MFDS - CH5019 - Project Report

Descriptive Answer Evaluation

Group No : 30

Shrivarshan K (MM20B058)
Elango L (MM20B018)
Gatkal Siddhesh Sarjerao (MM20B019)
Arjit Tripathi (MM20B008)
Pratham Khandelwal (MM20B047)

1st June 2022

1 Problem Statement

An instructor wants to grade answers to descriptive questions automatically. The instructor has a template best answer that she/he has developed and wants to use the same to grade descriptive answers of students. The comparison results between the template and the student provided answer could be categorical or continuous. The aim is to develop an AI algorithm that can do this comparison automatically. The algorithm can be developed through any appropriate concept and the results demonstrated on ten test cases. Any training approach can be used and test it on any 10 test cases that you feel are appropriate. The algorithm should take two paragraphs, one correct answer and one student provided answer and return a result.

2 Introduction

Thousands of exams all over the world are subjective in nature and Subjective paper evaluation is a tricky and tiresome task to do by manual labor. Insufficient understanding and acceptance of data are crucial challenges while analyzing subjective papers using Artificial Intelligence (AI). Though there have been wide advances in the field of Natural Language Processing (NLP), there are no fool-proof and 100% dependable algorithms which can do the job of subjective analysis.

There has been lots of existing work in the area as found in our literature survey of this problem statement (All references are mentioned at end of report) Most of these use traditional counts or specific words to achieve this task. The major problem that we faced was a lack of curated datasets. In our search for possible data, there was not a single dataset which contained relevant data in the form of model and student answers.

3 Formulation

3.1 Overview

Since the project is an open ended problem, we have formulated it as follows. Given a model answer and student answer, our model should grade the student's answer out of 5 in a discrete manner (i.e. the possible marks being 0,1,2,3,4 and 5 only). In this project, we aim to tackle only brief subjective answers (about 60-70 words long). Given that such answers contain only enough subjectivity and content to be graded in such a manner, we have gone ahead with the above approach. Our problem is now essentially a classification problem.11[11]1

3.2 Approach

Several preprocessing steps are performed on the answers, such as cleaning the data and tokenization before working on it. We then compare the student answers to the model answers. Our idea is to check for semantic similarity in both paragraphs in terms of similarity in language,keywords and structure. One possible approach we considered was directly correlating the similarity scores to grades.

However, it was our belief that semantic similarity by itself could not capture the closeness in the answers. We also felt that a human touch of subjectivity was lacking. Thus the work presented in this project involves the creation of a small dataset consisting of model answer,student answer and a score manually assigned by us for each pair. 3 different similarity algorithms (explained in detail) are run on this data. These scores are then treated as a feature vector to be fed into a supervised classification algorithm. For this purpose we make use of Random Forest. As many pre-existing ML libraries and models as possible are used.

4 Preprocessing

4.1 Lowercasing and Punctuation removal

For sake of uniformity all sentences are converted into lowercase and punctuation marks are removed.

4.2 Tokenization

Tokenization is a key (and mandatory) aspect of working with text data. Tokenization is a way of separating a piece of text into smaller units called tokens. Here we use sentence tokenization where the delimiter is the full stop.

4.3 Stop Word removal

For our tasks of grading, stop words are not needed as the other words present in the dataset are more important and give the general idea of the text. Infact they add noise to the sentences most of the time. Common stopwords include I, is, where, to, him, etc..

4.4 Lemmatization

Lemmatization is the grouping together of different forms of the same word.For example, to lemmatize the words “cats,” “cat’s,” and “cats’” means taking away the suffixes to bring out the root word cat. One major problem that we faced was to whether we had to lemmatize considering the fact the bringing down the word to the root results in loss of meaning especially in subjective evaluation.

5 Similarity Algorithms - to find match score

5.1 Cosine Similarity - I

Word Embedding - term used for the representation of words in terms of real-valued vectors such that the meaning of the word is mathematically encoded. In this method of finding similarity between two texts, we make use of the Bag of Words(BoW) model for Word Embedding. For cosine similarity, we have :

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

	data	digital	economy	is	new	of	oil	the
doc_1	1	1	1	1	0	1	1	2
doc_2	1	0	0	1	1	0	1	0

$$A \cdot B = \sum_{i=1}^{n} A_i B_i$$

$$= (1 * 1) + (1 * 0) + (1 * 0) + (1 * 1) + (0 * 1) + (1 * 0) + (1 * 1) + (2 * 0)$$

$$= 3$$

$$\sqrt{\sum_{i=1}^{n} A_i^2} = \sqrt{1 + 1 + 1 + 1 + 0 + 1 + 1 + 4} = \sqrt{10}$$

$$\sqrt{\sum_{i=1}^{n} B_i^2} = \sqrt{1 + 0 + 0 + 1 + 1 + 0 + 1 + 0} = \sqrt{4}$$

$$cosine\ similarity = cos\theta = \frac{A \cdot B}{|A||B|} = \frac{3}{\sqrt{10} * \sqrt{4}} = 0.4743$$

Cosine similarity is one of the metric to measure the text-similarity between two documents irrespective of their size. The Cosine similarity of two documents will range from 0 to 1 where values closer to 1 indicate more similarity. We have manually implemented the above algorithm

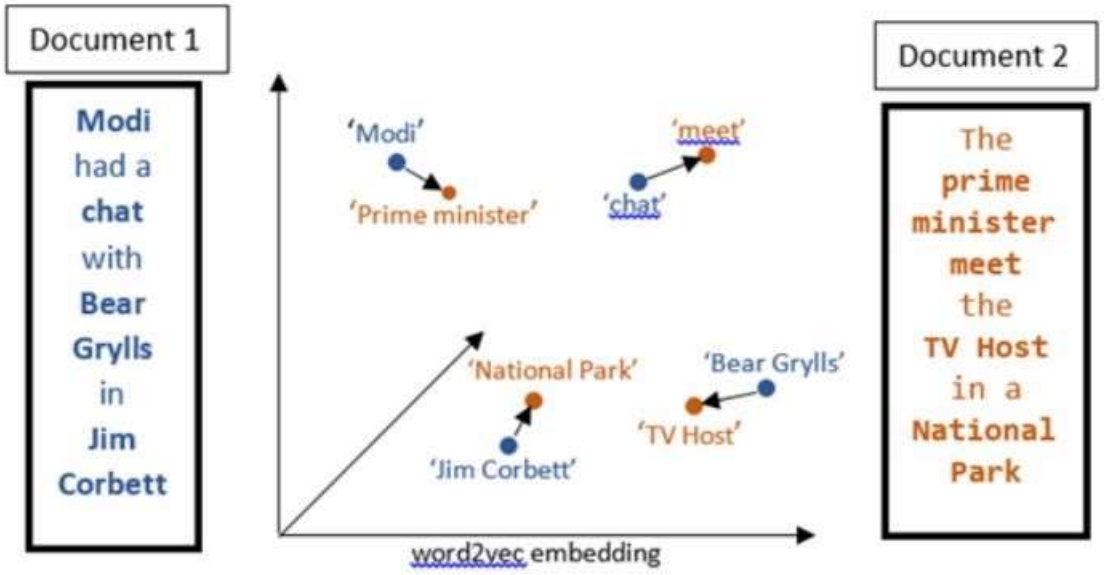
5.2 Cosine Similarity - 2

In this method of similarity, we use a state-of-the art pretrained model named BERT for sentence embeddings. BERT provides an easy method to compute dense vector representations for sentences and paragraphs. The above is encapsulated in the sentence transformers library which has its own cosine similarity method that we make use of.

5.3 Word Mover Distance

Word Mover's Distance (WMD), suggests that distances and between embedded word vectors are to some degree semantically meaningful. It utilizes this property of word vector embeddings and treats text documents as a weighted point cloud of embedded words. The distance between two text documents A and B is calculated by the minimum cumulative distance that words from the text document A needs to travel to match exactly the point cloud of text document B.

We make use of Word2Vec is a predictive word embedding technique which converts a word into a vector of numbers based on the context of the target word. It uses Neural Network whose hidden layer encodes the word representation.



The word mover distance which indicates how much we need to move to go from words of Text A to Text B is then calculated using the gensim library via the google-newsvectors-negative 300 pretrained model.

6 Custom Dataset

To capture the relation between the similarity scores obtained through three different approaches and the subjective grade out of 5, a very small dataset was created. The dataset contains of 60 data samples, each sample containing a model answer, a student answer and a class label. The answers were prepared by us ensuring a wide variety of topics and ensuring that synonyms, reordering and restructuring of sentences were used. These were then assigned class labels (grades out of 5) by averaging all of our 5 manually assigned scores.

Grade	number of answers
0	7
1	11
2	11
3	11
4	11
5	9

Table 1 : Distribution of class labels in the answers data

The final dataset was prepared by running our 3 similarity algorithms on these 60 pairs of answers and obtaining 3 scores. i.e. the dataset used for training consists of 60 data samples of 3 features each and a class label between 0 to 5.

Important note - In this project, we aim to create a model that grades the student's answer purely based on the model answer. i.e say if a student answers a question on a topic X with relevant, meaningful points, but such that none of his points match the ones on the model answer, the model will be forced to give him a 0.

7 Multi-Class Classifier

To solve our 6 class classification problem, we make use of the Random Forest algorithm. Our choice was due to us having us a very small dataset and needing a reasonable accuracy. Other models such as SVM and XGBoost were used but RF was found to give the best results.

7.1 Random Forest - The Algorithm

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. It is based on ensemble learning.

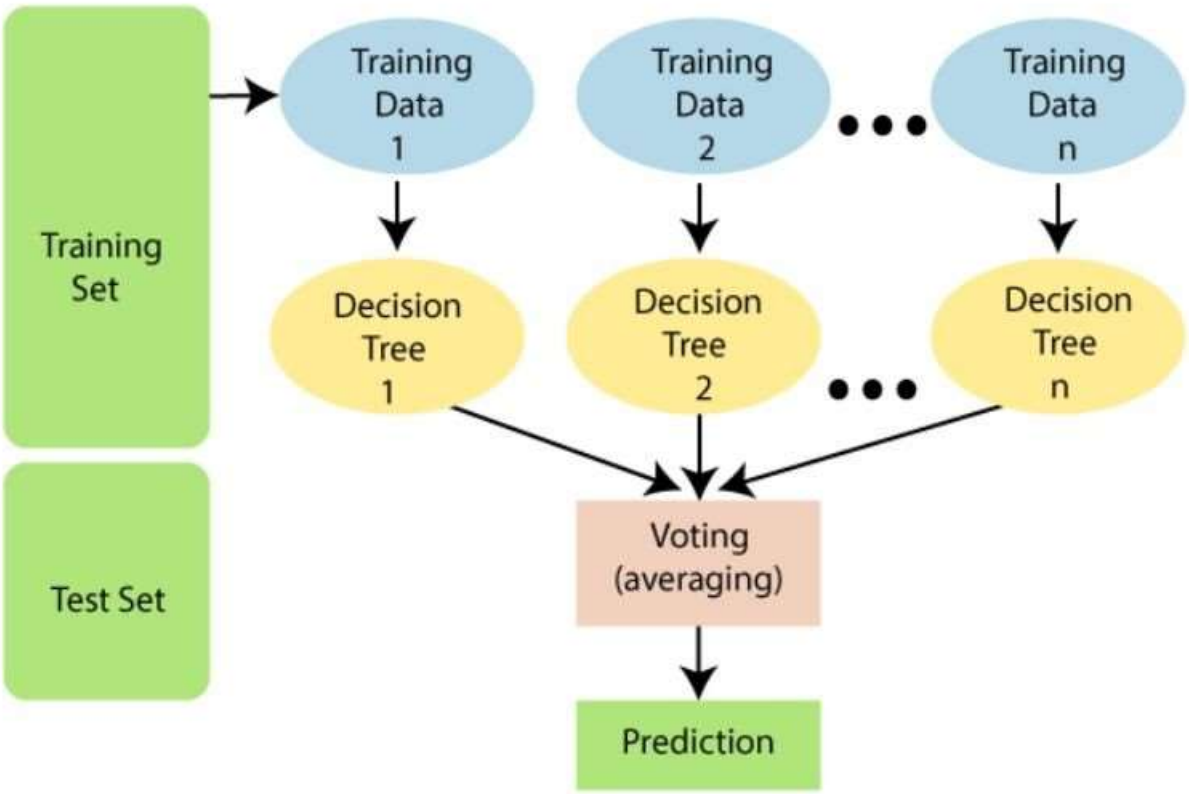


Fig 1 : Illustration for Random Forest(9)

7.2 Training

The dataset was split as 90-10 for training and testing. i.e. we have 54 training samples and 6 testing samples. The best results were obtained for the following hyperparameters : Due to the number of data samples being very very low, there is very to little significance for plotting the learning curve. Hence, we are not attaching any results related to the training (Loss metrics, accuracy, confusion matrix, etc..). The best parameters were determined subjectively only. However, the results on the 6 samples are more or less accurate.

Hyperparameter	Value
n_estimators	30
min_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0

The RandomForestClassifier class from the scikit learn library was used.

For predictions, when we get the input, the feature vector is constructed (by similarity algorithms) and the saved model is used for classification.

8 Grammar

For checking whether the language is used by the student is upto the certain mark, we make use of the textgear api to find number of grammar errors. If there are more than 5 errors and the student has scored more than 5, we deduce 1 mark.

9 Results

1. Model answer : Kerala is known as God's own country. It is the tagline of Kerala Tourism which was coined by Walter Mendez. The wealth of natural beauty in the form of placid backwaters, lush greenery, picturesque hill towns, and beautiful beaches has resulted in making it a stunning place to visit. Along with this, the rich cultural heritage with it's exotic musical and dance forms makes it a traveller's paradise.
   Student answer : God's own country is the state of Kerala. It is one of the most popular tourist destinations. Culture plays a major role in making it a paradise on earth. Nature's beauty is in full view here because we have exotic beaches, beautiful and calm backwaters and lush greenery throughout. Added to this, the beautiful architecture, the rich flora and fauna and the amazing weather makes it god's own country.

   Our model grades it as : 4

2. Model answer : Chicago is the most populous city in the U.S. state of Illinois. Famed for its bold architecture, it has a skyline punctuated by skyscrapers. It is home to the world's busiest airports in O'Hare International. Located on the shores of freshwater Lake Michigan, Chicago was incorporated as a city in 1837. It is an international hub for finance, culture, commerce, industry, education.
   Student Answer : Chicago is located in the Indian state of Tamil Nadu. It has home to beautiful temples, football stadiums as well as interesting museums. Some of the world renowned landmarks here include the Kapaleeshwar temple as well as the Marina Beach. There are no rivers in Chicago leading to water scarcity during the very hot summers. The entire population is centred in the downtown part of the city.

   Our model grades it as : 0

3. Model Answer : Adsorption is a mass transfer process that is a phenomenon of sorption of gases or solutes by solid or liquid surfaces. The adsorption on the solid surface is that the molecules or atoms on the solid surface have residual surface energy due to unbalanced forces. Adsorption is a surface phenomenon. There are two kinds of adsroption which are Physisorption and Chemisorption respectively.
   Student Answer : The accumulation of the molecular species at the surface rather

than in the bulk of a solid or liquid is known as adsorption. For example, water vapour are adsorbed by silica gel. When a gas is absorbed on the surface of a solid its entropy decreases. The two types of adsorption are Physisorption and Chemisorption. Adsorption is an endothermic process since we have to heat the material for it to occur.

Our model grades it as : 3

4. Model Answer : Rainforests are characterized by a closed and continuous tree canopy and a moisture-dependent vegetation. Rainforests can be mainly classified as tropical rainforest or temperate rainforest. The presence of epiphytes and lianas and the absence of wildfires are common features in rainforests. The world's largest rainforest is the Amazon rainforest. Rainforests cover less than 3 percent of the planet.
   Student Answer : Rainforests cover a small percent of the planet and are identified by the continous canopy of trees. The flora and fauna here are unique. Amazon is considered to be the world' largest rainforest. Rainforests are of two types which are Tropical and Temperate. Rainforests receive moderate amount of rainfall and are considered to be the home to several rare tribes and flowers.

   Our model grades it as : 4

5. Model Answer : Social media has become a part of our life. Most of the people have started having low self-esteem, therefore, they need social validation to feel good about themselves. Social capital has become important in today's world. People value themselves with the number of likes and comments they have got. In essence, we are obsessed with being accepted by people and social media helps with that.
   Student Answer : Because some people are more confident behind a screen than in front of others. These are the kind of people who want attention but are too uncomfortable to ask for it. So they grab the social media platform as an opportunity to do just that. Usually the more active people are online, the more quiet they are in real life. The facade of social media actually gives them a chance to show people what they have got without being uncomfortable in their presence.

   Our model grades it as : 1

6. Model Answer : Digital marketing, also called online marketing, is the promotion of brands to connect with potential customers using the internet and other forms of digital communication. This includes not only email, social media, and webbased advertising, but also text and multimedia messages as a marketing

channel. Several companies active in this field are Flipkart, Amazon, Myntra, etc.. which are online shopping portals.

Student Answer : Digital marketing is the marketing and advertising of a business, person, product, or service using online channels, electronic devices, and digital technologies. A few examples of digital marketing include social media, email, payper-click (PPC), search engine optimization (SEO), and more. Digital marketing can be a boon or a bane. In today's world, it has become a necessity by all means.

Our model grades it as : 2

7. Model Answer: Homeopathy is a pseudoscientific system of alternative medicine. Its practitioners, called homeopaths, believe that a substance that causes symptoms of a disease in healthy people can cure similar symptoms in sick people. It is an unconventional medical system that was developed in Germany more than 200 years ago. Homoeopathic medicines are prepared in a standardized, well-controlled and hygienic environment.

   Students Answer : Homeopathy is an unconventional style of medicine derived from pseudoscientific principles. The rise in popularity was due to them being safe and being prepared in a controlled and hygienically pure environment. This system of alternative medicine was developed in Germany more than a century ago. Homeopaths believe that substances which cause the disease can be used as cure of symptoms in sick people.

   Our model grades it as : 5

8. Model Answer : Probability is the branch of mathematics concerning numerical descriptions of how likely an event is to occur, or how likely it is that a proposition is true. The probability of an event is a number between 0 and 1.The higher the probability of an event, the more likely it is that the event will occur. Probability of zero need not imply zero possibility.

   Student Answer : Probability means probability. It is a branch of mathematics that deals with the occurrence of a random event. The value is expressed from zero to one. The probability formula is defined as the possibility of an event to happen is equal to the ratio of the number of favourable outcomes and the total number of outcomes. Probability is a unique chapter and has foundations in multiple areas like Data science.

   Our model grades it as : 2

9. Model Answer : Silk is a natural protein fiber, some forms of which can be woven into textiles. The protein fiber of silk is composed mainly of fibroin and is produced by certain insect larvae to form cocoons. The best-known silk is obtained from the cocoons of the larvae of the mulberry silkworm Bombyx mori

reared in captivity. Silk is known for its luster, shine, strength, and durability, and it has a long trading history.

Student Answer : Silk is one of the most sought after fibres in the world. Silk is obtained from certain insect cocoons. It is a very expensive and shiny material. Silkworms Lay Up To 300 Eggs and its a well known fact that China has a secret silk industry. Silkworms consume mulberry leaves and gives the textile its protein fiber like material.

Our model grades it as : 3

10. Model Answer : Dried fruit is fruit that has had almost all of the water content removed through drying methods. The fruit shrinks during this process, leaving a small, energy-dense dried fruit. Raisins are the most common type, followed by dates, prunes, figs and apricots.Rich in proteins, vitamins, minerals and dietary fibre, dry fruits make for a delicious and healthy snack.

    Student Answer : Any fruit which has its water content removed from it is called a dry fruit. A small fruit is obtained after the drying and shrinking process. The most common dry fruits include raisins,dates, figs, apricots and pistachios. Dry fruits often constitute a healthy snack and have a sweet yet sour taste. Some dry fruits are known to be very beneficial to the body and provide stamina.

    Our model grades it as : 4

10 Scope for improvement

- Increase in the number of data samples in the dataset. This will help increase model accuracy while also allowing us to use more number of discrete grades. • When there is a considerable discrepancy in the length of the student answer and the model answer, we find the model predictions are off.
- Further work can include on working on grading subjective answers by students which have different content but correct content as compared to model answers.

11 References

1. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (EMNLP2019)
2. Subjective Answers Evaluation Using Machine Learning and Natural LanguageProcessing : Muhammed Farrukh Bashir ,Hamza Arshad,Abdul Rehman Javed, (Member, IEEE), Natalia Kryviska , And Shahand S. Band

3. Subjective Answer Evaluation using Natural Language Processing and MachineLearning :Abhishek Girkar1, Mohit khambayat2, Ajay Waghmare3, Supriya Chaudhary4
4. https://www.sbert.net/docs/usage/semantic textual similarity.html
5. https://acadpubl.eu/hub/2018-118-24/3/577.pdf
6. Toma˜s Mikolov et al. Efficient Estimation of Word Representations in Vector Space, 2013.
7. https://studymachinelearning.com/cosine-similarity-text-similarity-metric/
8. https://medium.com/@nihitextra/word-movers-distance-for-text-similarity-7492aeca71b0
9. https://www.javatpoint.com/machine-learning-random-forest-algorithm