which they are presented. It is clear to a human interpreter that these two sentences vary in meaning, due to the order of the words. Thus, any effective sentence similarity algorithm must take into account word order as this will impact both the sentence meaning and the overall similarity rating.

B. BACKGROUND

This section provides a brief overview of sentence similarity measures (SSM) including the three main categories: ontological [7], knowledge-based approaches [7], corpus-based methods [7] and more recently a fourth category deep learning based [13]. Latent Semantic Analysis (LSA) [28] is a mathematical method for modelling words and paragraphs in order to understand natural language texts. The method is based on a corpus-based approach and calculates similarity between two paragraphs of text. To apply LSA to a domain, a large corpus [29] is required. A limitation of LSA is that it does not take into consideration word order and scholars argue that it is not grounded in human perception and intention [29]. STASIS is also a corpus based SSM, which measures the level of similarity between two utterances using an ontological approach based on a taxonomy of words [30]. STASIS calculates the distance between words in an ontology, using WordNet [6], as well as the distance of words to their closest subsumer. This algorithm was tested against two gold-standard datasets STSS-65 and STSS-131 and results showed a high correlation with human results [20]. Dandelion is a short sentence similarity measure which compares the semantic and syntactic similarity between two sentences and shows these results separately [21]. It uses a knowledge-based approach for short sentences between 5-20 words giving a rating of the similarity between the two sentences. It currently supports seven languages (English, Italian, French, German, Portuguese, Spanish and Russian) [21]. Some examples of where this algorithm have been used in research include webpage ranking [31] and automated assessment of short, structured questions [32]. One final example of a sentence similarity algorithm is SEMILAR (the SEMantic simILARity toolkit) [22]. SEMILAR is a corpus-based similarity measure which uses the word-to-word semantic similarity measures in the WordNet Similarity library [33] as well as using Latent Semantic Analysis (LSA) [28]. SEMILAR uses two annotation protocols: greedy and optimal annotation. The greedy method pairs a target word in one sentence with all the words in the other sentence and retains the matching word with the highest word-to-word similarity score to the target word regardless of how other words match each other. The optimal matching strategy is inspired from optimal matching methods proposed for tasks where a set of items must be matched against another set, while optimizing the overall matching score and not individual scores [22]. While in greedy matching the goal is for a target word to find a best matching word in the opposite sentence, in optimal matching the goal is to match items such that an overall optimal matching is achieved [34]. SEMILAR was tested on several datasets to help with paraphrasing, entailment, and elaboration [33]. For the purpose of comparison with FUSE_2.0, experiments presented in this paper utilised the greedy method in SEMILAR. Where there have been significant improvements in the development of SSM [35], the above-mentioned algorithms were not designed to capture human perception-based words within short texts through relation to the context in which they were used, therefore comparing their performance with FUSE_2.0 will build a better picture, as to why a FSSM is needed to cater for the uncertainty of fuzzy words in a sentence or utterance. The uncertainty that lies within perception-based words makes them difficult for machines to measure using standard SSM algorithms since words mean different things to different people as stated by Mendel [18]; therefore, a FSSM is needed to cater for the uncertainty of fuzzy words in a sentence or utterance.

C. CHALLENGES OF GATHERING HUMAN RATINGS

Typically, the only way to evaluate sentence similarity measures (SSM) is through using human subjective opinions. This is a resource intensive process. O'Shea developed a methodology where sentence pairs where taken [20] and 64 participants were asked to assess their similarity on a scale of [0-4]. This method has since become a gold standard for evaluation and two gold standard datasets were produced because of this research, STSS-65 and STSS-131 [20]. Each dataset contains 65 and 131 sentence pairs respectively. There are certain challenges that arise when creating a dataset which is to be used by an SSM. The first challenge is to find the correct domain to represent, in this instance, datasets containing short sentences. There is then the challenge of collecting valid human ratings for similarity between these sentence pairs. The research presented in this paper focused on short text sentence pairs, and human ratings were collected from native English speakers in the Northwest region of UK, to ensure that regional dialect did not interfere with the ratings and words did not have too vague of a meaning and reduce similarity, thus resulting in the distorting of results. The final challenge is to know what statistical measure to use when measuring the similarity, which, in this instance is the average human ratings (AHR). The Pearson's correlation coefficient [36] is a long-established measure of agreement used in semantic similarity that assumes a linear relationship between the two variables being compared - machine generated similarity and average human ratings across a sample size of at least 32 participants [20]. Pearson's correlation coefficient will be applied as the statistical measure in the research presented in this paper.

D. COMPUTING WITH WORDS

Zadeh first introduced the term Computing with Words (CWW) in 1996 [37]. CWW models words using fuzzy sets and is used when information is not precise enough to use numbers. This is often the case when applications involving humans are used, as humans tend to deal better with words than they do with numbers [38]. As explained by Zadeh [37], in crisp set theory, an object is either completely in a set, shown with the degree membership of 1, or completely outside the set, shown with the degree membership of 0. In fuzzy theory however, the membership degree is a range between [0-1]. Originally fuzzy sets were modelled using Type-1 fuzzy sets (FS). However, as Mendel explains in [39], using a Type-1 FS model for a fuzzy word is an incorrect scientific theory which follows from the following line of reasoning:

(i) A Type-1 fuzzy set (A) for a word is defined by its membership function $\mu A(x)(x \in X)$, where $\mu A(x)$ is the membership function for the fuzzy set A. X is referred to as the universe of discourse. The membership function associates each element ($x \in X$), with a value in the interval [0,1] that is totally certain once all of its parameters are specified;

(ii) Words mean different things to different people, and thus are uncertain;

(iii) It is a contradiction to say that something certain can model something that is uncertain.

Type-1 fuzzy sets are not able to directly model such uncertainties because their membership functions are totally crisp [40]. On the other hand, Type-2 fuzzy sets can model such uncertainties because their membership functions are themselves fuzzy. Membership functions of Type-1 fuzzy sets are two-dimensional, whereas membership functions of Type-2 fuzzy sets are three-dimensional, which in turn makes them more computationally difficult to draw and understand and so to help with this difficulty, Interval Type-2 (IT2) FS were created [40]. CWW uses linguistic uncertainty and so fuzzy sets are needed to model words. Mendel suggests using IT2 fuzzy sets to model these uncertainties using the Footprint of Uncertainty (FOU) [41]. IT2 fuzzy sets are the most widely used Type-2 fuzzy sets because they are simple to use and because, at present, it is very difficult to justify the use of any other kind for modelling fuzzy words [40]. When the Type-2 fuzzy sets are modelled as IT2 fuzzy sets, all secondary membership grades are equal to 1. In this case, embedded Type-2 fuzzy sets can be treated as embedded Type-1 fuzzy sets so that no new concepts are needed to derive the union, intersection, and complement of such sets [40]. After each derivation, interval secondary grades were merely appended to all the results in order to obtain the final formulas for the union, intersection, and complement of Interval Type-2 fuzzy sets [40].

V. FUSE - AN INTERVAL TYPE-2 FUZZY ONTOLOGICAL SIMILARITY MEASURE

This section describes the design of an Interval Type-2 (IT2) Fuzzy Ontological Similarity Measure known as FUSE. FUSE is a sentence similarity algorithm that takes two short sentences or utterances in English and uses Interval Type-2 modelling and a fuzzy ontology to calculate the semantic and syntactic similarity between the two sentences. The evolution of the FUSE algorithm from FUSE_1.0 to FUSE_2.0 has resulted in the creation of a fuzzy dictionary [Appendix A] containing 386 fuzzy words, that are broken down into nine separate fuzzy categories which is part of the novel contribution of this paper. To create the FUSE algorithm, first human perception-based words were modelled using IT2 FS. This required a series of experiments which enabled human participants to rate the similarity of words within the nine categories. Details of the experimental methodology are fully explained in [5]. Capturing the ratings from human participants on predefined categories of fuzzy words allowed the building of fuzzy category ontologies which are required to measure the distance between fuzzy words in the FUSE algorithm.

Fig. 1 shows a component diagram for the FUSE algorithm. It shows how two natural language utterances, *U1* and *U2* are fed into the FUSE algorithm and the steps involved in computing the overall sentence similarity rating. The FUSE algorithm will be formally defined in Section VII B.

VI. HUMAN MODELLING OF FUZZY WORDS

This section describes a series of experiments designed to

a) Capture human based perception words in nine categories;
b) Model the words using IT2 FS;
c) Produce a Fuzzy Dictionary;
d) Develop ontologies for nine fuzzy categories.

A. CAPTURING HUMAN BASED PERCEPTION WORDS

As explained in [5] the coverage of words in FAST was very limited with only 196 words in total for all six categories. Thus, the first stage in the development of the FUSE algorithm was to expand the words in the six original categories (Size/Distance, Temperature, Age, Frequency, Worth, Level of Membership) used in FAST. This expansion is referred to as FUSE_1.0. To do this, the Oxford English Dictionary [4] was used and all one-word synonyms for the existing words in the six categories were collected. This initial process increased the total number of words in the six categories to 309 words, giving a 60.07% increase over FAST and its existing fuzzy words. 32 native English participants from the Northwest region of UK were then used to rate each of the words in the six categories on a scale of [0,10]. FUSE_2.0 expanded the existing six categories to nine fuzzy categories. Three new categories (Strength, Speed, Brightness) were added to the existing fuzzy dictionary. These categories were based on Zadeh's [42] theory of perception-based words, where he states that measurements are crisp numbers whereas perceptions are fuzzy numbers. The Oxford English Dictionary [4] was again used to collect all one-word synonyms for each category and human experiments were conducted in two stages, the first stage was to see which words should stay in the three new categories as chosen by human participants and the second stage to determine what the

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3194510

ratings should be per word according to human participants. To carry out the first stage, human participants were asked to take part by being given the words in each proposed category. They were then asked to cross out any word they felt did not belong to that chosen category. Fig. 2 shows a snippet of one of the participants answers for the category *Brightness*. Each participant was asked to do this for all words in each of the three new proposed categories. A total of 17 participants successfully took part in the experiment. Although O'Shea [20] claimed that 32 participants is a significant number for participants, other studies have shown variations in the number of participants versus the number of words/sentences that the participants were asked to rate similarity of [20]. To filter out the results, two English language experts were consulted, and it was agreed that a threshold of 70% was set as a minimal acceptance rate for a word being kept in the chosen category. Any word that fell above the acceptance rate of 70% by all the participants was kept in the chosen category as a measure of quality control from this experiment. Table I shows the results of this experiment, the first column shows the category labels, the second column represents the original number of words that were collected per category using the Oxford English Dictionary [4], and the final column shows the number of words that were kept in each category as a result of the quality control following the experiment and applying the 70% threshold.

B. MODELING WORDS USING INTERVAL TYPE-2 FUZZY SETS

FUSE requires all words to be modelled using IT2 fuzzy sets. Following on from the experiment described in Section VI A, all words in all nine categories were modelled using this approach. The modelling of words is fully explained in [5] and the same method was also used to model the words in the three new proposed categories allowing all the fuzzy words in each category to be represented on a normalized scale of [-1, +1] to stay consistent with the other six categories present in the fuzzy dictionary.

C. PRODUCTION OF A FUZZY DICTIONARY

The IT2 modelling allowed the creation of the fuzzy dictionary for all nine categories. Table II shows the total number of fuzzy words present in each of the nine categories used for FUSE_2.0. Each category holds words that have a defuzzified value on a scale of [-1, +1] which is obtained using the IT2 FS model. The fuzzy dictionary which comprises of a full list of the words with their defuzzified values for each of the nine categories is available in Appendix A.

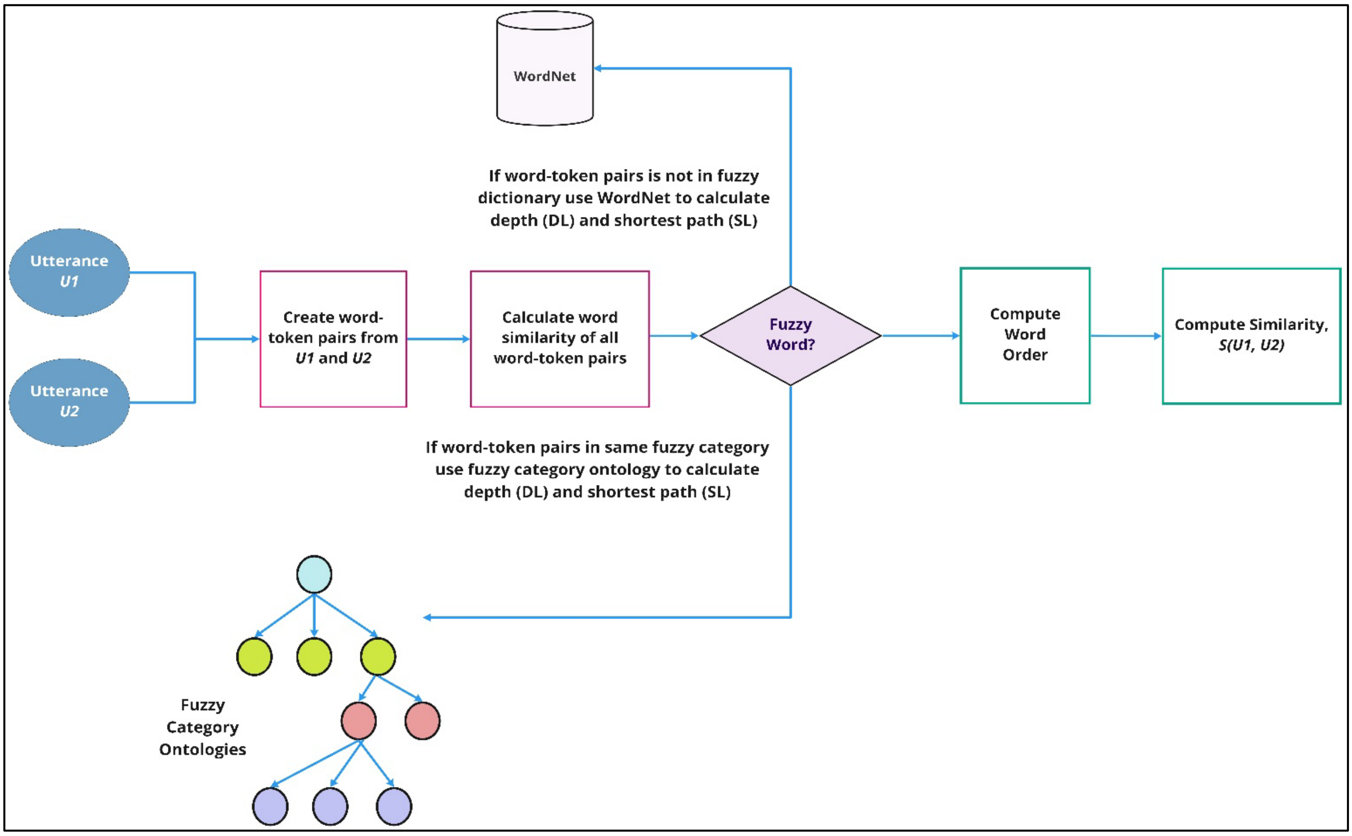
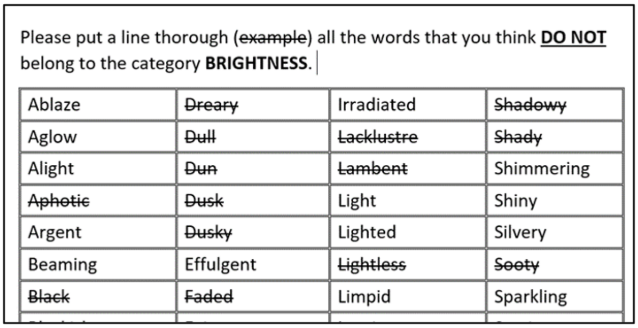


FIGURE 1. Component Diagram for FUSE Algorithm.


Please put a line thorough (~~example~~) all the words that you think **DO NOT** belong to the category **BRIGHTNESS.**

Ablaze	~~Dreary~~	Irradiated	~~Shadowy~~
Aglow	~~Dull~~	~~Lacklustre~~	~~Shady~~
Alight	~~Dun~~	~~Lambent~~	Shimmering
~~Aphotic~~	~~Dusk~~	Light	Shiny
Argent	~~Dusky~~	Lighted	Silvery
Beaming	Effulgent	~~Lightless~~	~~Sooty~~
~~Black~~	~~Faded~~	Limpid	Sparkling

FIGURE 2. Partial Participant Answer Sheet for Brightness Category.

TABLE I
FUZZY WORD FOR THREE NEW CATEGORIES

Categories \ No. of Words	Original No. of Words	Kept No. of Words
Brightness	107	27
Strength	109	24
Speed	81	26

D. DEVELOPMENT OF FUZZY ONTOLOGIES

To utilize the fuzzy dictionary in FUSE_2.0, fuzzy ontologies had to be created for each category. Each category is treated as a concept. Words within each concept are treated as instances. Each concept has a taxonomy that arranges the words as a binary tree so that the root node always takes the value 0. The defuzzified value of words are equally placed into nodes in intervals of ± 0.2, which was an empirically determined threshold. This approach allows calculation of the path length and depth of the Lowest Common Subsumer (LCS) to be calculated for fuzzy words in a category which could not be done using traditional resources such as WordNet, due to lack of coverage of fuzzy words [5] (see Section VII B). Fig. 3, shows the words in the category '*Speed*' represented in an ontological structure. The numbers next to each word represent the defuzzified value on a scale of [-1, 1] of that word, obtained from the human rating experiment and modelled using the IT2 FS approach described in Section VII. Each partition contains words up to a certain fixed value, with the negative values on one side and the positive values on the other, which allows path length to be calculated. The full methodology to develop the fuzzy ontologies can be found in [5]. When calculating the similarity between two sentences, if a word is present in the fuzzy dictionary, then the defuzzified value for that word will be used, granted that the words per sentence pair belong to the same fuzzy category. If this is not the case, then it will use WordNet [6] to obtain path length and depth. The fuzzy ontologies are used to derive the semantic and syntactic

TABLE II
NO. OF WORDS PER FUZZY CATEGORY

Categories	No. of Words
Size/Distance	91
Temperature	36
Age	42
Frequency	48
Worth	61
Level of Membership	31
Brightness	27
Strength	24
Speed	26

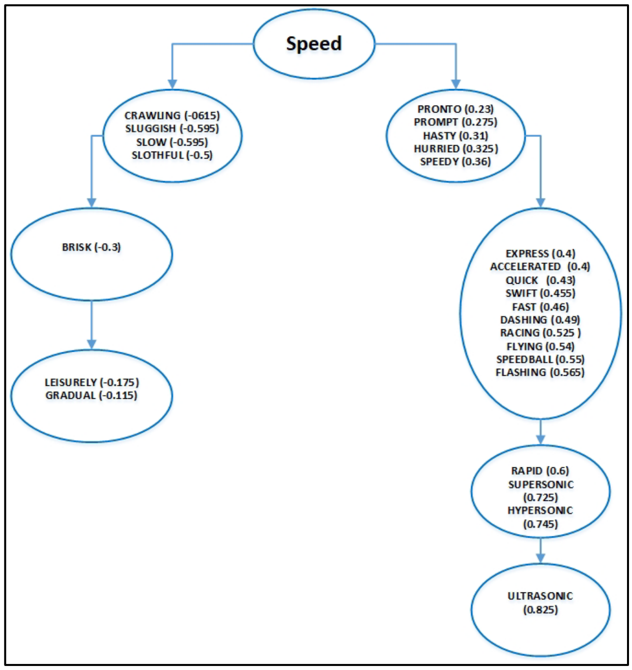


FIGURE 3. Ontology Structure for Speed Category.

values which are utilized in the final sentence similarity rating between the two sentences.

VII. EXPERIMENTAL METHODOLOGY

A. OVERVIEW

In 2016 Mendel [43] introduced the Footprint of Uncertainty [FOU] where he computes a FOU for a set of words by capturing 50 intervals from one participant. This was done by getting one participant to rate a word on a scale of *l*-*r*, with *l* being Left and *r* being Right, giving the left(x_l,y_l) and right(x_r,y_r) endpoints. Using this one rating from the one participant, Mendel then goes on to generate 100 random numbers (L_1, L_2,…,L_{50}; R_1, R_2,…,R_{50}) and used these to further generate 50 endpoint interval

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3194510

pairs [(L_1, R_1), (L_2, R_2),…,(L_{50},R_{50})], thus reducing the time required to collect ratings. Whereas Mendel used the one-person approach, this method was adapted, and 32 participants were used as opposed to one, to create a richer array of human results from 32 different people [20]. 32 native English-speaking participants from the Northwest region of UK were used to collect human ratings per fuzzy category for the nine categories of FUSE using the HM Approach [43]. Data was collected for the nine fuzzy categories using an online questionnaire and participants were asked to rate the words as a range in each category on a scale of [0-10]. The words in each category were presented in random order to not affect the results given by the participants. An example of how the questions were presented in the questionnaires is shown below. For example, given the word '*Baby*' belonging to the category '*Age*' the question was presented as follows:

"Rate the word BABY as a measure of Age on a scale of 0 to 10. (You can go up to one decimal place). PLEASE ONLY WRITE YOUR ANSWERS IN THE FORMAT "x to y" WHERE x AND y ARE THE NUMBERS YOU HAVE CHOSEN."

To not exhaust the participants and thus impact the results negatively each participant was asked to rate words belonging to just one category per sitting. Each question asked the participant to provide a range for a given word in the chosen category on a scale of [0-10] indicating where they believed this word fell from start to finish, i.e., the word *Baby* belonging to the category *Age* may have the range 2 to 2.7. A [0-10] measure ruler was also provided for reference as shown in Fig. 4. A generic example was provided at the start of the questionnaire not relating to that category to ensure the participants understood what was meant by range. An example of this is provided below:

"For example, the word COLD which belongs to the category TEMPERATURE. In my opinion I would say that on a scale of 0 to 10, Cold is between 2 - 3.5."

Once this experiment was conducted for all nine fuzzy categories, the data cleaning progress could begin. Each category had more than 32 participants taking part in the rating of words due to over subscription of volunteers,

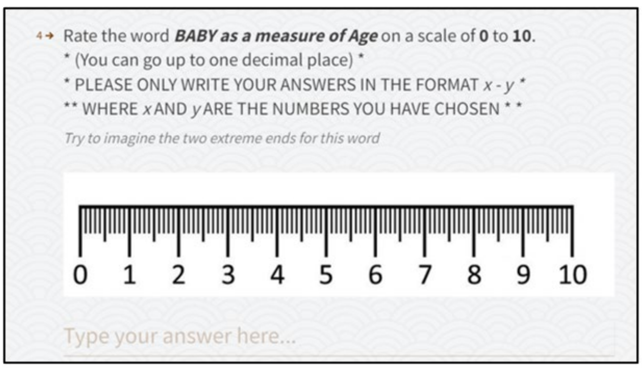


FIGURE 4. Questionnaire Example.

which was helpful when reducing noise and outliers. Mendel's statistic and probability theory [43] was used to remove noise, the steps of which are explained below:

1- The first step was to remove any potential bad data, so in this case it was any value that was outside of the proposed scale of [0-10];

2- The second step involved removing outliers from the results. The experiment was conducted with the use of a box and whisker test [44] to remove outliers simultaneously and the results were left with the data intervals that fell within an acceptable two-sided tolerance limit

3- The final step involved removing data intervals that had no overlap or very little overlap. This is due to the fact that while Mendel states *words mean different things to different people* [45], he also argues that *words should mean similar things to different people* [45], therefore if most participants rated a word between the intervals of [2-4] and a few rated the same word on an interval of [6-7] or [8-9] then the latter two would be considered to not have any overlap with the other results and will be removed.

On completion of these three steps, the results were left with $m \leq n$, where n is the original data ranges collected by all participants and m is the data intervals after conducting the above three steps, where m = 32 clean value ranges per category for all nine fuzzy categories. Each word per category was analysed to find the upper and lower FOU per word as proposed by Mendel [43], from this the COG (center of gravity) was obtained per word using:

$$COG = \frac{\left(\left(\frac{a+b}{2}\right)+\left(\frac{c+d}{2}\right)\right)}{2} \quad (1)$$

where a = upper left FOU, b = lower left FOU, c = lower right FOU and d = upper right FOU.

Table III shows a defuzzified example for the word '*Close*' from the category '*Size/Distance*' on a scale of [0-10]. The values are calculated using the triangular membership function. '*x*' is the scale of [0-10], '*lower*' represents the lower boundaries, and '*upper*' represents the upper boundaries. '*t-norm$_{(prod)}$*' is the multiplication of lower and upper, and '*t-norm$_{(min)}$*' is the minimum boundary from the lower or upper. Fig. 5 shows the Type-1 defuzzified graphical representation of the word '*Close*' in the category '*Size/Distance*' that has resulted from the triangular membership calculation. The values in the column '*t-norm$_{(min)}$*' have been used to plot the graph. The COG value was then normalised to a scale of [-1,+1] to give the defuzzified value per word per category, using equation (2):

$$y = a + \frac{(x-A)(b-a)}{B-A} \quad (2)$$

TABLE III
DEFUZZIFIED EXAMPLE FOR CLOSE

x	Lower	Upper	T-norm(prod)	T-norm(min)
0	0.00	0.00	0.00	**0.00**
1	0.25	0.27	0.07	**0.25**
2	0.52	0.53	0.27	**0.52**
3	0.79	0.80	0.63	**0.79**
4	0.93	0.95	0.88	**0.93**
5	0.62	0.71	0.44	**0.62**
6	0.31	0.47	0.15	**0.31**
7	0.00	0.24	0.00	**0.00**
8	0.00	0.00	0.00	**0.00**
9	0.00	0.00	0.00	**0.00**
10	0.00	0.00	0.00	**0.00**

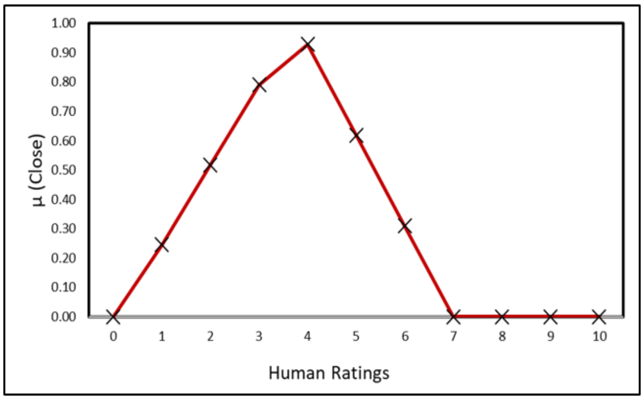


FIGURE 5. Triangular Membership for (Close).

where A = smallest number in dataset, B = largest number in dataset, a = minimum normalised value (-1), b = maximum normalised value (+1) and x = value we want to scale (in this case the COG). This was done for all the words in the nine fuzzy categories and thus ensured each word had a rating which would be used as part of the fuzzy dictionary [Appendix A] in both FUSE_1.0 and FUSE_2.0.

B. THE FUSE ALGORITHM

The FUSE_1.0 and FUSE_2.0 algorithms are designed to measure the similarity of two fuzzy utterances up to 25 words in length. A fuzzy utterance must contain at least one fuzzy word. A fuzzy word is a word that does not have a fixed meaning and can vary in meaning depending on the perspective of an individual [40]. The FUSE_2.0 algorithm can be defined as follows:

Given two fuzzy utterances, U_1 and U_2, their similarity $S(U_1, U_2)$ is computed. The FUSE algorithm builds upon the original STASIS approach [14], where the semantic similarity vectors and the word order similarity vectors for both the utterances are computed. These vectors are constructed using the information about the word pairs and their associated depth and shortest path length in the WordNet dictionary [6]. The extra information about the fuzzy words are included, and when applicable, the lowest common subsumer depth and shortest path length using the FUSE_1.0 approach [5] are computed. The information content measurements for the Brown Corpus [46] are included. Combining all this information allows the computation of the similarity between the two utterances. w_i is denoted as a single word in either of the utterances for $i \in I$, some indexing set. Let $U = U_1 \cup U_2$ be the set of all distinct words appearing in U_1 or U_2. Following Li's approach [14] $T :=$ {adjective, adposition, adverb, conjunction, determiner, noun, numeral, particle, pronoun, verb} is set, to be the set of all the possible tags to be assigned to each word w_i via the map $\tau : U_i \longrightarrow U_i \times T$, such that: $\omega_i := \tau(w_i) = (w_i, t)$.
This information is obtained from WordNet [6] and Brown's Corpus [46] . W_1 and W_2 were set to be the sets of all the word-token pairs (w_i, t) from $U_1 \times T$ and $U_2 \times T$ respectively. The first stage of this computation is shown in Fig. 6, which populates these sets. Let $\omega_{i,j} \in W_1 \times W_2$ be a pair of word pairs ω_i and ω_j, i.e. $\omega_{i,j} := (\omega_i, \omega_j)$. The set of all pairs of word-token couples were denoted by Ω. The function $f: W_1 \times W_2 \longrightarrow \{0,1\}$ on the elements $\omega_{i,j} \in \Omega$, was defined via:

$$f(\omega_1, \omega_2) = \begin{cases} 1 & \text{if both } \omega_1 \,\&\, \omega_2 \text{ are fuzzy words} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Let C denote the set of fuzzy categories, where $C :=$ {Size/Distance, Temperature, Age, Frequency, Worth, Level of Membership, Speed, Strength, Brightness}. The co-membership in a fuzzy category is determined by the function $c: W1 \times W2 \longrightarrow \{0,1\}$ such that:

$$c(\omega_1, \omega_2) = \begin{cases} 1 & \text{If } \omega_1 \,\&\, \omega_2 \text{ are in the same fuzzy category C} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

If two words are not in the same fuzzy category or neither are fuzzy words, the depth and shortest path length are calculated from the values obtained from WordNet. The depth of the word pair is computed via:

$$SD: \Omega \longrightarrow (0,1) \text{ such that } \quad \omega_{i,j} \longmapsto d_{i,j} \tag{5}$$

The path length via:

$$SL: \Omega \longrightarrow (0,1) \text{ such that } \quad \omega_{i,j} \longmapsto l_{i,j} \tag{6}$$

Algorithm 1 Create word-token pairs $\omega_i = (w_i, t)$

Require: U_1, U_2
$W_1 := (); W_2 := ()$
for $k \in \{1, 2\}$ **do**
  for $w_i \in U_k$ **do**
    $\omega_i := \tau(w_i)$
    if $\omega_i \notin W_k$ **then**
      $W_k := W_k \cup \{\omega_i\}$
    end if
  end for
end for
return W_1 and W_2

FIGURE 6. Algorithm 1 – Create Word-Token Pairs.

Word similarity $wordSim$ via:

$$wordSim: \Omega \times R \times R \longrightarrow R$$

$$wordSim(\omega_{i,j}, d_{i,j}, l_{i,j}) \longmapsto e^{-\alpha l} \cdot tanh\,(\beta d) \tag{7}$$

where $d := d_{i,j}$, $l := l_{i,j}$ and the parameters α and β were empirically determined as 0.15 and 0.85, respectively. However, if two fuzzy words come from the same fuzzy category $c \in C$, the lowest common subsumer depth and the shortest path length can be computed within this ontology. FD and FL were denoted by the functions analogous to SD and SL, coming from the FUSE ontology. These attributes, shown in Fig. 7 are used to compute the matrix of similarities of the word pairs $\omega_{i,j}$. Finally, Fig. 8 shows for each of the utterances U_k the semantic similarity vector s_k and the word order similarity vector r_k were computed. The angular distances between these determine the level of similarity, and thus:

1. The semantic similarity S_s is computed as the cosine of the angle γ between the vectors s_1 and s_2:

$$S_s := \frac{s_1 \cdot s_2}{||s_1||\,||s_2||} = cos(\gamma) \tag{8}$$

2. The word order similarity S_r is computed in terms of tan of half the angle μ between the word order vectors r_1 and r_2:

$$S_r := 1 - \frac{||r_1 - r_2||}{||r_1 + r_2||} = 1 - tan(\tfrac{1}{2}\mu) \tag{9}$$

3. The similarity of the two utterances S is determined to be a linear combination of S_s and S_r:

$$S(U_1, U_2) := \delta cos(\gamma) + (1 - \delta) tan(\frac{1}{2}\mu) \tag{10}$$

where $0 < \delta \leq 1$ decides the relative contributions of semantic and word order information to the overall similarity computation.

Algorithm 2 The matrix of word similarities $\check{S}$

Require: W_1 and W_2
$\Omega := W_1 \times W_2$
$\check{S} := []$, where $\check{S} \in Mat_{n_1 \times n_2}(\mathbf{R})$.
for all $\omega_{i,j} \in \Omega$ **do**
  if $f_{i,j} = 1$ **then**
    if $c_{i,j} = 1$ **then**
      $d_{i,j} := \mathcal{FD}(\omega_{i,j}$
      $l_{i,j} := \mathcal{FL}(\omega_{i,j})$
    else
      $d_{i,j} := \mathcal{SD}(\omega_{i,j}$
      $l_{i,j} := \mathcal{SL}(\omega_{i,j})$
    end if
  else
    $d_{i,j} := \mathcal{SD}(\omega_{i,j}$
    $l_{i,j} := \mathcal{SL}(\omega_{i,j})$
  end if
  $\check{s}_{i,j} := wordSim(\omega_{i,j}, d_{i,j}, l_{i,j})$
end for
return $\check{S}$

FIGURE 7. Algorithm 2 – Matrix of Word Similarities Š.

VIII. EVALUATION OF FUSE_2.0

A. OVERVIEW

In this section FUSE_2.0 is evaluated in terms of its correlation with average human ratings (AHR) across five datasets. The results are then compared against other established and appropriate SSM's such as STASIS [14], Dandelion Semantic [21], Dandelion Syntactic [21] and SEMILAR [22] with that of the AHR to see which algorithm gave a higher correlation value with the AHR. This evaluation method is the established approach in the field of SSM [14, 20].

B. EVALUATION METHODOLOGY

This experiment investigated the correlation of the FUSE_2.0 algorithm against the AHR, whilst also investigating the presence of the fuzzy dictionary to see if it helped with the correlations against the AHR. This investigation was ran on several datasets and compared with other sentence similarity measures (Section IV B). The aim of the experiments was to test the following null hypothesis:

H_0: FUSE_2.0 gives a higher correlation with human ratings compared to other SSM's.

To test H_0, FUSE_2.0 was ran against each of the five datasets (FUSE-62, SWFD [47], MWFD [47], STSS-65 [20] and STSS-131 [20]) and the sentence similarity results for each Sentence Pair [SP] was recorded. To be able to test the improvement of FUSE_2.0, all five datasets were also run with STASIS [14], Dandelion Semantic [21], Dandelion Syntactic [21] and SEMILAR [22] algorithms

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3194510

Algorithm 3 Similarity of utterances

Require: U_1, U_2 and the corresponding $\check{S}$
$s_1 := [], s_2 := [], r_1 := [], r_2 := [], T := \check{S}^T, U = U_1 \cup U_2$
for $i \in \{1, \ldots, n_1\}$ **do**
  $r_1[i] := i$
  if $\check{s}_i \neq \underline{0}$ **then**
    $idx := j$ such that $\check{s}_{i,j} = \max_j(\check{s}_{i,j})$
    $s_1[i] := \check{s}_{i,idx} \cdot I(w_i) \cdot I(w_{idx})$ where $w_{idx} \in W_2$.
    $r_1[index(w_{idx})$ in $U] := i$
  else
    $s_1[i] := 0$
  end if
end for
for all $k \in \{n_1 + 1, \ldots, m\}$ **do**
  if $r_1[k]$ is not defined **then**
    Set $r_1[k] := 0$.
  end if
end for
for $i \in \{1, \ldots, n_2\}$ **do**
  Compute s_2 and r_2 in the analogous way to the above, taking the transpose of $\check{S}, T$, as the argument.
end for
$S_s(s_1, s_2) := cos(\gamma)$
$S_r(r_1, r_2) := 1 - tan(\frac{1}{2}\mu)$
$S(U_1, U_2) := \delta cos(\gamma) + (1 - \delta) tan(\frac{1}{2}\mu)$
return $S(U_1, U_2)$

FIGURE 8. Algorithm 3 – Similarity of Utterances.

and the sentence similarity results for each SP was recorded.

Using Pearson's correlation coefficient [36], the correlation for each dataset was compared to the Average Human Ratings (AHR). Pearson's correlation provides statistical evidence for a linear relationship between two variables x and y and can be computed as follows [36]:

$$r_{xy} = \frac{cov(x,y)}{\sqrt{var(x)}.\sqrt{var(y)}} \tag{11}$$

where r_{xy} is the correlation coefficient, $cov(x, y)$ is the sample covariance of x and y; $var(x)$ is the sample variance of x; and $var(y)$ is the sample variance of y [36].

C. DATASETS

Five datasets were used in total containing both fuzzy sentence pairs, and non-fuzzy sentence pairs. A full breakdown of these datasets is given in Table IV. The reader's age is determined after examining the contents of each dataset and performing a feasibility test [48]. This feasibility is important because it influences how clearly a text can be understood by the reader. By making text as clear as possible to understand allows improved participant selection [49, 50].

TABLE IV
DATASET DESCRIPTION

Dataset	Description	Fuzzy / Non-Fuzzy	Readers Age
FUSE-62	62 sentence pairs specifically designed by English language experts to contain fuzzy words from all nine categories of FUSE_2.0	Fuzzy	14-15 yrs. old
SWFD	30 sentence pairs containing one fuzzy word per sentence	Fuzzy	10-11 yrs. old
MWFD	30 sentence pairs containing two or more fuzzy word per sentence	Fuzzy	College graduate
STSS-65	65 Gold standard sentence pairs	Non-Fuzzy	10-11 yrs. old
STSS-131	131 Gold standard sentence pairs	Non-Fuzzy	8-9 yrs. old

D. RESULTS AND DISCUSSION

Table V shows the results of the datasets for the different SSM algorithms. Pearson's correlation coefficient was calculated for each of the five datasets using the five algorithms, FUSE_2.0, STASIS [14], Dandelion Semantic [21], Dandelion Syntactic [21] and SEMILAR [22] compared to the AHR. It can be seen from the results in Table V that FUSE_2.0 gave a higher correlation for each dataset compared to all the other algorithms. Fig. 9 illustrates a graphical representation of the results from Table V showing FUSE_2.0 achieving the highest correlation with AHR for all datasets tested, compared to the other SSM's. It can be seen from the results in Table V that the dataset containing the greatest number of fuzzy words (MFWD) gave the highest correlation (0.768202) and the dataset with no fuzzy words STSS-131 gave the lowest correlation (0.518458). This strongly suggests that the more fuzzy words present in a short text or sentence pair, the better the FUSE_2.0 algorithm performs and further highlights the need to consider the presence of fuzzy words on sentence similarity.

Conduction of an Intra-Class Correlation Coefficient (ICC) [51] also produced some positive results as shown in Table VI. The ICC is important in a study as it represents

TABLE V
PEARSON'S CORRELATION VALUES ACROSS 5 DATASETS COMPARING FIVE SSM'S

Pearson Correlation of Results	STASIS	Dandelion Semantic	Dandelion Syntactic	SEMILAR	FUSE_2.0
FUSE-62	0.543	0.546	0.312	0.533	0.544
SFWD	0.645	0.433	0.577	0.627	0.688
MFWD	0.745	0.629	0.736	0.758	0.768
STSS-65	0.681	0.537	0.620	0.661	0.690
STSS-131	0.502	0.406	0.152	0.491	0.518

TABLE VI
A & P VALUE RESULTS ACROSS FIVE DATASETS

Inter-Rater Correlation Results	a-value	Cicchetti Measure	p-value	Accept or Reject
62 SP	0.872	Excellent	0.000	Accept
SFWD	0.911	Excellent	0.000	Accept
MFWD	0.947	Excellent	0.000	Accept
STSS-65	0.883	Excellent	0.000	Accept
STSS-131	0.104	Poor	0.199	Reject

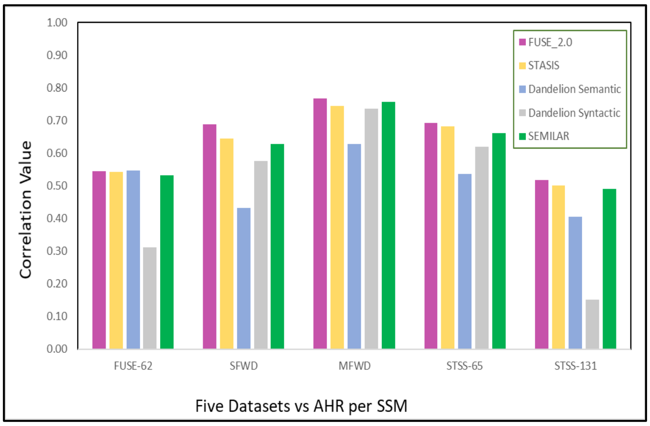


FIGURE 9. Results comparison for five datasets based on correlation value.

the extent to which the data collected in the study is correct and a good representation of the variables measured. Cicchetti gives the following guidelines for ICC measures (also referred to as the *a-value*) [52]:

- Less than 0.40 – Poor;
- Between 0.40 and 0.59 – Fair;
- Between 0.60 and 0.74 – Good;
- Between 0.75 and 1.00 – Excellent.

Looking at the *a-value* in Table VI, it can be seen that that four of the datasets (FUSE-62, SWFD [47], MWFD [47], STSS-65 [20]) show an *Excellent* rating based on the *a-value*. It can further be shown that the more fuzzy words present in a dataset, the higher the *a-value*. This can be seen in the MFWD dataset with the *a-value* being the highest of all datasets (*a-value* = 0.947); this is because the MFWD has two or more fuzzy words present per sentence pair. The *p-value* is the standard method that is used in statistics to measure the significance of empirical analyses [53].The *p-value* for four of the datasets (FUSE-62, SWFD [47], MWFD [47], STSS-65 [20]) is < 0.001 which is less than 0.05 and therefore, statistically significant. This result provides support for our research hypothesis H_0 which strongly suggests that the expansion of the fuzzy dictionary and the introduction of a fuzzy ontology affects the level of similarity. Looking at both the *a-value* (second column) and the *p-value* (fourth column) in Table VI, the dataset that held the highest number of non-fuzzy words (STSS-131) [20], is the dataset that gave the lowest *a-value* result (*a-value* = 0.104) which is deemed as *Poor* according to Cicchetti and the *p-value* was rejected. This shows that the more fuzzy words present in a dataset, the higher the *a-value*, which in turn means FUSE_2.0 performs better when more fuzzy words are present in a sentence or utterance. Most SSM's use WordNet [6], and since WordNet is constantly being improved, results can vary over time, therefore it is important to note that if this experiment was to be repeated, results may vary slightly. FUSE_2.0 shows that fuzzy words must be considered when looking at sentence similarity measures as they play a significant role in the similarity of sentences. Looking back at the experiments conducted on the five datasets using the five algorithms, FUSE_2.0, STASIS [14], Dandelion Semantic [21], Dandelion Syntactic [21] and SEMILAR [22] and the original null hypothesis presented in Section VIII:

H_0: FUSE_2.0 gives a higher correlation with human ratings compared to other SSM.

It can be concluded that H_0 can be accepted based on both the *a-value* and the *p-value* shown in Table VI for a confidence level of 95%.

IX. CONCLUSION AND FURTHER WORK

This paper described the creation of a fuzzy sentence similarity measure referred to as FUSE. Nine fuzzy categories were used to create a fuzzy ontology

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3194510

embedded within the FUSE algorithm. Experiments were conducted on FUSE_2.0 using five datasets, three of which were fuzzy datasets consisting of fuzzy words, and the remaining two did not contain fuzzy words. Results showed that considering a fuzzy measure in a sentence similarity measure improved the correlation when compared with the average human ratings when using the FUSE_2.0 algorithm. When comparing the FUSE_2.0 algorithm with other SSM algorithms that do not cater for the presence of fuzzy words, it has been shown that FUSE_2.0 gave a higher correlation to the average human ratings as opposed to all other SSM algorithms tested. This further emphasises the importance of taking into consideration the presence of fuzzy words in a sentence or utterance. Looking back at the original research question proposed in Section II:

Can Type-2 fuzzy sets be used to represent an individual's perception within a fuzzy semantic similarity-based measure?

This paper has successfully managed to answer this question via capturing the human perceptions of fuzzy words and producing representative models of these fuzzy words though using Interval Type-2 fuzzy sets. These models were then integrated successfully within FUSE_2.0 as evidenced by the results in Section VIII D. The main advantages of using this FSSM is achieving a higher correlation with the AHR by considering the presence of perception-based (fuzzy words) in sentences or utterances. The limitation however as discussed in Section VIII D, is that the FUSE algorithm may produce a lower correlation with the AHR when using sentences or utterances that have little or no fuzzy words present as shown in the results of Table VI.
Further work on the FUSE algorithm will take into consideration the impact of negation words such as '*not*' on fuzzy words and evaluate the effect on sentences or utterances. Further work will also cater for the overall similarity of fuzzy words from different fuzzy categories which may be present in sentence pairs. This will aim to further improve the correlation of fuzzy sentences compared with the average human ratings. Currently FUSE_2.0 ignores this scenario and only provides a fuzzy measure if fuzzy words are from the same fuzzy category. If this is not the case, the FUSE algorithm uses WordNet to calculate similarity if fuzzy words present in a sentence do not belong to the same fuzzy category. Additional further work can also be conducted to adapt to other languages especially on low resource languages through investigating lexical resources and designing fuzzy dictionaries. The development and integration of this algorithm into such applications will allow for a richer modelling of human perception-based words.

APPENDIX

APPENDIX A – THE FUSE FUZZY DICTIONARY

1 - SIZE/DISTANCE

MICROSCOPIC	-1	AVERAGE	0.029762
MINUSCULE	-0.88095	MEAN	0.029762
DINKY	-0.86905	ACCESSIBLE	0.035714
TEENY	-0.85714	HALFWAY	0.035714
TITCHY	-0.7381	ISOLATED	0.047619
LITTLE	-0.70833	CENTRAL	0.065476
SMALL	-0.70833	GOODLY	0.065476
WEE	-0.70833	MIDWAY	0.065476
INSIGNIFICANT	-0.70238	MIDPOINT	0.066667
PETITE	-0.64286	CENTRE	0.066667
DIMINUTIVE	-0.58333	MEDIAN	0.083333
NEAREST	-0.58333	MIDDLE	0.083333
PIDDLING	-0.58333	MID	0.089286
TINY	-0.55952	REMOTE	0.178571
MINUTE	-0.55357	METHODICAL	0.184524
SHORT	-0.52381	ABUNDANT	0.214286
UNIMPORTANT	-0.52381	CONSIDERABLE	0.309524
PALTRY	-0.51191	LOADS	0.333333
TRIVIAL	-0.5	THICK	0.333333
NEAR	-0.47619	FAR	0.363095
MESIAL	-0.44048	SIZEABLE	0.392857
CONJOINING	-0.43452	LARGE	0.482143
BESIDE	-0.41071	PRINCELY	0.482143
ADJOINING	-0.38095	BOUNDLESS	0.535714
THIN	-0.36364	DISTANT	0.541667
TOKEN	-0.35714	WHACKING	0.541667
NEARBY	-0.35119	SUBSTANTIAL	0.60119
QUALITY	-0.35119	BIG	0.660714
MOMENT	-0.32143	GREAT	0.660714
NORM	-0.29167	FARAWAY	0.666667
CLOSE	-0.28571	HEFTY	0.678571
ALONGSIDE	-0.27976	LONG	0.684211
ADJACENT	-0.26191	JUMBO	0.720238
ORDINARY	-0.22619	EPIC	0.75
MEDIUM	-0.20238	MASSIVE	0.75
PROXIMATE	-0.20238	OVERSIZED	0.754386
EQUIDISTANT	-0.14286	IMMENSE	0.754386
TIDY	-0.14286	GIANT	0.809524
USUAL	-0.1131	HUGE	0.827381
AWAY	-0.10119	ENORMOUS	0.833333

NORMAL	-0.10119	MEGA	0.839286
PROXIMAL	-0.05357	COLOSSUS	0.869048
REGULAR	-0.05357	GIGANTIC	0.892857
STANDARD	-0.05357	MAMMOTH	0.894
BONNY	-0.02381	GARGANTUAN	1
MEDIAL	0.011905		

2 - TEMPERATURE

FROZEN	-1	BALMY	0.134948
SUB-ZERO	-1	TEMPERATE	0.204152
ARCTIC	-0.93772	LUKEWARM	0.231834
FREEZING	-0.89619	WARM	0.480969
ICY	-0.7301	HUMID	0.550173
FROSTY	-0.70934	PERSPIRING	0.550173
CHILLY	-0.6955	SPICY	0.550173
BRISK	-0.6263	BAKING	0.619377
COLD	-0.57786	HOT	0.619377
BITTER	-0.55709	SWEATY	0.688581
BITING	-0.45329	SCALDING	0.750865
COOL	-0.45329	HEATED	0.757785
BRACING	-0.31488	STEAMING	0.757785
NIPPY	-0.28028	SWELTERING	0.792388
TEPID	-0.24568	ROASTING	0.861592
MILD	-0.23875	BOILING	0.889273
BODY-TEMPERATURE	0	SCORCHING	0.930796
FRIGID	0.100346	BURNING	1

3 - AGE

BABY	-1	GROWNUP	0.078014
NEW	-0.963768	PRIMORDIAL	0.0797101
LATEST	-0.93939	PREHISTORIC	0.33333
BABYISH	-0.891304	JUVENILE	0.4565217
CHILDISH	-0.804347	AGED	0.6449275
EARLIEST	-0.789855	PRIMEVAL	0.7028985
INFANTILE	-0.789855	ADULT	0.7173913
VULNERABLE	-0.768115	ANTIQUATED	0.7898550
UNDERAGE	-0.659420	DECREPIT	0.7898550
RECENT	-0.623188	OLDER	0.789855
CHILD	-0.586956	EXPERIENCED	0.8260869
YOUNG	-0.586956	OLD	0.8478260
ADOLESCENT	-0.514492	MATURE	0.8623188
YOUTHFUL	-0.514492	PRIMITIVE	0.8695652
PUBESCENT	-0.442028	SENIOR	0.8913043
IMMATURE	-0.333333	PRIMAL	0.8985507
CHILDLIKE	-0.33333	ELDERLY	0.9275362
PREPUBESCENT	-0.29078	ARCHAIC	0.9347826
TEENAGE	-0.144927	ANTIQUE	0.9710144
MIDDLEAGED	0.049645	PENSIONABLE	0.9710144
FULL-GROWN	0.06383	ANCIENT	1

4 - FREQUENCY

NEVER	-0.68	REGULARLY	0.25
HARDLY	-0.425	ESPECIALLY	0.3
BARELY	-0.4	PERIODICALLY	0.3
SOMEWHAT	-0.4	COMMONLY	0.325
SCARCELY	-0.39	CUSTOMARILY	0.35
SELDOM	-0.365	NATURALLY	0.35
FAINTLY	-0.35	TYPICALLY	0.35
NARROWLY	-0.335	CONSISTENTLY	0.4
RARELY	-0.33	ORDINARILY	0.4
INFREQUENTLY	-0.325	FREQUENTLY	0.405
SLIGHTLY	-0.325	OFTEN	0.405
NOTABLY	-0.3	REPEATEDLY	0.405
UNPREDICTABLY	-0.255	CONSTANTLY	0.425
CONVENTIONALLY	-0.245	CONTINUOUSLY	0.425
UNUSUALLY	-0.23	DAILY	0.425
OCCASIONALLY	-0.2	INEVITABLY	0.425
UNCOMMONLY	-0.165	GENERALLY	0.45
ON-OCCASION	-0.14035	NORMALLY	0.45
USUALLY	-0.005	CONTINUALLY	0.5
HABITUALLY	0	ROUTINELY	0.5
FAIRLY	0.085	ALWAYS	0.575
INVARIABLY	0.135	EXTREMELY	0.625
EXCEPTIONALLY	0.15	PERSISTENTLY	0.645
MODERATELY	0.15	EVERYTIME	1

5 - WORTH

APPALLING	-1	FAIR	-0.137931
DIRE	-1	ADEQUATE	-0.068965
DREADFUL	-1	PERMISSIBLE	-0.068965
HORRENDOUS	-1	ALRIGHT	-0.048275
INSUFFERABLE	-1	MIDDLING	-0.034482

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

INTOLERABLE	-1	SATISFACTORY	0
USELESS	-0.95862	NORMAL	0.0344827
UNSATISFACTORY	-0.93103	ORDINARY	0.0344827
UNBEARABLE	-0.91724	PASSABLE	0.0344827
POOR	-0.89655	AVERAGE	0.1034482
UNACCEPTABLE	-0.87586	NICE	0.2068965
BAD	-0.83448	PLEASANT	0.2068965
DISAPPOINTING	-0.82758	DELIGHTFUL	0.3793103
TERRIBLE	-0.82758	ENJOYABLE	0.4137931
AWFUL	-0.79310	GOOD	0.4827586
PATHETIC	-0.79310	GREAT	0.5448275
ROTTEN	-0.75862	SUBLIME	0.5517241
UNPLEASANT	-0.75862	LOVELY	0.5862068
DISSATISFYING	-0.72413	WONDERFUL	0.6896551
TEDIOUS	-0.69655	SPLENDID	0.7172413
BORING	-0.68965	BRILLIANT	0.7241379
UNDESIRABLE	-0.68965	FANTASTIC	0.7379310
NASTY	-0.66667	AMAZING	0.7931034
INADEQUATE	-0.65517	TREMENDOUS	0.8275862
SUBSTANDARD	-0.58620	ASTONISHING	0.8620689
FINE	-0.41379	SUPERB	0.8965517
MEDIOCRE	-0.41379	EXCELLENT	0.9310344
OK	-0.27586	MAGNIFICENT	0.9379310
REASONABLE	-0.20689	MARVELLOUS	0.9655172
SUITABLE	-0.20689	GLORIOUS	1
ACCEPTABLE	-0.13793		

6 - LEVEL OF MEMBERSHIP

BARELY	-1	SUITABLE	0.2
HARDLY	-0.968	AVERAGE	0.24
LITTLE	-0.92	APPROPRIATE	0.36
SCARCELY	-0.88	MOSTLY	0.36
BIT	-0.76	AMPLE	0.4
SCRAPING	-0.76	GENERALLY	0.4
FRACTIONALLY	-0.648	USUALLY	0.4
SLIGHTLY	-0.64	ALMOST	0.44
PARTIALLY	-0.48	SUFFICIENT	0.44
JUST	-0.216	MAINLY	0.64
SOMEWHAT	-0.16	SERIOUSLY	0.672
ADEQUATE	-0.088	SUBSTANTIALLY	0.712
ENOUGH	0.12	SIGNIFICANTLY	0.72
RATHER	0.12	LARGELY	0.76
HALFWAY	0.128	GREATLY	1

MIDDLING	0.184	SUITABLE	0.2

7 - STRENGTH

WEAK	-0.738	ENERGETIC	0.285
POWERLESS	-0.645	STURDY	0.305
DELICATE	-0.57	FIRM	0.35
FEEBLE	-0.525	HEAVY	0.375
PUNY	-0.525	ATHLETIC	0.375
ABLE	0.01	VIGOROUS	0.375
CAPABLE	0.1	HARDY	0.375
DURABLE	0.1	TOUGH	0.4
ROBUST	0.21	MUSCULAR	0.465
STABLE	0.23	SOLID	0.48
REINFORCED	0.255	MIGHTY	0.575
HEARTY	0.28	STRONG	0.645

8 - BRIGHTNESS

LIGHTLESS	-0.64	ILLUMINATED	0.4
MOONLIT	-0.38	FLASHING	0.45
BURNISHED	-0.35	GLARING	0.45
AGLOW	-0.2	LIGHT	0.5
TWINKLING	0.02	GOLDEN	0.55
BURNING	0.1	SHINY	0.55
BEAMING	0.25	SPARKLING	0.55
ALIGHT	0.35	SUNNY	0.55
ILLUMINED	0.35	BLAZING	0.55
LIGHTED	0.35	RADIANT	0.55
GLITTERING	0.35	GLISTENING	0.55
LUMINOUS	0.38	BRIGHT	0.57
SHIMMERING	0.4	DAZZLING	0.6
SUNLIT	0.4	NOT-BRIGHT	-0.14

9 - SPEED

CRAWLING	-0.615	ACCELERATED	0.4
SLUGGISH	-0.595	QUICK	0.43
SLOW	-0.595	SWIFT	0.455
SLOTHFUL	-0.5	FAST	0.46
BRISK	-0.3	DASHING	0.49
LEISURELY	-0.175	RACING	0.525

GRADUAL	-0.115	FLYING	0.54
PRONTO	0.23	SPEEDBALL	0.55
PROMPT	0.275	FLASHING	0.565
HASTY	0.31	RAPID	0.6
HURRIED	0.325	SUPERSONIC	0.725
SPEEDY	0.36	HYPERSONIC	0.745
EXPRESS	0.4	ULTRASONIC	0.825

REFERENCES

[1] Banerjee, (2020, Apr. 14) Natural Language Processing (NLP) Simplified : A Step-by-step Guide [Online]. Available: https://datascience.foundation/sciencewhitepaper/natural-language-processing-nlp-simplified-a-step-by-step-guide, Accessed on: 15th Dec. 2020.

[2] Alexa Voice Service Overview (v20160207) (2021) Alexa Voice Service [Online]. Available : www.developer.amazon.com, Accessed on: 14th Sept. 2017.

[3] S. Singh and H. K. Thakur, "Survey of Various AI Chatbots Based on Technology Used," In 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2020, pp. 1074-1079, doi: 10.1109/ICRITO48877.2020.9197943.

[4] Oxford English Dictionary (2021) [Online]. Available: https://www.oed.com/, Accessed on: 18th Oct. 2017.

[5] N. Adel, K. Crockett, A. Crispin, D. Chandran and J.P. Carvalho. "FUSE (Fuzzy Similarity Measure)-A measure for determining fuzzy short text similarity using Interval Type-2 fuzzy sets," In 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1-8, Jul. 2018.

[6] Princeton University, "About Wordnet." [Online] Available: http://wordnet.princeton.edu/, Accessed on: 21st Oct. 2017.

[7] D. Lin, "An information-theoretic definition of similarity," Icml, vol. 98, no. 1998 pp. 296-304, Jul. 1998.

[8] P. Sunilkumar, and A.P. Shaji. "A Survey on Semantic Similarity." In 2019 International Conference on Advances in Computing, Communication and Control (ICAC3), pp. 1-8, IEEE, 2019.

[9] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," Journal of artificial intelligence research, vol. 11, pp. 95-130, Jul. 1999.

[10] A.W. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, "Content-based image retrieval at the end of the early years," IEEE Transactions on pattern analysis and machine intelligence, vol. 22, no. 12, pp. 1349-1380, Dec. 2000.

[11] R.K. Srihari, Z. Zhang and A. Rao, A. "Intelligent indexing and semantic retrieval of multimodal documents," Information Retrieval, vol. 2, no. 2, pp. 245-275, May 2000.

[12] S.J. Green, "Building hypertext links by computing semantic similarity," IEEE Transactions on Knowledge and Data Engineering, vol. 11, no. 5, pp. 713-730, Sept. 1999.

[13] M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun and C Gao. "A survey on the techniques, applications, and performance of short text semantic similarity," Concurrency and Computation: Practice and Experience, vol. 33, no. 5, e5971, Mar. 2021.

[14] Y. Li, D. McLean, Z.A. Bandar, J.D. O'shea and K. Crockett. "Sentence similarity based on semantic nets and corpus statistics," IEEE transactions on knowledge and data engineering, vol. 18, no. 8, pp. 1138-1150, Jun. 2006.

[15] P. Achananuparp, X. Hu and X. Shen. "The evaluation of sentence similarity measures." In International Conference on data warehousing and knowledge discovery, pp. 305-316, Springer, Berlin, Heidelberg, Sep. 2008.

[16] D. Chandran, K. Crockett, D. Mclean and Z. Bandar. "FAST: A fuzzy semantic sentence similarity measure," In 2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1-8, Jul. 2013.

[17] L.A. Zadeh. "Fuzzy logic = computing with words," In Computing with Words in Information/Intelligent Systems 1, pp. 3-23. Physica, Heidelberg, 1999.

[18] J.M. Mendel and R.B. John. "Type-2 fuzzy sets made simple," In IEEE Transactions on fuzzy systems, vol. 10, no. 2, pp. 117-127, 2002.

[19] M. Hao and J.M. Mendel. "Encoding words into normal interval Type-2 fuzzy sets: HM approach," In IEEE Transactions on Fuzzy Systems, vol. 24, no. 4, pp. 865-879, 2015.

[20] J.D. O'Shea, Z.A. Bandar and K. Crocket. "A new benchmark dataset with production methodology for short text semantic similarity algorithms," In ACM Transactions on Speech and Language Processing (TSLP), vol. 10, no. 4, pp. 1-63, Jan. 2014.

[21] SpazioDati, "Dandelion API", [Online], Available: https://dandelion.eu/, Accessed on: 24th Jan. 2020.

[22] V. Rus, M. Lintean, R. Banjade, N.B. Niraula and D. Stefanescu. "Semilar: The semantic similarity toolkit," Proceedings of the 51st annual meeting of the association for computational linguistics: system demonstrations, pp. 163-168, Aug. 2013.

[23] P. Resnik and D. Yarowsky. "Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation," Natural language engineering, vol. 5, no. 2, pp. 113-133, Jun. 1999.

[24] A.W. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain. "Content-based image retrieval at the end of the early years," IEEE Transactions on pattern analysis and machine intelligence, vol. 22, no. 12, pp. 1349-1380, Dec. 2000.

[25] R.K. Srihari, Z. Zhang, and A. Rao. "Intelligent indexing and semantic retrieval of multimodal documents," Information Retrieval, vol. 2, no. 2, pp. 245-275, May 2000.

[26] S.J. Green. "Lexical semantics and automatic hypertext construction," ACM Computing Surveys (CSUR), vol. 31, no. 4es, pp. 22-es, Dec. 1999.

[27] D.L. Medin, R.L. Goldstone and D. Gentner. "Respects for similarity." Psychological review, vol. 100, no.2, p. 254, 1993.

[28] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer and R. Harshman. "Indexing by latent semantic analysis," Journal of the American society for information science, vol. 41, no. 6, Sept. 1990, pp. 391-407.

[29] T.K. Landauer and S. Dumais. "Latent semantic analysis," Scholarpedia, vol. 3, no. 11, 2008, 4356.

[30] Y. Li, Z.A. Bandar and D. McLean. "An approach for measuring semantic similarity between words using multiple information sources," IEEE Transactions on knowledge and data engineering, vol. 15, no. 4, pp. 871-882, Jul. 2003.

[31] N.T. Bhuvan and M.S. Elayidom. "A supervised multimodal search re-ranking technique using visual semantics," International Journal of Intelligent Enterprise, vol. 7, no. 1-3, pp. 279-290, 2020.

[32] T. Luchoomun, M. Chumroo and V, Ramnarain-Seetohul. "A knowledge based system for automated assessment of short structured questions," 2019 IEEE Global Engineering Education Conference (EDUCON), pp. 1349-1352, Apr. 2019.

[33] T. Pedersen, S. Patwardhan and J. Michelizzi. "WordNet:: Similarity-Measuring the Relatedness of Concepts," In AAAI, vol. 4, pp. 25-29, Jul. 2004.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3194510

[34] V. Rus, M. Lintean, C. Moldovan, W. Baggett, N. Niraula, and B. Morgan. "The similar corpus: A resource to foster the qualitative understanding of semantic similarity of texts," Semantic Relations II: Enhancing Resources and Applications, The 8th Language Resources and Evaluation Conference (LREC 2012), pp. 23-25, May 2012.

[35] D. Chandrasekaran and V. Mago. "Evolution of semantic similarity—a survey," ACM Computing Surveys (CSUR), vol. 54, no. 2, pp. 1-37, 2021.

[36] Kent State University, "SPSS Tutorials: Pearson Correlation", [Online] Available: https://libguides.library.kent.edu/SPSS/PearsonCorr, Accessed on: 18th Sept. 2020.

[37] L.A. Zadeh. "Fuzzy logic= computing with words," In Computing with Words in Information/Intelligent Systems 1, pp. 3-23. Physica, Heidelberg, 1999.

[38] V.V. Cross and T.A. Sudkamp, "Similarity and compatibility in fuzzy set theory: assessment and applications," vol. 93. Springer Science & Business Media, 2002.

[39] J.M. Mendel, "Computing with words: Zadeh, Turing, Popper and Occam," In IEEE computational intelligence magazine, vol. 2, no. 4, pp.10-17, Nov. 2007.

[40] J.M. Mendel and R.B. John. "Type-2 fuzzy sets made simple," In IEEE Transactions on fuzzy systems, vol. 10, no. 2, pp. 117-127, Aug 2002.

[41] P.K. Muhuri, P.K. Gupta and J.M. Mendel. "Person Footprint of uncertainty-based CWW model for power optimization in handheld devices," In IEEE Transactions on Fuzzy Systems, vol. 28, no. 3, pp. 558-568, Apr. 2019.

[42] L. Zadeh.. "From computing with numbers to computing with words. From manipulation of measurements to manipulation of perceptions." In IEEE Transactions on circuits and systems I: fundamental theory and applications, vol. 46, no. 1, pp. 105-119, 1999.

[43] M. Hao and J.M. Mendel. "Encoding words into normal interval Type-2 fuzzy sets: HM approach," In IEEE Transactions on Fuzzy Systems, vol. 24, no. 4, pp. 865-879, Oct. 2016.

[44] R.E. Walpole, R H. Myers, S.L. Myers and K. Ye. "Probability and statistics for engineers and scientists," Macmillan New York, 1993.

[45] F. Liu and J.M. Mendel. "Encoding words into interval Type-2 fuzzy sets using an interval approach." In IEEE transactions on fuzzy systems, vol. 16, no. 6, pp. 1503-1521, 2008.

[46] W.N. Francis and H. Kucera. "Brown Corpus Manual (Revised and Amplified)." Department of Linguistics, Brown University, [Online] Available: http://korpus.uib.no/icame/manuals/brown/index.html. Assessed on: 17th Oct. 2017.

[47] D. Chandran, "The development of a fuzzy semantic sentence similarity measure" (Doctoral dissertation, Manchester Metropolitan University), 2013.

[48] Readability Formulas, "Automatic Readability Checker", [Online] Available: https://readabilityformulas.com/free-readability-formula-tests.php, Accessed on: 15th Jan. 2018.

[49] Text Inspector, "What is Readability and How Does It Work?" [Online] Available: https://textinspector.com/what-is-readability-and-how-does-it-work/, Accessed on: 15th Jan. 2018.

[50] Siteimprove, "3 Reasons Readability Should Matter to Content Editors", [Online] Available: https://siteimprove.com/en/blog/what-is-readability-why-should-content-editors-care-about-it/, Accessed on: 15th Dec. 2018.

[51] T.K. Koo and M.Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research." Journal of chiropractic medicine, vol. 15, no. 2, pp.155-163, 2016.

[52] D.V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," In Psychological Assessment, vol. 6, no. 4, p. 284, Dec. 1994.

[53] N. Fenton, and M. Neil, "Risk assessment and decision analysis with Bayesian networks." CRC Press, 2018.

NAEEMEH ADEL has completed her BSc in Computer Science from the University of Salford and MSc in Information systems from Manchester Metropolitan University. She is currently in her final year of her PhD entitled "Fuzzy Natural Language Similarity Measures Through Computing with Words" which she is studying part-time at Manchester Metropolitan University. She is a Tutor teaching Information Systems and Computing Fundamentals at Manchester Metropolitan University as well as unit leader and tutor on the Skills Bootcamps in Digital project funded by the Department of Education . She is a student member of IEEE, member of IEEE CIS and a member IEEE WIE UKI and actively participates in STEM activities and public outreach.

KEELEY CROCKETT (M'06–SM'14) has over 21 years' experience of research and development in computational intelligence algorithms and applications, including adaptive psychological profiling, fuzzy systems, dialogue systems, and educational tutoring systems. She is currently a Professor in Computational Intelligence at Manchester Metropolitan University and co-academic lead for the ERDF funded Greater Manchester AI Foundry. She is active in the development of responsible and trustworthy artificial intelligence, from both a business and educational perspective. She is current Chair of the IEEE Taskforce on Ethical and Social Implications of Computational Intelligence and co-chair of the IEEE Women in Engineering Educational Outreach and a UK STEM Ambassador.

DARIA LIVESEY is a lecturer in Mathematics in the Department of Computing and Mathematics at Manchester Metropolitan University. Her interests are in computer algebra, linear algebra and mathematics education.

JOAO PAULO CARVALHO has a PhD (2002), a MSc (1996) and an Electrical and Computer Engineer (1992) degree from Instituto Superior Técnico, University of Lisbon, Portugal (Técnico Lisboa), where he is currently a Tenured Associate Professor at the Department of Electrical Engineering and Computers. He has lectured courses on Computational Intelligence, Distributed Systems, Entrepreneurship and Technology Transfer, Computer Architectures and Digital Circuits since 1999. He is a researcher and Director of INESC-ID, where he has coordinated 5 funded research projects and participated in more than a dozen national and European funded projects. His current main research interests involve developing and applying new Computational Intelligence techniques to natural language processing, text mining, social network analysis, social sciences and earth sciences. He has authored over 140 papers in international scientific Journals, books and peer-reviewed conferences. He is Area Editor of Fuzzy Sets and Systems and Associate Editor of 2 other international Journals. He was the General Chair of IPMU2020, Program Co-Chair and organizer of IFSA-EUSFLAT2009, Web chair of the 2010 IEEE World Congress on Computational Computation, Publicity-chair of FUZZ-IEEE2015 and FUZZ-IEEE2017, Program-Cahir of IPMU2016, PR and Publicity-chair of IEEE-WCCI2017, and is a PC member of more than 30 international conferences.