Text Based Deception Detection Using a Hashing Algorithm and Machine Learning Techniques

Fahad Abdulridha[*], Baraa M. Albaker[**]

[*]College of Engineering, Al-Iraqia University, Saba'a Abkar Complex, Baghdad, Iraq
Email: fahad.m.abdulridha@aliraqia.edu.iq
https://orcid.org/0009-0007-1828-3108

[**]College of Engineering, Al-Iraqia University, Saba'a Abkar Complex, Baghdad, Iraq
Email: baraamalbaker@ymail.com
https://orcid.org/0000-0002-6030-3121

Abstract

One of the most challenging goals in the fields of law enforcement and the court rooms is the ability to detect deception and false information, this is due to the major role it has on national security and the justice system. One of the most important ways recent literatures has explored to overcome this challenge, is detecting deception using artificial intelligence techniques to find patterns in verbal and nonverbal features. Text based analysis has been one of the most important modals for this task since written text can be found in transcribed audio, emails, online messaging services, news articles, and many more. In this work, a combination of machine learning techniques and data processing using hashing algorithms is used applied to n-gram feature representation on two of the largest datasets in deception detection field. Together, results of up to %94.59 accuracy were achieved. The paper reviews the most common techniques used in recent literature, it also details the methodology followed for data processing and model training to achieve these results.

Keywords- Deception Detection; Lying Detection; Machine Learning; text classification.

I. Introduction

Detecting deception has been one of the most widely researched problems for as long as judicial systems have existed in human history, attempts as early as the ancient Greeks and Indians 2000 years ago, with the earliest recorded attempt found around 600-900 B.C.E. in the works of the Hindu Dharmasastra of Gautama [1]. The first attempts however to produce a reliable system which can detect deception was in the early 20th century with the invention of the polygraph in 1921 [2]. Moreover, the human capacity to identify deception, demonstrated to be no more reliable than chance [3], [4], has spurred the growing dependence on technology, particularly artificial intelligence, in recent times.

In the early stages of automated deception detection, the approach relied on devices that measured physiological changes in individuals, such as respiration, heart rate, brain activity, and body temperature. However, these methods proved to be highly error-prone due to their dependence on human interpretation and decision-making for the outcome. Additionally, relying solely on physiological reactions could lead to false positives, as the interviewee's anxiety, stress, or fear could significantly impact the results [5].

With the advent of artificial intelligence, new avenues for deception detection emerged that significantly reduced the need for human intervention and physiological measurements. This breakthrough enabled the entire process of deception detection to become fully automated, encompassing data collection, analysis, and classification.

One of the most promising modals that has been explored in recent years is the text-based deception detection, which includes written text in internet messaging services, emails, social media, news articles, as well as transcriptions of audio conversation recordings. These relied on detecting patterns in the text to indicate false information, spam, emotion, or deception attempts. Text based deception detection was especially useful because it did not suffer from the same weak points that audio or video-based modals had such as the quality of the recordings, the setting at which the recording was done, lighting, noise, among others.

In the pursuit of AI-driven deception detection, the pivotal factor influencing performance, particularly in terms of classification accuracy, revolves around the quantity and quality of the available data. However, in deception detection, datasets are severely lacking, especially concerning research focused on audio-oriented applications. There are currently only 8 publicly available datasets, out of which merely 2 are specifically designed to support audio and acoustic features without requiring additional preprocessing and feature extraction. The remaining datasets consist of raw video clips, necessitating further processing. Furthermore, these datasets suffer from a limited number of participants, with an average of 51 participants per dataset, ranging from 101 participants at the highest to just 26 participants at the lowest [6].

In this work, two of the largest datasets in the field of deception detection are experimented with by attempting to classify the audio transcriptions of each recording using machine learning techniques paired with text processing and analysis tools. Transforming the text into a matrix of token occurrences using a hashing algorithm. The text is also preprocessed to reduce sparsity in its vocabulary and remove non-impactful and noisy words while picking the most impactful ones for the classification task.

II. Related work

In computational linguistics, deception detection is commonly approached as a text classification problem. The main objective is to create a system capable of categorizing an unseen document as either truthful or deceptive. To accomplish this, the system is initially trained on known instances of deception. One of the pioneering studies that adopted this methodology was conducted by Newman et al. in 2003 [7]. They demonstrated that by employing supervised machine learning techniques and quantitative text parameters as features, automatic classification of texts as deceptive or truthful could be achieved. The authors achieved a correct classification rate of 67% for identifying liars and truth-tellers when the topic remained constant, and an overall rate of 61% for all cases.

Deception detection in computational linguistics relies on token unigrams and the LIWC lexicon, introduced by Newman's paper. LIWC, a text analysis program [8], categorizes words into psychologically meaningful groups across four main dimensions: standard linguistic, psychosocial processes, relativity, and personal concerns. LIWC 2015 computes up to 88 output variables per text, covering linguistic dimensions, psychological constructs, relativity aspects, personal concerns, and punctuation information. Numerous studies have demonstrated the effectiveness of LIWC and machine learning in automatic deceit identification, outperforming baseline accuracy.

Limited attention has been given to deception identification based on demographic data using computational approaches due to resource scarcity [9] Only two other resources for deception detection with demographic data have been identified [9]. In addition, researchers have explored the connection between deception and personality traits through linguistic expression [10]. They found that certain personality traits can improve machine learning models' ability to distinguish deceptive statements based on communication style. However, the study's small sample size limited the exploration of various personality types. Furthermore, Levitan et al. [11] investigated oral speech deception detection and achieved

improved classifier accuracy by including binned NEO-scores, gender, and language alongside prosodic and LIWC features. This yielded a 65% accuracy, representing a substantial increase over the majority class baseline. In summary, LIWC and machine learning methods have proven valuable in deception detection, while research exploring deception with demographic data and personality traits shows promising potential.

III. Datasets used

The first dataset utilized in this study is called "The Miami University Deception Detection Database (MU3D)" [12]. It consists of 320 publicly available videos, involving 80 participants (40 males and 40 females) with ages ranging from 18 to 26 (refer to Fig. 1). Each participant recorded four videos, recounting four different topics. “Describe a person you truly lie” representing positive truth, “describe a person you truly dislike” for negative truth, “describe a person you falsely like” for positive lie, and “describe a person you falsely dislike” for negative lie scenarios. The duration of the video ranges from 24 to 57 seconds, with an average duration of 35 seconds (as shown in Fig. 2).

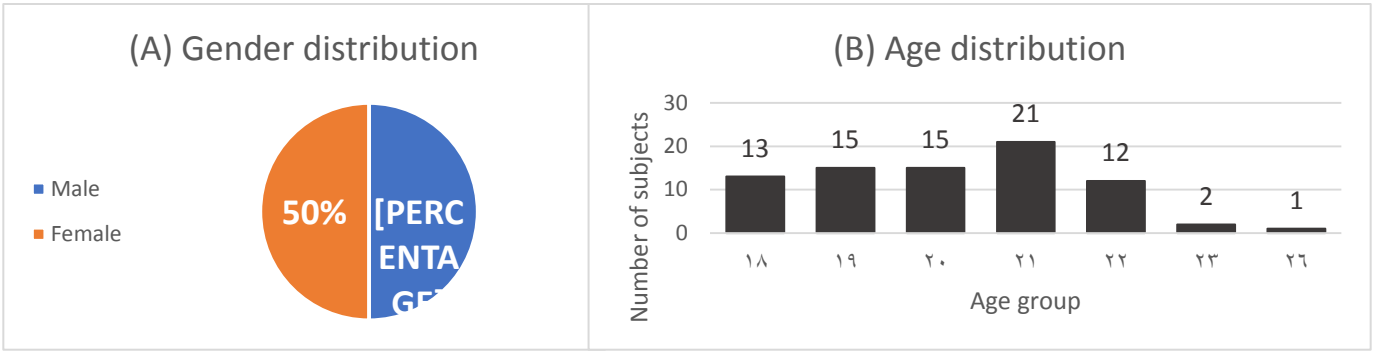

Figure 1. (A): gender ratio of MU3D database; (B): Number of participants for each age group of MU3D database

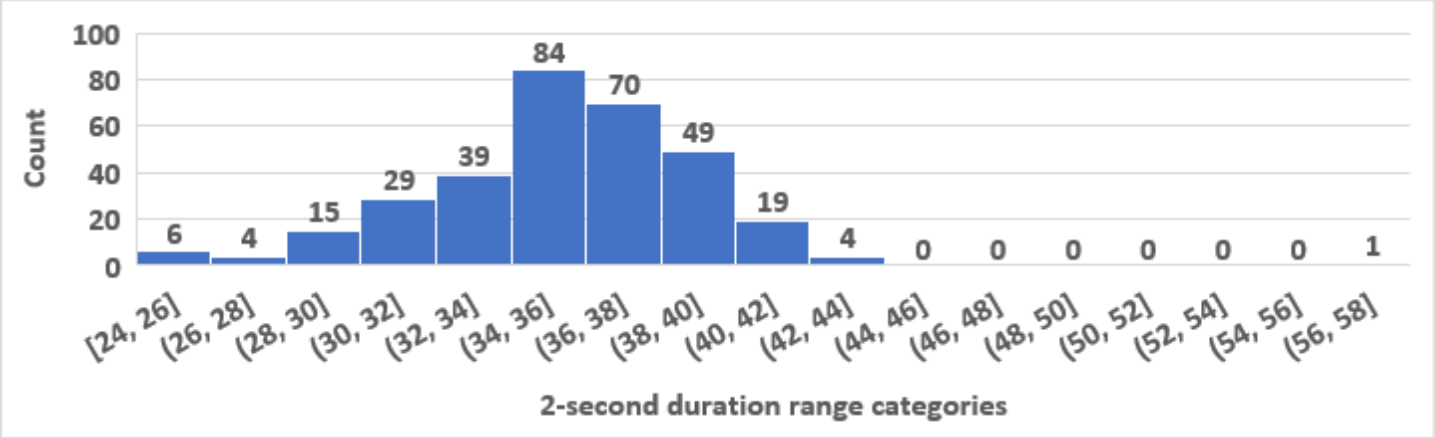

Figure 2. Audio clips duration distribution for MU3D dataset.

The second database utilized in this study is called the "Real Life Deception Detection Database" [13]. It comprises 121 publicly available videos recorded during trials and courtrooms, featuring witnesses or defendants. The dataset includes 61 deceptive clips and 60 truthful clips, involving 56 subjects (21 females and 35 males) aged between 16 and 60 years. The duration of the video clips ranges from 4 seconds to 1 minute and 11 seconds, with an average duration of 27.7 seconds (as depicted in Fig. 3).

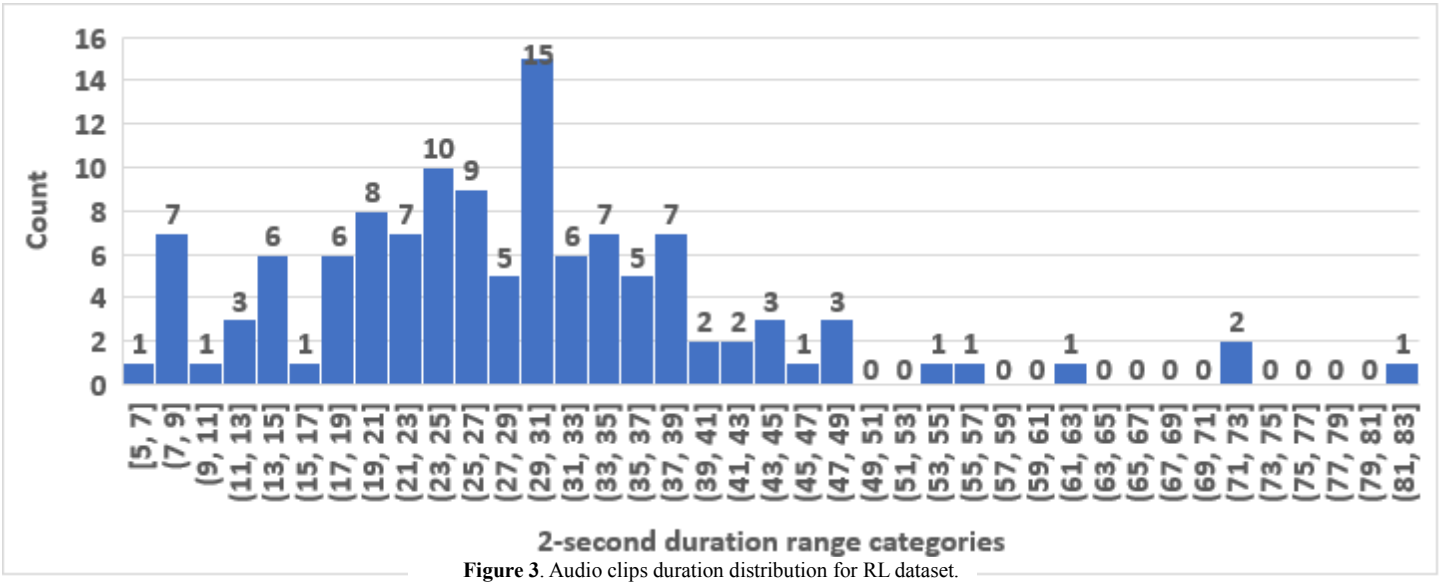

Figure 3. Audio clips duration distribution for RL dataset.

IV. Audio transcription and text processing

The transcriptions for each of the two datasets were produced using Deepgram Nova speech to text model. Each raw audio recording was transcribed individually and conversations where t wo or more people were involved were separated, only the target subject speech was kept for processing and model training. MU3D dataset produced 32,718 words with approximately 171 average words per minute. RL dataset produced 8,055 words with approximately 66 words per minute.

For the preprocessing part, two larger datasets were produced from each original dataset containing all the transcriptions and their respective classification. Each dataset was then lemmatized to reduce the number of similar words by reducing each word to its root and remove duplicates, stop words are also removed from the dataset and all words are turned lower case for a uniform and informative final dataset. The lemmatized dataset is then vectorized using a hashing algorithm and each word was given a token occurrence. The vectorized transcriptions, now a matrix, its values are scaled down using a term-frequency times inverse document-frequency algorithm to reduce the impact of tokens that have high frequency which makes them less informative (see Fig. 4).

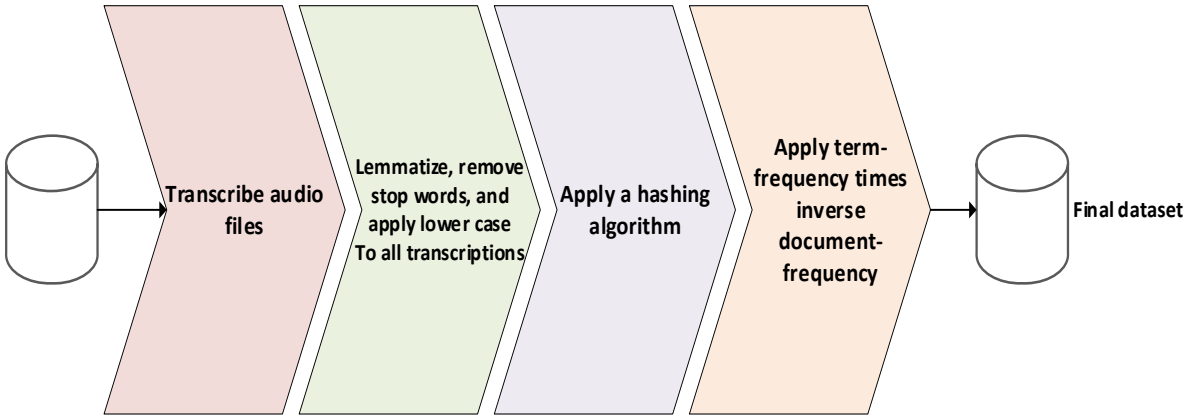

Figure 4. Audio transcription and text processing steps.

V. Classification and results

Random Forest (RF) is used for the classification task due to its out of the box resilience to overfitting, high dimensional data, and noise [14], [15]. This is especially useful for the datasets produced in this work due to its high dimensionality nature and potential for overfitting. RF is also one of the most successful machine learning algorithms for real world problem solving [16].

The cuML python library by RAPIDS [17] implementation of RF is used due to its utilization of the GPU to accelerate computation speed when working with such large datasets. Hyperparameter optimization was also performed using DASK [18], another python library that allows the utilization of the GPU for accelerated and distributed computation. The Randomized Search Cross Validation implementation is used with a 5-fold cross validation for optimizing the number of classifiers/decision trees in the ensemble (*n_estimators*), maximum number/percentage of features to be considered by each spawned decision tree (*mtry*), the maximum number of splits that each decision tree is allowed to make (*max_depth*). The values in Table 1 were found.

Table 1. Estimation of optimized values for hyperparameters of the model

Hyperparameter	Best value
n_estimators	1500
mtry	0.8
max_depth	20

Furthermore, the final datasets were split into train and test subsets with a 70 to 30 ratio and the results illustrated below are strictly based on classification accuracy of the unseen test subset. Before the classification process takes place, a feature selection algorithm was performed on the datasets to consider the most impactful keywords to be used for training the Random Forest classifier. The algorithm selected the top 100 keywords from each dataset using a Chi-Square scoring algorithm. This value was chosen after a

number of manual iterations to find the lowest number of features that have minimum impact on performance. This allowed the system to maintain a high classification accuracy while minimizing model complexity and reducing overfitting. The classification process was conducted on several n-gram variations and the results are illustrated in table 2.

Table 2. Classification accuracy for each n-gram variation.

n-gram (top 100 keywords)	MU3D	RL
Unigram	%71.87	%75.67
Bigram	%81.25	**%94.59**
Trigram	%75.0	%67.56
Unigram + Bigram	%73.95	%75.67
Unigram + Bigram + Trigram	**%84.37**	%81.08

The results above have consistent trends for both datasets, combining all n-gram values in for the vectorizer seems to produce reliable results, with MU3D achieving its best accuracy with this combination at %84.37. RL too gets a high accuracy with this combination, but its best accuracy comes from bigrams alone at %94.59, a state-of-the-art result for this dataset.

VI. Conclusion and future work

The aim of this work has been to improve the classification accuracy of deception detection on two of the largest datasets in the field by analyzing the performance of different combinations of n-gram features using machine learning and hashing algorithms. State-of-the-art results were achieved with RL dataset reaching 94% on the bigram feature level and %81.08 on the combined unigram, bigram and trigram combination which is also an improvement over earlier attempts. MU3D achieved similar results with same latter combination with an accuracy of %84.37, another major improvement in performance for this dataset.

A major limitation for the developed system is its lack of ability to classify text in real-time. This is due to the way the extracted features are preprocessed before classifying it as deceptive or truthful. The usefulness of real-time classification lies in its ability to provide immediate feedback in real life applications which can aid in decision-making on the spot. The authors also recognize that these results can be further improved by experimenting with other artificial intelligence algorithms and techniques in future work. As it's worth noting that deep learning has been historically successfully at producing high performing systems for this modal.

VII. Ethical considerations

A major concern for AI-driven deception detection are the ethical issues that accompany its use, while AI has made large contribution towards improving the performance of such systems, it also made it more accessible to the public. Many of the datasets mentioned above require a signed agreement to license the use of these datasets which details the privacy and usage requirements of the available data. Furthermore, any developed deception detection system relative in performance to state-of-the-art poses risk of misuse. These systems can be of great help in the judicial system when used by the appropriate authorities; they can also be used by unauthorized parties to cause harm by means of false positive classification. The authors of this work have signed and agreed to all terms of use of the datasets employed and prohibit use or distribution of the developed system to parties outside legal authorities concerned with homeland security.

References

[1] P. V. Trovillo, "A history of lie detection," *J. Crim. Law Criminol.*, vol. 29, 30, pp. 848–881, 104–119, 1939.

[2] J. Synnott, D. Dietzel, and M. Ioannou, "A review of the polygraph: history, methodology and current status," *Crime Psychol. Rev.*, vol. 1, no. 1, pp. 59–83, Jan. 2015, doi: 10.1080/23744006.2015.1060080.

[3] P. Ekman and M. O'Sullivan, "Who can catch a liar?," *Am. Psychol.*, vol. 46, pp. 913–920, 1991, doi: 10.1037/0003-066X.46.9.913.

[4] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo, "Deception detection using a multimodal approach," in *Proceedings of the 16th International Conference on Multimodal Interaction*, Istanbul Turkey: ACM, Nov. 2014, pp. 58–65. doi: 10.1145/2663204.2663229.

[5] L. Saxe, D. Dougherty, and T. Cross, "The validity of polygraph testing: Scientific analysis and public controversy," *Am. Psychol.*, vol. 40, pp. 355–366, 1985, doi: 10.1037/0003-066X.40.3.355.

[6] A. Omirali, A. Shoiynbek, K. Kozhakhmet, and N. Sultanova, “A Review of Deception Detection Databases,” 2022, doi: DOI : 06.2016-67962946/2022.7658.

[7] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, “Lying Words: Predicting Deception from Linguistic Styles,” *Pers. Soc. Psychol. Bull.*, vol. 29, no. 5, pp. 665–675, May 2003, doi: 10.1177/0146167203029005010.

[8] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, “The Development and Psychometric Properties of LIWC2007”.

[9] V. Pérez-Rosas and R. Mihalcea, “Gender Differences in Deceivers Writing Style,” in *Human-Inspired Computing and Its Applications*, A. Gelbukh, F. C. Espinoza, and S. N. Galicia-Haro, Eds., in Lecture Notes in Computer Science, vol. 8856. Cham: Springer International Publishing, 2014, pp. 163–174. doi: 10.1007/978-3-319-13647-9_17.

[10] T. Fornaciari, F. Celli, and M. Poesio, “The Effect of Personality Type on Deceptive Communication Style,” in *2013 European Intelligence and Security Informatics Conference*, Aug. 2013, pp. 1–6. doi: 10.1109/EISIC.2013.8.

[11] S. I. Levitan *et al.*, “Identifying Individual Differences in Gender, Ethnicity, and Personality from Dialogue for Deception Detection,” in *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 40–44. doi: 10.18653/v1/W16-0806.

[12] K. Hugenberg, A. R. McConnell, J. W. Kunstman, E. P. Lloyd, J. C. Deska, and B. Humphrey, “Miami University Deception Detection Database,” Mar. 2017, Accessed: Mar. 23, 2023. [Online]. Available: http://sc.lib.miamioh.edu/handle/2374.MIA/6067

[13] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, “Deception Detection using Real-life Trial Data,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, in ICMI ’15. New York, NY, USA: Association for Computing Machinery, Nov. 2015, pp. 59–66. doi: 10.1145/2818346.2820758.

[14] T. G. Dietterich, “Ensemble Methods in Machine Learning,” in *Multiple Classifier Systems*, in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2000, pp. 1–15. doi: 10.1007/3-540-45014-9_1.

[15] R. Caruana, N. Karampatziakis, and A. Yessenalina, “An empirical evaluation of supervised learning in high dimensions,” in *Proceedings of the 25th international conference on Machine learning*, in ICML ’08. New York, NY, USA: Association for Computing Machinery, Jul. 2008, pp. 96–103. doi: 10.1145/1390156.1390169.

[16] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, “Do we need hundreds of classifiers to solve real world classification problems?,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3133–3181, Jan. 2014.

[17] S. Raschka, J. Patterson, and C. Nolet, “Machine Learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence.” arXiv, Mar. 31, 2020. Accessed: Jun. 03, 2023. [Online]. Available: http://arxiv.org/abs/2002.04803

[18] M. Rocklin, “Dask: Parallel Computation with Blocked algorithms and Task Scheduling,” presented at the Python in Science Conference, Austin, Texas, 2015, pp. 126–132. doi: 10.25080/Majora-7b98e3ed-013.