iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** I **Month of publication:** January 2024

DOI: https://doi.org/10.22214/ijraset.2024.58009

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue I Jan 2024- Available at www.ijraset.com

Exploring the State of Natural language Processing: A Survey of Recent Advances, Challenges and Future Scope

Rakhi Krishna C R[1], Pallavi N R[2], Dr. Shashikala SV[3]
[1]*Dept. of CSE, BGS Institute of Technology, Adichunchanagiri University*
[2]*Assistant Professor, Dept. of CSE, BGS Institute of Technology, Adichunchanagiri University*
[3]*Professor & Head, Dept. of CSE, BGS Institute of Technology, Adichunchanagiri University*

Abstract: ***Recent developments in deep learning have largely been responsible for the excellent performance on benchmark datasets of natural language processing algorithms. These advancements have greatly enhanced the capabilities of NLP systems utilised in programmes like sentiment analysis, speech recognition, and virtual assistants. However, the vulnerability of these systems to adversarial attacks has revealed the limitations in their robustness and language understanding abilities, which poses challenges when deploying NLP systems in real-world scenarios. This study provides a thorough analysis of studies on NLP robustness, providing a structured overview across multiple dimensions. We explore a variety of robustness-related topics, including as methodologies, measures, embedding, and benchmarking. Moreover, it emphasizes the need for a multi-dimensional perspective on robustness and offer insights into ongoing research efforts while identifying gaps in the literature. We propose potential directions for future exploration aimed at addressing these gaps and enhancing the robustness of NLP systems.***

Keywords: Natural language processing (NLP), Virtual assistants, Sentiment analysis.

I. INTRODUCTION

Artificial intelligence (AI)'s field of natural language processing (NLP) is concerned with how computers and human language interact. It aims to enable computers to accurately comprehend, interpret, and produce human language that is meaningful and useful. It is essential for bridging the divide between human communication and computer comprehension.

Human language is intricate and nuanced, creating it challenging to be understood and processed by computers. NLP combines various techniques, algorithms, and models to tackle this complexity and extract meaning from text or speech data. It draws upon disciplines such as linguistics, computer science, and statistics to develop computational models that produce and comprehend words. NLP's scope is vast and includes a variety of jobs and applications.

Some of the key areas within NLP include:

1) *Language Understanding:* NLP algorithms analyze and interpret the meaning of written or spoken text, extracting relevant information and identifying patterns and relationships. This comprises activities like question answering, sentiment analysis, named entity recognition, and text categorization.
2) *Language Generation:* NLP models may produce text that resembles that of people, producing coherent and contextually relevant responses. This includes text summarization, machine translation, conversation systems, and text production are examples of tasks for creative purposes.
3) *Language Processing:* NLP involves various techniques for processing and manipulating language data. This includes tasks such as tokenization (segmenting text into meaningful units), part-of-speech tagging (assigning grammatical tags to words), syntactic parsing (analyzing sentence structure), and semantic role labeling (identifying the roles of words or phrases in a sentence).

NLP techniques have widespread applications in numerous domains. They are used in customer service Virtual assistants, chatbots, information retrieval systems, social media sentiment analysis, language translation services, and more. It has also found applications in healthcare, finance, education, law, and other fields where large amounts of text data need to be analyzed and processed.

In recent years, advancements in deep learning as well as the accessibility of massive datasets have propelled NLP to new heights. Pretrained language models, such as GPT (Generative Pretrained Transformer) and BERT (Bidirectional Encoder Representations from Transformers), have demonstrated remarkable language understanding and generation capabilities. These models have opened possibilities for more accurate and context-aware NLP applications.

However, NLP still faces several challenges. Ambiguity, context understanding, and handling out-of-vocabulary words are ongoing research areas. Ethical considerations, biases in data, and privacy concerns are also important factors to address in NLP development.

As NLP continues to evolve, researchers and practitioners are exploring new techniques and approaches to enhance language understanding, generation, and processing. The future of NLP holds promising opportunities for further advancements, enabling machines to communicate with humans in more natural and meaningful ways.

The benefits of Natural Language Processing (NLP) stem from its ability to enable computers to do extensive textual data analysis. While it may appear as a recent technology, its origins can be traced It was originally employed for machine translation (MT) in the early 1950s. The amount of text data created each day has increased tremendously as we observe a surge of technology developments producing innovative and disruptive applications. As a result, there is now a need for strong technologies like NLP to effectively process vast volumes of data. Numerous industries are widely adopting NLP to derive meaningful insights from data and overcome various challenges. Everyday examples of NLP applications include Google Translate, Google Assistant, Alexa from Amazon, and Cortana from Microsoft. The financial sector, companies like Prudential and Bank of America utilize NLP in their chatbots, such as Erica. In the business world, systems are extensively employed for tasks like spam detection, intrusion detection, and malware analysis.

However, machine learning techniques, including NLP, evolving field. are vulnerable to attacks. Adversaries can exploit these techniques to go against the goals and promises of applications. For instance, in smart speaker applications, minimal modifications to input can be used by adversaries to trigger incorrect or Voice squatting for malicious device activation. Similarly, adversaries can target NLP models used for spam detection and deceive them into allowing spam to pass through email filters. To trick intrusion detection systems and categorize malware as innocuous software, malware developers may target NLP-based models. adversaries may even try to manipulate models used in the finance industry to deceive loan application systems and improperly qualify or disqualify customers for loans.

This research aims to contribute a comprehensive and in-depth analysis of recent advancements in NLP robustness. It brings value to the field through the following key aspects:

a) *Expanded Taxonomy:* An enriched and comprehensive taxonomy is introduced, covering various dimensions of NLP robustness derived from a broad range of NLP applications. This facilitates a more holistic understanding of the subject.
b) *Study Classification:* Research done recently on the resilience of NLP are categorized and classified based on several variables, including among other things, models, embedding methods, measurements, and defense mechanisms. This structured overview provides insights into the research landscape.
c) *Comparative Analysis:* Different approaches to NLP robustness are compared, showcasing each party's own advantages and disadvantages. This analysis provides insightful information on the efficiency and limitations of various techniques and methods.
d) *Identification of Research Gaps:* Gaps in the existing literature are identified, and a roadmap for future research directions is provided. By pinpointing areas requiring further investigation, researchers are encouraged to address these gaps and advance the field of NLP robustness.

In conclusion, this work empowers researchers to explore NLP robustness from various perspectives, encompassing learning techniques/models, embedding techniques, datasets, defense mechanisms, and robustness metrics. It is a beneficial resource for enhancing the understanding and development of robust NLP systems.

II. BACKGROUND STUDY

While previous surveys have touched upon the topic of NLP robustness, they have been limited in scope, focusing on specific aspects of robustness and failing to provide a comprehensive review of the entire spectrum. These surveys have not adequately covered areas such as benchmark datasets, embedding methods, evaluation metrics, and defence tactics current adversarial assaults, all of which are crucial contributors to NLP robustness. Additionally, there is a dearth of surveys in the NLP domain that offer researchers insights into perspectives and emerging trends within this rapidly evolving field.

In this distinct survey, we aim to fill in this gap presenting a systematic framework that reviews, categorizes, summarizes, and integrates existing literature while also identifying research topics that warrant further exploration. Our goal is to equip researchers who possess the requisite skills and novel ideas to gain a deeper understanding of the explore new research themes and the NLP robustness research landscape.

Feng et al. [14] conducted a thorough investigation into data augmentation for enhancing NLP robustness. Their survey delved into various techniques, such as rule-based and model-based approaches, aimed at fortifying NLP models against adversarial attacks. However, it is important to note that their work solely focuses on a specific facet of robustness, leaving out other crucial defense mechanisms employed to enhance the robustness of NLP systems. For instance, apart from data augmentation, there exist additional defense mechanisms that contribute to achieving robustness in NLP. Strong data training and data training methods that can be certified are noteworthy examples of such mechanisms. These approaches, along with data augmentation, collectively play a pivotal role in fortifying NLP systems against adversarial challenges. While Feng et al.'s survey provides valuable insights into data augmentation techniques, it is essential to acknowledge the broader landscape of defense mechanisms and strategies that contribute to overall NLP robustness. By considering a comprehensive range of defense mechanisms, we can achieve a more comprehensive understanding of the multifaceted nature of NLP robustness and explore avenues for further research and development.

Yoo et al. [15] conducted a review specifically focus on resilience through algorithms for producing adversarial instances. While their work provides valuable insights into this aspect of robustness, it is important to note that their scope is limited. Robustness in NLP encompasses a multitude of variables beyond adversarial examples.

Embedding techniques, robustness metrics, and robustness techniques are essential elements within the wider framework of NLP robustness. These aspects, omitted in Yoo et al.'s survey, have substantial contributions to strengthening NLP models' robustness. In a similar vein, [29] conducted a research survey with a specific focus on transformers, conditioned NLP models. Their study primarily investigates robustness from a model perspective, highlighting the contributions of learning models like BERT and RoBERTa. However, the survey lacks coverage of other vital aspects, such as defenses, attacks, and techniques, which are crucial considerations in the domain of NLP robustness.

For a complete grasp of NLP robustness, it is essential to consider a broad array of factors and dimensions, including, but not limited to, adversarial examples, embedding techniques, robustness metrics, and diverse defense and attack strategies. By incorporating these varied elements, researchers can foster a comprehensive outlook and contribute to the progress of robust NLP systems.

Zhang et al. performed a survey [17] centered on using adversarial assaults to gauge how resilient NLP systems are to disturbances. While the study offers valuable insights into the dimension of adversarial attacks, it is vital to acknowledge because the idea of robustness is multifaceted. Attacks from the other side are only one component of resilience, and there are various other essential elements that contribute to the overall robustness of NLP systems, including defense mechanisms, evaluation metrics, and embedding techniques, among others.

As far as our knowledge goes, there is currently no all-encompassing study in the existing literature that consolidates advancements in understanding robustness by considering a holistic pipeline encompassing the essential steps involved in implementing an NLP system. This research intends help bridge this gap by providing a thorough summary of the diverse dimensions of robustness, considering the entire pipeline of an NLP system. Through this comprehensive approach, we aim to contribute to a deeper understanding of NLP system robustness and offer valuable insights to further the field's study and development.

III. An Inclusive Overview On NLP

To provide a succinct summary of the developments in NLP during the last ten years, particularly focusing on resilience against assaults and defences, we propose a simplified system flow that incorporates common elements found in systems. By employing these elements, we can effectively depict the various advancements that have been achieved.

Let's investigate utilising a natural language processing model, particularly for natural language production, to demonstrate this system flow. Preprocessing is the first stage in a typical NLP pipeline, as seen in Figure 1 at a high level. This stage raw linguistic input and prepares it for further processing by the model through various mapping techniques. Once the initial preprocessing and mapping are complete, the next step involves embedding. The embedding step transforms the initial representation of the input into a suitable format that the NLP model may easily accept. This transformation enhances the model's ability to effectively process and understand the input data. Afterward, the NLP model progresses into a training phase, where it adapts and fine-tunes various parameters essential for language generation. This training process empowers the model to attain the essential knowledge and proficiency for generating coherent and contextually appropriate language.

Throughout this depiction, our objective is to emphasize the interplay between the various components within the NLP system and how they collectively enhance robustness. By presenting this systematic overview, we can develop a deeper comprehension of the progress made in NLP and its implications for achieving robustness in language generation tasks.

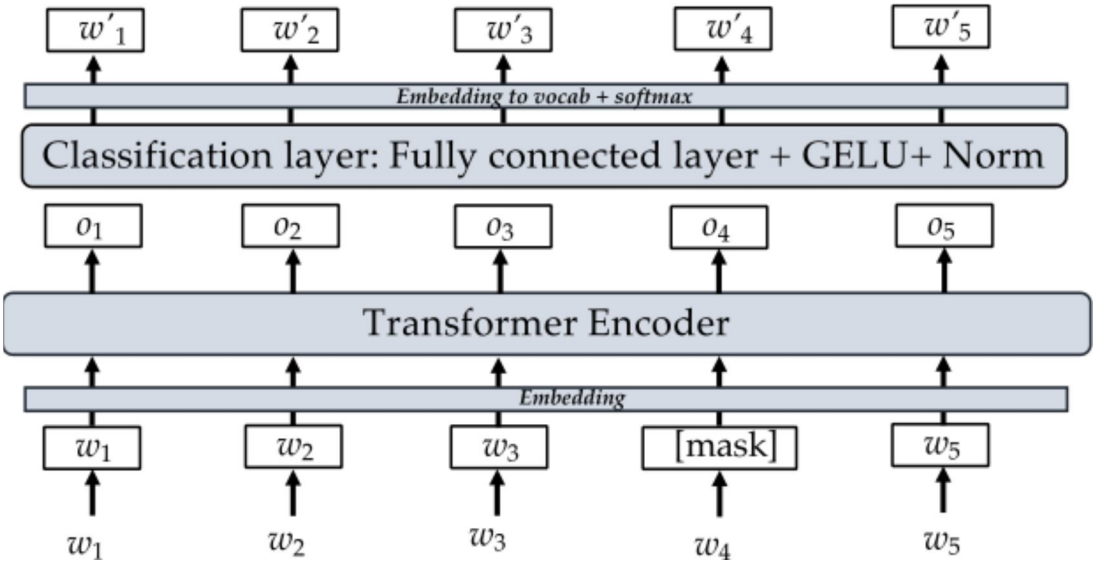


Fig. 1. A BERT Embedding technique-based NLP pipeline

A. Taxonomy: Robustness in NLP

Figure 2 illustrates a clear taxonomy that showcases the wide array of research efforts in the written word focused a discussion of NLP and the related robustness analysis. This taxonomy covers various components of the pipeline, methodologies, embedding, evaluation benchmarks (datasets), evaluation metrics, attack space (threat model and granularity), and defence mechanisms, among others. To fully comprehend the objectives of each paper and systematically grasp their contributions, must be arranged in a standardized manner according to the pipeline we have described. As a result, in the subsequent sections, we delve into the specific endeavors dedicated to each element of the pipeline, examining the advancements and insights presented in the literature.

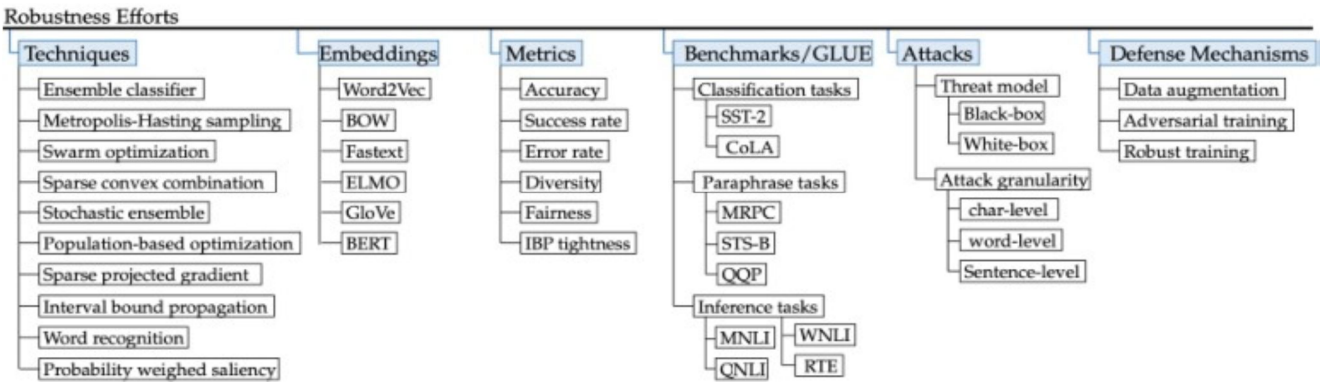


Fig. 2. An overview of the robustness analysis

Through the implementation of this methodical approach, our goal is to offer a comprehensive comprehension of the research landscape, accentuating the efforts made in each element of the NLP pipeline. This dissection enables us to explore the contributions in a structured fashion and attain profound insights into the progress of NLP robustness research.

IV. TOOLS & METHODS FOR ROBUSTNESS

In Section I it is made clear that NLP models are undergoing growing utilization in performing an extensive array of human-like tasks. Given the increasing use of NLP systems in several applications in the real world, the significance of constructing strong NLP models cannot be overstated. The effects of inaccurate projections made by these models can be severe, and in certain cases, even life-threatening (as seen in medical imaging diagnosis systems).

Regrettably, the absence of robustness in NLP models has led to numerous instances of failure after their deployment in real-world scenarios. Recent studies have indicated that around two-thirds of NLP systems in use today, including applications like essay grading and Twitter analysis by Microsoft, have encountered failures attributed to their lack of robustness [18]. To provide further illustration, consider the case of Amazon's NLP-based recruiting tool, which had to be discontinued due to its prejudice towards female candidates was evident [30].

These examples underscore the critical necessity for robustness in NLP models. Addressing the vulnerabilities and limitations that can arise when deploying NLP systems in real-world contexts is essential. By focusing on enhancing the robustness of NLP models, we can mitigate the risks associated with false predictions and ensure their reliability and effectiveness in practical applications.

A. Robustness Analysis and Testing Tools

In response to the shortcomings of different the scientific community has undertaken considerable investigations on the resilience of NLP systems, underscoring the importance of thorough testing before deploying models. One notable approach is CheckList introduced in [18], which is an unbiased technique for assessing NLP models. CheckList facilitates thorough testing by using a matrix of broad language abilities and test kinds. This approach is applicable to both commercial and research NLP models, revealing weaknesses even after internal testing, though it does not provide specific solutions for the identified weaknesses.

Another contribution is Robustness Gym (RG) by Goel et al. [25], overcoming obstacles in the evaluation of NLP systems. RG serves as a straightforward yet flexible evaluation toolbox, providing a standard method for evaluation. It enables NLP professionals to contrast outcomes from various frameworks and create new techniques by utilising built-in abstractions. While promising, the RG framework lacks a comprehensive understanding of NLP model behavior and fails to pinpoint the causes of model degradation. extending the technology to identify model errors and supply insights into performance would be valuable.

Rychalska et al. [24] developed WildNLP, a framework for in-situ assessment of model stability, in their paper. The primary goal of WildNLP is to correct text corruptions such typos and misspellings. By comparing the performance of models in four well-known NLP tasks (QA, NLI, NER, and SA), the authors assess the resilience of the models. They discover that even if modern embedding approaches might assist to increase resilience, excellent model performance is not a guarantee of adequate robustness. Improving model robustness requires considering multiple factors beyond adversarial attacks, such as underlying model characteristics, test information, and suitable metrics.

The impact of datasets on the effectiveness of highlighting algorithms and revealing robustness has also been examined. Hendrycks et al. [31] For seven NLP datasets, rigorously examine and quantify out-of-distribution (OOD) generalisation. The generalisation of various models, such as Bag-of-Words (BoW) models, CNNs, and LSTMs, is measured using a robustness benchmark that considers realistic distribution changes. The performance of pretrained transformers degrades noticeably less, according to the authors. Additionally, they investigate characteristics that impact robustness and find that while more diverse pretraining data might help, larger models are not always more resilient than smaller models. The SST-2 dataset and the IMDB dataset, two well-known benchmarks for sentiment analysis, are utilised to train and test a model, respectively, using the generalisation benchmark. For models based on BoW, CNN, and LSTM, accuracy results are presented.

B. Robustifying NLP Models: Techniques

Numerous techniques have been introduced in the literature to improve the robustness of NLP models. These methods include word recognition algorithms, stochastic ensembles, interval bound propagation, ensemble classifiers with randomised smoothing, and more. We will outline a few commonly employed robustness strategies and highlight their applicability in the discussion that follows.

1) Randomized Smoothing for Ensemble Classifiers

In NLP tasks, relying solely on a single classifier can be problematic since manipulating input features can have a big effect on the classification results. In the machine learning literature, ensemble classifiers have been offered as a solution to this problem. These classifiers consist of multiple independently built classifiers that are aggregated to produce the final classification result. Incorporating ensemble classifiers in NLP models helps reduce bias in the training data and enhances robustness.

In the context of ensemble classifiers, robustness refers to the ability of a smoothed classifier to provide the correct prediction the class label y and any input x. All inputs in the perturbation text are accurately categorised thanks to a robust model. A method is called randomised smoothing that transforms a classifier into a new smoothed classifier, offering robustness in a specific setting. This method may validate an NLP model's resistance to different adversarial approaches, such as word-substitution attacks.

For instance, Dirichlet Neighborhood Ensemble (DNE), a randomised smoothing approach that trains a resilient model by mitigating substitution-based assaults, was developed by Zhout et al. [22]. DNE creates fictitious sentences by supplementing the training data with a group of a word and its synonyms and sampling the embedding vectors for each word in an input phrase. Without losing efficiency, our sampling approach guarantees resilience against adversarial assaults.

2) Randomized Smoothing for Stochastic Ensemble

Stochastic ensemble classifiers demonstrate randomness and uncertainty in their underlying models. This stochastic characteristic of NLP models can be effectively utilized to comprehend and characterize their behavior [19]. In their research, Ye et al. [20] provided a new randomised smoothing approach that is proven to be reliable. Using random word swaps, this approach creates a stochastic ensemble from the input phrases. Their technique may demonstrably guarantee the robustness of the model by using the statistical features of the ensemble.

What sets this method apart is its simplicity and generalizability? It does not rely on any specific model structure and Black-box queries on the model outputs are all that are needed. It may therefore be used with any pre-trained model, such BERT, and at various levels of granularity, such as word-level or sub word-level. Even though it has been demonstrated that robust training increases overall resilience against adversarial word-level perturbations, it is worth noting that re-executing Starting the training process from scratch is necessary for robust training on a different model. This limitation, pertains to certifiably robust training, is not addressed in the study mentioned.

However, it is crucial to acknowledge that the randomized smoothing technique is task-specific. If one applies the technique to a different task, like NLI, it requires restarting the robust training procedure from scratch, resulting in significant overhead.

3) Bound to an Interval Propagation

Machine learning classifiers are built using the Interval Bound Propagation (IBP) method with certifiable robustness. By employing interval arithmetic, IBP defines a loss function that minimizes an upper limit on any pair of logits' maximum difference when the input is shifted inside of a norm-bounded ball [25]. IBP has found widespread and successful application in the field of computer vision to achieve robustness guarantees [25-28].

A notable advantage of IBP in the context of NLP models is its capability to handle in addition to the continuous perturbations frequently utilised in computer vision applications, there are discrete perturbations [14]. However, it is crucial to consider robustness in conjunction with models trained on diverse datasets, regardless of the specific evaluation metric. An apparent limitation of this approach is the lack of utilization of a comprehensive benchmark like GLUE, which includes six benchmark datasets. These datasets might have been used to evaluate the classification accuracy further and show the effectiveness of the robustness technique.

4) Word Recognition

Several studies have explored the use of word recognition models in addressing adversarial attacks and strengthening the stability of NLP systems [20-24]. However, it remains uncertain whether this approach can be transferred and generalized to different network architectures and linguistic tasks, presenting a limitation that warrants further investigation. Nonetheless, word recognition remains a valuable technique in NLP, contributing to the understanding and representation of various forms of words in linguistic tasks.

Pruthi et al., for instance, suggested integrating a word recognition model as a defence mechanism against adversarial spelling errors that target a BERT model used for sentiment analysis. Combining both attack and defense strategies in a single study holds promise. Their findings indicated that the accuracy was drastically lowered from 90.3% to 45.8% by a single adversarially-crafted character assault. However, by employing their defense approach, the accuracy was restored to 75%. This method effectively strengthens NLP models for combating spelling errors and can also be utilized to detect words corrupted by random keyboard errors, providing defense against word perturbation attacks. Notably, this approach stands out as it not only identifies vulnerabilities in modern NLP models but also suggests a robustness method to mitigate these attacks.

V. ROBUSTNESS METRICS

A comprehensive evaluation of NLP model robustness requires the use of well-defined and relevant metrics that offer insights into the resilience of the model to hostile assaults. Metrics measure the performance of machine learning systems during both training and testing, playing a dual function in their design. Similarly, Machine learning model performance is evaluated and tracked using resilience measures in antagonistic conditions.

Our literature survey revealed the employment of various metrics in different research studies to assess robustness, with their Figure 2 highlights the context. Some of these measures will be reviewed in the sections that follow.

A. *Attack Percentage*

Researchers have extensively employed the attack success rate as a gauge of antagonistic assaults' effectiveness on NLP models. While the success rate of attacks is a straightforward and easily interpretable metric, it has limitations. It quantifies the efficiency of adversarial attacks by measuring the ratio of successful attempts to the total number of attack attempts. Different hostile instances have different qualities, and although some may be quickly discovered or mitigated using straightforward heuristics, others offer more difficult problems significant challenges for detection and mitigation using the same heuristics.

For example, Alzantot et al. utilized the attack success rate to evaluate the efficiency of their adversarial assaults using genetic algorithms, providing insights into the robustness of NLP models against contrasting instances.

The success rate of attacks is a commonly used and straightforward metric for assessing the stability of NLP models. Successful attempts refer to adversarial examples that meet predefined conditions, such as perturbation size and the desired effect on the classifier's output, such as reducing confidence or altering the categorization designation. It overlooks the elements of the produced adversarial instances' quality that increased the success rate.

B. *Rate of Error*

The error rate, commonly referred to as the robustness error, is a statistic that counts the instances in which an NLP model categorises input texts erroneously. For instance, according to Goodfellow et al. [13], a variety of models, including those based on neural networks, routinely misclassify adversarial cases. These instances are created by giving input examples from a dataset minor, purposefully worst-case perturbations, which encourage the models to confidently provide inaccurate results. The authors discovered that classifiers with different designs frequently misclassify adversarial instances. Various research investigations have used the error rate as a measure. While the error rate is a simple and straightforward metric for evaluating NLP models' resistance to hostile assaults should not be the determining factor in sole measure used to evaluate how well machine learning models perform. It has been extensively utilised in research projects to evaluate the resistance of NLP models against hostile assaults. A lower mistake rate, as opposed to the attack success rate, suggests a more robust NLP model against adversarial assaults. It fails to consider the intrinsic and often significant differences among the examples that contribute to the error rate.

C. *Bounds Tightness of IBP*

According to section IV-C3, Interval Bound Propagation (IBP) is a technique employed to achieve robustness. The advantage of using the IBP tightness metric is its ability to assess the Verifiable resistance to word substitution assaults of NLP models. If a model cannot surpass the boundary, regardless of how adversaries create adversarial examples, it attains provably guaranteed robustness. Numerous research projects have used the IBP tightness metric, such as [27]. [24], on the other hand, employed a vast family of label-preserving transformations, where each word in the input text can be changed to a comparable one, are taken into consideration, using the same measure to investigate verified resilience to word replacements. It is not recommended to rely simply on the IBP tightness metric as a measure of verified robustness. Researchers have investigated a metric for explicitly proving the degree of model resilience against adversarial assaults is the tightness of IBP's upper and lower limits. However, it is important to note that ideally, other evaluation metrics IBP tightness should be utilised in conjunction with normal accuracy and training accuracy to provide a more comprehensive assessment.

D. *Correct Classification*

Classification accuracy is a statistic that measures an NLP model's ability to accurately categorise input texts under various attack methods, such as word-level and character-level substitution assaults, as well as white-box and black-box attacks [12]. While classification accuracy is simple to compute and understand, it does not consider the effectiveness of specific instances on overall accuracy. Researchers have looked at the tightness of IBP's upper and lower bounds as a metric to formally evaluate the model's resistance to adversarial assaults. Interval Bound Propagation (IBP) is a technique for achieving resilience [18]. Many research studies have utilized classification accuracy as a performance measure, such as [14-27]. For instance, [29] used They used classification accuracy to assess the Metropolis-Hastings Sampling Algorithm (MHA) they had devised its superior attacking capability compared to the baseline model. If a model does not cross the barrier, regardless of how adversaries provide adversarial cases, it achieves provably-guaranteed resilience [20].

IBP tightness has been used as an employed in various research works, such as [27]. For example, [27] used IBP tightness to investigate the robustness for transformers verification problem, while [24] applied investigating certified robustness to word replacements using the same metric, considering a several label-preserving modifications. The benefit of IBP tightening is its ability to Verify the NLP models' provable resistance to word substitution assaults. However, it should not be solely relied upon as proof of authentic verified robustness. Alternative assessment measures are ideal, IBP tightness should be utilised in conjunction with normal accuracy and training accuracy to provide a more comprehensive assessment.

E. Diversity

Diversity is an important concept in training data, which means that examples from one class should be as different as possible from examples of another class to ensure better training [27]. However, the authors do not provide an explanation of how they measure the runtime of their algorithm, as there can be a trade-off between performance and accuracy in search algorithms.

Researchers have used the diversity metric in several studies [26-27] to evaluate the NLP models' classification precision and their robustness against adversarial attacks. The diversity metric has a strength in because it encourages a quality evaluation across different NLP domains, which allows for assessing models in ways that other metrics may not capture. For instance, [11] propose an FreeLB is an adversarial training technique that modifies input words to reduce the likelihood of antagonistic attacks and increase invariance in the embedding space.

On transformer-based models in NLU and reasoning problems, they use this strategy and observe improvements in test scores for models like BERT and RoBERTa. Various search algorithms, such as greedy search, genetic algorithms, and particle swarm optimization, have been extensively studied and could have been considered for this research [15]. It is not apparent, nevertheless, how this statistic may be applied to quantify diversity using recall and accuracy.

VI. FUTURE SCOPE & DISCUSSIONS

The datasets used in evaluating NLP models vary greatly and lack consistency, which is evident. Achieving impressive results on these benchmarks does not guarantee exceptional performance when deploying NLP models in real-world scenarios. Research has indicated that models can become adept at solving the dataset itself rather than truly understanding the language [28]. Hence, there are several gaps in the benchmark dataset landscape that require attention.

Firstly, initial efforts have been made within the NLP community to create unbiased benchmark datasets. However, the existing datasets have limitations, and techniques need to be developed to identify and remove biases from the data. This step would enable the evaluation of how NLP models perform in real-world deployments and address spurious correlations in datasets.

Secondly, relying on Using the same dataset for training and testing might give a misleading impression of a model's accuracy and durability. Examining NLP models on various datasets, such example by implementing the super GLUE, which presents a more challenging set of tasks compared to the traditional GLUE benchmark.

Thirdly, there is the requirement for a uniform evaluation system that allows thorough and equitable evaluations across varied language tasks. Developing a standardized and unified benchmark dataset for evaluation purposes would be highly valuable. Additionally, establishing strong baselines trained on domain-specific data could serve as a valuable test bed for comparison.

By addressing these gaps, we can enhance the reliability and meaningfulness of NLP model evaluations, enabling their deployment in real-world applications with improved performance and understanding.

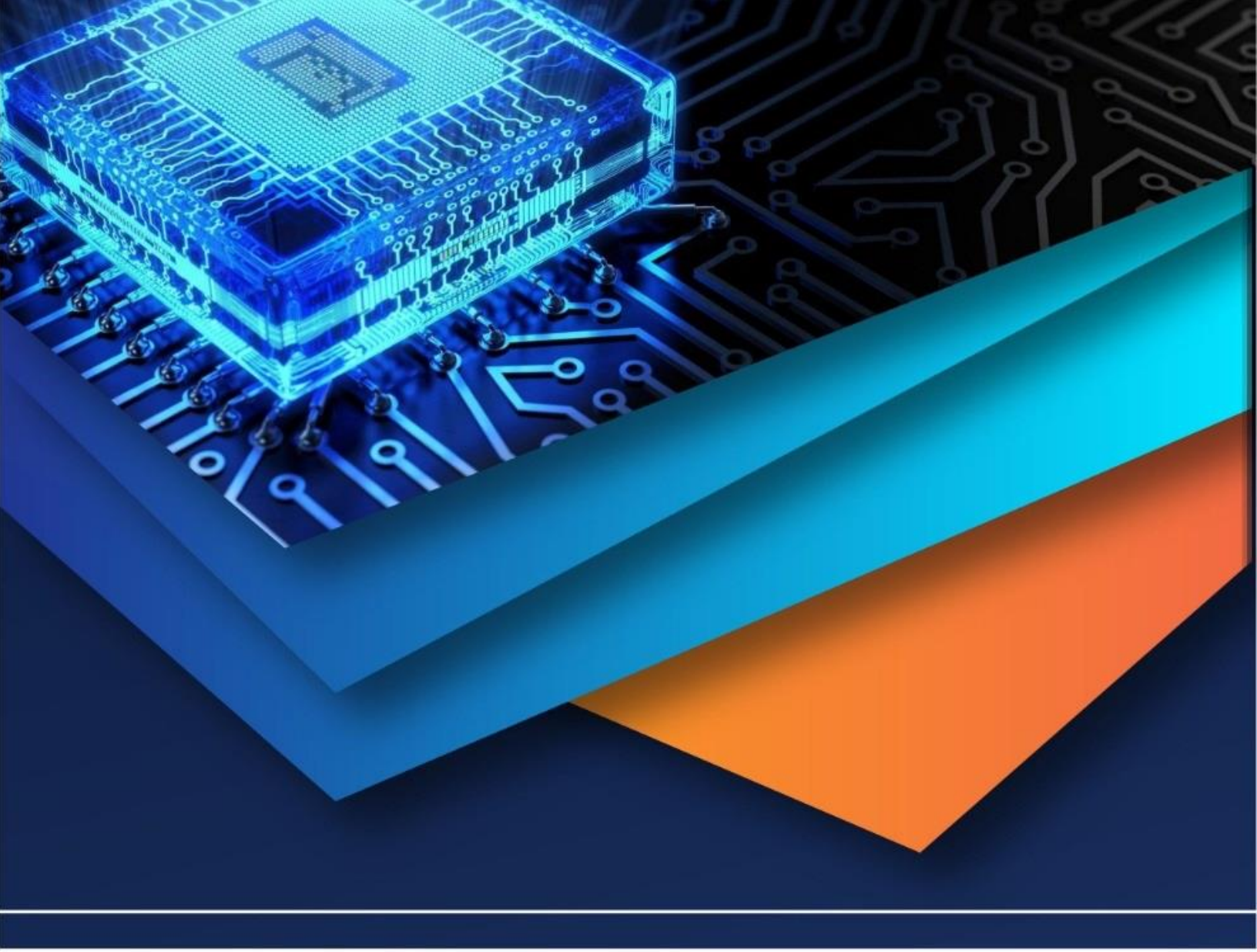
VII. CONCLUSION

This paper presents a comprehensive and well-structured survey of NLP robustness research, identifying gaps in the existing literature and proposing future research directions across different stages of the pipeline. With various NLP projects in the real-world facing failure due to a lack of robustness, it becomes crucial to explore robustness as a multifaceted idea and devise novel techniques to address its challenges effectively. These techniques should target spurious correlations and accomplish excellent out-of-distribution precision to handle perturbations and ensure accurate text classification in practical scenarios. The goal of this research is to provide a valuable resource for the research community, offering insights into appropriate techniques, metrics, and datasets, while also inspiring further exploration to address the identified gaps in this field.

REFERENCES

[1] S. Mannarswamy and S. Roy, ``Evolving AI from research to real life_Some challenges and suggestions," in Proc. 27th Int. Joint Conf. Artif. Intell., Jul. 2018, pp. 5172_5179.

[2] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, ``Natural language processing: An introduction," J. Amer. Med. Inform. Assoc., vol. 18, no. 5, pp. 544_551, 2011.
[3] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge, U.K.: Cambridge Univ. Press, 2020.
[4] A. Abusnaina, A. Khormali, H. Alasmary, J. Park, A. Anwar, and A. Mohaisen, ``Adversarial learning attacks on graph-based IoT malware detection systems," in Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst. (ICDCS), Jul. 2019, pp. 1296_1305.
[5] N. Zhang, X. Mi, X. Feng, X.Wang, Y. Tian, and F. Qian, ``Understanding and mitigating the security risks of voice-controlled third-party skills on Amazon Alexa and Google Home," 2018, arXiv:1805.01525.
[6] L. Blue, L. Vargas, and P. Traynor, ``Hello, is it me you're looking for? Differentiating between human and electronic speakers for voice interface security," in Proc. 11th ACM Conf. Secur. Privacy Wireless Mobile Netw., Jun. 2018, pp. 123_133.
[7] S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su, and A. Mohaisen, ``You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones," in Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst. (ICDCS), Jun. 2017, pp. 183_195.
[8] M. Abuhamad, A. Abusnaina, D. Nyang, and D. Mohaisen, ``Sensorbased continuous authentication of smartphones' users using behavioural biometrics: A contemporary survey," IEEE Internet Things J., vol. 8, no. 1, pp. 65_84, Jan. 2021.
[9] A. Abusnaina, A. Khormali, D. Nyang, M. Yuksel, and A. Mohaisen, ``Examining the robustness of learning-based DDoS detection in software defined networks," in Proc. IEEE Conf. Dependable Secure Comput.(DSC), Nov. 2019, pp. 1_8.
[10] H. Alasmary, A. Abusnaina, R. Jang, M. Abuhamad, A. Anwar, D. Nyang, and D. Mohaisen, ``Soteria: Detecting adversarial examples in control graph-based malware classi_ers," in Proc. IEEE 40th Int. Conf. Distrib. Comput. Syst. (ICDCS), Nov. 2020, pp. 888_898.
[11] E. Buber, B. Diri, and O. K. Sahingoz, ``Detecting phishing attacks from URL by using NLP techniques," in Proc. Int. Conf. Comput. Sci. Eng. (UBMK), Oct. 2017, Art. no. 337342.
[12] J. X. Morris, E. Li_and, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, ``TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP," 2020, arXiv:2005.05909.
[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, ``Explaining and harnessing adversarial examples," 2014, arXiv:1412.6572.
[14] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, ``A survey of data augmentation approaches for NLP," 2021, arXiv:2105.03075.
[15] J. Y. Yoo, J. X. Morris, E. Li_and, and Y. Qi, ``Searching for a search method: Benchmarking search algorithms for generating NLP adversarial examples," 2020, arXiv:2009.06368.
[16] J. Li, T. Tang, W. X. Zhao, and J.-R. Wen, ``Pretrained language models for text generation: A survey," 2021, arXiv:2105.10311.
[17] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, ``Adversarial attacks on deep-learning models in natural language processing: A survey," ACM Trans. Intell. Syst. Technol., vol. 11, no. 3, pp. 1_41, May 2020.
[18] M. T. Ribeiro, T.Wu, C. Guestrin, and S. Singh, ``Beyond accuracy: Behavioral testing of NLP models with CheckList," 2020, arXiv:2005.04118.
[19] J. Zeng, X. Zheng, J. Xu, L. Li, L. Yuan, and X. Huang, ``Certified robustness to text adversarial attacks by randomized [MASK]," 2021, arXiv:2105.03743.
[20] M. Ye, C. Gong, and Q. Liu, ``SAFER: A structure-free approach for certified robustness to adversarial word substitutions," 2020, arXiv:2005.14424.
[21] X. Dong, A. T. Luu, R. Ji, and H. Liu, ``Towards robustness against natural language word substitutions," in Proc. Int. Conf. Learn. Represent., 2020,pp. 1_14.
[22] A. Kumar, A. Levine, S. Feizi, and T. Goldstein, ``Certifying confidence via randomized smoothing," 2020, arXiv:2009.08061.
[23] P.-S. Huang, R. Stanforth, J. Welbl, C. Dyer, D. Yogatama, S. Gowal, K. Dvijotham, and P. Kohli, ``Achieving verified robustness to symbol substitutions via interval bound propagation," 2019, arXiv:1909.01492.
[24] B. Rychalska, D. Basaj, A. Gosiewska, and P. Biecek, ``Models in the wild: On corruption robustness of neural NLP systems," in Proc. Int. Conf. Neural Inf. Process. Cham, Switzerland: Springer, 2019, pp. 235_247.
[25] K. Goel, N. Rajani, J. Vig, S. Tan, J. Wu, S. Zheng, C. Xiong, M. Bansal, and C. Ré, ``Robustness gym: Unifying the NLP evaluation landscape," 2021, arXiv:2101.04840.
[26] M. Cheng, W. Wei, and C.-J. Hsieh, ``Evaluating and enhancing the robustness of dialogue systems: A case study on a negotiation agent," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., vol. 1, 2019, Art. no. 33253335.
[27] Z. Shi, H. Zhang, K.-W. Chang, M. Huang, and C.-J. Hsieh, ``Robustness verification for transformers," 2020, arXiv:2002.06622.
[28] J. Li, T. Du, S. Ji, R. Zhang, Q. Lu, M. Yang, and T. Wang, ``TextShield: Robust text classi_cation based on multimodal embedding and neural machine translation," in Proc. 29th USENIX Secur. Symp., 2020, pp. 1381_1398.
[29] T. Lin, Y. Wang, X. Liu, and X. Qiu, ``A survey of transformers," 2021, arXiv:2106.04554.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY