***Introduction to Adversarial Attacks and Defenses in Natural Language Processing:***

In recent years, the field of Natural Language Processing (NLP) has witnessed remarkable advancements, fueled by the rapid growth of deep learning techniques. These techniques have enabled machines to comprehend, generate, and manipulate human language with unprecedented accuracy and fluency. However, as NLP models become more capable and pervasive, they are also susceptible to a new breed of challenges: adversarial attacks. Adversarial attacks refer to carefully crafted input data designed to exploit vulnerabilities in machine learning models, leading them to produce erroneous or unintended outputs.

Natural language processing (NLP) models have recently achieved remarkable success in a variety of tasks, such as text classification, machine translation, and question answering. However, these models have also been shown to be vulnerable to adversarial attacks. Adversarial attacks are designed to cause models to make incorrect predictions by introducing carefully crafted perturbations to the input data. These perturbations are often imperceptible to humans, but they can have a significant impact on the model's output.(**Li et al. (2019))**

Much like their counterparts in computer vision, adversarial attacks in NLP have emerged as a critical concern. These attacks can range from subtle word substitutions that mislead sentiment analysis, to sophisticated manipulations that deceive machine translation systems. The underlying premise of these attacks is to expose the brittleness of NLP models, often highlighting their over-reliance on superficial cues rather than genuine semantic understanding.

Adversarial attacks on NLP models are a serious threat to the security and reliability of these models. These attacks can be used to manipulate the output of NLP models in a variety of ways, such as changing the sentiment of a text, the meaning of a sentence, or the factuality of a claim. Adversarial attacks can also be used to bypass security measures, such as spam filters and CAPTCHAs.

Understanding and mitigating adversarial attacks in NLP is of paramount importance due to their real-world implications. Consider the implications of a misleading sentiment analysis in product reviews, or the potential chaos arising from manipulated news articles in automated news summarization systems. As NLP models are increasingly integrated into decision-making pipelines, their vulnerability to adversarial attacks could have wide-ranging consequence

Adversarial training is a promising defense against adversarial attacks on NLP models. Adversarial training involves training the model on both genuine and adversarial examples. This helps the model to learn to identify and resist adversarial perturbations. Adversarial training has been shown to be effective against a variety of adversarial attacks on NLP models, such as word substitution attacks and grammatical error attacks. **(Miyato et al. (2016))**

The study of adversarial defenses in NLP is equally crucial. Defenses aim to bolster the resilience of NLP models against adversarial attacks by enhancing their robustness and generalization capabilities. Just as the arms race between attackers and defenders has been a hallmark of cybersecurity, a similar dynamic is emerging in the realm of NLP.( **Madry et al. (2017))**

This research domain raises several intriguing questions. How do we measure the robustness of NLP models? What are the different types of adversarial attacks specific to language, and how can we generate effective adversarial examples? Can we develop defenses that enhance the security of NLP models without sacrificing their performance on genuine inputs? Furthermore, understanding the relationship between adversarial attacks and model interpretability is a pressing concern, as transparent models are often more resilient to adversarial perturbations.

In this context, this survey aims to provide a comprehensive overview of the landscape of adversarial attacks and defenses in NLP. We delve into the various attack strategies that manipulate text inputs and investigate the different defense mechanisms proposed to counteract them. By exploring the evolution of this field, we seek to shed light on the challenges that lie ahead and inspire future research directions to build more robust and trustworthy NLP models.

In the subsequent sections, we will discuss the foundational concepts of adversarial attacks, highlight prominent attack techniques, delve into existing defense strategies, and conclude with a reflection on the open challenges and opportunities in this dynamic and critical research area.`

References:

* **Li et al. (2019)** showed that NLP models can be vulnerable to adversarial attacks.
* **Miyato et al. (2016)** found that adversarial attacks can be used to manipulate the output of NLP models in a variety of ways.
* **Madry et al. (2017)** proposed adversarial training as a promising defense against adversarial attacks on NLP models.