Fuzzy Influence in Fuzzy Semantic Similarity Measures

Naeemeh Adel, Keeley Crockett
Department of Computing and Mathematics
Manchester Metropolitan University, Chester Street,
Manchester, M1 5GD.
United Kingdom
N.Adel@mmu.ac.uk;
K.Crockett@mmu.ac.uk

Joao P. Carvalho
INESC-ID
Instituto Superior Tecnico, Universidade de Lisboa.
Portugal
joao.carvalho@inesc-id.pt

Valerie Cross
Computer Science and Software Engineering
Miami University,
Oxford, OH.
USA
crossv@miamioh.edu

***Abstract*: The field of Computing with Words has been pivotal in the development of fuzzy semantic similarity measures. Fuzzy semantic similarity measures allow the modelling of words in a given context with a tolerance for the imprecise nature of human perceptions. In this work, we look at how this imprecision can be addressed with the use of fuzzy semantic similarity measures in the field of natural language processing. A fuzzy influence factor is introduced into an existing measure known as FUSE. FUSE computes the similarity between two short texts based on weighted syntactic and semantic components in order to address the issue of comparing fuzzy words that exist in different word categories. A series of empirical experiments investigates the effect of introducing a fuzzy influence factor into FUSE across a number of short text datasets. Comparisons with other similarity measures demonstrates that the fuzzy influence factor has a positive effect in improving the correlation of machine similarity judgments with similarity judgments of humans.**

Keywords: computing with words, natural language processing, FUSE, semantic similarity

I. Introduction

Similarity measures combine semantic and syntactic features of natural language to determine a similarity measure of two short texts. Short texts are typically 25 words or less in length [1] and include structured (sentences) and unstructured (tweets) [2, 3, 4, 5]. Substantial research has been undertaken in the field of traditional semantic similarity [6], with methods typically grouped into corpus-based [7], string-based [8], knowledge-based [9], and hybrid [1]. Applications cover a wide area, including tweet similarity [3 and 4], fake news detection [10], spam email classification [11], Radicalization Detection Based [12] and determining effective shilling attack strategies in recommendation systems [8]. Traditional similarity measures did not calculate the impact of fuzzy words in the content of the short text.

Zadeh's early work on Computing with Words (CWW) looked at the "exploitation of the tolerance for imprecision" [13] through a methodology designed to bridge the gap between human natural language and logical computation and reasoning. More recent work by Mendel recommended that since "words mean different things to different people", Type-2 and Interval Type-2 fuzzy sets should be used to model their meaning [14] in order to capture word uncertainties. Within natural language processing applications, such as dialogue systems [15], Type-2 and Interval Type-2 fuzzy sets have allowed for improved understanding of how humans use words in different contexts to elicit better machine responses to human utterances. In this work, we define a fuzzy word as a word that has a subjective meaning, is often considered ambiguous, and is based on an individual's perception, within a given context and at a given time. Adopting a hybrid approach based on crisp and fuzzy ontologies and a corpus, FAST [16] was the first fuzzy similarity measure to be developed and evaluated specifically on datasets containing fuzzy words [16]. In FAST, human perception based words were modelled using Type-1 fuzzy sets. The FUSE measure tackled the issue of uncertainty of human judgement [17] by modelling fuzzy words using Interval Type-2 fuzzy sets, originally proposed by Hao and Mendel [17]. FUSE was successfully evaluated and extended to include hedge words in [18].

A weakness of FUSE was that to obtain a similarity measurement of a fuzzy word within short texts, the words had to be within the same fuzzy category in order to determine their distance within the fuzzy category ontology. The category ontologies that were used catered for synonyms of the English language and were extensive, it was found through empirical experimentation that it was not able to measure word similarity directly between words like '*hot*' in the Temperature category and '*large*' in the Size/Distance category. In this case, the word pair (*hot* and *large*) were passed to the generalised WordNet ontology to compute the word pair similarity. Effectively, the fuzziness of the words was lost and not included in the final short text similarity calculation.

The contribution of this paper is to propose an extension to the FUSE measure [19] by the inclusion of a Fuzzy Influence (FI) factor (defined in Section III) into the short text overall similarity calculation. The aim is to ensure that each fuzzy word has an impact on each text, regardless of whether it has a matched pair word in the same fuzzy ontology or not.

Typically, semantic measures usually comprise of two weighted elements; the semantic part and the syntactic part which are optimised against human ratings of similarity. The aim on a training dataset is to obtain a machine based method with the highest correlation to human ratings. In this paper we report the summarised results from a number of empirical experiments where we determine the effect and interaction of FI with the semantic and syntactic features in short texts. We examine the effects across three datasets containing fuzzy words, were human similarity ratings have been obtained. We show that the introduction of a fuzzy influence factor can have a positive effect on the overall sentence similarity of fuzzy sentence similarity measures (FSSM) leading to better

correlation of human ratings when using the Pearson's correlation coefficient.

This paper is organised as follows: Section II briefly describes relevant work in fuzzy semantic similarity measures. Section III describes the Fuzzy Influence factor and how the FUSE algorithm was extended. The experimental methodology along with datasets is described in Section IV. Results and analysis are presented in Section V.

II. Related Work

A. Fuzzy Semantic Similiarty Measures

Fuzzy semantic similarity measures calculate the semantic and syntactic similarity of a short text pair through combining both the syntactic and semantic features of a short text which are weighted. Fuzzy human perception based words were first incorporated into semantic similarity measures in the FAST algorithm. Words (selected through human experimentation) were first selected for 6 categories originally proposed by Zadeh, and were modelled using Type-1 fuzzy sets [16]. Whilst FAST captured the fuzziness of words in a sentence, the modelling of them was still subjective and opinion based. Since FAST, research in the field of Computing with Words, first advocated the use of Type-2 [20] and then later the use of Interval Type-2 fuzzy models in order to model first-order word uncertainties [21].

FUSE_1.0, a more recent fuzzy measure [19], models words using Mendel's Hao-Mendel Approach (HMA) using Interval Type-2 fuzzy sets [17]. Utilising the same 6 categories as FAST, each category was firstly expanded with the number of fuzzy words and 32 English speaking participants were used to score the words in each category on a scale of [0-10]. The data was then cleaned [17], and the footprint of uncertainty (FOU's) for each word was determined. Fuzzy ontologies where then constructed for each category of fuzzy words before being applied in the FUSE measure. These category ontologies were used to compute the similarity of fuzzy word pairs. Non-fuzzy word pairs were passed to the Princeton WordNet – a lexical database of English words, comprising of sets of cognitive synonyms, each related to a distinct concept [22]. FUSE_1.0 was extended further (FUSE_2.0) to include 9 fuzzy categories and applied within a dialogue system [15]. These categories are Size/Distance, Temperature, Age, Frequency, Worth, Level of Membership, Strength, Brightness, Speed. An issue with WordNet is that it is continually updated, and this can effect results generated by any short text similarity measure that uses it. Thus, FUSE versions have evolved over the years. In this paper, we incorporate the proposed Fuzzy Influence Factor into FUSE_4.0 which models words in 9 fuzzy categories and uses the December 2020 version of WordNet [15]. The full pseudo code for the FUSE_1.0 algorithm can be found in [19] and a revised version of the algorithm is currently under review.

B. Evaluation of Semantic Similiarty

Measures that compute semantic similarity of short texts usually require correlations with ratings of similarity given by humans. Over the years, a number of datasets have been published [2, 7] which have adopted methodologies designed to capture unbiased human ratings [2, 7]. Semantic similarity measure results can also be compared against other measures. In this work, we evaluate FUSE_4.0 against 3 other measures, STASIS [23], SEMILAR [24] and the commercial Dandelion API [25]. STASIS measures similarity using an ontological approach based on a taxonomy of words achieved by calculating the distance between words in an ontology, using WordNet, as well as the distance of words to their closest subsumer. SEMILAR [24] (SEMantic simILARity toolkit) utilises the word-to-word semantic similarity measures in the WordNet Similarity library [26] as well as using Latent Semantic Analysis [27]. Dandelion API is a commercial sentence similarity measure which computes the semantic and syntactic components separately [25]. One successful use of Dandelion API is in an Automated Short Answer Scoring within knowledge-based systems [28].

III. Fuzzy Influence

Currently the FUSE_1.0 algorithm calculates the semantic and syntactic similarity of a sentence pair through a weighted combination of analysis on both the syntactic and semantic elements of a short text. A weakness of the approach used in FUSE_1.0 is that it does not take into consideration sentence pairs where fuzzy words are not in the same category; for example comparing the word "*slow*" to "*normal*". While both these words do belong to fuzzy categories (*Speed* and *Worth* respectively), they do not fall in the same fuzzy category and so WordNet is used to derive their values. Several variants of FUSE have been developed; for example, FUSE_3.0 uses 9 categories of fuzzy words and the WordNet 2019 version [15].

A. Fuzzy Infuence

In this work, we propose the addition of a fuzzy influence factor (FI) within the FUSE algorithm. FI overcomes a weakness of FUSE by ensuring fuzzy words not in the same fuzzy categories but within the same sentence have a human associated impact on determining the sentence's similarity. The *FI* for a sentence pair *sn*, can be defined as:

$$FI_{sn} = \frac{1}{n-i} \tag{1}$$

where n is the number of all the words in the sentence pair sn; and $n > 0$, and i is the count of all the fuzzy words in sn. If all the words in the sentence pair are fuzzy, i.e. $n = i$, we set $FI_{sn} \coloneqq 1$, and so FI_{sn} takes values between 0 and 1. FI is applied to all sentence pair calculations within FUSE, regardless of whether fuzzy words are in the same category or not. In [19], the FUSE algorithm was first proposed to calculate the overall similarity between two fuzzy utterances, U_1 and U_2, through the weighted addition of syntactic and semantic components. In FUSE_4.0, the overall similarity of S(U1, U2) is then calculated as:

$$S(U_1, U_2) = sem\_sim * w1 + syn\_sim * w2 + FIsn * w3 \tag{2}$$

where $w1, w2, w3 \in [0..1]$ *and* $\sum w1..w3 = 1$, and sem_sim and syn_sim are calculated using pairs of semantic and syntactic similarity vectors which were determined by a word similarity measure and a short joint word vector set comprising of word frequency information and word order. See [19] for full definitions.

IV. Experimental Methodology

To investigate the relationships between the semantic, syntactic and FI components, an empirical experiment was conducted for FUSE_2.0 to see if the introduction of a fuzzy influence factor will affect the overall sentence similarity rating to give a value closer to that of the human ratings (HR). The hypothesis for this experiment is given below:

H_0 = The inclusion of a fuzzy influence factor (FI) in the calculation of the overall semantic similarity of a sentence improves the overall correlation when compared to human ratings.

A. Metrics

In each set of experiments, three metrics (semantic, syntactic and fuzzy influence factor) are used to measure the effectiveness of variants in the FI factor within FUSE_4.0. The Pearson's correlation coefficient is used to show statistical evidence for a linear relationship between two variables *x* and *y* in this work between the human ratings and those generated by FUSE_4.0 and is defined as [29]:

$$r_{xy} = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)}.\sqrt{\text{var}(y)}} \quad (3)$$

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*.

The reliability of inter-rater agreement of human ratings of short text pairs across a population is conducted using Cicchetti's approach [30] which uses the intra-class correlation coefficient (*a-value*). The following guidelines are followed for the interpretation of inter-rater agreement measure (*a-value*) [30]:

- *a-value* < 0.40 - Poor.
- 0.40 >= *a-value* <= 0.59 - Fair.
- 0.60 >= *a-value* <= 0.74 - Good.
- 0.75 >= *a-value* <= 1.00 - Excellent.

The *a-value* is important as it shows the extent to which the data, that is collected for this study, is a correct representation of the variables measured; therefore, the aim is to achieve an *excellent* rating to maximise reliability of the human ratings of the short text pairs [31, 32].

B. DataSets

In this work three datasets, FI-25, 62-SP, and the Multiple Fuzzy Word Dataset (MFWD) were used to investigate the fuzzy influence factor. Initial work was undertaken on a dataset known as FI-25 which comprised of a set of 25 test sentences (with inclusion criteria defined below) which have been randomly sampled from 3 existing datasets, STSS-131 [2], 62-SP [15] and MFWD [34]. SP1 – SP15 of the Sentence Pairs (SP) in FI-25 consisted of poor human rating correlations when run on FUSE_1.0. These poor human rating correlations imply that automated semantic similarity measurement of FUSE_1.0 were far different than that of the average human ratings. Ideally we would like similarity values derived from the measure to be as close to the human ratings as possible. The remaining 10 pairs (SP16 – SP25) gave high correlations with human ratings by FUSE_1.0. This means that the ratings were close to that given by the human raters. This dataset was created to ensure that the impact of the fuzzy influence factor was assessed against both high and low correlations. The general methodology for collecting human ratings can be found in [19]. The 62-SP dataset was specifically designed by English language experts to contain fuzzy words from all 9 categories from the FUSE fuzzy dictionary. The origin of the sentences came from a gold standard dataset STSS-131 [2] which contained 131 crisp sentence pairs. 62 random sentence pairs were extracted from this dataset and fuzzy words from each of the 9 fuzzy categories were placed in each sentence pair using English language experts to ensure the sentences were still meaningful. A constraint on the randomisation was to ensure there were an equal number of sentence pairs in the low, medium and high categories, as identified by human participants in previous published studies [2]. This meant that each sentence had at least 2 fuzzy words. The reader's age for this dataset has been calculated as 14-15 years old (Ninth to Tenth graders) using the Automatic Readability Checker [33]. Finally, the Multiple Fuzzy Word Dataset (MFWD) [34] contains 30 sentence pairs where each sentence contains more than one fuzzy word.

V. Experimental Results and Discussion

For each of the experiments in this section the following experimental methodology was followed. The semantic, syntactic and FI weights were each separately changed using increments of 0.05 between the ranges of 0 and 1. In each case one of the weights was fixed, whilst the other pair were changed to ensure the sum of all weights was always 1. At each iteration, Pearson's correlation was recorded each time to see which values gave the best results.

A. Experiment 1 - FI on FI-25

The FI factor within FUSE_4.0 was used with a range of different empirical weighting values for the semantic, syntactic and fuzzy influence factor to see which gave the sub-optimal results. Due to space, only a range of empirical values are reported in this paper. Optimal results are calculated by comparing Pearson's correlation (*r-value*) with human ratings. The higher the *r-value,* the closer the ratings to those of humans. In F1-25, the correlation was calculated for both the "bad" performing sentence pairs (NPW) (Table I) as well as the "good" performing sentence pairs (PW) (Table III), where bad and good results were generated by FUSE_1.0. Pearson's correlation of FUSE_FI is also compared with those of several earlier versions of FUSE as well as 4 other similarity algorithms that do not cater for fuzzy words: STASIS, which is a similarity measure using WordNet [23], SEMILAR [24], Dandelion API Semantic [25] and Dandelion API syntactic [25].

Table I shows the correlation findings for experiments 1.1-1.5 ran on the sentences that were not performing well under FUSE_1.0. Results from Table I show that the measures (Sem 0.5, Syn 0.2, FI 0.3) from experiment 1.5 gave the best overall correlation and the highest correlation, beating the other algorithm measures with the exception of API Syn for SP's that did not perform well originally with FUSE_1.0 as shown in Table II. The higher the correlation, the closer the similarity ratings are to those of the human ratings (HR).

Figure 1 shows a scatter plot for the relationship between the two variables [35]. In this instance, the two variables are the human ratings (HR) and the correlation following the fuzzy influence factor (FI) experiment. Each dot on the scatter plot shows the values for each sentence pair on the X and Y axis, with *x* being FI and *y* being HR. The scatter plot in Figure 1 shows the positive correlation of the human ratings (HR) with the fuzzy influencer factor (FI), each on a scale of [0..1], where 0 represents no similarity and 1 represents maximum similarity. For experiment 1.5 for the 15 sentence pairs that did not perform well under FUSE_1.0, the line of best fit shows the mathematically best fit for the data; also referred to as the 'trendline'. This line shows the behaviour of a set of data, when the line goes up, this shows a positive linear relationship between the variables.

SP15 where SP15a = *"The little village of Resina is also situated near the spot"* and SP15b = *"He seems an excellent man and I think him uncommonly pleasing"*, is a clear outlier with the average human rating being 0.075, where FUSE_4.0 coming close to 0.206. SP15 contains fuzzy words {*little*, *near*, *excellent* and *uncommonly*}, *little* and *near* belong to the *Size/Distance* category, *excellent* belongs to *Worth* category and *uncommonly* belongs to *Frequency* category.

Table III shows the correlation findings for experiments 1.6-1.10 ran on the sentences that performed well under FUSE_1.0. Results from Table III show that the component weightings (Sem 0.7, Syn 0.05, FI 0.25) from experiment 1.9 gave the best overall correlation and the highest correlation, beating the other algorithm measures for SP's that performed well originally with FUSE_1.0 as shown in Table IV. The higher the correlation, the closer the similarity ratings are to those of the human ratings (HR). Figure 2 shows the scatter plot for the positive correlation of the human ratings (HR) with the fuzzy influencer (FI) for experiment 1.9 for the 10 sentence pairs that performed well under FUSE_1.0. The trendline shows a positive linear relationship between the variables. The results from these experiments on the FI-25 dataset gave positive indicators that H_0 would be accepted.

B. Experiment 2 - FI on 62-SP

FI-25 was a limited dataset, so a series of further empirical experiments were undertaken using a similar range of semantic, syntactic and FI factor weights using the 62-SP dataset. 62-SP consisted of 62 sentence pairs specifically designed by English language experts to contain at least 2 fuzzy words per sentence from all 9 categories [15]. Table V shows the correlation findings for experiments 2.1-2.5 ran on the 62-SP dataset. Results from Table V show that the measures (Sem 0.5, Syn 0.2, FI 0.3) from experiment 2.5 gave the best overall correlation with human ratings and also higher than competing measures as shown in Table VI. The scatter plot in Figure 3 shows the positive correlation of the human ratings (HR) with the fuzzy influencer (FI) for experiment 2.5 for the 62-SP dataset.

C. Experiment 3 - FI on MFWD

The same 5 experiments were also conducted on the published MFWD dataset [34]. This dataset consisted of 30 sentence pairs specifically designed by English language experts to contain at least 2 fuzzy words per sentence. Table VII shows the correlation findings for experiments 3.1-3.5 for the MFWD dataset. Results shown in Table VII show that the measures (Sem 0.8, Syn 0.1, FI 0.1) from experiment 3.1 gave

TABLE I. RESULTS FROM SELECTED HYPER-PARAMETER OPTIMISATION FOR FI-25 SP'S NOT PERFORMING WELL UNDER FUSE_1.0

Pearson Correlation	**r Value** **Exp. 1.1** **Sem 0.8** **Syn 0.1** **FI 0.1**	**r Value** **Exp. 1.2** **Sem 0.7** **Syn 0.1** **FI 0.2**	**r Value** **Exp. 1.3** **Sem 0.75** **Syn 0.15** **FI 0.1**	**r Value** **Exp. 1.4** **Sem 0.7** **Syn 0.05** **FI 0.25**	**r Value** **Exp. 1.5** **Sem 0.5** **Syn 0.2** **FI 0.3**
HR vs FUSE_4.0	0.695292	0.706837	0.717356	0.687299	0.771050

TABLE II. COMPARISON OF SSM BEST RESULTS FROM TABLE I

SSM	**r Value**
HR vs FUSE_4.0	0.771050
HR vs FUSE_2.0	0.681673
HR vs FUSE_3.0	0.706030
HR vs STASIS	0.712598
HR vs API Semantic	0.495320
HR vs API Syntactic	0.883992
HR vs SEMILAR	0.765862

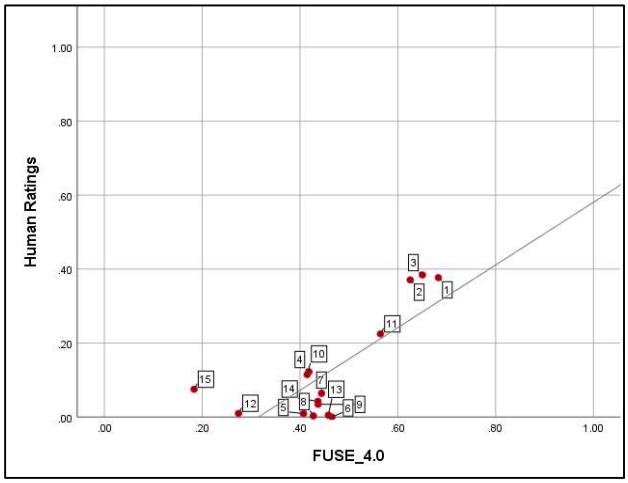

Fig. 1. FI-25 Scatter Plot NPW **(Sem 0.5, Syn 0.2, FI 0.3)**

TABLE III. RESULTS FROM SELECTED HYPER-PARAMETER OPTIMISATION FOR FI-25 SP'S PERFORMING WELL UNDER FUSE_1.0

	r Value	r Value	r Value	r Value	r Value
Pearson Correlation	**Exp. 1.6 Sem 0.8 Syn 0.1 Fl 0.1**	**Exp. 1.7 Sem 0.7 Syn 0.1 Fl 0.2**	**Exp. 1.8 Sem 0.75 Syn 0.15 Fl 0.1**	**Exp. 1.9 Sem 0.7 Syn 0.05 Fl 0.25**	**Exp. 1.10 Sem 0.5 Syn 0.2 Fl 0.3**
HR vs FUSE_4.0	0.249668	0.233447	0.187771	0.299713	0.082649

TABLE IV. COMPARISON OF SSM BEST RESULTS FROM TABLE III

SSM	r Value
HR vs FUSE_4.0	0.299713
HR vs FUSE_2.0	0.191413
HR vs FUSE_3.0	0.205204
HR vs STASIS	0.167745
HR vs API Semantic	0.051874
HR vs API Syntactic	0.073051
HR vs SEMILAR	0.128564

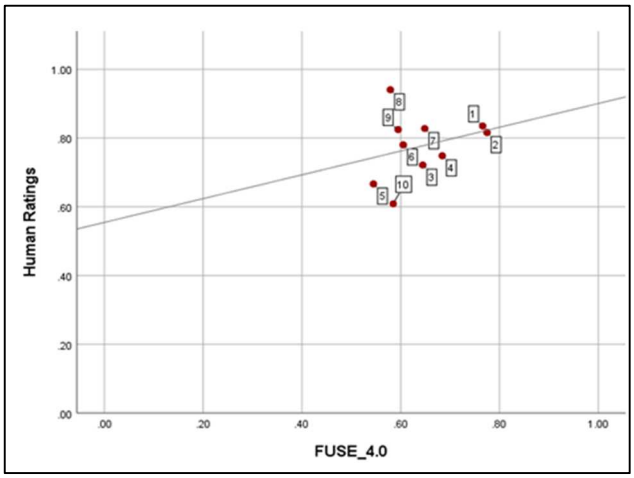

Fig. 2. FI-25 Scatter Plot PW **(Sem 0.7, Syn 0.05, Fl 0.25)**

TABLE V. RESULTS FROM SELECTED HYPER-PARAMETER OPTIMISATION FOR 62-SP

	r Value	r Value	r Value	r Value	r Value
Pearson Correlation	**Exp. 2.1 Sem 0.8 Syn 0.1 Fl 0.1**	**Exp. 2.2 Sem 0.7 Syn 0.1 Fl 0.2**	**Exp. 2.3 Sem 0.75 Syn 0.15 Fl 0.1**	**Exp. 2.4 Sem 0.7 Syn 0.05 Fl 0.25**	**Exp. 2.5 Sem 0.5 Syn 0.2 Fl 0.3**
HR vs FUSE_4.0	0.622094	0.642160	0.646160	0.625525	0.702729

TABLE VI. COMPARISON OF SSM BEST RESULTS FROM TABLE V

SSM	r Value
HR vs FUSE_4.0	0.702729
HR vs FUSE_2.0	0.555268
HR vs FUSE_3.0	0.626043
HR vs STASIS	0.592999
HR vs API Semantic	0.526305
HR vs API Syntactic	0.671170
HR vs SEMILAR	0.664572

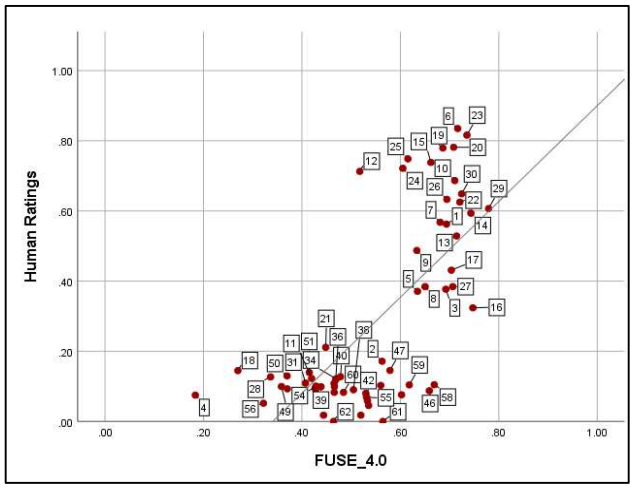

Fig. 3. 62-SP Scatter Plot **(Sem 0.5, Syn 0.2, Fl 0.3)**

TABLE VII. RESULTS FROM SELECTED HYPER-PARAMETER OPTIMISATION FOR MFWD

Pearson Correlation	r Value Exp. 3.1 Sem 0.8 Syn 0.1 FI 0.1	r Value Exp. 3.2 Sem 0.7 Syn 0.1 FI 0.2	r Value Exp. 3.3 Sem 0.75 Syn 0.15 FI 0.1	r Value Exp. 3.4 Sem 0.7 Syn 0.05 FI 0.25	r Value Exp. 3.5 Sem 0.5 Syn 0.2 FI 0.3
HR vs FUSE_4.0	0.758884	0.741024	0.755944	0.734019	0.693317

TABLE VIII. COMPARISON OF SSM FROM BEST RESULTS FROM TABLE VII

SSM	r Value
HR vs FUSE_4.0	0.758884
HR vs FUSE_2.0	0.753772
HR vs FUSE_3.0	0.768331
HR vs STASIS	0.745248
HR vs API Semantic	0.700868
HR vs API Syntactic	0.393033
HR vs SEMILAR	0.730265

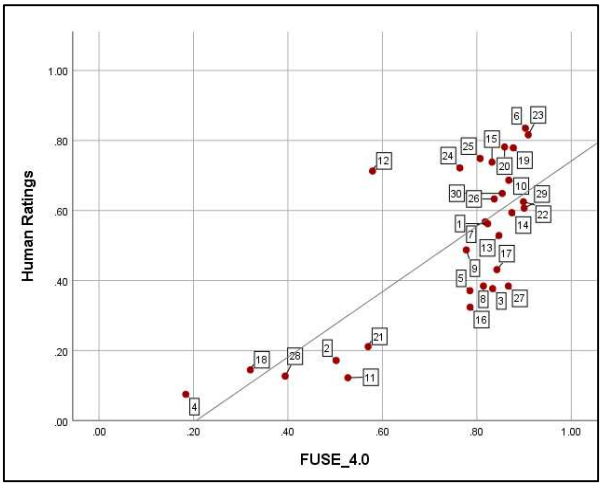

Fig. 4. MFWD Scatter Plot **(Sem 0.8, Syn 0.1, FI 0.1)**

TABLE IX. (A-VALUE) AND (P-VALUE) FOR EACH DATASET

Datasets	FI25_NPW	FI25_PW	FI25	62-SP	MFWD
a-value	0.998	0.953	0.997	0.987	0.999
p-value	.000	.000	.000	.000	.000

the best overall correlation and the highest correlation beating the other algorithm measures with the exception of FUSE_3.0 which was slightly higher as shown in Table VIII. The scatter plot in Figure 4 shows the positive correlation of the human ratings (HR) with the fuzzy influencer (FI) for experiment 3.1 for the MFWD dataset.

VI. DISCUSSION

Table IX shows information with regards to the datasets that were used in the FI experiment. The *a-value* shows the Intra-class Correlation Coefficient (ICC) for each of the datasets that we experimented on across the different algorithms. Since the *a-value* results are between 0.75 and 1.00 for each dataset, it is deemed that the inter-rater agreement of human ratings are *excellent* according to Cicchetti [30]. Table IX also shows that the *p-value* for each dataset is less than 0.05 for a confidence level of 95% and thus provides support for our research hypothesis H_0.

The snapshot of empirical experiments conducted on several datasets indicated that the inclusion of a FI factor in a FSSM can improve the performance of the algorithm in terms of its correlation with human ratings. The interaction of the FI factor with both the semantic and syntactic components of FUSE_4.0 must be kept to a minimum, in order to preserve the importance of the word order and ontological path length in calculating the overall similarity. This work, whilst accepting H_0, recognizes that more work needs to be done in determining a more generalizable FI factor.

VII. CONCLUSION AND FUTURE WORK

In closing, this work has shown that a fuzzy influence factor has a positive impact on the correlation of human ratings in a FSSM. Experimental results in this paper have shown that the FI factor must be empirically determined. The results across 3 datasets have shown an *excellent* rating for ICC. Although this FI is relatively simple, it has to a degree been able to model the uncertainty of human perception-based words which have already been modelled using Interval Type-2 fuzzy sets. The FUSE algorithm can show distinct benefits over crisp semantic similarity algorithms only when there is at least one fuzzy word in the short text pair. Therefore, the FUSE algorithm is recommended when it is important to assess the similarity of fuzzy words in a given context.

Further work includes investigating the generalisability of the FI factor and modelling fuzzy logic operators, such as NOT within the context of fuzzy short text similarity.

ACKNOWLEDGMENT

This work was partially supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020.

REFERENCES

[1] Y. Li, Z.A. Bandar, J.D. O'Shea, D. Mclean 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, 2006.

[2] J.D. O'Shea, Z.A. Bandar and K. Crockett, "A new benchmark dataset with production methodology for short text semantic similarity algorithms", *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 10, no. 4, p.19, 2013.

[3] P. Zhang, X. Huang and L. Zhang, "Information mining and similarity computation for semi- / un-structured sentences from the social data", *Digital Communications and Networks*, 2020.

[4] C. Little, D. Mclean, K. Crockett and B. Edmonds, "A Semantic and Syntactic Similarity Measure for Political Tweets". *IEEE Access*, vol. 8, pp.154095-154113, 2020.

[5] N. Alnajran, K. Crockett, D. McLean and A. Latham, "An Empirical Performance Evaluation of Semantic-Based Similarity Measures in Microblogging Social Media" In *2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)*, pp.126-135, Dec. 2018.

[6] D.W. Prakoso, A. Abdi, and C. Amrit, "Short text similarity measurement methods: a review", Soft Computing, pp.1-25, 2021.

[7] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 2. no. 2, pp.1-25, 2008.

[8] V.W. Anelli, Y. Deldjoo, T. Di Noia, E. Di Sciascio and F.A. Merra, Sasha: Semantic-aware shilling attacks on recommender systems exploiting knowledge graphs. In *European Semantic Web Conference* ,pp. 307-323, May. 2020.

[9] H.J.P.J. Dong-hong, "Convolutional Network-Based Semantic Similarity Model of Sentences", *Journal of South China University of Technology (Natural Science)*, vol. 45, no. 3, p.68, 2017.

[10] X. Zhang, and A.A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion" *Information Processing & Management*, vol. 57, no. 2, p.102025, 2020.

[11] S. Venkatraman, B. Surendiran and P.A.R. Kumar, "Spam e-mail classification for the Internet of Things environment using semantic similarity approach", *The Journal of Supercomputing*, vol. 76, no. 2, pp.756-776, 2020.

[12] O. Araque and C.A. Iglesias, "An Approach for Radicalization Detection Based on Emotion Signals and Semantic Similarity", *IEEE Access*, vol. 8, pp.17877-17891, 2020.

[13] L.A. Zadeh, "Fuzzy logic = computing with words", *Fuzzy Systems, IEEE Transactions on*, vol. 4, no. 2, pp.103-111, 1996.

[14] J.M. Mendel, "Type-2 Fuzzy Sets as Well as Computing with Words", In *IEEE Computational Intelligence Magazine*, vol. 14, no. 1, pp.82-95, Feb. 2019.

[15] N. Adel, K. Crockett, D. Chandran and J.P. Carvalho, "Interpreting Human Responses in Dialogue Systems using Fuzzy Semantic Similarity Measures", *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp.1-8, 2020.

[16] K. Crockett, D. Chandran and D. Mclean, "On the Creation of a Fuzzy Dataset for the Evaluation of Fuzzy Semantic Similarity Measures", *IEEE WCCI – FUZZ*, China, pp.752-759, 2014.

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

[18] N. Adel, K. Crockett, A. Crispin, J.P. Carvalho and D. Chandran, Human Hedge Perception–and its Application in Fuzzy Semantic Similarity Measures. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1-7, 2019.

[19] N. Adel, K.A. Crockett, A. Crispin, D. Chandran and J. Carvalho, "FUSE (Fuzzy Similarity Measure) - A Measure for Determining Fuzzy Short Text Similarity Using Interval Type-2 Fuzzy Sets", *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp.1-8, 2018.

[20] A. Bilgin, H. Hagras, A. Malibari, M.J. Alhaddad, and D. Alghazzawi, "Towards a general type-2 fuzzy logic approach for computing with words using linear adjectives", *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE),* pp.1-8, 2012.

[21] J.M. Mendel and D. Wu, "Perceptual Computing: Aiding People in Making Subjective Judgments", John Wiley & Son, 2010.

[22] Princeton University, "About Wordnet", [Online], Available: https://wordnet.princeton.edu/ [Accessed 13 Jun. 2014].

[23] 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, 2003.

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

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

[26] T. Pedersen, S. Patwardhan and J. Michelizzi, "WordNet:Similarity - Measuring the Relatedness of Concepts", In *The Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, pp.1024-1025, Jul. 2004.

[27] 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, pp.391-407, 1990.

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

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

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

[31] M.L. McHugh, Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3), pp.276-282, 2012.

[32] Statistics How To, "Intraclass Correlation", [Online] Available https://www.statisticshowto.com/intraclass-correlation/#:~:text=Intraclass%20correlation%20measures%20the%20reliability,groups%20or%20sorted%20into%20groups.&text=A%20high%20Intraclass%20Correlation%20Coefficient,values%20from%20the%20same%20group, [Accessed 11 Jan. 2021].

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

[34] D. Chandran, "The development of a fuzzy semantic sentence similarity measure", *Doctorate of Philosophy*, School of Computing, Maths and Digital Technology, Manchester Metropolitan University (MMU), 2013.

[35] Chartio | Data Tutorials, "A Complete Guide to Scatter Plots", [Online] Available https://chartio.com/learn/charts/what-is-a-scatter-plot/, [Accessed 11 Jan. 2021].