740-1720: Richa Sharma, Simrat Deol, Udit Kaushish, Prakher Pandey, Vishal Maurya

DWAEF: a deep weighted average ensemble framework harnessing novel indicators for sarcasm detection

RESPONSES TO REVIEWERS' COMMENTS

We the co-authors of the paper #740-1720 thank the learned editors and reviewers for their insightful feedback. All comments of the reviewers have been written in BLACK, the responses of the authors in GREEN and the text from the manuscript in BLUE font colours.

Responses to comments of reviewer 1:

Abstract
"The encouraging findings produced after applying DWAEF demonstrate that the proposed method surpasses the outcomes of previous research that made use of primitive features" should more clearly state what exactly is the improvement in terms metrics used for evaluation or appropriate statistics.

Response: The abstract has been modified as follows:

"A comparative report between the results acquired using primitive features and those obtained using a combination of primitive features and proposed indicators is provided. The highest accuracy of 92% was achieved after applying DWAEF, the proposed framework which combines the primitive features and novel indicators together as compared to 78.58% obtained using Support Vector Machine (SVM) which was the lowest among all classifiers."

What dataset is the GNN framework pre-trained on?

Response: In section 4.4.1, in the graph construction module, the authors have mentioned about the pre-training of the GNN framework as follows:

"The graph construction module focuses on building a syntactic dependency tree-based static graph for each of the texts in the dataset. All of the dependency trees are built using Stanford NLP Group's CoreNLP server. Dependency relations from the dependency parsing trees are converted into dependency graphs. 3000 sentences consisting of both similes and non-similes were collected for pre-training the GNN framework. Pre-training is done to improve the framework's ability to recognize similes in the combined dataset."

The proposed data is combination of Twitter data, News headlines and SARC dataset. Did the authors check the results for verifying that this did not affect the results? The dataset used is from different sources like Twitter, New headlines and SARC dataset. Are these under the same domain or same topics? Does domain have effect on the results? What is the authors view on using different data set for the same experiments?

Response: The subsection 4.1 (Dataset preparation) has been modified to address these comments as follows:

"The dataset utilized in this study was sourced from various platforms, including Twitter, News Headlines, and the SARC dataset. These sources do not belong to the same domain or topic. Twitter data, for instance, encompasses a broad range of subjects, whereas news headlines may concentrate on specific areas, such as politics, sports, or entertainment. The domain can significantly affect the outcomes of sarcasm detection because the use of sarcasm can vary depending on the context and subject matter. For example, the prevalence of sarcasm in political discussions may differ from its occurrence in conversations about sports or entertainment. The motivation for combining data from various online sources such as Twitter, Reddit, and News Headlines is to expose the model to different writing styles as well as myriad domains prevalent online. Moreover, the proposed features in this study operate independently of the domain. Hence, there are no visible effects on the results obtained. This approach is therefore expected to help the model make accurate predictions for a wide range of text messages."

Comparison made between only using primitive features and combination of primitive and proposed features:

1. Are the experiments using the same settings and parameters?

Response: The subsection 5.4 (Comparison of results obtained using primitive features vs. DWAEF) has been modified as follows:

"The hyperparameter settings in both experiments were kept the same to ensure that any observed differences in performance were due to the changes in the feature set rather than differences in hyperparameters. In each case, combining primitive features with the proposed features yielded superior outcomes."

2. The base experiments reference is not provided. Provide references for the corresponding papers for readers.

Response: Though many studies exist in the literature that used varied subsets of primitive features in their experiments, that have been cited in section 2 (Related work), yet the authors were unable to find a single study that incorporated all primitive features in their experiments. This motivated the authors of this study to combine most of these features while conducting their experiment. The results of this experiment were then compared with the results obtained by the experiment conducted using DWAEF, the proposed framework.

The dataset size is approx. 2889 instances. What would be the impact on the system using a big dataset?

Response: We thank the reviewer for enlightening us about this aspect. The following has been added to section 6 (Conclusion and future work).

"In future, the researchers plan to investigate the use of larger datasets for sarcasm detection in deep ensemble framework, while considering potential drawbacks such as increased time and complexity for training and a lack of annotation resources. Efficient resource management strategies such as unsupervised learning techniques and pre-trained models, and crowd sourcing model for annotation shall be explored to mitigate these challenges. The optimal dataset size will depend on the specific application and available resources, and further

research should identify the trade-offs and limitations associated with using larger datasets for sarcasm detection."

Responses to comments of reviewer 2

Response: No comments to be addressed

Responses to comments of reviewer 3

-- the related work on sarcasm is a bit narrow, some foundational work is completely missing

Response: Section 2 (Related Work) has been modified and more work along with the references added to the section.

-- uncertainty about the dataset used in this research (how was it collected? In which period? Which keywords were used? Which data is used? What are the characteristic of the dataset, beside its genre and source?)

Response: The subsection 4.1 (Dataset preparation) has been modified to address the above concerns:

"The researchers of this work prepared a dataset of 2,891 sentences written in English. The dataset utilized in this study was sourced from various platforms, including Twitter, News Headlines, and the SARC datasets. TextBlob [32] was used to check whether the text was written only in English. Out of these, 1538 were sarcastic and were compiled from various sources- i) 520 sentences were extracted from Twitter with hashtags- #sarcasm, #not, #sarcastic, #irony, #satire between the time period of June 2022-October 2022; ii) 520 were taken from the News Headlines dataset curated by [33]; iii) remaining 498 were taken from the SARC dataset curated by [34]. The 1,353 non-sarcastic sentences were compiled from Twitter and the News Headlines dataset."

-- the fact that sarcasm is a language/culture dependent phenomenon is never mentioned in the paper (which is a fundamental characteristic)

Response: We thank the learned reviewer for making us aware about this crucial aspect of sarcasm. The following additions have been made to Section 1 (Introduction) and Section 6 (Conclusion and Future Work).

Introduction

"Moreover, sarcasm is closely tied to linguistic and cultural norms and practices, and its use can vary significantly between different languages and cultures. While sarcasm has clear definitions in the literature, its interpretation may be greatly affected by the cultural background and contextual knowledge of the reader [2] [3]. Sarcasm can be considered rude or disrespectful in some cultures while being accepted as a form of humour in others."

Conclusion and Future Work

"While working in the domain of sarcasm detection, the authors came across other crucial perspectives regarding sarcasm. It was sensed that the interpretation of sarcasm can also be influenced by cultural and linguistic factors, and people from different cultures may interpret sarcasm differently. Different languages may have their own unique styles and patterns of sarcasm, which may not be easily translatable or understandable to those unfamiliar with the language. Therefore, cultural and linguistic contexts are crucial when studying or detecting

sarcasm. Sarcasm detection models that work well in one language or culture may not perform well in another, highlighting the importance of considering these factors in model design and evaluation. Hence, in future, the authors plan to incorporate more advanced tools in the DWAEF framework and fine-tune it to perform cross-lingual, cross-cultural and multimodal predictions."

Further comments:

-- In the first page, since the authors all belong to the same affiliation, its name could be written only once

Response: The LaTex template settings restricted us to write in this manner. The editorial team may guide us in changing the format of the affiliation.

-- technical details for the implementation of the models could be moved to the appendix

Response: The initialization preprocessing steps of GNN framework have been moved to appendix B1. The fuzzy rules set have been moved to appendix B2.

-- In section 5.1 the authors mention a dataset of 3000 similes used for pre-training models. More details or a footnote with a link/URL are needed.

Response: The details and the link to the dataset have been included in appendix B1.

-- the paper could be further revised for some typos (even though not in a preponderant measure)

Response: The authors have taken utmost care to proofread the manuscript and rectify all typos and grammatical mistakes.

-- the figures need to be adjusted for inclusivity (readability colorblind measures) --> https://www.ascb.org/science-news/how-to-make-scientific-figures-accessi...

Response: The authors have modified all 10 images to ensure inclusivity, (readability colorblind measures).

Responses to comments of reviewer 4:

Reasons to reject:

* The literature review is not systematic in a useful way. (See my notes below.)

Response: The authors have revised and modified the organization of the literature review.

* It is assumed the reader is highly familiar with fuzzy logic, and I strongly doubt many will be. I was simply unable to follow this section. The authors need to provide much more background to make this work for the average reader.

Response: The following additions have been made to sub-subsection 4.4.3 for further clarity on fuzzy logic.

"Fuzzy logic is a form of multi-valued logic that deals with reasoning that is approximate rather than precise. In fuzzy logic, a concept can possess a degree of truth (degree of membership) denoted by 'μ' anywhere between 0.0 and 1.0, unlike in standard logic where a concept can only be completely true (1.0) or false (0.0). Fuzzy logic is useful for reasoning about inherently

vague concepts, such as "tallness". In traditional set theory, an element either belongs to a set or not, but in fuzzy logic, an element can partially belong to a set, with a degree of membership ranging betwwen 0 and 1. In general, the degree of membership is used to represent the uncertainty or fuzziness in a concept or variable and is a key component in fuzzy logic's ability to reason with imprecise or uncertain information. In fuzzy logic systems, a membership function is used to map this degree of membership of an element to a set, based on a specified set of fuzzy rules. Trapezoidal membership functions are one way to define fuzzy sets. They are a type of membership function that allows for more flexibility in defining the shape of the fuzzy set.

A trapezoidal membership function is a piecewise linear function that has four parameters: a, b, c and d, where a $\le$ b $\le$ c $\le$ d. The function starts at 0 for x $\le$ a, increases linearly from 0 to 1 from a to b, stays at 1 from b to c, decreases linearly from 1 to 0 from c to d, and stays at 0 for x $\ge$ d. By using trapezoidal membership functions, non-polygonal fuzzy sets can be defined that have smooth transitions between membership degrees. The shape of the membership function can be controlled by adjusting its parameters, allowing for the creation of fuzzy sets that match the intended intuition. For instance, a trapezoidal membership function can be used to define the fuzzy set "tall person" with parameters such as a=170cm, b=175cm, c=185cm and d=190cm. The resulting function would have a membership degree of 0 for heights below 170cm and above 190cm, a membership degree of 1 for heights between 175cm and 185cm, and smoothly increasing and decreasing membership degrees for heights between 170cm and 175cm and between 185cm and 190cm. Trapezoidal membership functions allow for the generation of non-polygonal fuzzy sets, which enable more precise and flexible modelling of linguistic variables in fuzzy logic.”

* Discussion of the ensembling method is very sketchy and unsourced. I have no clue how the ensembling procedure works from the description, nor do I have any hint how I'd go about learning more about it.

Response: The subsection 4.5 (The Deep Weighted Average Ensemble Framework) has been modified to include a detailed explanation on how the weighted average ensembling procedure works. The modified version is as follows:

“The framework is implemented with the help of DeepStack Library [47]. Initially, the curated dataset is used to pretrain the three models. During training, each of the base learners receives the outputs of the GNN-based simile detection framework, the BERT-based metaphor detection framework, the fuzzy based polarity shift detection framework as well as the primitive features as its inputs. Next, the predictions produced by each module of the ensemble are weighed based on their performance on a hold-out validation dataset. This is represented by a Dirichlet ensemble object. To achieve more accurate solution, a weight optimization search is used in conjunction with randomized search based on the Dirichlet distribution on the validation dataset. Randomized search based on the Dirichlet distribution involves randomly sampling the weights from a Dirichlet distribution, which generates a probability distribution over the weights that is continuous and flexible. Weight optimization search iteratively adjusts the ensemble model weights to maximize accuracy on the validation dataset. Through this process the base learners, a TabNet [46], a 1-D CNN and an MLP, are assigned optimal weights. These weights determine the contribution of each model to the

final prediction. Subsequently, the predictions of each model are combined through a weighted average. No meta-learner is used in this ensemble method. Lastly, the ensemble model is assessed on a separate test set to determine how well it can perform on new data."

* Discussion of the basic facts of the data set, including even what language it is written in (!), are missing. No information are provided about how it was annotated except by "four expert linguists": did they double-annotate and perform consensus? What's the interannotator agreement? This sort of information seems necessary to even consider that the authors' data set might be valid.

Response: The authors have modified subsection 4.1 (Dataset Preparation) to include the basic facts of the data set. This includes- the language in which the dataset is written, measures taken to ensure it was predominantly monolingual English and information on how it was annotated. The modified subsection is as follows:

"Description of the dataset: The authors of this work prepared a dataset of 2,891 sentences written in English. The dataset utilized in this study was sourced from various platforms, including Twitter, News Headlines and the SARC datasets. TextBlob [36] was used to check whether the text was written only in English. Out of these, 1,538 were sarcastic and were compiled from various sources- i) 520 sentences were extracted from Twitter with hashtags- #sarcasm, #not, #sarcastic, #irony, #satire between the time period of June 2022-October 2022; ii) 520 were taken from the News Headlines dataset curated by [37]; iii) remaining 498 were taken from the SARC dataset curated by [38]. The 1,353 non-sarcastic sentences were compiled from Twitter and the News Headlines dataset. These sources do not belong to the same domain or topic. Twitter data, for instance, encompasses a broad range of subjects, whereas News Headlines may concentrate on specific areas, such as politics, sports, or entertainment. The domain can significantly affect the outcomes of sarcasm detection because the use of sarcasm can vary depending on the context and subject matter. For example, the prevalence of sarcasm in political discussions may differ from its occurrence in conversations about sports or entertainment. The motivation for combining data from various online sources such as Twitter, Reddit and News Headlines is to expose the model to different writing styles as well as myriad domains prevalent online. Moreover, the proposed features in this study operate independently of the domain. Hence, there are no visible effects on the results obtained. This approach is therefore expected to help the model make accurate predictions for a wide range of text messages.

Annotation: Double annotation was performed by four expert linguists who independently performed annotation on the dataset. Once all four annotators had completed their annotations, the annotations were compared with each other to identify any discrepancies or disagreements. A consensus-based approach was followed to resolve these discrepancies or disagreements. This involved having the annotators discuss and come to a consensus on the correct annotation together. The agreed-upon annotations from all four annotators were used to create a final, consensus annotation for the entire dataset. Further, a series of preprocessing measures were taken. After preprocessing the dataset was reduced to 2,889 sentences."

Response: The split of the dataset into training, testing and validation has been included in subsection 5.1 as given below.

"The GNN framework described in the .section 4.4.1 was pretrained on a dataset of roughly 3000 sentences, out of which roughly 50% were similes and the rest 50% were non-similes. The entire dataset was split into three groups: 60% of the data was used for training, 20% was used for testing and the rest 20% was used for validation. This pretrained framework was then deployed on the main collated sarcasm dataset to extract the presence of a simile as one of the features. With a batch size of 32, the model was executed for 100 epochs."

* There is no meaningful error analysis, so there are few clues how one might go about improving their system. Indeed, there are few future directions to be found.

Response: Subsection 5.4 (Analysis of results obtained using DWAEF) has been modified to include a confusion matrix along with its inference to highlight the error analysis.

"Based on the presented confusion matrix, the performance of the model can be inferred as follows. The model achieved a high accuracy rate of 92.01%, indicating that it correctly predicted the majority of cases. Additionally, the precision of the model was found to be 92.98%, indicating that when it predicted a positive class, it was correct 92.98% of the time. Furthermore, the recall rate of the model was 89.59%, indicating that the model correctly identified 89.59% of actual positive cases. The F1 score of the model was also found to be high at 91.62%, indicating that the model performed well in terms of both precision and recall.

However, the model exhibited some limitations, such as a number of false positives (FP) and false negatives (FN). Specifically, the model produced 91 false positive predictions, which could lead to unnecessary actions or decisions being taken. Additionally, the model produced 140 false negative predictions, which could result in missed opportunities or errors in decision-making.

In conclusion, while the model performed well with high accuracy, precision, recall, and F1 score, there is still room for improvement in reducing the number of false positives and false negatives to further enhance its performance. These findings may have implications for decision-making processes in applications that utilize this model, and can inform future improvements in its development."

I think most of these issues could be fixed on resubmission without a great deal of effort from the authors, and crucially without the need to rerun any experiments, probably. For this reason, I have indicated "revise and resubmit" in my overall recommendation.

Nanopublication comments:

Further comments:

Below, I have some brief notes, mostly about style, for the paper.
The authors have a tendency to capitalize noun phrases which are not proper names (and thus should not be capitalized). I would submit that "DWAEF" is not a proper name (and thus should be written "deep weighted average ensemble-based framework"); certain "graph neural network" is not a proper name, and few people even explain what "BERT" stands for (it's obvious the authors decided it was called "BERT" before they decided what it stood for; the acronym being spelled out gives no real insights into what it is; in fact it's not "bidirectional" at all because it's not an RNN). "Fuzzy Logic" is absolutely not a proper name. Anyways, don't do that. "Python", the name of a programming language, is however a proper name, so it should be written in titlecase.

Response: The authors have diligently revised and edited the document to rectify such errors.

Usually "vs" is written with a period after.

Response: The authors have taken care to correct this error.

I am not familiar with the use of "viz", but the authors write it with and without a following period; pick one style and use it throughout.

Response: The authors have carefully combed through each sentence and paragraph, scrutinizing the text for any such inconsistency. The authors have replaced the occurrences of “viz./viz” with a more commonly used word.

Most of the figures have text that is a lot smaller than the body. I find a lot of them hard to read. The text should be made to match (roughly) that of the body text, and ideally you'd also match the fonts too.

Response: The authors have revised all figures to cater to such discrepancies.

The literature review reports accuracy numbers from previous work. These numbers are only meaningful if the studies are working on exactly the same corpora, but they are not, and so they should be omitted. For instance, the authors' [11] used a new, crowd-sourced corpus; whereas the authors' [7] used two annotators to build their own corpus. [11], which appears to be published later, doesn't mention [7]. One corpus is surely much easier than the other, and that, rather than the methods, explains much of the difference in accuracy numbers across studies.

Response: The authors have meticulously included more relevant information in the literature review to ensure their literature review section is more comprehensive and relevant.

The literature review does not have much of a logical structure. It is rather brief, which is unfortunate because there is a big literature here and the comparisons across data sets and methods are not common as they should be. I don't understand the distinctions drawn between "machine learning methods", "deep learning and transformer-based methods" (transformers are a type of deep learning method, and both deep learning and transformers in particular are examples of machine learning), and "graph neural network-based methods" (which are also arguably deep learning too, and certainly machine learning). Also, the authors mention that some of the work in the "deep learning" category used SVMs (from the "machine learning" category).

Response: The authors have conducted a thorough revision of the existing literature to guarantee that their related work section is inclusive, pertinent, and current. More important studies and publications have been included in a lucid and well-organized manner, serving to establish clear context, scope, and significance.

Emoticons, smileys, etc. being labeled "pragmatic features" is a strange notion to me. Pragmatics is concerned with the role of context in linguistic discourse; smileys are not "context" in the relevant sense. I think the term you may want would be something like "paralinguistics", perhaps.

Response: The authors have done the necessary changes.

"#sarcasme": is this English or, say, French? I found that misspelling confusing. May just be a typo though.

Response: The authors admit that it was an inadvertent typo that has been rectified.

"fallen short of the speaker's affection": this is not idiomatic English to me. To "fall short" is to disappoint someone, which is not a likely sarcastic interpretation of that expression.

Response: The authors appreciate this review and have made the necessary corrections to the text. The modified text is:

"...the hearer is an unpleasant addition to the speaker's life"."

What is the third type of metaphor discussed by study [20]? It is mentioned that this study has a third type, and that it is not handled in this paper. What is it, and why can't it be handled here?

Response: Study [20] identifies three types of metaphors: nominal, verbal, and adjectival. The third type of metaphor (adjectival) takes the form of an *adjective acting on a noun*, for example, *"He has a fertile imagination".* While, our paper focuses on nominal and verbal metaphors, we acknowledge the existence of adjectival metaphors. However, due to their different linguistic structure, they require a separate analysis and methodology, which is beyond the scope of our present study.

The word called "separator" for subordinate classes is usually known as a "subordinating conjunction" in both traditional and modern grammar.

Response: The authors agree that the term "subordinating conjunction" is a more widely recognized and accepted term for the type of word used to introduce a subordinate clause. Therefore, the authors have revised the terminology in the manuscript to improve clarity and accuracy.

"Ensemble learning strategies combine multiple machine learning algorithms to produce poor predictive outcomes. These results are then fused together to generate more accurate solutions.": This is a strange description of ensembling to me. The idea is not to find "poor learners", but to combine multiple models whose errors are no more than weakly correlated. What's written here has the unfortunate suggesting that the models that make up the ensemble ought to be bad.

Response: The authors admit the inconsistency of their writing style and have carefully paraphrased the above to convey meaning in a better way,

"Ensemble learning strategies enhance the performance and accuracy of a predictive model by merging multiple models. This approach involves training several models, each with unique algorithms or hyperparameters, and then fusing their predictions in a manner that maximizes the final outcome. Any ensemble framework comprises a collection of base learners and meta-learners. Base learners, also known as weak learners, are machine learning classifiers whose predictions are combined with those of other weak learners to compensate for their weaknesses. The meta learner or strong learner is the combined learnt model."

Figure 2 should just be a table. There is a lot of (justified, IMO) hatred for the use of pie charts. In this case there is really very little information being conveyed here and it would be better in tabular form. It would also be better to have (in addition to percentages), raw counts of the data.

Response: The authors have converted figure 2 to a table.

"viz, simile... is this a typo? I am not sure how to read this.

Response: The authors have revised the manuscript for a better and more lucid word choice.

I recommend that large counting numbers like "2891" be written with comma separators. Without these, they are much harder to read and it is uncommon to see them written in this comma-free form in published text.

Response: The authors have made the necessary changes to make large counting numbers easier to read and understand. The authors also appreciate the reviewer's keen eye to details and his worthy inputs on improving the clarity of the written work.

"Twitter is full of redundancy due to the rampant usage of slang, hashtags, emoticons, alterations in spelling, loose usage of punctuation, and so forth": I do not understand what these features of informal text have to do with is "slang" (a rather poorly defined notion in the first place) redundant? I found this an insightful discussion of what these social media features "mean" and how they ought to be handled: https://aclanthology.org/N13-1037/

Response: The authors have added the following paragraph to the text in subsection 4.2. (Data preprocessing) to give an insight regarding the use to slang in the preprocessing task of the experiment.

"The use of slang, hashtags, emoticons, alterations in spelling, and loose punctuation are not inherently redundant and serve various purposes, such as expressing emotions or emphasizing a point. However, they can pose challenges for tasks related to NLP and text analysis, as they often deviate from standard grammar and syntax. The study in [40] provides a more nuanced understanding of how these features of informal text can be modelled and analyzed to improve the quality of NLP frameworks and crowdsourced annotation tasks. As a result, the authors have accordingly taken these features into account before completely removing them."

"overdone": this seems like a value judgment that does not belong here.

Response: The authors have paraphrased the sentence in question.

"aforesaid": extremely archaic word choice, at least in my variety of English.

Response: The authors have revised the manuscript accordingly to use more modern and widely understood language that accurately conveys the intended meaning.

Could the authors put the list of intensifiers into an appendix? That way, this work can still be replicated when (as is certain) the Wikipedia page is edited in the future.

Response: The list of interjections has been moved to appendix A1.

Same thing with interjections: that ought to be an appendix, I think.

Response: The list of interjections has been moved to appendix A2.

"syntactical": ungrammatical for me, should just be "syntactic".

Response: The authors apologize for such grammatical inconsistencies. The manuscript has been scanned and rectified for such errors.

How was a GNN used to collect syntactic patterns? It seems like the author just ran an off-the-shelf parser. I would just say that.

Response: The authors have modified subsection 4.4.1 (GNN Framework for Simile Detection) to include more details on this.

"For the purpose of this research, simile is detected on the basis of its syntactic pattern using a GNN. The process typically involves first parsing the sentence using a syntactic parser to identify the syntactic relationships between the words. In this case, this was done using Stanford NLP Group's CoreNLP server [40]. Fig. 2 presents dependency trees of two sentences containing similes created using Stanford NLP Group's CoreNLP server. The resulting parse tree is then converted into a graph representation. To capture more complex syntactic relationships between the words in a sentence, the authors of this research implemented Bi-Fuse GraphSAGE [41]. To create a representation of each node, GraphSAGE aggregates information from its neighbouring nodes in the graph. This representation is then refined by Bi-Fuse GraphSAGE, which encodes the graph structure in a bi-directional manner by passing messages between nodes in both forward and backward directions, enabling more complex relationships between nodes to be captured. Subgraphs corresponding to specific syntactic structures of a simile are identified by analyzing these refined node representations. Additionally, a graph-level representation is created by combining these refined node embeddings, summarizing the properties of the entire graph. The class label of a sentence is predicted using these graph-level representations. A GNN-based text classification model is used to learn the dependency structure of similes. The entire set-up for the simile classification model consists of a graph construction module, a graph embedding module and a prediction module. Each of the modules is implemented using Graph4NLP library [42]. The modules are elaborated thoroughly below and the entire framework is summarized in Fig. 3."

I am not interested in the names of the "steps" on page 10. These are implementation details that have no bearing on the science here.

Response: The authors added step names on page 10 for transparency and technical clarity. To enhance text readability and fluency, the authors have moved some of the information such as the initialization preprocessing steps to appendix B1.

How were the authors able to confirm that the text in question was, say, monolingual English (as it appears to be assumed but never discussed)?

Response: The authors have used TextBlob to ensure this. The same has been discussed in subsection 4.1 (Dataset Description).

"Description of the dataset: The researchers of this work prepared a dataset of 2,891 sentences written in English. The dataset utilized in this study was sourced from various platforms, including Twitter, News Headlines, and the SARC dataset. TextBlob \cite{loria2018textblob} was used to check whether the text was written only in English."

The series of equations on page 11 are meaningless because none of the terms have been defined yet. It might be useful to define them first (instead of later).

Response: The meanings of the variables used in the equations have been include as per the kind suggestion.

"upto": should be "up to".

Response: The authors have taken care to rectify such typographical and grammatical inconsistencies.

It is a little difficult to understand how the BERT-based metaphor detection system works. I don't think I could recreate it from this description. Was fine-tuning used or not? I am not sure. The description should be made much more clear.

Response: The description of the proposed BERT-based metaphor detection system has been made clearer and more understandable. The sub-subsection 4.4.2 (BERT-based Structure for Metaphor Detection) has modified to include the changes.

“Word2Vec and GloVe are two types of architectures used commonly for word embedding, but they have limitations in certain tasks such as representing terms that are not in their vocabulary and distinguishing between opposites. For example, the words "good" and "bad" are often very close to each other in the vector space created by these models, which can be problematic for NLP applications like sentiment analysis. On the other hand, BERT is a pretraining method that uses a self-supervised approach to learn from masked text sections. Developed by a team at Google Research, BERT is based on the Transformer architecture and is designed to learn deep bidirectional representations from unlabeled text by conditioning on both left and right contexts. While Word2Vec and GloVe are unidirectional models that can only understand context in one direction, BERT can move sentences both to the left and right to fully comprehend the context of the target word or group of words. As a result, the detection of metaphors is achieved by generating embeddings using a BERT-based network. The basic BERT model, without any fine-tuning, can generate embeddings for the phrases that can be used to calculate cosine similarity. Fig. 5 illustrates the framework used for detecting the presence of a metaphor and Fig. 6 illustrates the BERT-based network used for generating the embeddings with the hidden layer representations in red. For the BERT base, each encoder layer outputs a set of dense vectors.”

The list of clause separators (which are a mix of prepositions and conjunctions) are given in Python form on page 14; just write it like normal text, without the [and]. Or, move it to the appendix.

Response: The authors appreciate the reviewer’s suggestion and have moved the list of subordinate conjunctions (clause separators) to appendix A3.

"moulded": I am not sure what this is supposed to mean here.

Response: The authors apologize for any confusion caused by the use of the term "moulded" in the manuscript. Upon reviewing the context in which the term was used, the authors have realized that it may not have been the most appropriate word choice. The authors have altered the text to use more precise and accurate language to convey the intended meaning.

I do not have enough context to understand the use of fuzzy logic as described on the bottom of page 14. It seems to assume more knowledge of this software than the reader is likely to have. For instance, I haven't the foggiest idea what a "trapezoidal membership function" would mean, nor a "non-polygonal fuzzy set". I also don't follow the fuzzy rules on page 15 because once again, I don't know as much as the authors about fuzzy logic, which is not a major part of NLP research in general; it requires extensive introduction if it's an important idea.

Response: The following additions have been made to sub-subsection 4.4.3 for further clarity on fuzzy-logic, trapezoidal membership function and non-polygonal fuzzy set.

“Fuzzy logic is a form of multi-valued logic that deals with reasoning that is approximate rather than precise. In fuzzy logic, a concept can possess a degree of truth (degree of membership) denoted by ’μ’ anywhere between 0.0 and 1.0, unlike in standard logic where a concept can only be completely true (1.0) or false (0.0). fuzzy logic is useful for reasoning about inherently vague concepts, such as "tallness". In traditional set theory, an element either belongs to a set or not, but in fuzzy logic, an element can partially belong to a set, with a degree of membership ranging between 0 and 1. In general, the degree of membership is used to represent the uncertainty or fuzziness in a concept or variable and is a key component in fuzzy logic’s ability to reason with imprecise or uncertain information. In fuzzy logic systems, a membership function is used to map this degree of membership of an element to a set, based on a specified set of fuzzy rules. Trapezoidal membership functions are one way to define fuzzy sets. They are a type of membership function that allows for more flexibility in defining the shape of the fuzzy set.

A trapezoidal membership function is a piecewise linear function that has four parameters: a, b, c, and d, where $a \leqslant b \leqslant c \leqslant d$. The function starts at 0 for $x \leqslant a$, increases linearly from 0 to 1 from a to b, stays at 1 from b to c, decreases linearly from 1 to 0 from c to d, and stays at 0 for $x \geqslant d$. By using trapezoidal membership functions, non-polygonal fuzzy sets can be defined that have smooth transitions. between membership degrees. The shape of the membership function can be controlled by adjusting its parameters, allowing for the creation of fuzzy sets that match the intended intuition. For instance, a trapezoidal membership function can be used to define the fuzzy set "tall person" with parameters such as a=170cm, b=175cm, c=185cm, and d=190cm. The resulting function would have a membership degree of 0 for heights below 170cm and above 190cm, a membership degree of 1 for heights between 175cm and 185cm, and smoothly increasing and decreasing membership degrees for heights between 170cm and 175cm and between 185cm and 190cm. Trapezoidal membership functions allow for the generation of non-polygonal fuzzy sets, which enable more precise and flexible modelling of linguistic variables in fuzzy logic.”

I don't know what a "Dirichlet ensemble object" ("Dirichlet" is a person's name so it's capitalized, the rest is not a proper noun though so it should be in lowercase) is, so I can't make sense of the ensembling. No discussion or citations are provided.

Response: The subsection 4.5 (The Deep Weighted Average Ensemble Framework) has been modified to include a detailed explanation on how the weighted average ensembling procedure works along with what a Dirichlet ensemble object represents. The modified version is as follows:

“The framework is implemented with the help of DeepStack Library [47]. Initially, the curated dataset is used to pretrain the three models. During training, each of the base learners receives the outputs of the GNN-based simile detection framework, the BERT-based metaphor detection framework, the fuzzy based polarity shift detection framework as well as the primitive features as its inputs. Next, the predictions produced by each module of the ensemble are weighed based on their performance on a hold-out validation dataset. This is represented by a Dirichlet ensemble object. To achieve more accurate solution, a weight optimization search is used in conjunction with randomized search based on the Dirichlet

distribution on the validation dataset. Randomized search based on the Dirichlet distribution involves randomly sampling the weights from a Dirichlet distribution, which generates a probability distribution over the weights that is continuous and flexible. Weight optimization search iteratively adjusts the ensemble model weights to maximize accuracy on the validation dataset. Through this process the base learners, a TabNet [46], a 1-D CNN and an MLP, are assigned optimal weights. These weights determine the contribution of each model to the final prediction. Subsequently, the predictions of each model are combined through a weighted average. No meta-learner is used in this ensemble method. Lastly, the ensemble model is assessed on a separate test set to determine how well it can perform on new data"

It is claimed on page 15 that the model is neither undertrained nor overfit. As far as I can tell I was never told about any dev/validation or testing set though, nor do I know how large they are, so I can't really assess that.

Response: The details about the split of the dataset into training, testing and validation has been included in subsection 5.1 as follows.

"The GNN framework described in the section 4.4.1 was pretrained on a dataset of roughly 3,000 sentences, out of which roughly 50% were similes and the rest 50% were non-similes. The entire dataset was split into three groups: 60% of the data was used for training, 20% was used for testing and the rest 20% was used for validation. This pretrained framework was then deployed on the main collated sarcasm dataset to extract the presence of a simile as one of the features. With a batch size of 32, the model was executed for 100 epochs."

Figure 9 is superfluous. Yes, the accuracy goes up and the loss goes down. Of course, it does. Just report the final test accuracy, the only statistic that actually matters. These loss and accuracy curves are useful for the developer, but not useful for the reader who is trying to evaluate your proposal.

Response: The authors feel that loss and accuracy curves presented in the figure are essential to understanding the proposed approach's performance during the training process. They provide readers additional insight beyond the final test accuracy. While we recognize that these curves may be more beneficial for the developer, we still believe that they are valuable for readers who seek to understand the model's performance in detail.

In table 5 the hyperparameters are reported like "avg_pooling": write that in English ("average pooling"), not code.

Response: All tables in the manuscript have been meticulously modified to eliminate any inconsistencies. The authors have taken great care to ensure that the tables are presented in a consistent and accurate manner, and that any errors have been rectified.

"drop out": this is traditionally called "dropout", one word.

Response: The necessary corrections have been done throughout the manuscript.

"learning rate reduce factor": I don't know what this is and no context is given.

Response: The learning rate reduce factor is a hyperparameter in machine learning algorithms, especially in those using stochastic gradient descent optimization. It determines the extent to which the learning rate (step size) should be reduced during training if the algorithm does not improve or overshoots the optimal solution.

The authors give no error analysis. However, they do give a few correctly classified examples. I would prefer some useful error analysis instead.

Response: Subsection 5.4 (Analysis of results obtained using DWAEF) has been modified to include a confusion matrix along with its inference to highlight the error analysis.

"Based on the presented confusion matrix, the performance of the model can be inferred as follows. The model achieved a high accuracy rate of 92.01%, indicating that it correctly predicted the majority of cases. Additionally, the precision of the model was found to be 92.98%, indicating that when it predicted a positive class, it was correct 92.98% of the time. Furthermore, the recall rate of the model was 89.59%, indicating that the model correctly identified 89.59% of actual positive cases. The F1 score of the model was also found to be high at 91.62%, indicating that the model performed well in terms of both precision and recall.

However, the model exhibited some limitations, such as a number of false positives (FP) and false negatives (FN). Specifically, the model produced 91 false positive predictions, which could lead to unnecessary actions or decisions being taken. Additionally, the model produced 140 false negative predictions, which could result in missed opportunities or errors in decision-making.

In conclusion, while the model performed well with high accuracy, precision, recall, and F1 score, there is still room for improvement in reducing the number of false positives and false negatives to further enhance its performance. These findings may have implications for decision-making processes in applications that utilize this model, and can inform future improvements in its development."

The bibliography is full of typos, punctuation errors, and capitalization errors. Unfortunately, one cannot simply use the auto-generated bibliography entries without a bit of editing.

Response: The authors have given meticulous attention to free the bibliography from all typos, punctuations errors and capitalization errors.