**LITERATURE SURVEY**

**Essay Grader using Ensemble Strategies and Word2Vector method**

1. **Introduction:**

This paper focuses on it evaluates people's essay and it keep track of their scores. Our project gives linguistic support to all the users and check the authenticity of the essay. For example, if a GRE or IELTS candidate looking for someone to evaluate their essay, our project will help them with grading of their essay and provide suggestion to improve their writing skill. This certainly reduces time and risk of falsified evaluation of the passage. To integrate the primary feature in other software, project’s primary service is hosted in cloud architecture and use API keys to access it securely. This also provides an attractive dashboard to the users to keep track of their scores and their essays.

1. **Overview of Studies:**

We have found some studies related to the essay grader. Some of the studies are listed below:

1. Paper Title: Text Classification using Different Feature Extraction Approach.

Publication: IEEE

DOI Number: [10.1109/eStream.2019.8732167](https://doi.org/10.1109/eStream.2019.8732167)

Description: It uses Term Frequency Inverse Document Frequency (TF-IDF) weighing scheme to extract feature’s advantage and uses Latent Semantic Analysis (LSA) and Linear Discriminant Analysis (LDA) to overcome TF-IDF’s limitation.

1. Paper Title: Analysis of Text Categorization Represented with Word Embeddings Using Homogeneous Classifiers.

Publication: IEEE

DOI Number: [10.1109/INISTA.2019.8778329](https://doi.org/10.1109/INISTA.2019.8778329)

Description: It uses Word embeddings and ensemble strategies to classify text with much higher accuracy. Word embedding dual neural network word2vector model to map words to real time integers according to syntax and semantic relativity. Ensemble strategies uses voting system to choose a model which provides much more accuracy and less training time.

1. Paper Title: Automatic Text categorization and summarization using rule reduction.

Publication: IEEE

INSPEC Accession Number: 12818753

Description: In this we use an automatic text categorization and summarization approach to analyze the structure of input text. In this work a text analyzer is developed to derive the structure of the input text using rule reduction technique in three stages namely, Token Creation, Feature Identification and Categorization and Summarization. This analyzer is tested with sample input texts and gives noteworthy results.

1. Paper Title: A Fuzzy Approach to Text Classification with Two-Stage Training for Ambiguous Instances.

Publication: IEEE

DOI: 10.1109/TCSS.2019.2892037

In this method we focus on cyberhate classification, since the spread of hate speech via social media can have disruptive impacts on social cohesion and lead to regional and community tensions. Automatic detection of cyberhate has thus become a priority research area. In particular, we propose a modified fuzzy approach with two-stage training for dealing with text ambiguity and classifying four types of hate speech, namely, religion, race, disability, and sexual orientation-and compare its performance with those popular methods as well as some existing fuzzy approaches.

1. Paper Title: Two-Level Hierarchical Hybrid SVM-RVM Classification Model

Publication: IEEE

DOI Number: [10.1109/ICMLA.2006.52](https://doi.org/10.1109/ICMLA.2006.52)

Description: The proposed model first level uses an RVM to determine the less confident classified examples and the second level then makes use of an SVM to learn and classify the tougher examples. We show the benefits of the hierarchical approach on a text classification task, where the two-levels outperform both learning machines.

1. Paper Title: Improving classification in data mining using hybrid algorithm

Publication: IEEE

DOI Number: [10.1109/IICIP.2016.7975380](https://doi.org/10.1109/IICIP.2016.7975380)

Description: Data mining is a powerful concept with great potential to predict future trends and behavior. It refers to the extraction of hidden knowledge from large datasets using techniques like statistical analysis, machine learning, clustering, neural networks and genetic algorithms. Hybrid algorithms for data mining are a logical combination of multiple pre-existing techniques to enhance performance and provide better results. The hybrid algorithm proposed in this paper uses the concept of clustering and decision tree induction to classify the data samples. When the proposed approach is tested on real life datasets, the results obtained show improved accuracy in most cases.

1. Paper Title: Noisy text categorization

Publication: IEEE

DOI Number: [10.1109/TPAMI.2005.248](https://doi.org/10.1109/TPAMI.2005.248)

Description: This work presents categorization experiments performed over noisy texts. By noisy, we mean any text obtained through an extraction process (affected by errors) from media other than digital texts (e.g., transcriptions of speech recordings extracted with a recognition system). The performance of a categorization system over the clean and noisy (word error rate between /spl sim/ 10 and /spl sim/ 50 percent) versions of the same documents is compared. The noisy texts are obtained through handwriting recognition and simulation of optical character recognition. The results show that the performance loss is acceptable for recall values up to 60-70 percent depending on the noise sources. New measures of the extraction process performance, allowing a better explanation of the categorization results, are proposed.

1. Paper Title: LDA based topic modeling of journal abstracts

Publication: IEEE

DOI Number: [10.1109/ICACCS.2015.7324058](https://doi.org/10.1109/ICACCS.2015.7324058)

Description: In this work, a dataset with 200 abstracts fall under four topics are collected from two different domain journals for tagging journal abstracts. The document models are built using LDA (Latent Dirichlet Allocation) with Collapsed Variational Bayes and Gibbs sampling. Then the built model is used to extract appropriate tags for abstracts. The performance of the built models is analyzed by the evaluation measure perplexity and observed that Gibbs sampling outperforms CV B0 sampling. Tags extracted by two algorithms remains almost the same.

1. Paper Title: Exploring Feature Selection and Support Vector Machine in Text Categorization.

Publication: IEEE

DOI Number: [10.1109/CSE.2013.160](https://doi.org/10.1109/CSE.2013.160)

Description: This paper explored several methods of feature selection that can be used to reduce high dimensionality of feature space in text documents such as Information Gain, Gain Ratio, CHI-Squares, Mutual Information and Document frequency. Next, the study adopted text categorization using Support Vector Machines. The results showed that Support Vector Machines perform well and very fast both in training and testing datasets.

1. Paper Title: HDLTex: Hierarchical Deep Learning for Text Classification

Publication: IEEE

DOI Number: [10.1109/ICMLA.2017.0-134](https://doi.org/10.1109/ICMLA.2017.0-134)

Description: This paper approaches this problem differently from current document classification methods that view the problem as multi-class classification. Instead we perform hierarchical classification using an approach we call Hierarchical Deep Learning for Text classification (HDLTex). HDLTex employs stacks of deep learning architectures to provide specialized understanding at each level of the document hierarchy.

1. Paper Title: Graph-Based Semantic Learning, Representation and Growth from Text: A Systematic Review.

Publication: IEEE

DOI Number: [10.1109/ICOSC.2019.8665592](https://doi.org/10.1109/ICOSC.2019.8665592)

Description: The Vector Space Model (VSM), is the main technique to model the semantics from the text. However, the VSM model suffers from notable limitations. The main alternative model for VSM model is a graph-based model. This paper presents a systematic review on the graph-based processes of Semantic Learning, Representing and Growth (SLRG) from the text. Then it describes a new branch in graph-based SLRG modeling, inspired from the cognitive-semantics.

1. Paper Title: Substation Transformer Failure Analysis Through Text Mining

Publication: IEEE

DOI Number: [10.1109/ISCAIE.2019.8743719](https://doi.org/10.1109/ISCAIE.2019.8743719)

Description: This study mainly focuses on bag-of-word text analysis approaches, which means that only word frequencies per text are used and word positions are ignored. Although this simplifies text content dramatically, research and many applications in the real world show that word frequencies alone contain adequate information for many types of analysis. As a result of analysis, keywords like "leak", "lightning", "animal", "cable" and "temperature" are identified as the main causes of transformer failures based on the number of word frequency in the tripping dataset. Further enhancement could be made in the future to predict the failure beforehand using predictive analytics approaches.

1. Paper Title: Deep Learning Model used in Text Classification

Publication: IEEE

DOI Number: [10.1109/ICCWAMTIP.2018.8632592](https://doi.org/10.1109/ICCWAMTIP.2018.8632592)

Description: Common text classification applications include spam identification, news text classification, information retrieval, emotion analysis, and intention judgment, etc. Traditional text classifiers based on machine learning methods have defects such as data sparsity, dimension explosion and poor generalization ability, while classifiers based on deep learning network greatly improve these defects, avoid cumbersome feature extraction process, and have strong learning ability and higher prediction accuracy. For example, convolutional neural network (CNN)[I]. This paper introduces the process of text classification and focuses on the deep learning model used in text classification.

1. Paper Title: Application of text classification and clustering of Twitter data for business analytics

Publication: IEEE

DOI Number: [10.1109/MINTC.2018.8363162](https://doi.org/10.1109/MINTC.2018.8363162)

Description: It uses machine learning techniques and tools in determining patterns and trends to gain actionable insights. This paper selected a popular food brand to evaluate a given stream of customer comments on Twitter. Several metrics in classification and clustering of data were used for analysis. A Twitter API is used to collect twitter corpus and feed it to a Binary Tree classifier that will discover the polarity lexicon of English tweets, whether positive or negative. A k-means clustering technique is used to group together similar words in tweets in order to discover certain business value.

1. Paper Title: A comparative approach for multiclass text analysis.

Publication: IEEE

DOI Number: [10.1109/ISDFS.2018.8355325](https://doi.org/10.1109/ISDFS.2018.8355325)

Description: This paper presents multiclass text analysis for the classification problem in Spanish documents. Even if Spanish language is considered as one the most spoken language, text classification problem has not yet been carried out for different problems in multiclass analysis. Two different approaches; Naive Bayes and Maximum Entropy were used as machine learning techniques. The corpus was created with 10 different categories. Smoothing parameters and three different document models were integrated to the study. During the comparative analysis, optimal parameters were determined using their sensitivity on the accuracy, the precision and the recall. Consequently, Maximum Entropy was found as the best technique even if both techniques were relevant in multiclass classification.

1. Paper Title: A review on text mining

Publication: IEEE

DOI Number: [10.1109/ICSESS.2015.7339149](https://doi.org/10.1109/ICSESS.2015.7339149)

Description: Text mining is the process of extracting previously unknown, understandable, potential and practical patterns or knowledge from the collection of text data. This paper introduces the research status of text mining. Then several general models are described to know text mining in the overall perspective. At last we classify text mining work as text categorization, text clustering, association rule extraction and trend analysis according to applications.

1. Paper Title: Research on energy-efficient text classification

Publication: IEEE

DOI Number: [10.1109/ICITEC.2014.7105614](https://doi.org/10.1109/ICITEC.2014.7105614)

Description: In this paper, we evaluate the energy cost of different classifiers and reduce energy cost by parallelization, trying to find a classifier that performs best on both aspects - effectiveness and efficiency. Several classifiers are obtained by using existing libraries or implementing classification algorithms. Comprehensive experiments on three real datasets show that an improved version of Naive Bayes can have competitive precision compared to SVM while has low energy costs. Parallelization can further reduce its energy cost by a factor of 10 for RCV1 dataset.

1. Paper Title: Research on text categorization based on machine learning

Publication: IEEE

DOI Number: [10.1109/ICAMS.2010.5552917](https://doi.org/10.1109/ICAMS.2010.5552917)

Description: In recent years, text categorization based on machine learning is a widely used technology in the field of information retrieval and text mining and has gained many advances. This paper presents a survey on the up-to-date development in text categorization based on machine learning, including model, algorithm and evaluation and introduces some machine learning methods applied in text categorization.

1. Paper Title: Text analysis based on time series

Publication: IEEE

DOI Number: [10.1109/ICSTCC.2013.6688932](https://doi.org/10.1109/ICSTCC.2013.6688932)

Description: The paper presents a text mining application for searching and computing the correlations between the rhythmicities of terms with high frequency appearance, using the time series model. Some theoretical basics of the model are presented, including details about preprocessing the observed texts using natural language processing techniques, followed by implementation details and graphical results.

1. Paper Title: Efficient Deep Learning Model for Text Classification Based on Recurrent and Convolutional Layers

Publication: IEEE

DOI Number: [10.1109/ICMLA.2017.00009](https://doi.org/10.1109/ICMLA.2017.00009)

Description: In this paper, we propose a neural language model that relies on Convolutional Neural Network (CNN) and Bidirectional Recurrent Neural Network (BRNN) over pre-trained word vectors. We utilize bidirectional layers as a substitute of pooling layers in CNN in order to reduce the loss of detailed local information, and to capture long-term dependencies across input sequences. We validate the proposed model on two benchmark sentiment analysis datasets, Stanford Large Movie Review (IMDB), and Stanford Sentiment Treebank (SSTb). Our model achieves a competitive advantage compared with neural language models on the sentiment analysis datasets.

1. Paper Title: Sentiment analysis and classification based on textual reviews

Publication: IEEE

DOI Number: [10.1109/ICICES.2013.6508366](https://doi.org/10.1109/ICICES.2013.6508366)

Description: Mining is used to help people to extract valuable information from large amount of data. Sentiment analysis focuses on the analysis and understanding of the emotions from the text patterns. It identifies the opinion or attitude that a person has towards a topic or an object and it seeks to identify the viewpoint underlying a text span. Sentiment analysis is useful in social media monitoring to automatically characterize the overall feeling or mood of consumers as reflected in social media toward a specific brand or company and determine whether they are viewed positively or negatively on the web. So analyzing sentiment using Multi-theme document is very difficult and the accuracy in the classification is less. The document level classification approximately classifies the sentiment using Bag of words in Support Vector Machine (SVM) algorithm. In proposed work, a new algorithm called Sentiment Fuzzy Classification algorithm with parts of speech tags is used to improve the classification accuracy on the benchmark dataset of Movies reviews dataset.

1. Paper Title: Searching research papers using clustering and text mining

Publication: IEEE

DOI Number: [10.1109/CONIELECOMP.2013.6525763](https://doi.org/10.1109/CONIELECOMP.2013.6525763)

Description: This works purposes a better classification of research papers, the architecture works with a database of knowledge related with the topics of programming, databases and operating systems. That's the initial work of a classification using text mining techniques to search into the documents with natural language contained and get the best words of their content to get a database knowledge, that's the first step to get the desired knowledge also the proposed work use the same engine to make searches classifying the information introduced by the final user and searching in the correct cluster.

1. Paper Title: Convolution-Based Neural Attention With Applications to Sentiment Classification

Publication: IEEE

DOI Number: [10.1109/CONIELECOMP.2013.6525763](https://doi.org/10.1109/CONIELECOMP.2013.6525763)

Description: It uses convolution operation to simulate attentions and give a mathematical explanation of our neural attention model. We then introduce a new network architecture, which combines a recurrent neural network with our convolution-based attention model and further stacks an attention-based neural model to build a hierarchical sentiment classification model.

1. Paper Title: A Single Attention-Based Combination of CNN and RNN for Relation Classification

Publication: IEEE

DOI Number: [10.1109/ACCESS.2019.2891770](https://doi.org/10.1109/ACCESS.2019.2891770)

Description: Proposed a novel Att-RCNN model to extract text features and classify relations by combining recurrent neural network (RNN) and convolutional neural network (CNN). This network structure utilizes RNN to extract higher level contextual representations of words and CNN to obtain sentence features for the relation classification task. In addition to this network structure, both word-level and sentence-level attention mechanisms are employed in Att-RCNN to strengthen critical words and features to promote the model performance.

1. Paper Title: Long Document Classification From Local Word Glimpses via Recurrent Attention Learning

Publication: IEEE

DOI Number: [10.1109/ACCESS.2019.2907992](https://doi.org/10.1109/ACCESS.2019.2907992)

Description: we propose to tackle the long document classification task by incorporating the recurrent attention learning framework, which can produce the discriminative features with significantly less words. Specifically, the core work is to train a recurrent neural network (RNN)-based controller, which can focus its attention on the discriminative parts. Then, the glimpsed feature is extracted by a typical short text level convolutional neural network (CNN) from the focused group of words. The controller locates its attention according to the context information, which consists of the coarse representation of the original document and the memorized glimpsed features. By glimpsing a few groups, the document can be classified by aggregating these glimpsed features and the coarse representation.

1. Paper Title: Text Classification using Different Feature Extraction Approaches

Publication: IEEE

DOI Number: [10.1109/eStream.2019.8732167](https://doi.org/10.1109/eStream.2019.8732167)

Description: We examine the results of applying three different text feature extraction approaches while classifying short sentences and phrases into categories with a neural network in order to find out which method is best at capturing text features and allows the classifier to achieve highest accuracy. The examined feature extraction methods include a plain Term Frequency Inverse Document Frequency (TF-IDF) approach and its two modifications by applying different dimensionality reduction techniques: Latent Semantic Analysis (LSA) and Linear Discriminant Analysis (LDA).

1. Paper Title: Analyzing Objective and Subjective Data in Social Sciences: Implications for Smart Cities

Publication: IEEE

DOI Number: [10.1109/ACCESS.2019.2897217](https://doi.org/10.1109/ACCESS.2019.2897217)

Description: we investigate a novel way of analyzing data from social sciences studies by employing machine learning and data science techniques. This enables us to maximize the insight gained from this type of studies by fusing both objective (sensor information) and subjective data (direct input from the users). The pilot study is concerned with better understanding the interactions between citizens and urban green spaces.

**3.Conclusion:**

From the different strategies and implementation, we will be using the Word2Vector method. The Word2vector method is a collection of models which will be helpful in producing word embeddings. The Word2vector model is basically two-layered neural network. The word2vector works by getting input text and converts into a vector. The Word2vector method provides high accuracy by using the choice of model architecture (CBOW or Skip-gram), increasing the training data set, increasing the number of vector dimensions, and increasing the window size of words considered by the algorithm. Each of these improvements comes with the cost of increased computational complexity and therefore increased model generation time. Accuracy increases overall as the number of words used increases, and as the number of dimensions increases.

Advantages of using Word2Vector:

* The idea is very intuitive, which transforms the unlabeled raw corpus into labeled data (by mapping the target word to its context word), and learns the representation of words in a classification task.
* The data can be fed into the model in an online way and needs little preprocessing, thus requires little memory.

**4.Reference:**

1. **Robert Dzisevič, Dmitrij Šešok -** Text Classification using Different Feature Extraction Approaches
2. **Zeynep Hilal Kilimci, Selim Akyokuş -** The Analysis of Text Categorization Represented with Word Embeddings Using Homogeneous Classifiers
3. **C. Lakshmi Devasena, M. Hemalatha -** Automatic Text categorization and summarization using rule reduction
4. **Han Liu, Pete Burnap** - A Fuzzy Approach to Text Classification With Two-Stage Training for Ambiguous Instances
5. **Catarina Silva, Bernardete Ribeiro** - Two-Level Hierarchical Hybrid SVM-RVM Classification Model
6. **Akanksha Ahlawat, Bharti Suri** - Improving classification in data mining using hybrid algorithm
7. **A. Vinciarelli -** Noisy text categorization
8. **P. Anupriya, S. Karpagavalli -** LDA based topic modeling of journal abstracts
9. **Shuzlina Abdul-Rahman, Sofianita Mutalib, Nur Amira Khanafi, Azliza Mohd Ali -** Exploring Feature Selection and Support Vector Machine in Text Categorization
10. **Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. Gerber, Laura E. Barnes** - HDLTex: Hierarchical Deep Learning for Text Classification
11. **Ismael Ali, Austin Melton -** Graph-Based Semantic Learning, Representation and Growth from Text: A Systematic Review
12. **Nanthiine Nair Ravi, Sulfeeza Mohd Drus, Prajindra Sankar Krishnan, Nur Laila Abdul Ghani -** Substation Transformer Failure Analysis Through Text Mining
13. **Jingjing Cai, Jianping Li, Wei Li, Ji Wang -** Deeplearning Model Used in Text Classification
14. **Alrence Santiago Halibas, Abubucker Samsudeen Shaffi, Mohamed Abdul Kader Varusai Mohamed** - Application of text classification and clustering of Twitter data for business analytics
15. **Semuel Franko, Ismail Burak Parlak** - A comparative approach for multiclass text analysis.
16. **Yu Zhang, Mengdong Chen, Lianzhong Liu** - A review on text mining
17. **Hao Lin** - Research on energy-efficient text classification
18. **Yu Wanjun, Song Xiaoguang** - Research on text categorization based on machine learning
19. **Iulia Badea, Stefan Trausan-Matu** - Text analysis based on time series
20. **Abdalraouf Hassan, Ausif Mahmood** - Efficient Deep Learning Model for Text Classification Based on Recurrent and Convolutional Layers.
21. **K. Mouthami, K. Nirmala Devi, V. Murali Bhaskaran** - Sentiment analysis and classification based on textual reviews
22. **E. Alan Calvillo, Alejandro Padilla, Jaime Muñoz, Julio Ponce, Jesualdo T. Fernandez** - Searching research papers using clustering and text mining
23. **Jiachen Du, Lin Gui, Yulan He, Ruifeng Xu, Xuan Wang** - Convolution-Based Neural Attention With Applications to Sentiment Classification
24. **Xiaoyu Guo, Hui Zhang, Haijun Yang, Lianyuan Xu**- A Single Attention-Based Combination of CNN and RNN for Relation Classification
25. **Jun He, Liqun Wang, Liu Liu, Jiao Feng, Hao Wu** - Long Document Classification From Local Word Glimpses via Recurrent Attention Learning
26. **Robert Dzisevič, Dmitrij Šešok** - Text Classification using Different Feature Extraction Approaches
27. **Laura Erhan, Maryleen NDubuaku, Enrico Ferrara, Miles Richardson, David Sheffield, Fiona J.Ferguson, Paul Brindley, Antonio Liotta** - Analyzing Objective and Subjective Data in Social Sciences: Implications for Smart Cities