Impact Factor (SJIF): 6.092 E-ISSN: 2583-1615

International Journal of Research in Academic World

Received: 16/July/2023 IJRAW: 2023; 2(8):204-214 Accepted: 27/August/2023

Enhancing Subjective Valuation using Artificial Intelligence

*1Babu S and 2Parameswari P

*1Research Scholar, Department of Computer Science, Karruppannan Mariyappan College, Mutthur, Tirupur, Tamil Nadu, India.

2Principal, Palanisamy College of Arts, Perundurai, Erode, Tamil Nadu, India.

Abstract

Subjective valuation is the process of assessing the worth, desirability, or importance of something based on individual preferences, experiences, and personal judgment. It is a fundamental aspect of decision-making and plays a crucial role in various domains such as economics, psychology, marketing, and finance. However, subjective valuation is inherently complex and can be influenced by various biases and inconsistencies, leading to suboptimal decision-making outcomes. Large number of students attend subjective type exam. For evaluation of such large number of papers manually required hard efforts. Sometimes quality of evaluation may change according to mood of evaluator. The evaluation work is very lengthy and time consuming. Competitive and entrance exams typically contain objective or multiple-choice questions. These exams are evaluated on machine as they conducted on machine and therefore their evaluation is easy. It also saves multiple resources and human interaction and hence it is errorless. There is multiple system are available for evaluation objective (MCQ) type question but there is no provision for subjective (Descriptive) type question. It will be very helpful for educational institutions if the process of evaluation of descriptive answers is automated to capably assess student's exam answer sheets.

Keywords: Subjective answer, artificial intelligence, human intelligence, expert systems, machine learning, deep learning

1. Introduction

Subjective questions and answers can survey the exhibition and capacity of an understudy in an unconditional way. The answers, normally, are not bound to any limitation, and understudies are allowed to think of them as indicated by their mentality and under-remaining of the idea. So, a few other imperative contrasts separate emotional responses from their objective partner. As far as one might be concerned, they are significantly longer than the goal questions. Also, they find opportunity to compose. Besides, they convey considerably more setting and take a ton of focus and objectivity from the educator assessing them.

Assessment of such inquiries utilizing PCs is a precarious undertaking, fundamentally on the grounds that regular language is equivocal. A few preprocessing steps should be performed, like cleaning the information and tokenization prior to dealing with it. Then, at that point, the text-based information can measure up utilizing different methods, for example, record likeness, inactive semantic designs, idea charts, ontologies. The last score can be assessed in light of Closeness, watchwords presence, structure, language [11, [12]. A few endeavors have been made in the past to tackle this issue [13, [14, 15], however there is still space for enhancements, some of which is examined in this paper.

Subjective tests are viewed as additional intricate and frightening by the two understudies and instructors because of their one essential component, setting. An emotional response requests the checker check each expression of the solution for scoring effectively, and the checker's psychological wellness, exhaustion, and objectivity assume a gigantic part in the general outcome. Thusly, it is significantly more time and asset effective to allow a framework to deal with this monotonous and fairly basic errand of assessing emotional responses. Assessing objective responses with machines is extremely simple and plausible. A program can be taken care of with questions and single word responds to that can rapidly plan understudies' reactions. All things considered; emotional responses are considerably more testing to handle. They are shifted long and contain a tremendous measure of jargon. Moreover, individuals will quite often utilize equivalent words and helpful truncations, which makes the interaction that much precarious.

Much work has been finished on the subject of emotional responses assessment in some structure, for example, measuring Comparability between various texts, words, and even archives, finding the setting behind the text and planning it with the arrangement's unique circumstance, including the thing expression in the reports, matching watchwords in the responses, etc. In any case, issues, for example, Tf-Idf loosing semantic setting [16], absence of hyper-boundaries tuning [17], expensive preparation [18], and need for better datasets [15] still exist.

*Corresponding Author: Babu S

In this paper, we investigate an AI and natural language handling-based approach for emotional answers assessment. Our work depends on normal dialects handling methods, for example, tokenization, lemmatization, message addressing procedures like TF-IDF, Pack of Words, word2vec, closeness estimating strategies like cosine similitude, and word mover's distance, arrangement techniques like multinomial Innocent Bayes. We utilize different assessment measures, for example, F1-score, Exactness, and Review to consider the exhibition of different models in contrast to one another. We additionally examine different procedures utilized in the past for abstract responses assessment or text closeness assessment overall.

Following is a portion of the significant constraints while managing emotional responses:

- Existing investigations will generally have equivalent words.
- Existing investigations will generally have a broad scope of potential lengths.
- Existing investigations will generally be haphazardly requested among their sentences.

This paper proposes a better than ever approach to assessing spellbinding inquiry responds to consequently utilizing AI and regular language handling. It utilizes 2 stage way to deal with tackling this issue. To start with, the responses are assessed utilizing the arrangement and gave watchwords utilizing different Likeness based strategies, for example, word mover's distance. Then the outcomes from this step are then used to prepare a model that can assess replies without the requirement for arrangements and catchphrases. For instance, an emotional inquiry "What is the capital city of Pakistan and what is it renowned for?" can have a right response of "Islamabad is the capital city of Pakistan and it is well known for Mountain View". Prior to assessing the understudy's solution to the inquiry, both the inquiry, the response, and furthermore a catch phrases fundamental for the response are taken care of into the framework (for this situation, watchwords will be Islamabad and mountain view), and the framework assesses the understudy's response by looking at both the comparability (remembering setting) of modular response and understudy's reaction as well as the presence or nonattendance of any watchwords. So, an understudy's response of "Karachi is the capital of Pakistan, it is renowned for mountain landscape" could get half stamps, "Islamabad and mountain view" could get 30% imprints since the principal watchwords are available even though setting is missing and "Islamabad is the capital and its popular for mountain landscape" could get 100 percent marks since it fulfills both logical similitudes as well as catchphrases presence corresponding to the right response.

a) **Motivation:** This type of assessment by machines is a major forward-moving step in supporting the instructive area to play out their different obligations productively and decrease the difficult work in unimportant errands like contrasting the responses and a right arrangement for this situation. This prompts educator investing more energy showing understudies, setting up a superior educational plan, and assessing their tests with less human mistakes and more straightforwardness.

b) **Contribution:** This paper contributes by tackling the issue of emotional responses assessment utilizing AI and regular handling methods, it concentrates on different edges of sentence closeness estimating lattices and proposes a method for preparing an AI model, which can thus assist with building up trust in assessment score pushing ahead. Different commitments incorporate a pre-arranged informational index with solution's, replies, and catchphrases cautiously organized by instructors.

c) **Paper Association:** The remainder of the paper is coordinated as follows: Area II presents the foundation of the issue and the writing audit. Area III gives the proposed approach. Segment IV presents the exploratory examination and results. Area V closes the paper.

2. Background and Literature Review

As referenced previously, the assessment of emotional responses is definitely not a groundbreaking insight, and it has been worked upon for very nearly twenty years. Different methods have been executed to take care of this issue, for example, enormous information Normal Language Handling, Idle Semantic Examination, Bayes hypothesis, K-closest classifier, and, surprisingly, formal procedures like Conventional Idea Investigation. They are classified into three categories: Measurable, Data Extraction, and Full Normal Language Handling.

a) Technical Background

i). **Statistical Procedure:** It depends on watchword coordinating and is viewed as poor as it can't handle issues like equivalent words or consider the unique situation. A few works have been finished on emotional paper assessment utilizing this approach [19, 20].

ii). **Information Extraction (IE) Method:** Data Extraction methods rely upon getting a structure or an example from the text so the text can be broken into ideas and their connections [1]. The conditions found to assume a critical part in delivering scores and should be affirmed from a specialist in space [2, 3].

iii). **Natural Language Processing (NLP):** These methods include utilizing normal language apparatuses to parse the text and find its semantic importance [4, 5]. That significance can then measure up to the importance got from the answer for relegate the last score.

Text reports should be handled and prepared for the machine; this step is called preprocessing and includes different regular language procedures, for example, Tokenization, Stopword Evacuation, Grammatical features Labeling, Lemmatization, Stemming, Case Collapsing. A portion of these methods are momentarily made sense of beneath. Nitin *et al.* [06] examined computerized scoring frameworks and the utilization of Regular Language Handling and AI in them. Zhiwei *et al.* [7] utilized Normal Language Handling to gauge tree likeness.

iv). **Tokenization:** Tokenization is the method involved with isolating information into more modest parts, like passages, sentences, words, and characters. Tokenization is fundamental while managing regular language on the grounds that each word should be handled independently to get its actual importance. In this work, we tokenize information into sentences and words in view of blank areas and period signs. Tokenization is one of the earliest strides during normal language supportive [8]. Kairat *et al.* [9] talks about assessment consequences of three existing sentence division and word tokenization frameworks on the Estonian web dataset.

v). **Stopword Evacuation:** Normal Language has a huge jargon, and most components are there for the simplicity of human comprehension, for example, 'the', 'in', 'on', 'is, etc. These words play next to zero job in most AI errands and could frustrate the cycle by aiding the model gets prepared on the different information. Each language has some realized stop words, which are normally taken out from the corpus to make the dataset more thick and one of a kind. Alexandra *et al.* [10] contends that the utilization of stop word expulsion is shallow and that point induction helps little from the act of eliminating stopwords beyond general terms. Mustafa *et al.* [21] notes that the impact of stopword expulsion affects the genuine outcomes also. Notwithstanding, it ought to be noticed that incessant words with minor significance for the AI model ought to be eliminated to work on the model.

vi). **Parts of Discourse Labeling:** Grammatical forms labeling is the handling of labeling each word in the information to its connected piece of speed, like a thing, action word, qualifier, descriptor. Grammatical form labeling should be possible by different devices, for example, NLTK pos tagger and grasps the construction of the sentence. It has energizing applications, for example, finding thing phrases in the sentence, lessening words to their lemma, etc. Divya *et al.* [22] utilized grammatical feature labeling for productive nostalgic examination of Twitter.

vii). **Lemmatization:** Words found in regular language have a place in many structures, like different tense structures. For instance, the words 'go', 'going', 'went' all have a place with a similar root word 'go' however have various structures. Lemmatization is the most common way of decreasing every one of the words in the dataset to their root structures. Lemmatization requires an itemized word reference of the words to relate them to their lemma, likewise called the root. It likewise utilizes part of speed data to relate the words to their particular root in the word reference. Francesco *et al.* [23] utilized Lemmatization and backing vector machines to arrange Italian text.

viii). **Stemming:** Stemming is an approach to diminishing words to their stems, and it depends on the possibility that each language has a conventional sentence structure of some sort or another, and the jargon is framed by continuously remembering those standards of punctuation. In this way, by utilizing those equivalent principles, we can lessen every one of the comparable words back to their stems by eliminating their postfixes that make them unique. For instance, stemming plurals into singulars (words into words), stemming finishing characters, etc. There are different stemming calculations always in each language, for example, Potter's calculation for English word stemming. Jabbar *et al.* [24] talks about different stemming calculations used to stem text-based information.

ix). **Case Collapsing:** The normal language contains words in various cases, frequently copying the specific words in light of their case. Consequently, it is normal to diminish every one of the information into a similar case, generally lower case, so the machine can decipher each word in a similar way.

After the preprocessing has been finished on the information according to prerequisites, printed information is changed over in a mathematical structure since machines just comprehend numbers and under-stand them quite well. This cycle is called word installing, and a portion of the strategies utilized include Sack of Words, TF-IDF, word2vec.

x). **Bag of Words (BoW):** Pack of Words is a gullible strategy that includes addressing the jargon of the text-based information as a vector. That vector contains the list number addressing either the count or the specific word at that file in the text. BoW keeps count of frequencies of words however loses the setting of words. One illustration of BoW is a one-hot vector. Sunil *et al.* [25] utilized pack of-words (BoW) vector portrayal to quantify the closeness of two archives concerning each term happening in the records.

xi). Term Recurrence Opposite Record Recurrence (TF-IDF) TF-IDF is like BoW, where it counts the frequencies of all words present in the report, however it additionally monitors the number of various sentences that have those words. Along these lines, it gives data about the count and the worth of a word in the document. Sammut *et al.* [26] discusses Tf-Idf in detail. Havrlant *et al.* [27] gives a probabilistic explanation of TF-IDF approach. Ankit *et al.* [28] used TF-IDF to predict stock trends.

xii). **Word2Vec:** Word2vec is a strategy that utilizes a brain network model to gain word relationship from an enormous dataset. It tends to be prepared for high aspects, for example, 300, which helps keep the words' semantic amounting to anything unblemished. After the preparation is finished, a word2vec model can identify equivalent words or recommend different words in view of the sentence. One illustration of a pre-prepared word2vec model is Google News' 300-dimension word2vec model that contains around 100 Billion words.

After the text has been changed over into mathematical structure, otherwise known as vectors, the time has come to think about those vectors and track down the Comparability of disparity between them. A portion of the significantly involved techniques for this undertaking are Cosine Closeness, Jacquard comparability, and Word Mover's Distance. Figure 1 outline Word2Vec installing. Jin *et al.* [29] concentrated on a semantic likeness calculation technique in view of Word2vec.

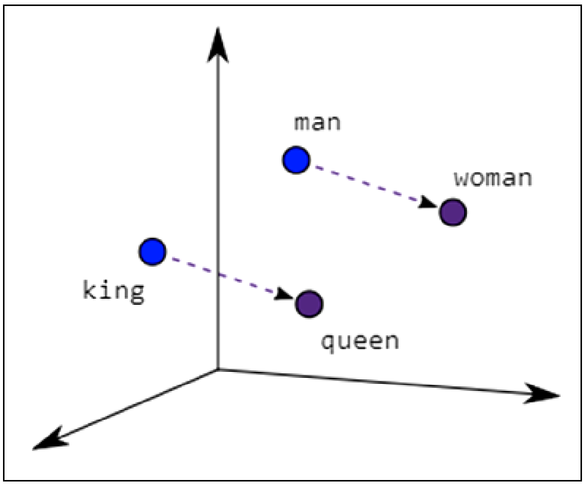

Fig 1: Illustration of Word2Vec embedding

xiii). **Cosine Likeness:** "Cosine likeness is a proportion of Similitude between two non-no vectors of an inward item space that actions the cosine of the point between them. A cosine point between two vectors is estimated, and it's worth lies somewhere in the range of 0 and 1, 1 addressing a full match. Park *et al.* [30] presented a cosine comparability-based way to deal with working on the presentation of customary classifiers like MNB, SVM, and CNN. The cosine of 0° is 1, and it is under 1 for some other point in the stretch." this strategy is utilized broadly in the errand of text handling.

xiv). **Jacquard Similitude:** The jacquard similitude is the proportion of convergence to association in regards to common words. It tracks down the association of two texts and tracks down their converging terms. Then, at that point, partition the convergence by its association. The higher the outcome, the more normal words and the greater the crossing point.

xv). **Word Mover's Distance (WDM):** Word Mover's Distance attempts to gauge the semantic distance of two records, and word2vec embeddings bring the semantic estimation. In particular, word2vec is used in their trials. When the word embeddings are gotten, the semantic distance among archives is characterized by the accompanying three sections: record portrayal, similarity metric, and a (sparse) flow grid. It has been displayed to beat a significant number of the cutting-edge strategies in k-closest neighbors grouping [7]. Sato *et al.* [31] notes that Weapon of mass destruction ought to be better than BOW since Weapon of mass destruction can consider the hidden math, while BOW can't. The likenesses got from these strategies are essentially what we really want to assess an emotional response.

b) Literature Audit

Hu *et al.* [6] proposed an Idle Semantic Ordering approach for the evaluation of emotional inquiries on the web. They utilized Chinese programmed division strategies and emotional ontologies to make a k-layered LSI space grid. The responses were introduced in TF-IDF implanting grids, and afterward Particular worth Deterioration (SVD) was applied to the term-report framework, which shaped a semantic space of vectors. LSI assumed the part of decreasing issues with synonym and polysemy. Finally, the Likeness between answers was determined utilizing cosine comparability. Dataset comprised of 35classes and 850 occasions set apart by educators, and the outcomes showed a 5% contrast in evaluating done by educator and the proposed framework.

Kusner *et al.* [7] introduced a clever idea of utilizing Word Mover's Distance (Weapon of mass destruction) to track down the disparity between two texts. The framework utilized no hyper-boundaries and utilized a casual Weapon of mass destruction way to deal with relax the vector space limits. Dataset included eight true sets, including Twitter opinion information and BBC sports articles. Word2vec model from google news was utilized, and two other custom models were prepared. KNN characterization approach was utilized to group the testing information. Subsequently, loosened up Weapon of mass destruction decreased the blunder rates and prompted 2 to multiple times quicker characterization.

Kim *et al.* [32] proposed a technique to grade short descriptive responses lexico-semantic example (LSP) because of its great exhibition with morphologically complex Korean language. LSP can structure the semantic of the response to assist with grasping the client's goals. An equivalent rundown was likewise used to assist with extending the catchphrases, so they match different response styles. Dataset was acquired from 88 understudies and changed over completely to LSP, which was subsequently contrasted with the arrangement LSP with score the response. Subsequently, the framework performed better compared to the current framework by 0.137

Orkphol *et al.* [34] utilized the word2vec way to deal with repeated words on a fix-sized vector space model and afterward estimated the Likeness of sentences utilizing a cosine comparability measure. Word2vec from google was utilized, and the sentence vector was gotten because of a normal of words in the sentence. The score was acknowledged whether it passed a predefined edge for similitude results, somewhere in the range of 0 and 1. Assessment proportion of review and precision was utilized, and subsequently, the framework's exhibition was 50.9% with and 48.7% without the likelihood of sense circulation.

Oghbaie *et al.* [33] proposed a couple wise Likeness measure to gauge the similitude between two records in light of the watchwords which show up in no less than one of the documents. The work proposed another comparability measure called PDSM (match wise record similitude measure), a changed form of the best properties approach. The proposed similitude measure was applied to message mining applications, for example, reports detection, K Closest Neighbors (KNN) for single-name characterization, and K-implies grouping. An assessment proportion of exactness was utilized, and thus, the PDSM technique delivered improved results than different measures like the Jaccard coefficient by 0.08 review.

Xia *et al.* [8] consolidated the word2vec approach with the authoritative archive corpus to recognize likenesses between various regulation records. Cosine similitude was utilized to gauge the Comparability between various sentence vectors. Thus, word2vec worked on the precision by 0.2 contrasted with the Sack of Words approach, which could additionally be expanded by 0.05-

0.10 via preparing the word2vec model on regulation reports.

Wagh *et al.* [35] proposed a multi-rules dynamic point of view to track down the Closeness between authoritative reports. The work included utilizing Man-made consciousness and accumulation procedures like arranged weighted normal (OWA) for getting the closeness esteem between various archives. Dataset was gotten from Indian High Legal dispute decisions from years going from 1950 to 1993. Assessment proportions of F1score and review were utilized. Subsequently, an idea based closeness approach, for example, the one proposed in work performed better compared to different strategies like TF-IDF, getting a F1-score of up to 0.8.

Alian *et al.* [36] concentrated on different variables influencing sentence comparability and summarizing recognizable proof utilizing different word installing models, bunching calculations, and weighting techniques to track down the setting of sentences. Pre-prepared embeddings included AraVec and FastTex, both prepared for the Arabic language. The Arabic preparation dataset included around 77,600,000 tweets. Accordingly, pre-prepared install ding with marked information from specialists gave better review and accuracy of 0.87 and 0.782 for K-implies and agglomerative grouping.

Muangorathub *et al.* [5] proposed an original methodology of literary theft location utilizing formal idea examination (FCA). The work showed formal setting in FCA, beginning with two sets containing components for certain traits that a how relate the component to its set. The reports and their common watchwords shaped a gathering set in FCA whose qualities are normally yet not restricted to 0 and 1. The moved toward utilized a many-esteemed setting. The work likewise presented another closeness idea that utilizes both the item degree and property aim. The methodology utilized isn't regularly used in comparability examination and positions comparative reports since they have comparable item and characteristic aims. The proposed framework distinguished counterfeiting in reports with 94% precision.

Jain *et al.* [37] proposed an original methodology for emotional inquiries assessment utilizing idea charts. Idea charts were made for both the arrangement and the response, and the score was assessed utilizing different diagram likeness techniques. Montes *et al.* [38] cleared up different strategies for find Likenesses between idea charts and data recovery from such diagrams.

Bahel *et al.* [39] introduced a design for assessment of abstract inquiries utilizing text rundown, text words spasms, and catchphrases outline and contrasted the outcomes and existing methodologies. The outcomes showed a blunder of 1.372 contrasted with 1.312 blunder from Jaccard's comparability approach. The methodology, in any case, neglected to process non textual information like graphs, pictures, and different arrangements.

Table 1 shows the outline of the writing survey

Table 1: Summary of Literature Review

Ref.	Model	Contribution	Evaluation Matrix	Limitations
[16]	LSI, SVD	LSI reduced problems with polysemy	Exactness	TF-IDF loses semantic
[17]	WMD, KNN	Relaxed WMD better than WMD	Exactness, Recall	No hyper-parameters tuning
[32]	LSP	LSP better for morphologically complexity	Precision, Recall	Need Dictionary-based LSP
[33]	PDSM, KNN	PDSM method than Jaccard coefficient	Accuracy	Domain Dependent Performance
[34]	Word2vec, Cosine Similarity	Fix-sized vector space model	Accuracy, Recall	Need better lexicon resources
[8]	Word2vec, Cosine Similarity	word2vec 0.2 better than BoW	Accuracy	high training cost
[40]	Jaccard & Cosine, Word2vec	Cosine Similarity better than Jaccard	Accuracy, Recall	Could use Programmable G-Arrays
[35]	AI Aggregation, OWA	Concept based similarity better than TF-IDF	Recall, F1-Score	Average weighting scheme
[36]	Word Embedding, Clustering	Pre-trained embedding better than K-means	Recall, Precision	massuve training time
[15]	FCA	FCA uses both object extent & attribute intent	Exactness	New approach needs datasets

a) **Solution:** The arrangement is an emotional response that is being utilized to plan understudies' reactions. This arrangement should contain every one of the catchphrases and settings talked about in the responses in isolated lines/sections. The educator/evaluator commonly readies the answer for the inquiry.

b) **Answer:** The response is an emotional reaction from the understudy that will be assessed. It generally contains some or the catchphrases as a whole and ranges 1 to a couple of sentences relying upon the sort of inquiry and the understudy's composing style. It quite often contains equivalent words contrasted with the arrangement and, thusly, requires considerably more semantic consideration while handling.

c) **Data Collection:** To prepare and test the proposed model, there is a requirement for an enormous measure of corpus containing emotional inquiry responds to, yet there is no openly accessible marked abstract inquiry addresses corpus as far as I could possibly know. In this work, we make emotional responses named corpus. For producing corpus, the significant thing is to focus on those sites and websites where emotional inquiries and answers exist. We creep different sites and gather an emotional inquiry responds to corpus. The slither information has a place with different spaces like software engineering and general information.

d) **Data Annotation:** In the wake of getting crept information, there is further requirement for annotation of information since that slithered information is unlabeled. To annotate information, a gathering of various workers is chosen, which have a place with the space of our emotional inquiry responds to corpus. We enlist 30 distinct annotators from various col-leges and colleges and live in Pakistan's various urban communities. The majority of them are understudies and instructors. The typical period of annotators is in the 21-25 territory, though a few annotators are in the age scope of 27-51. We task annotators to best score the emotional inquiry responds to as indicated by the responses given by understudies.

i). **Keyword Age:** Toward the starting period of explanation, the information contains downright responses and no particular catchphrases. We task annotators to recognize the fundamental terms from the arrangement which can represent the moment of truth the general score of that inquiry. These catchphrases assist with concluding whether an understudy has referenced pertinent data in their emotional responses or not.

ii). **Data Explanation Quality Approval:** Information approval is urgent to acquiring precise execution. We perform explanation from three unmistakable annotators of a solitary model. We keep the larger part casted a ballot score as the last clarified name for a specific model.

iii). **Corpus Insights:** Our explained corpus contained north of 1,000 short emotional inquiries, each containing a right response (arrangement) and 20 understudies' solutions to the inquiry, which were all commented on. The corpus additionally contained fundamental catchphrases in regards to emotional inquiries which were separated from the arrangements.

e) **Preprocessing Module:** Subsequent to taking contributions from the client, both the arrangement and the response go through some preprocessing steps, which include tokenization, stemming, lemmatization, stop words evacuation, case collapsing, finding, and joining equivalents to the text. Note that stop words are not taken out while passing the information to word2vec in light of the fact that word2vec contains a huge jargon and can use those stop words to comprehend the text. In any case, stop words are taken out prior to passing to an AI model, for example, Multinomial Credulous Bayes since they thwart the machine's capacity to become familiar with the examples.

f) **Similarity Estimation Module:** This module comprises of WDM and Cosine Comparability functions which take two sentences or word vectors and return their Likeness. WDM lets us know the disparity while Cosine.

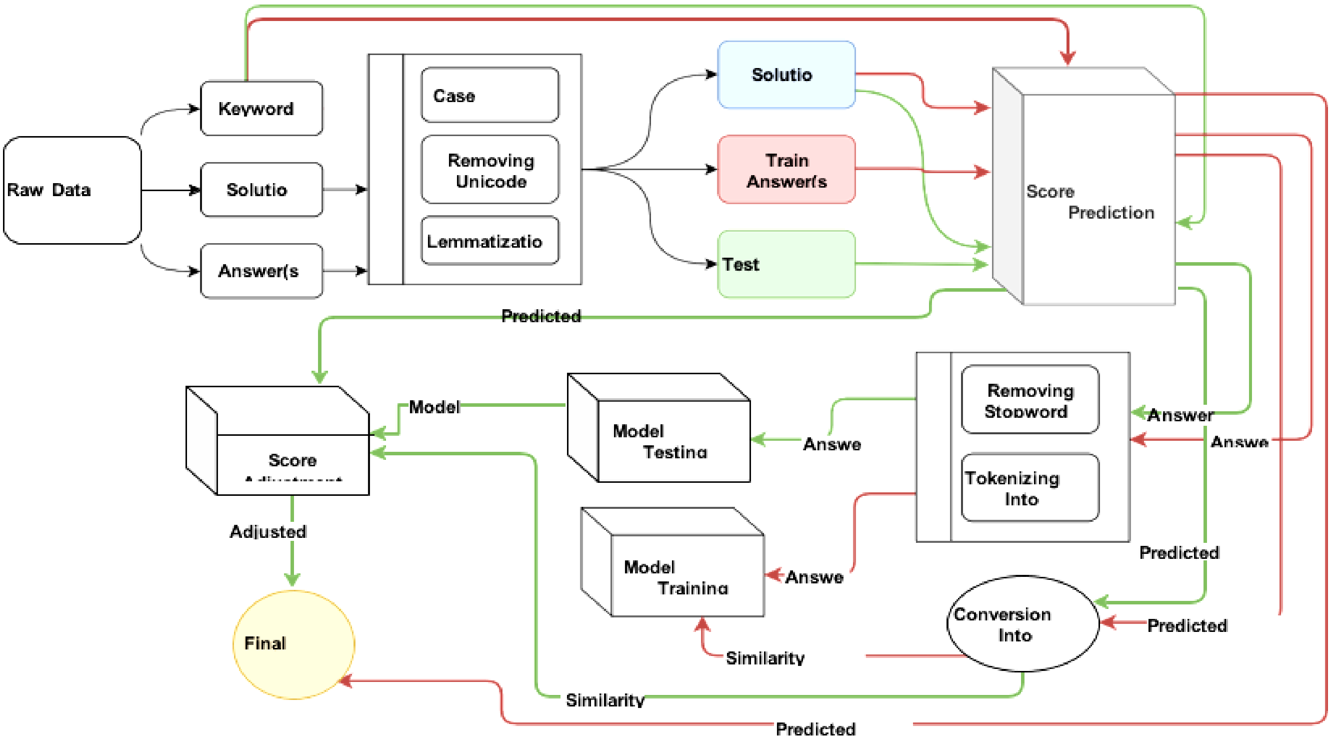

Fig 2:

Closeness estimates Likeness. Our methodology utilizes both of these likeness estimates each in turn and thinks about the outcomes toward the end. Different closeness (or uniqueness) thresholds utilized are given in Table 2.

1. **Limits Examination:** Different limits utilized in this paper have been experimentally concluded to create the ideal outcome, WDM edges of WDM_LOWER and WDM_UPPER address the dis-similitude between two sentences, where greater divergence addresses high likeness. 0.7 limit for WDM_LOWER was tentatively seen to address semantically very much like sentences, and 1.6 edges for WDM_UPPER were seen to address semantically fewer comparative sentences. Anything past 1.6 is thought to be excessively not at all like consider suitable for examination.

   Essentially, Cosine similitude limits COS_LOWER and COS_UPPER address the closeness between two sentences, it ought to be noticed that cosine comparability doesn't consider the setting of two sentences while estimating likeness rather than WDM, subsequently the use of both of this comparability (or difference) estimating approaches.

f) Result Predicting Module

Result Foreseeing Module is the center of this work. Figure 3 shows the working of this module. It works on the Following Algorithm 1:

We currently have the general score determined by our module utilizing either WDM or Cosine Comparability while thinking about the greatest matched arrangement/answer sentence matches.

This outcome can measure up to a real score or taken care of into an AI model to be prepared.

1. Machine Learning Model Module

This model comprises of AI models prepared on the information acquired from the outcome expectation module. Its working is as per the following:

- Input information from Result Expectation Module.
- Preprocess the arrangement and reply, eliminating stop words, and use Count vectorizer to address them in one or the other Pack of Words or TF-IDF structure.
- Convert the general score acquired from Result Prediction Module into some classification. Four classes A, B, C, and D, are utilized in the paper, addressing first, second, third, and fourth quarter of a 100. For instance, an addresses marks from 0 to 25, and B addresses 26 to 50.
- The quantity of classes is kept to a base in view of the inaccessibility of the real dataset. Basically, these classifications can be stretched out to cover more modest score ranges.
- An AI model, for example, Multinomial Guileless Bayes, which performs well for multi-class classification is chosen

Table 2: Similarity Thresholds

Threshold Notation	Threshold Value	Threshold Description
WDM_LOWER	0.7	Dissimilarity between two sentences using WDM is <= 0.7 meaning sentences are semantically very similar.
WDM_UPPER	1.6	Dissimilarity between two sentences using WDM is <= 1.6 meaning sentences are semantically a little bit similar.
COS_LOWER	0.2	Similarity between two sentences using Cosine Similarity is >= 0.2 meaning sentences are semantically a little bit similar.
COS_UPPER	0.7	Similarity between two sentences using Cosine Similarity is >= 0.7 meaning sentences are semantically very similar.

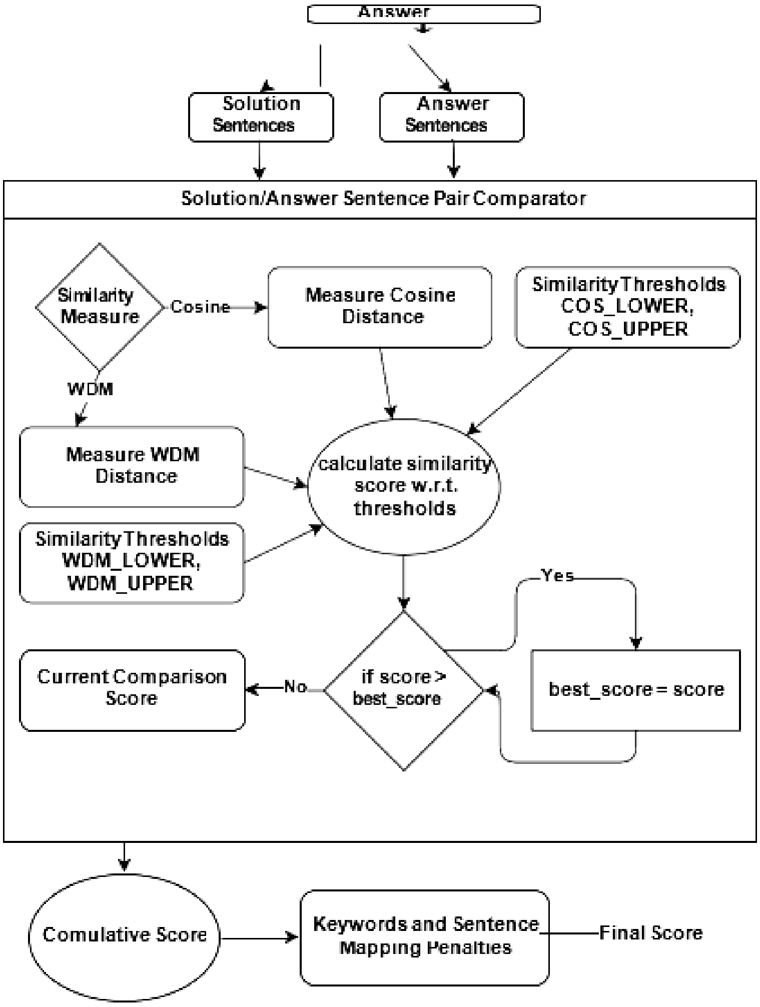
Answer
Solution Sentences
Answer Sentences
Solution/Answer Sentence Pair Comparator
Similarity Measure
Cosine
Measure Cosine Distance
Similarity Thresholds COS_LOWER, COS_UPPER
WDM
Measure WDM Distance
calculate similarity score w.r.t. thresholds
Similarity Thresholds WDM_LOWER, WDM_UPPER
Yes
Current Comparison Score
No
if score > best_score
best_score = score
Comulative Score
Keywords and Sentence Mapping Penalties
Final Score

Fig 3: Flow Chart of Result Prediction Module

1. We created a corpus consisting of sentences and the synonyms of sentences present in the corpus.
2. Sentences are composed of solution sentences and answer sentences. We defined them as: S ∈ *Ssen*1, *Asen*1, *Ssen*2, *Asen*2, *Ssenn,Asenn*
3. We tokenize each solution sentence and answer sentence present in the corpus. Sentence and corresponding tokens are define by: Sen ∈
4. After that we calculate comparison score CScurrent for every sentence Ssen in solution-sentences.
5. Compare it to every sentence Asen in answer-sentences and calculate current comparison score CCS in the following manner.
6. Keywords-weight Kw calculation:
7. Keep all keywords K present in Ssen as Ssenk.
8. Keep all the K present in both Ssen and Asen as Asenk
9. Calculate percentage of number of keywords K% i
10. Asenk w.r.t. Ssenk, this shows how many K current Asen contains w.r.t. current Ssen.
11. Calculate keywords-weight Kw by dividing keywords-percentage by 100, obtaining a value between 0 and 1.
12. Similarity distance calculation:
13. Calculate similarity distance Sd and similarity weight Sw between Ssen and Asen using either one of the following methods.
14. Word Movers Distance WM D method:
15. if Sd <=WMDlower then
16. $S_w = 1 - S_d$
17. $C_{CS} = S_w + K_w$
18. else if S_d <=$W_{MDupper}$ then
19. $S_w = 1.6 \quad S_d$
20. if K_w >= 0.3 (30% keywords present) then
21. $C_{CS} = S_w + K_w$
22. else if S_w == *null*(*nokeywordspresentin*S_{sen}) then
23. else if K_w < 0.3 (less than 30% keywords present) then

24. $C_{CS} = 0$
25. end if
26. end if
27. $CosineSimilarityC_{Sim}Method$:
28. if $S_d >= C_{SimUpper}$ then
29. $S_d = S_d$
30. CCS = Sw + Kw
31. else if Sdis >= CSimLower then
32. if Kw < 0.3(30%keywordspresent) then
33. CCS = Sw + Kw
34. else if Kw == null (no keywords present in Ssen then
35. CCS = Sw
36. else if Kw < 0.3 (less than 30% keywords present then
37. CCS = 0
38. end if
39. end if
40. if CCS > CScurrent then
41. CScurrent = CCS
42. end if
43. Calculate overall score (0S) by taking average of CScurrent of all Ssen.
44. Calculate missing keywords penalty (MKp) as percentage of keywords found is Ssen but not in Asen for it's highest $CS_{current}$.
45. Reduce 0_S by $M_{Kp}/1.6$.
46. Calculate unmapped S_{sen} penalty (Um_{Ssen}) which is percentage of S_{sen} that couldn't be mapped to any A_{sen}.
47. Reduce O_S by Um_{Ssen}.
48. Calculate unmapped A_{sen} penalty Um_{Asen} which is percentage of AS that could not be mapped to any S_{sen}.
49. Reduce O_S by Um_{Asen}.
50. Return O_S as overall score of that answer.
    - The preprocessed answer is used as testing data with the machine learning model to predict its class/category, and that category is checked with the result obtained from Result Prediction Module. This gives us confidence in the predicted result from the model.
    - The preprocessed answer is fed into the machine learning model along with its label. Moreover, the model is updated according to new data.
    - The predicted class is sent to the Final Score Prediction Module along with the solution, answer, and the *overall score*.

The advantage of the model is that it acts as a confidence booster for the Result Prediction Module, provided it has been trained on enough data. Furthermore, it can stand for its own and can be used to predict the grades/class of an answer once it has been trained on enough data. This eliminates the need for word2vec or Result Prediction Module discussed before and produces a model that can be used as a standalone evaluator for that particular question.

It also helps deal with the abnormal cases where the Result Prediction Model fails to predict the correct result for a particular answer due to semantic dissimilarity on behalf of the less trained word2vec model.

g) Final Score Prediction Module

This module is shown in Figure 4; it takes input from the machine learning module and validates the overall score with the class obtained from the machine learning module. Suppose the class matches the score. The score is considered finalized. If the class does not match the score, then the addition or deduction of half the number of values in that range is made based on whether the model suggested score is greater or lesser than the Similarity equivalent score.

It is either assumed that the machine learning model is not trained enough and the score is considered true or if the model has been extensively trained, adjusted score after the model suggestion is considered final, Accepting some inaccuracy from both the Score Prediction the Machine Learning Module.

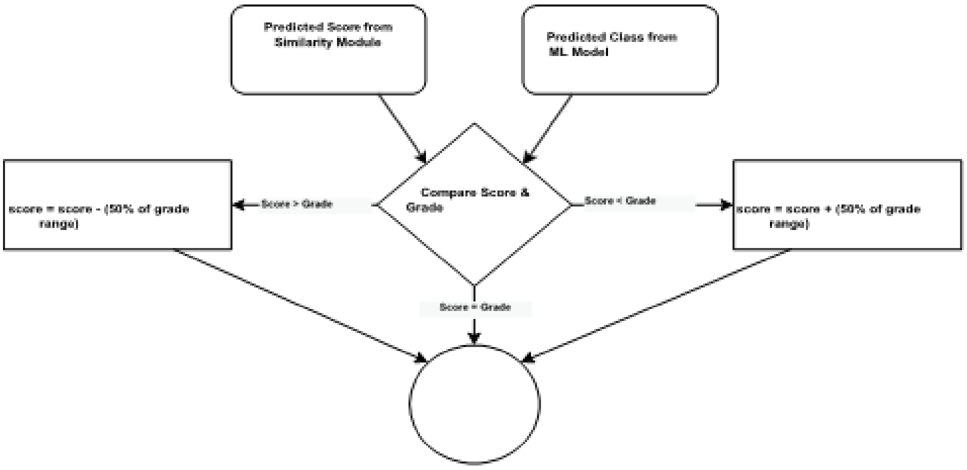
Predicted Score from Similarity Module
Predicted Class from ML Model
Compare Score & Grade
Score > Grade
Score < Grade
score = score - (50% of grade range)
score = score + (50% of grade range)
Score = Grade

Fig 4: Flow Chart of Final Score Prediction Module

Experimentation and Results

The experiment setup consists of a python notebook running on a web-based Google Co lab portal with a RAM of 12 GB and an HDD of 100+ GB. No GPU is turned on for this experiment.

A pre-trained word2vec model from Google consisting of 300 dimensions of around 100 Billion words vocabulary is used for this experiment. Corpus was divided into 8:2 ratio representing test and train data, respectively. Train data was used to calculate initial scores from the score prediction modules and train the machine learning model. Afterward, testing data was fed to the system one by one, updating the machine learning model.

The results are obtained using cosine similarity and word mover's distance combined with a Multinomial Naive Bayes model. Both the approaches with and without the model produced results in under a minute at Google Colab. The results are as follows. Table 3 shows the comparison of the first ten answers used for training purposes. The score prediction module is working fairly accurately, achieving 88% accuracy. This much accuracy is significant because of word2vec in this case, and it can capture the semantic meaning of answers so well that it gives us very well Similarity among answers. Furthermore, if word2vec lacks inconsistent answers, key-word mapping and unmapped sentences thresholds still give a satisfactory score to the answers.

Table 3: Score Prediction Using WDM before Model Suggestion

Human Score	WDM Approach Score	Error (%)
23	33	10
74	51	23
10	1	9
5	0	5
0	0	0
46	32	14
60	67	7
80	52	28
20	11	9
70	83	13

Table 4: Score Prediction Using WDM with Model Suggestion

Human Score	Error without Model	Error with Model
46	22	9.5
46	17	4.5
27	22	9.5
0	0	12.5
77	40	27.5
27	26	13.5
60	13	25.5
60	14	26.5
55	9	3.5
55	25	12.5

Table 4 shows the error when evaluating subjective answers with and without involving the model. It shows that the average errors decrease from 15.6% to 13.94% when using model suggestions for this small data set. The model's confidence level is likely to increase from 64% as the model keeps training more and more on the answers. This is a good feature of the proposed system, which leverages machine learning models to give confidence and suggestion to Similarity induced scores. Table 5 shows the errors in scores

Table 5: Score Prediction Using Cosine Similarity Before Model Suggestion

Human Score	Cosine Approach Score	Error % Age
23	33	10
74	72	2
10	17	7
5	0	5
0	9	9
46	34	12
60	79	19
80	72	8
20	34	14
70	95	25

Evaluated using the cosine similarity approach without any model suggestion. The results show an accuracy of 87%, primarily attributed to the proposed algorithm where keywords and sentence mapping play a massive role in the end. Cosine similarity performs poorly semantic-wise compared to WDM but can make some pretty good estimates where semantics are unnecessary.

Table 6 shows the difference in errors resulting from the machine learning model correction. It shows that the model's

Table 6: Score Prediction Using Cosine Similarity with Model Suggestion

Human Score	Error Without Model	Error With Model
46	13	0.5
46	13	0.5
60	18	30.5
60	18	30.5
55	9	3.5
55	24	11.5
27	19	6.5
0	13	25.5
77	27	14.5
27	1	13.5

Accuracy decreased by 1.54% when using cosine similarity along with classification models. This is because the results obtained by cosine similarity are semantically weak, and the model cannot get trained on the correct data as it does for the WDM case. Cosine similarity paired with a machine learning model yields 86% accuracy for this short dataset. Figure 5 shows the comparison of accuracy obtained from various combinations.

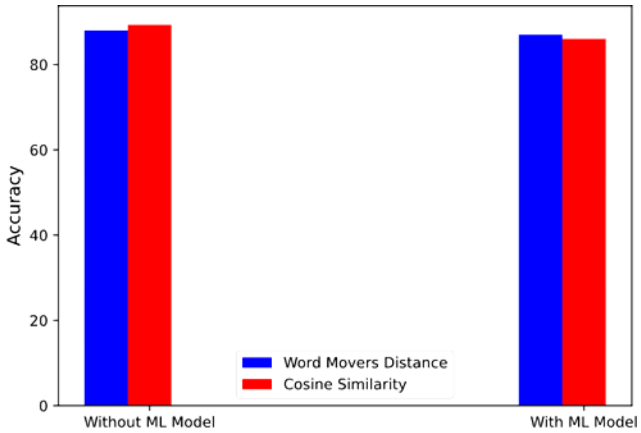

Fig 5: Accuracy Comparison of different models

1. **Tool for Evaluating Subjective Answers using AI(TESA)(2021):** Authors: Shreya Singh,Omkar Manchekar,Ambar Patwardhan All the studies which have been reviewed show that there are various different techniques for the evaluationof subjective answer sheets. The advantage of the system lies in the fact that it uses a weighted average of the closest to accurate techniques to provide the most optimized result.TESA is a systematic and reliable system which eases the role of evaluators and provides faster and more efficient outputs.
2. ***Assess-Automated subjective answer evaluation using Semantic Learning:*** Authors: Nidhi Dedhia,Kunal Bohra,Prem Chandak This automated approach is beneficial when students need to be assessed online for self-improvement. This system gives special emphasis to the specially-abled by providing various speech-based

usability features, where the gaps are filled by providing audio facilities like listening to the questions and answering them verbally. The advantage of this system is that it is near completion, has improved performance and caters to a very large audience.

3. **Automated Answer-Checker:** Authors: Vasu Bansal, M.L. Sharma, Krishna Chandra Tripathi the proposed system could be of great utility to the educators whenever they need to take a quick test for revision purposes, as it saves time and the trouble of evaluating the bundle of papers.

   This System would be beneficial for the universities, schools and colleges for academic purpose by providing ease to faculties and the examination evaluation cell

Conclusion

This paper proposed a novel way to deal with emotional responses assessment in light of AI and regular language handling strategies. Two score expectation calculations are proposed, which produce up to 88% precise scores. Different similitude and difference limits are contemplated, and different measures, for example, the watchword's presence and rate planning of sentences are used to defeat the strange instances of semantically free responses. That's what the trial-and-error results show, on normal word2vec approach performs better compared to conventional word inserting methods as it watches out for the semantics. Moreover, Word Mover's Distance performs better compared to Cosine Likeness generally speaking

However, subjective questions and answers are not required as the evaluation process is complex and efficient. Automated answer-checking applications that check written answers and mark weights just like humans do are more useful in today's modern world. Therefore, software applications designed to check subjective responses are even more useful in assigning grades to users after checking their responses in online exams

References

1. Srivastava G, Maddikunta PKR, Gadekallu TR. “A two-stage text feature selection algorithm for improving text classification,” 2021.
2. Mangassarian H, Artail H. “A general framework for subjective information extraction from unstructured english text,” Data Knowl. Eng. 2007; 62(2):52-367.
3. Oral B, Emekligil E, Arslan S, Eryigit G. “Information extraction from text intensive and visually rich banking documents,” Inf. Process. Manag. 2020; 57(6):102-361.
4. Khan H, Asghar MU, Asghar MZ, Srivastava G, Maddikunta PKR, Gadekallu TR. “Fake review classification using supervised machine learning,” in Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, Proceedings, Springer International Publishing, 2021.
5. Afzal S, Asim M, Javed AR, Beg MO, Baker T, “Urldeepdetect: A deep learning approach for detecting malicious urls using semantic vector models,” Journal of Network and Systems Management. 2021; 29(3):1-27.
6. Madnani N, Cahill A. “Automated scoring: Beyond natural language processing,” in Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26 2018 (E. M. Bender, L. Derczynski, and P. Isabelle, eds.), Association for Computational Linguistics, 2018, 1099-1109.
7. Lin Z, Wang H, McClean SI. “Measuring tree similarity for natural language processing based information retrieval,” in Natural Language Processing and Information Systems, 15th International Conference on and helps train the machine learning model faster. With enough training, the model can stand on its own and predict scores without the need for any semantics checking.
8. Grefenstette G, “Tokenization,” in Syntactic Wordclass Tagging, Springer, 1999, 117-133.
9. Sirts K, Peekman K. “Evaluating sentence segmentation and word tokenization systems on estonian web texts,” in Human Language Technologies-The Baltic Perspective- Proceedings of the Ninth International Conference Baltic HLT 2020,
10. Kaunas, Lithuania, September 22-23, 2020 (U. Andrius, V. Jurgita, K. Jolantai, and K. Danguole, eds.), vol. 328 of Frontiers in Artificial Intelligence and Applications, IOS Press, 2020, 174-181.
11. Schofield M, Magnusson, Mimno DM. “Pulling out the stops: Rethinking stop word removal for topic models,” in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers (M. Lapata, P. Blunsom, and A. Koller, eds.), Association for Computational Linguistics, 2017, 432-436.
12. Wang J, Dong Y. “Measurement of text similarity: A survey,” Inf. 2020; 11(9):421.
13. Han M, Zhang X, Yuan X, Jiang J, Yun W, Gao C. “A survey on the techniques, applications, and performance of short text semantic similarity,” Concurr. Comput. Pract. Exp., 2021,33(5).
14. Patil MSM, Patil MS. “Evaluating student descriptive answers using natural language processing,” *International Journal of Engineering Research & Technology (IJERT)*. 2014; 3(3):1716-1718.
15. Patil P, Patil S, Miniyar V, Bandal A. “Subjective answer evaluation using machine learning,” *International Journal of Pure and Applied Mathematics*. 2018; 118(24):1-13.
16. Muangprathub J, Kajornkasirat S, Wanichsombat A, “Document plagiarism detection using a new concept similarity in formal concept analysis,” Journal of Applied Mathematics, 2021.
17. Hu X, Xia H, “Automated assessment system for subjective questions based on lsi,” in 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 2010, 250-254.
18. Kusner M, Sun Y, Kolkin N, Weinberger K. “From word embed-dings to document distances,” in International conference on machine learning, PMLR, 2015, 957-966.
19. Xia C, He T, Li W, Qin Z, Zou Z. “Similarity analysis of law documents based on word2vec,” in 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C), IEEE, 2019, 354-357.
20. Mittal H, Devi MS. “Subjective evaluation: A comparison of several statistical techniques,” Appl. Artif. Intell. 2018; 32(1):85-95.
21. Cutrone LA, Chang M. “Automarking: Automatic assessment of open questions,” in ICALT 2010, 10th IEEE International Conference on Advanced Learning Technologies, Sousse, Tunisia, IEEE Computer Society, 2010, 143-147.
22. M. Çagatayli and E. Çelebi, “The effect of stemming and stop-word-removal on automatic text classification in turkish language,” in Neural Information Processing-22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9-12, 2015, Proceedings, Part I (S. Arik,T. Huang, W. K. Lai, and Q. Liu, eds.), of Lecture Notes in Computer Science, Springer, 2015; 9489:168-176.
23. Divyapushpalakshmi M, Ramar R. “An efficient sentimental analysis using hybrid deep learning and optimization technique for twitter using parts of speech (POS) tagging,”*Int. J Speech Technol*. 2021; 24(2):329-339.

24. Camastra F and Razi. “Italian text categorization with lemmatization and support vector machines,” in Neural Approaches to Dynamics of Signal Exchanges (A. Esposito,
25. M. Faúndez-Zanuy, F. C. Morabito, and E. Pasero, eds.), vol. 151 of Smart Innovation,
26. Systems and Technologies. Jabbar, S. Iqbal, M. Ilahi, S. Hussain, and A. Akhunzada, “Empirical evaluation and study of text stemming algorithms,” Artif. Intell. Rev, Springer. 2020, 47-54, 53(8):5559-5588,
27. 2020.
28. Aryal S, Ting KM, Washio T, Haffari G. “A new simple and ef-fective measure for bag-of-word inter-document similarity measurement,” CoRR, vol. abs/1902.03402,2019.
29. “TF-IDF,” in Encyclopedia of Machine Learning (C. Sammut and G. I. Webb, eds.), Springer, 2010, 986-987.
30. Havrlant L, Kreinovich V. “A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation),” *Int. J Gen. Syst.* 2017; 46(1):27-36.
31. Thakkar K. Chaudhari, “Predicting stock trend using an integrated term frequency-inverse document frequency-based feature weight matrix with neural networks,” Appl. Soft Comput. 2020; 96:106684.
32. Jin X, Zhang S, Liu J. “Word semantic similarity calculation based on word2vec,” in 2018 International Conference on Control, Automation and Information Sciences, ICCAIS 2018, Hangzhou, China, 12-16, IEEE, 2018, 24-27.
33. Park K, Hong JS, Kim W. “A methodology combining cosine similarity with classifier for text classification,” Appl. Artif. Intell. 2020; 34(5):396-411.
34. Sato R, Yamada M, Kashima H, “Re-evaluating word mover’s distance,” CoRR, 2021, abs/2105.14403.
35. Kim JE, Park K, Chae JM, Jang HJ, Kim BW, Jung SY. “Automatic scoring system for short descriptive answer written in korean using lexico-semantic pattern,” Soft Computing. 2018; 22:13):4241-4249.
36. Oghbaie M, Zanjireh MM. “Pairwise document similarity measure based on present term set,” Journal of Big Data. 2018; 5(1)1-23.
37. Orkphol K, Yang W. “Word sense disambiguation using cosine sim-ilarity collaborates with word2vec and wordnet,” Future Internet. 2019; 11(5):114.
38. Wagh RS, Anand D. “Legal document similarity: a multi-criteria decision-making perspective,” PeerJ Computer Science. 2020; 6:e262.
39. Alian M, Awajan A. “Factors affecting sentence similarity and paraphrasing identification,” *International Journal of Speech Technology*. 2020; 23(4):851-859.
40. Jain G, Lobiyal DK. “Conceptual graphs based approach for subjec-tive answers evaluation,” *Int. J Concept. Struct. Smart Appl.* 2017; 5(2):1-21.
41. Montes M, Lopez-Lopez A, Gelbukh A. “Information retrieval with conceptual graph matching. 2000; 1873:312-321.
42. Bahel V, Thomas A. “Text similarity analysis for evaluation of descriptive answers,” CoRR, vol. abs/2105.02935, 2021.
43. Qurashi W, Holmes V, Johnson AP. “Document processing: Methods for semantic text similarity analysis,” in 2020 International Conference on Innovations in Intelligent Systems and Applications (INISTA), 2020, 1-6.
44. Jagadamba G, C Shree G. ”Online Subjective answer verifying system Using Artificial Intelligence,” 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2020, 1023-1027.
45. Bashir MF, Arshad H, Javed AR, Kryvinska N, Band SS, “Subjective Answers Evaluation Using Machine Learning and Natural Language Processing,” in IEEE Access. 2021; 9:158972-158983.
46. Singh S, Manchekar O, Patwardhan A, Rote U, Jagtap S, Chavan H. ”Tool for Evaluating Subjective Answers using AI (TESA),” 2021 International Conference on Communication information and Computing Tech-nology (ICCICT), 2021.
47. Johri, Era and Dedhia, Nidhi and Bohra, Kunal and Chandak, Prem and Adhikari, Hunain, ASSESS-Auto-mated Subjective Answer Evaluation Using Semantic Learning (May 7, 2021). Proceedings of the 4th International Conference on Advances in Science Technology (ICAST2021).
48. Vasu Bansal M.L. Sharma and Krishna Chandra Tripathi, ”Automated Answer-Checker” December 2020In-ternational Journal for Modern Trends in Science and Technology 6(12):152-155, DOI:10.46501/IJMTST061229
49. Wang J, Dong Y. “Measurement of text similarity: A survey,” Inf. 2020; 11(9):421.
50. Han M, Zhang X, Yuan X, Jiang J, Yun W, Gao C. “A survey on the techniques, applications, and performance of short text semantic similarity,” Concurr. Comput. Pract. Exp, 2021, 33(5).
51. Patil MSM, Patil MS. “Evaluating student descriptive answers using natural language processing,” *International Journal of Engineering Research Technology (IJERT)*. 2014; 3(3):1716-1718.
52. Patil P, Patil S, Miniyar V, Bandal A. “Subjective answer evaluation using machine learning,” *International Journal of Pure and Applied Mathematics*, 2018, 118, 24:1-13
53. Muangprathub J, Kajornkasirat S, and Wanichsombat A. “Document plagiarism detection using a new concept similarity in formal concept analysis,” Journal of Applied Mathematics, 2021.
54. Hu X, Xia H. “Automated assessment system for subjective questions based on lsi,” in 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 2010, 250-254, IEEE,.
55. Kusner M, Sun Y, Kolkin N, Weinberger K. “From word embedding’s to document distances,” in International conference on machine learning, PMLR, 2015, 957-966.
56. A new model for evaluating subjective online ratings with uncertain intervals FJ Santos-Arteaga, M Tavana, D Di Caprio-Expert Systems with …, 2020-Elsevier
57. Effects of attribute and valence of e-WOM on message adoption: Moderating roles of subjective knowledge and regulatory focus
58. KT Lee, DM Koo-Computers in human behavior, 2012-Elsevier
59. Massive online crowdsourced study of subjective and objective picture quality D Ghadiyaram, AC Bovik-IEEE Transactions on Image …, 2015-ieeexplore.ieee.org