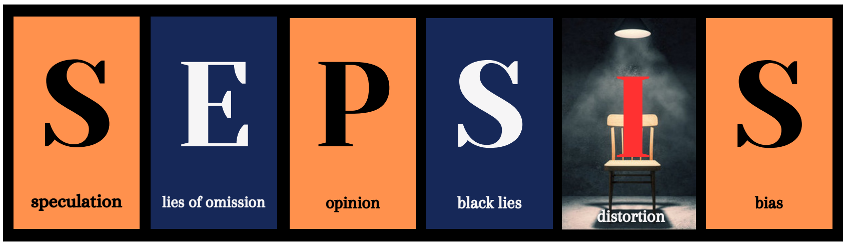


SEPSIS: I Can Catch Your Lies – A New Paradigm for Deception Detection

Anku Rani[1*] **Dwip Dalal**[2] **Shreya Gautam**[3] **Pankaj Gupta**[4]
Vinija Jain†[5,6] **Aman Chadha**†[5,6] **Amit Sheth**[7] **Amitava Das**[7]
[1] Massachusetts Institute of Technology [2]IIT Gandhinagar, India
[3]Politecnico di Milano, Italy [4]DTU, India [5]Stanford University, USA
[6]Amazon AI, USA [7]University of South Carolina, USA
ankurani@mit.edu

Abstract

Deception is the intentional practice of twisting information. It is a nuanced societal practice deeply intertwined with human societal evolution, characterized by a multitude of facets. This research explores the problem of deception through the lens of psychology, employing a framework that categorizes deception into three forms: *lies of omission*, *lies of commission*, and *lies of influence*. The primary focus of this study is specifically on investigating only *lies of omission*. We propose a novel framework for deception detection leveraging NLP techniques. We curated an annotated dataset of $876,784$ samples by amalgamating a popular large-scale fake news dataset and scraped news headlines from the Twitter handle of "*Times of India*", a well-known Indian news media house. Each sample has been labeled with four layers, namely: (i) the type of omission *(speculation, bias, distortion, sounds factual,* and *opinion)*, (ii) colors of lies *(black, white, grey,* and *red)*, and (iii) the intention of such lies *(to influence, gain social prestige,* etc*)* (iv) topic of lies *(political, educational, religious, racial,* and *ethnicity)*. We present a novel multi-task learning [MTL] pipeline that leverages the dataless merging of fine-tuned language models to address the deception detection task mentioned earlier. Our proposed model achieved an impressive F1 score of 0.87, demonstrating strong performance across all layers including the *type*, *color*, *intent*, and *topic* aspects of deceptive content. Finally, our research aims to explore the relationship between *lies of omission* and *propaganda* techniques. To accomplish this, we conducted an in-depth analysis, uncovering compelling findings. For instance, our analysis revealed a significant correlation between *loaded language* and *opinion*, shedding light on their interconnectedness. To encourage further research in this field, we are releasing the SEPSIS dataset and code at https://huggingface.co/datasets/ankurani/deception.

1 Defining Deception – Inspiration from Psychology

According to (Schuiling, 2004), deception is a behavior observed in various species and is considered an evolutionary adaptive trait. (DePaulo and Kashy, 1998) assert that deception is an integral part of social interactions, with the majority of humans engaging in deceptive acts at least once or twice a day. While most instances of deception are relatively minor, there is a frequent association between deception and egregious norm violations, such as theft, murder, and attempts to evade punishment for such crimes. Consequently, researchers have long been interested in identifying behaviors that can differentiate between truthful and deceitful communications.

Numerous studies have delved into describing the behavioral indicators of deceit. However, no sin-

[*] Work was done when the author was at the University of South Carolina

[†]Work does not relate to position at Amazon.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 97–128
July 28-29, 2025 ©2025 Association for Computational Linguistics

gle behavior or combination of behaviors has been found to possess the definitive ability to accurately determine deceptive communication. The empirical evidence supporting the significance of specific individual behaviors in deception often presents conflicting findings (DePaulo, 1985; Kraut, 1980; Vrij, 2000). One possible explanation for these contradictions in the literature regarding deception cues is the insufficient differentiation made by researchers between distinct subtypes of deception.

In the realm of psychology research, a consensus has yet to be reached regarding the classification of various types of deception. Nevertheless, we discovered that the framework outlined in Hample's work (Hample, 1982), visually described in fig. 1, provides a viable foundation for constructing NLP models. (Hample, 1982) categorizes deception into three distinct forms: *lies of omission*, *lies of commission*, and *lies of influence*. For the purpose of our study, we focus solely on investigating *lies of omission*. It is worth noting that the NLP community has extensively explored the fact verification problem, which is primarily associated with *lies of commission*. Conversely, *lies of omission* have received comparatively less attention. In this paper, we present a comprehensive study on lies of omission, which, to the best of our knowledge, is the first of its kind.

Our Contributions: SEPSIS dataset, MTL framework utilizing dataless LLM merging, unveiling the relationship between deception and propaganda.

- This paper presents a pioneering study on the phenomenon of lies of omission.
- It introduces the SEPSIS corpus (876,784 data points) and four layers of annotation, including type, color, intention, and topic.
- The paper introduces an MTL pipeline for SEPSIS classification.
- The MTL pipeline leverages the dataless merging of fine-tuned Language Models (LMs).
- It incorporates a tailored loss function specific to each layer, addressing different subproblems.
- Finally, the paper reveals a significant correlation between deception and propaganda techniques.

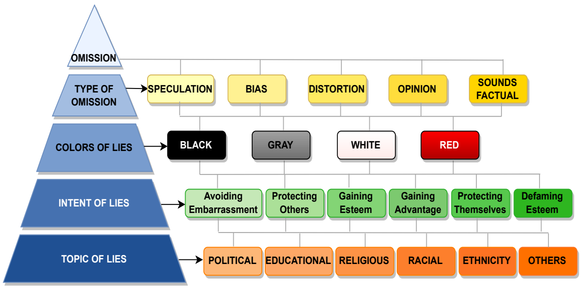


Figure 1: The figure represents the categorization of the SEPSIS corpus across all layers. The 1^{st} layer represents *type of omission* and its respective categories, 2^{nd} layer represents colors of lies, 3^{rd} layer represents the intent of lies, and 4^{th} layer represents the topic of lies.

2 Introducing SEPSIS: A novel corpus on lies of omission

We are delighted to introduce the **SEPSIS** corpus (**S**p**E**culation o**P**inion bia**S** d**IS**tortion), explicitly curated for *lies of omission*. This novel resource will significantly enhance the study and analysis of deceptive communication by focusing on the deliberate exclusion of information. Figure 1 offers a concise visual depiction that effectively summarizes the categorization we present in the SEPSIS. In the subsequent paragraphs, we present a collection of scientific inquiries along with their corresponding answers, which serve as the driving force behind our research. Furthermore, we delve into the influence of these questions on the development of our annotation schema, which lays the groundwork for our research framework.

Is there a specific dialogue act that individuals employ for lies of omission? Within the classical switchboard corpus (Godfrey et al., 1992), there exist 42 well-defined dialogue acts. Following extensive deliberation and analysis, we have reached the conclusion that individuals often utilize dialogue acts such as *speculation, opinion, bias, and distortions* when engaging in deceptive behavior.

These dialogue acts function as figurative communication techniques employed by individuals to mask their deceit through encryption (Elaad, 2003), particularly when they desire to disclose certain information selectively.

- **Speculation** entails conjecturing without ample evidence.
- **Opinion** is a subjective viewpoint formed without relying on factually accepted knowledge.
- **Bias** refers to unfair prejudice towards a particular individual or group.
- **Distortion** is the act of twisting something away from its genuine, inherent, or initial condition.

  , we define **sounds factual** as a statement that seems factual but may not be true.

1st level: type of omission

Speculation: Biden warned the US does not have 'resources to win WW3' as tensions rise in the Middle East.

Opinion: Poll: Trump receives low overall approval rating but praise for strong economy.

Bias: Russia lauds India for following own interests on energy issue.

Distortion: Republic TV: Jama Masjid in dark due to non-payment of electricity bills over four crores.

Sounds Factual: A US government study confirms most face recognition systems are racist.

What has been omitted? In the study of lies of omission, it is crucial to determine what information has been deliberately omitted. To address this, we draw inspiration from journalism, where the use of the 5W framework is common. The 5W framework consists of the questions *who, what, when, where, and why* which are considered fundamental in information gathering and problem-solving. These questions are frequently utilized in journalism and police investigations (Mott, 1942; Stofer et al., 2009; Silverman, 2020; Su et al., 2019; Smarts, 2017; Wikipedia, 2020). As an example:

{Hillary Clinton}$_{\text{who}_1}$ announces {Global Climate Resilience Fund}$_{\text{what}}$ for {women}$_{\text{who}_2}$ to{tackle climate change}$_{\text{why}}$

What is the vulnerability of the uttered lie? In the realm of deception research, it is of utmost importance to comprehend and quantify the susceptibility of lies. One approach involves categorizing lies into different colors, namely *black, red, white, and gray* (Ratliff, 2011; DePaulo, 2004). Each color represents a distinct type of lie with varying levels of vulnerability, as detailed below:

- **Black lie** is about simple and callous selfishness. Typically uttered when there is no benefit to others, its sole intention is to extricate oneself from trouble.
- **White lie** prioritizes others' welfare over personal interests, reflecting an altruistic nature.
- **Gray lies** exhibits dual behavior, partially benefiting others and partially benefiting oneself depending on the viewpoint.
- **Red lies** are spoken from a hatred and revenge perspective against individuals or groups.

2nd level: colors of lie

Red: Donald Trump's congratulatory post for North Korea's WHO membership sparks outrage and controversy.

Black: FTX collapse: Former CEO Sam Bankman-Fried urges court to toss charges.

White: An apple a day slashes frailty risk by 20 percent, but Study points otherwise.

Gray: Hillary Clinton Announces Global Climate Resilience Fund For Women To Tackle Climate Change.

What is the intent of the lie? Studying the intent of lies helps to comprehend deceptive language's objectives. We have thus categorized lies into different intents as shown below.

3rd level: intent of lie

Gaining Advantage: Elizabeth Holmes ordered dinners for Theranos staff but made sure they weren't delivered until after 8 p.m. so they worked late: book.

Protecting Themselves: ChatGPT creator Sam Altman testifies to US Congress on AI risks.

Avoiding Embarrassment: Trump's Suggestion That Disinfectants Could Be Used to Treat Coronavirus Prompts Aggressive Pushback, was Sarcastic?

Gaining Esteem: Sasan Goodarzi, the CEO of software giant Intuit, which has avoided mass layoffs, says tech firms axed jobs because they misread the pandemic.

Protecting Others: Nobel Laureate Malala Urges U.S. To Bolster Support For Afghan Girls, Women!

Defaming Esteem: Taiwan war would be 'devastating,' warns US Defense Secretary Lloyd Austin as he criticizes China at Shangri-La security summit.

- **Intent of Gaining Advantage** can be used as an act of intentionally providing false information or misleading others to gain an unfair advantage over them.
- **Intent of Protecting Themselves** can be used as a means of self-preservation or self-defense when an individual feels threatened or vulnerable.
- **Intent of Avoiding Embarrassment** can be employed to evade situations that may lead to embarrassment, humiliation, or social discomfort.
- **Intent of Gaining Esteem** can be utilized to enhance one's reputation, social status, or personal image.

- **Intent of Protecting Others** can be used as a means of preservation for others when a group or community feels threatened or vulnerable.
- **Intent of Defaming Esteem** intends to damage reputation by spreading false information or rumors.

What is the topic of lie? To study deception further and to understand its topical influence, this research categorizes different topics of lies such as political, educational, etc.

- **Political** deception occurs by the deliberate use of statements by political entities to manipulate public opinion.
- **Educational** deception occurs by the deliberate use of statements by academic entities to manipulate opinion, directed especially towards the younger population.
- **Racial** deception occurs when individuals intentionally misrepresent their racial identity or engage in deception driven by racial motives.
- **Religious** deception involves the act of deceiving others by misrepresenting one's religious beliefs.
- **Ethnic** deception refers to the act of intentionally manipulating one's ethnic identity by targeting specific ethnic groups.

4^{th} level: topic of lie

Political: No elections safe from AI, deep fake photos, videos of politicians to become common, warns former Google boss.

Educational: Hundreds gather at Florida school board meeting over Disney movie controversy: 'Your policies are not protecting us from anything.

Religious: Pope: Christianity, Islam share common commitment to good life.

Racial: Why shouldn't a mixed-race actress play Egyptian queen Cleopatra?

Ethnicity: Egyptians complain over Netflix depiction of Cleopatra as black.

3 SEPSIS: Data Sources, Annotation, and Agreement

At the outset, we engaged in the manual annotation of 5,100 sentences through four co-authors, employing four layers of deception. Subsequently, we applied data augmentation techniques as detailed in Section 4, culminating in a total of 8,76,784 data points.

3.1 Data Sources

In terms of data sources, we have identified two distinct categories of interest. The first category focuses on the presence of omissions in factual data, specifically news data. The second category examines the involvement of omissions in fake news data. To address these categories, we have selected data sources from two prominent outlets: (a) Times of India (The Times of India, 2022) Twitter handle, the renowned news agency in India, and (b) Information Security and Object Technology (ISOT) fake news dataset (University of Victoria, 2022). More information on these sources can be found in the appendix B.1. A detailed analysis of the SEPSIS corpus and the results can be found in Appendix B.4.

3.2 Data Annotation

We chose to leverage our four co-authors for annotation purposes, which provides a knowledgeable and reliable solution for annotating sensitive deception datasets, ensuring high-quality expert judgment throughout the process. To maintain annotation consistency, we implemented rigorous checks and measures throughout the entire annotation process. The dataset was annotated at the sentence level using a multi-class annotation approach, allowing each individual feature to be assigned multiple categories during the annotation process. For instance, a statement could be tagged as both speculative and sounding factual, recognizing the possibility for it to either be a verifiable fact or contain speculative elements that satisfy both possibilities. A comprehensive account of the overall annotation process is provided in Appendix B.2. Notably, during the initial layer of annotation, if a particular text appeared to be factual, we refrained from annotating the specific type, intent, and influence of the lie since it was treated as a fact.

3.3 Inter Annotator Agreement and Quality

To ensure quality control in the co-author annotations, we performed cross-validation annotation on 1000 data points. This validation dataset was utilized to assess the consistency of annotations provided by individual co-authors. Based on this assessment, we established annotation guidelines

	Lies of omission					Color of lies				Intent of Lies					
	Speculation	Bias	Distortion	Opinion	Sounds Factual	Black	White	Grey	Red	Gaining Advantage	Protecting Themselves	Avoiding Embarrassment	Gaining Esteem	Protecting Others	Defaming Esteem
Tweet	0.678	0.632	0.619	0.62	0.759	0.831	0.807	0.771	0.846	0.790	0.752	0.692	0.744	0.637	0.609
Fake News	0.719	0.661	0.683	0.603	0.727	0.878	0.845	0.811	0.892	0.759	0.81	0.738	0.677	0.709	0.681

Table 1: Kappa score representation for layer 1: *type of omission* layer 2: *colors of lies*, and layer 3: *Intent of lies*. Kappa score for the layer 4 topic of lies can be found in Appendix B.3.

and conducted calibration sessions among the co-author team. For the annotation task, each co-author contributed their expertise across all four layers of the annotation process. We obtained four annotations per sentence and subsequently consolidated the data using an improved voting technique, as suggested in (Hovy et al., 2013), which has been empirically shown to outperform majority voting. To assess the level of agreement in the annotated corpus, we also calculated the Cohen Kappa score (Cohen, 1960). Since there are multiple categories for a given sentence, we report class-wise agreement scores. The overall agreement score is presented in Table 1. An overview of data points is presented in Table 2. To understand how features across these four layers are dependent on each other, we present six heatmaps in Appendix B.4.

Data Source	Sentences	+ Paraphrasing	+ Mask Infilling
Tweets	2495	12475	389105
Fake News	2605	13025	487829
Total	5100	25500	876784

Table 2: Number of original sentences and augmented sentences using *paraphrasing* and *mask infilling*.

4 Data Augmentation

It is widely acknowledged that neural network-based techniques have a high demand for data. To address this data requirement, data augmentation has almost become a standard practice in the AI community (Van Dyk and Meng, 2001; Shorten et al., 2021; Liu et al., 2020). We have utilized three methods for data augmentation here: (i) paraphrasing, (ii) 5W masking followed by infilling (Gao et al., 2022).

4.1 Paraphrasing Deceptive Datapoints

The motivation for paraphrasing deceptive data stems from the diverse manifestations of textual deceptive content in real-world scenarios, often influenced by variations in writing styles among different news publishing outlets. It is vital to incorporate these variations in order to establish a robust benchmark that facilitates comprehensive evaluation and analysis (cf. Figure 8 in Appendix C.1 for examples).

Undoubtedly, manual generation of possible paraphrases is ideal; however, this process is time-consuming and labor-intensive. On the other hand, automatic paraphrasing has garnered significant attention recently (Niu et al., 2020; Nicula et al., 2021; Witteveen and Andrews, 2019; Nighojkar and Licato, 2021). We used GPT-3.5 (Brown et al., 2020) (specifically the *text-davinci-003* variant) (Brown et al., 2020) model as it generates linguistically diverse, grammatically correct, and a maximum number of considerable paraphrases, i.e., 5 in this case. This is the best-performing model for data augmentation using paraphrasing (Rani et al., 2023). Additionally, we conducted experiments with Pegasus (Zhang et al., 2020) and T5 (T5-Large) (Raffel et al., 2020) models, but GPT-3.5 (`text-davinci-003` variant) (Brown et al., 2020) outperformed them, as indicated in Appendix C.1. We gathered a total of 25,500 unique paraphrased deceptive data points through this method.

At this stage, several important questions arise: (i) *What is the accuracy of the paraphrases generated?* (ii) *How do they differ from or distort the original content?* To address these questions, we have conducted extensive experiments and obtained empirical answers. However, due to space limitations, please refer to Appendix C.1 for details of our experiments and conclusions. We have evaluated the paraphrase modules based on three key dimensions: *(i) Coverage: number of consid-*

erable paraphrase generations, (ii) Correctness: correctness of these generations, and (iii) Diversity: linguistic diversity in these generations.

4.2 Synthetic Data Augmentation using 5W Specific Mask Infilling

As mentioned previously in section 2, our hypothesis revolves around the possible omission of the 5W (who, what, when, where, and why) for deceits. With this in mind, we developed a pipeline to detect the presence of the 5W and subsequently replace them with deceptive/null information generated from a generative LM. In the subsequent subsections, we will present our methodology for designing 5W semantic role labeling and mask filling techniques to address 5W omission.

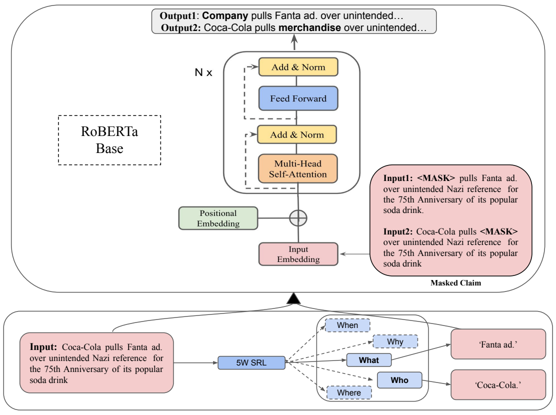

Output1: **Company** pulls Fanta ad. over unintended…
Output2: Coca-Cola pulls **merchandise** over unintended…
N x
Add & Norm
Feed Forward
Add & Norm
Multi-Head Self-Attention
RoBERTa Base
Positional Embedding
Input Embedding
Input1: <MASK> pulls Fanta ad. over unintended Nazi reference for the 75th Anniversary of its popular soda drink.
Input2: Coca-Cola pulls **<MASK>** over unintended Nazi reference for the 75th Anniversary of its popular soda drink
Masked Claim
Input: Coca-Cola pulls Fanta ad. over unintended Nazi reference for the 75th Anniversary of its popular soda drink
5W SRL
When
Why
What
Who
Where
'Fanta ad.'
'Coca-Cola.'

Figure 2: Architecture representation for the process of leveraging mask infilling using RoBERTa (Liu et al., 2019) for creating the deception dataset.

5W Semantic Role Labeling: Identification of the functional semantic roles played by various words or phrases in a given sentence is known as semantic role labeling (SRL). SRL is a well-explored area within the NLP community. There are quite a few off-the-shelf tools available: (i) Stanford SRL (Manning et al., 2014), (ii) AllenNLP (AllenNLP, 2020), etc. A typical SRL system initially identifies the verbs in a given sentence and subsequently associates all the related words/phrases with the verb through relational projection, assigning them appropriate roles. Thematic roles are generally marked by standard roles defined by the Proposition Bank (generally referred to as PropBank) (Palmer et al., 2005), such as: *Arg0, Arg1, Arg2*, and so on. We propose a mapping mechanism to map these PropBank arguments to 5W semantic roles (look at the conversion table 8, in appendix).

5W Slot Filling: Building upon our hypothesis, it is plausible for individuals to deliberately omit any of the given W to transform a statement into a lie of omission. Therefore, once we detect the presence of the Ws, our objective is to generate variations of the original statement by selectively omitting specific Ws. For this purpose, we train a masked LLM as depicted in the Figure 2. For the 5W slot-filling task we have experimented with five models: (i) MPNet (Song et al., 2020) , (ii) ELECTRA (Clark et al., 2020), (iii) RoBERTa (Liu et al., 2019), (iv) ALBERT (Lan et al., 2019), and (v) BERT (Devlin et al., 2018).

RoBERTa (Liu et al., 2019), a language model that leverages large-scale pre-training and removes the next sentence prediction objective, significantly enhancing language understanding. With its transformer architecture and fine-tuning, it predicts the original masked tokens in an *input sequence X* by maximizing the likelihood of the true masked tokens given the predicted *probabilities P*. Considering the scenario where all the Ws are present in a sentence, it is feasible to generate five variations. At this juncture, a crucial question arises: is there a high likelihood that the generated sentences deviate substantially from the original deceptive input? To substantiate we have calculated BLEU (Papineni et al., 2002) score between the original input and all the perturbed generations, reported in Table 3.

Model	BLEU Score
RoBERTa-base	0.7457
MPNet-base	0.7329
ELECTRA-large-generator	0.7225
BERT-base-uncase	0.7222
ALBERT-large-v2	0.7116

Table 3: BLEU Score for various models for mask infilling. RoBERTa performed the best.

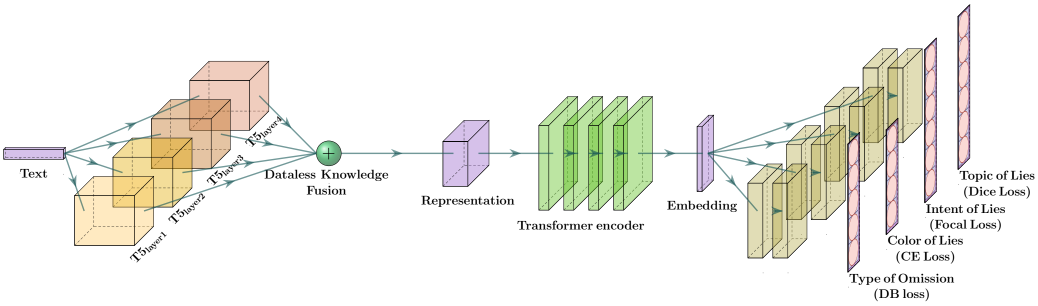

Text
T5layer1
T5layer2
T5layer3
T5layer4
Dataless Knowledge Fusion
Representation
Transformer encoder
Embedding
Topic of Lies (Dice Loss)
Intent of Lies (Focal Loss)
Color of Lies (CE Loss)
Type of Omission (DB loss)

Figure 3: Multi-task learning architecture delineating the process of an input text going through labeling along four dimensions: (i) types of omission, (ii) colors of lie, (iii) intention of lie, and (iv) topic of lie. Here, DB Loss stands for Distribution-Balanced Loss and CE loss stands for Cross Entropy loss (cf. Appendix D.2).

5 Designing the SEPSIS Classifier

SEPSIS, by its design, is a multitask-multilabel problem requiring the application of Multitask Learning (MTL) techniques. In general MTL framework utilizes a shared representation for all the tasks. It has been observed by several researchers (Parisotto et al., 2015; Rusu et al., 2015; Yu et al., 2020; Fifty et al., 2021) that shared representation has its own limitations and further effects on learning task-specific loss functions. In our approach, we introduced two specific innovations, detailed in subsequent sections. Using the MTL model (Fig. 3), we achieved a score of 0.81 F1 score on the human-annotated dataset (5000 samples) and 0.87 F1 score on the SEPSIS dataset (0.8M data points). Fig. 4 shows the F1 score across deception classes on the SEPSIS dataset (cf. Appendix D).

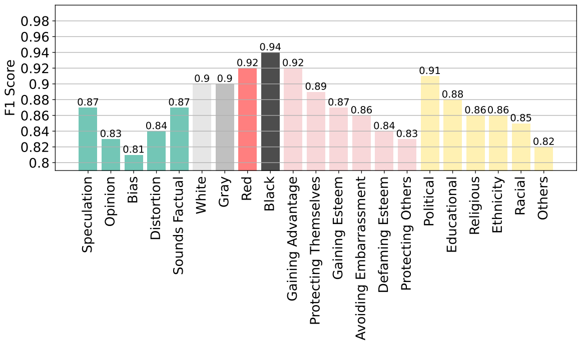


Figure 4: SEPSIS's F1 score for all classes of deception.

5.1 Merging Finetuned LLMs Brings Power!

Drawing inspiration from (Jin et al., 2022), we incorporated techniques for merging multiple fine-tuned LLMs, a process referred to as *dataless merging*. During our experimentation with various LLMs, we found that T5 performed exceptionally well for our specific case, and was also the best LM for dataless merging as emphasized in (Jin et al., 2022). For the four layers of deception, we fine-tuned four T5 models using the data outlined in Table 2. These models are denoted as T5$_{\text{layer1}}$, T5$_{\text{layer2}}$, T5$_{\text{layer3}}$, and T5$_{\text{layer4}}$. By leveraging the methodology proposed in (Jin et al., 2022), we merged these fine-tuned T5 models to achieve a better-shared representation tailored to our specific objectives. Figure 3 visually depicts the merging process via an architecture diagram.

5.2 Tailored Loss Function

During our exploration for suitable sub-task loss functions, we experimented with several available options, including (i) cross-entropy loss, (ii) focal loss (Lin et al., 2017), (iii) dice loss (Li et al., 2019), and (iv) distribution-balanced loss (DB) (Huang et al., 2021a). After a thorough evaluation, we observed that distribution-balanced loss yielded the best performance for layer 1, cross-entropy loss was most effective for layer 2, focal loss performed well for layer 3, and dice loss was the optimal choice for layer 4. For a comprehensive overview of the results and an in-depth discussion of different loss functions, please refer to the Appendix D.2.

6 Dissecting Propaganda through the Lens of Deception

As mentioned earlier, numerous studies have explored the behavioral indicators of lying, but there is hardly any consensus on categorization. However, the focus of this paper specifically revolves around investigating *lies of omission* and their connection to related research within the scientific community. Notably, there are works that have extensively examined the analysis of *propaganda* through language (Da San Martino et al., 2019; Martino et al., 2020).

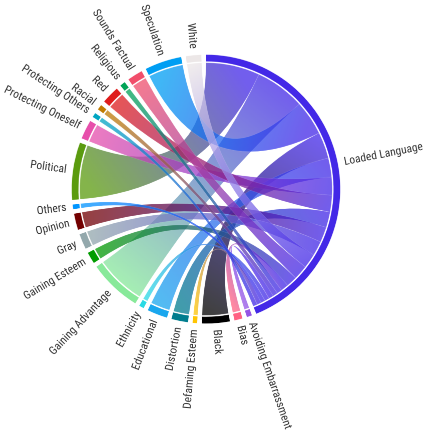


Figure 5: The Circos presents the co-occurrence of all the layers of deception with a propaganda technique named *loaded language*.

Our scientific curiosity led us to further investigate the specific types of *lies of omission* employed in strategizing particular propaganda, such as *exaggeration* and/or *red herring*. To conduct this study, we utilized the propaganda datasets introduced by (Da San Martino et al., 2019) and applied the SEPSIS classifier, as discussed in section 5 on the data. Through the analysis of these experiments, we made intriguing discoveries, including: (i) *the prevalence of political topic in loaded language compared to other propaganda types*, (ii) *the close association between the intention of gaining advantage and Name Calling*, and (iii) *the complexity underlying causal simplification as a form of speculation.* A Circos (Flourish, 2023) example is presented in Fig. 5 for a propaganda technique named *loaded language* (cf. Appendix E for Circos diagrams corresponding to propaganda techniques). Therefore, we firmly believe that our research on SEPSIS not only stands on its own but also acts as a bridge, facilitating a deeper understanding of deception.

7 Related Works

Deception detection has been explored on a wide range of applications, such as online dating services (Toma and Hancock, 2010) (Guadagno et al., 2012), social networks (Ho and Hollister, 2013), consumer reviews (Li et al., 2014) (Ott et al., 2011), and court transcripts (Fornaciari and Poesio, 2013) (Pérez-Rosas et al., 2015). Significant research findings have demonstrated a correlation between gender and deceit (Pérez-Rosas and Mihalcea, 2015), as well as a connection between deception and cultural factors (Pérez-Rosas and Mihalcea, 2014). The majority of conducted experiments are predicated on a binary classification approach for analyzing input text, specifically distinguishing between deceptive and non-deceptive instances as explored by (Mbaziira and Jones, 2016) and (Mihalcea and Strapparava, 2009). To the best of our knowledge, there is currently no computational study that comprehensively defines and categorizes deception by drawing insights from psychology. In our paper, we introduce SEPSIS, which presents a novel definition and dataset aimed at tackling the issue of *lies of omission* in language. We firmly believe that SEPSIS holds the potential for establishing a connection between deception and fake news, and we intend to explore this further.

8 Conclusion and Future Avenues

In conclusion, this research makes several key contributions. First, we have introduced SEPSIS, a novel multi-layered corpus focused on lies of omission. Second, our MTL framework leverages recent advances in language model fine-tuning and dataless merging to optimize deception detection, achieving 0.87 F1 score. Finally, we have uncovered compelling relationships between propaganda techniques and lies of omission through empirical analysis. The public release of our dataset and models will catalyze future research on this complex societal phenomenon.

9 Discussion and Limitations

In this section, we self-criticize a few aspects that could be improved and also detail how we (tentatively) plan to improve upon those specific aspects-

9.1 Categorization of deception

We have considered the four layers and categories based on our understanding of the psychological framework and going manually through multiple samples to understand the type, intent, topic, and colors of lie. However, this list may not be exhaustive. This is the reason for us to have put an *others* category in the topic of lies. Categories could increase when categorizing deception in real life.

9.2 Data Augmentation

We used paraphrasing and mask infilling for building the sepsis corpus. However, we understand that a few generations might not be deceptive and could have generated non-deceptive texts. However, we have done extensive manual testing, and believe such cases are nominal.

9.3 SEPSIS Classifier

One of the limitations of the SEPSIS Classifier is the computational heaviness associated with finetuning the T5 model for each specific layer. This process requires considerable computational resources and time. As the T5 models need to be finetuned for each task head, so total computational time increase significantly with an increase in the number of task head. It is important to consider these computational limitations when implementing multi-task learning architectures, as they can impact the feasibility and scalability of the approach, particularly in scenarios with limited computational resources or a large number of output tasks.

10 Ethical Considerations

Through this framework, we propose models to classify deception. We also developed a large augmented deceptive dataset. However, we must address the potential misuse of the dataset and models by entities who may exploit the framework to generate deceptive texts such as creating fake news by manipulating the content. The deliberate dissemination of deceptive news, spreading propaganda techniques to shape public opinion, is also a significant concern. We vehemently discourage such misuse and strongly advise against it.

References

AllenNLP. 2020. Allennlp semantic role labeling. https://demo.allennlp.org/semantic-role-labeling. [Online; accessed 2023-01-02].

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28:41–75.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Bella M DePaulo. 1985. Deceiving and detecting deceit. *The self and social life*, pages 323–370.

Bella M DePaulo. 2004. The many faces of lies. *The social psychology of good and evil*, pages 303–326.

Bella M DePaulo and Deborah A Kashy. 1998. Everyday lies in close and casual relationships. *Journal of personality and social psychology*, 74(1):63.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eitan Elaad. 2003. Effects of feedback on the overestimated capacity to detect lies and the underestimated ability to tell lies. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 17(3):349–363.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. 2021. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516.

Flourish. 2023. Chord diagram.

Tommaso Fornaciari and Massimo Poesio. 2013. Automatic deception detection in italian court cases. *Artificial intelligence and law*, 21(3):303–340.

Jun Gao, Changlong Yu, Wei Wang, Huan Zhao, and Ruifeng Xu. 2022. Mask-then-fill: A flexible and effective data augmentation framework for event extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4537–4544, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Rosanna E Guadagno, Bradley M Okdie, and Sara A Kruse. 2012. Dating deception: Gender, online dating, and exaggerated self-presentation. *Computers in Human Behavior*, 28(2):642–647.

Dale Hample. 1982. Empirical evidence for a typology of lies.

Shuyuan Mary Ho and Jonathan M Hollister. 2013. Guess who? an empirical study of gender deception and detection in computer-mediated communication. *Proceedings of the American Society for Information Science and Technology*, 50(1):1–4.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021a. Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8153–8161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021b. Balancing methods for multi-label text classification with long-tailed class distribution. *CoRR*, abs/2109.04712.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. 2022. Dataless knowledge fusion by merging weights of language models. *arXiv preprint arXiv:2212.09849*.

Robert Kraut. 1980. Humans as lie detectors. *Journal of communication*, 30(4):209–218.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. 2014. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1566–1576.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2019. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Pei Liu, Xuemin Wang, Chao Xiang, and Weiye Meng. 2020. A survey of text data augmentation. In *2020 International Conference on Computer Communication and Network Security (CCNS)*, pages 191–195. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. *arXiv preprint arXiv:2007.08024*.

A Mbaziira and J Jones. 2016. A text-based deception detection model for cybercrime. In *Int. Conf. Technol. Manag*, pages 1–8.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 309–312.

Frank Luther Mott. 1942. Trends in newspaper content. *The Annals of the American Academy of Political and Social Science*, 219:60–65. (Accessed on Jan 10 2023).

Bogdan Nicula, Mihai Dascalu, Natalie Newton, Ellen Orcutt, and Danielle S McNamara. 2021. Automated paraphrase quality assessment using recurrent neural networks and language models. In *International Conference on Intelligent Tutoring Systems*, pages 333–340. Springer.

Animesh Nighojkar and John Licato. 2021. Improving paraphrase detection with the adversarial paraphrasing task. *arXiv preprint arXiv:2106.07691*.

Tong Niu, Semih Yavuz, Yingbo Zhou, Nitish Shirish Keskar, Huan Wang, and Caiming Xiong. 2020. Unsupervised paraphrasing with pretrained language models. *arXiv preprint arXiv:2010.12885*.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. 2015. Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342*.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 59–66.

Verónica Pérez-Rosas and Rada Mihalcea. 2014. Cross-cultural deception detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 440–445.

Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in open domain deception detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1120–1125.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Anku Rani, S. M Towhidul Islam Tonmoy, Dwip Dalal, Shreya Gautam, Megha Chakraborty, Aman Chadha, Amit Sheth, and Amitava Das. 2023. Factify-5wqa: 5w aspect-based fact verification through question answering. *Preprint*, arXiv:2305.04329.

Brianna Ratliff. 2011. *Behavioral cues associated with lies of omission and of commission: An experimental investigation*. The University of Southern Mississippi.

Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. 2015. Policy distillation. *arXiv preprint arXiv:1511.06295*.

GA Schuiling. 2004. Deceive, and be deceived! *Journal of Psychosomatic Obstetrics & Gynecology*, 25(2):170–174.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8:1–34.

Craig Silverman. 2020. Verification handbook: Homepage. (Accessed on Jan 11 2023).

Media Smarts. 2017. How to recognize false content online - The new 5 Ws. (Accessed on Jan 11 2023).

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Kathryn T Stofer, James R Schaffer, and Brian A Rosenthal. 2009. *Sports journalism: An introduction to reporting and writing*. Rowman & Littlefield Publishers.

Jing Su, Xiguang Li, and Lianfeng Wang. 2019. The Study of a Journalism Which Is almost 99% Fake. *Lingue Culture Mediazioni-Languages Cultures Mediation (LCM Journal)*, 5(2):115–137.

The Times of India. 2022. Twitter profile - the times of india.

Catalina L Toma and Jeffrey T Hancock. 2010. Reading between the lines: linguistic cues to deception in online dating profiles. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 5–8.

University of Victoria. 2022. The isot fake news dataset. Online Academic Community.

David A Van Dyk and Xiao-Li Meng. 2001. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50.

Aldert Vrij. 2000. *Detecting lies and deceit: The psychology of lying and implications for professional practice*. Wiley.

Robert A Wagner and Michael J Fischer. 1974. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173.

Wikipedia. 2020. Five Ws. (Accessed on Jan 2023).

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with large language models. *arXiv preprint arXiv:1911.09661*.

Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 162–178. Springer.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Frequently Asked Questions (FAQs)

✻ **What were the specific instructions provided to the annotators and the criteria used for selecting them in the crowd annotation process of 5000 sentences through AMT?**

➠ The annotation pipeline outlines a step-by-step approach to deception detection based on different layers, as shown in Figure 1. To ensure reliable annotations, the dataset source was kept undisclosed from the annotators. Notably, for sentences categorized as "Sounds Factual," no additional annotations were made apart from missing W's.

✻ **How were the loss functions determined, specifically for each task head?**

➠ The selection of loss functions for each task head was based on the characteristics of the class distribution for that specific task. If the class distribution was imbalanced, loss functions designed to handle such scenarios were chosen. Detailed explanations and experimental results supporting the choice of each loss function can be found in the appendix section D.

✻ **Why RoBERTa was finally chosen as our baseline model for the Mask Infilling task?**

➠ Our experimentation in comparison to other state-of-the-art language models like RoBERTa-base, MPNet-base, ELECTRA-large-generator, BERT-base-uncase, and ALBERT-large-v2 revealed a higher Bilingual Evaluation Understudy (BLEU) score using RoBERTa. The selection of RoBERTa as the preferred model for the mask infilling task, based on its highest BLEU score, implies that RoBERTa's generated outputs exhibited a greater resemblance to the desired reference outputs. This characteristic of RoBERTa's performance is particularly advantageous for generating deceptive sentences that closely resemble reference sentences. By leveraging RoBERTa's capabilities, the task of producing deceptive sentences can be effectively achieved with a higher degree of fidelity to the reference sentences.

✻ **Why was the T5 base model chosen for model merging, and how was its performance evaluated?**

➠ The selection of the T5 base model for model merging involved extensive experimentation and evaluation of various language models (LLMs), such as RoBERTa, T5, and DeBERTa. Our evaluation aimed to identify the LLM that would deliver the best performance for our specific case. Initially, we assessed the individual performance of each LLM by utilizing them in the architecture to generate word embeddings, without employing model merging or fine-tuning. However, there was no significant improvement in scores observed for RoBERTa and DeBERTa when compared to using the LLM as-is (without merging) or with model merging. In contrast, the T5 model demonstrated an additional 4-5% improvement after applying Dataless Knowledge Fusion.

✻ **What are the details of the train-test validation split and other hyperparameters used for replicating the experiments?**

➠ The dataset was divided into an 80-20 train-test split, where 80% of the data was used for training and 20% for testing. To assess the model's performance, we employed 5-fold cross-validation.The train-test split was meticulously crafted to ensure that each sentence and its augmented versions are exclusively present in either the train set or the test set, but never in both. This careful arrangement guarantees the absence of any sentence overlap (i.e. sentence "S" present in train split and paraphrased version of sentence "S" present in test spilt), maintaining the integrity of the data and enhancing the overall quality of the split. The train-test split of the dataset would be made available along with all the hyperparameters of the code on GitHub for replication of the results.

Appendix

This section provides supplementary material in the form of additional examples, implementation details, etc. to bolster the reader's understanding of the concepts presented in this work.

A Lies of omission – across cultures

Instances of lies of omission can be discovered in ancient literature from diverse cultures across the globe. In order to stimulate further discussion and provide motivation, we will present (in the appendix - due to obvious space limitation) two specific examples—one from the Western tradition and another from the Eastern tradition. These examples serve to highlight the prevalence and significance of lies of omission in literature and emphasize the need for deeper exploration of this phenomenon.

The merchant of Venice: In Shakespeare's play, Antonio, an antisemitic merchant, borrows money from the Jewish moneylender Shylock in order to assist his friend in pursuing a relationship with Portia. Antonio can't repay the loan, and without mercy, Shylock demands a pound of his flesh as collateral. At this critical moment, Portia, who is now married to Antonio's friend, disguises herself as a lawyer and intervenes to save Antonio. Though the agreement allows Shylock to claim a pound of flesh, he must ensure that not a single drop of blood is shed, as causing harm to a Christian is strictly forbidden by law.

Mahabharata - *Ashwathama hatho, naro va kunjaro va*: This story is derived from an ancient Indian epic *"The Mahabharta"*. In this excerpt, *Ashwathama* is an elephant. *Ashwathama* was also the name of the son of Guru Dronacharya. Yudhishtir, one of the Pandavas and *Dharmraj* (which means he would never lie), faces the daunting task of confronting his unbeatable mentor, Guru Dronacharya, from whom he and his brothers had learned the art of warfare. Reluctant to engage in direct combat against his beloved teacher, Yudhishtir follows the advice of Lord Krishna and employs a strategy of omission. He announces the death of Ashwathama, but discreetly adds the words "naro va kunjaro va," indicating that it is actually a question whether the deceased Ashwathama is a human or an elephant. While Yudhishtir technically did not prevaricate, the news of his son's supposed demise deeply affects Guru Dronacharya, causing him to lose his will to fight and making it easier for Yudhishtir to overcome him. The story highlights Yudhishtir's adherence to his principles of truthfulness while employing a clever tactic of omission to gain an advantage in the battle.

B Dataset Curation

This contains additional information on data sources, data cleaning, annotation, and Inter annotator agreement

B.1 Data Sources

The dataset contains two types of articles fake and real news. This dataset was collected from real-world sources; the truthful articles were obtained by crawling articles from Reuters.com (News website). As for the fake news articles, they were collected from different sources. The fake news articles were collected from unreliable websites that were flagged by Politifact (a fact-checking organization in the USA) and Wikipedia. For this research, the fake news dataset is leveraged. The data source has a file named "Fake.csv" which contains more than 12,600 articles from different fake news outlet resources. Each article contains the following information: article title, text, type, and the date the article was published on. We chose 2500 data points randomly from this set for this research.

B.2 Data Cleaning and Annotation Quality check

Data cleaning involves two iterations, data set preparation, and a human-level review of the manual annotations. The process involved the removal of URLs and unnecessary internet taxonomy with the aim to increase data quality. To further increase the quality of data for human understanding, we reviewed the annotations manually by following the below-mentioned steps:

- Accounting for multiple annotations against a single field by the same annotator by getting rid of one of the two annotations along the lines of the definitions formulated at the start of the process.
- Filling in for fields annotated by the first entity and missed by the second entity by accounting for the gaps by building along the lines of definitions established earlier. Correcting typographical errors implicating a similar meaning.
- Overriding annotations for a couple of data items where the reviewer found them overwhelmingly wrong.

B.3 Inter Annotator Agreement

In the section 3.3 we have reported inter-annotator scores for all the 3 layers in table 1. In addition, here we are reporting inter-annotator agreement for the topic of lie in the appendix B.3.

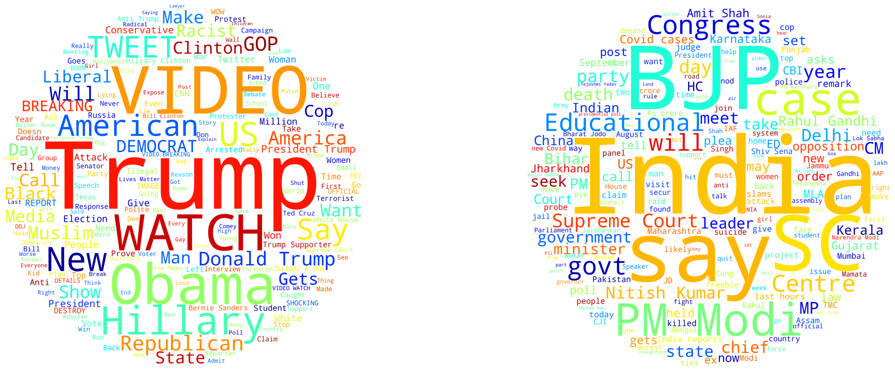
	Political	Educational	Religious	Ethnicity	Racial	Others
Twitter	0.82	0.78	0.81	0.73	0.76	0.72
Fake News	0.87	0.84	0.85	0.77	0.82	0.79

Table 4: Inter Annotator Agreement score for Topic of Lies.

B.4 Data Analysis of SEPSIS Corpus and Insights

This section contains a thorough analysis of the entire corpus.

Word representation of the sepsis corpus: We have utilized two different data sources to understand the frequency of words, we present the word clouds in fig 6a and fig 6b. An interesting insight is figure 6a represents US news and figure 6b represents the Indian media house.



(a) Word cloud of data collected from ISOT fake news.

(b) Word cloud of data collected from Times of India.

Statistics on categories across entire corpus: We further present the percentage of each feature across the entire dataset as represented in table 5.

Layers of Deception	**Categories within the layer**	**Number of datapoints**	**Percentage**
Layer 1: **Type of Omission**	Speculation	311754	35.56%
	Bias	72268	8.24%
	Distortion	150249	17.14%
	Opinion	154590	17.63%
	Sounds Factual	187923	21.43%
Layer 2: **Colors of Lies**	Black	322634	45.31%
	White	90019	12.64%
	Gray	182161	25.58%
	Red	117245	16.47%
Layer 3: **Intent of Lies**	Gaining Advantage	332661	47.73%
	Protecting Themselves	202395	29.04%
	Gaining Esteem	124197	17.96%
	Avoiding Embarrasment	24505	3.52%
	Defaming Esteem	6938	1.00%
	Protecting Others	5236	0.75%
Layer 4: **Topic of Lies**	Political	546780	72.36%
	Educational	109596	14.50%
	Ethnicity	29343	3.88%
	Religious	27575	3.64%
	Racial	27354	3.61%
	Others	15250	2.01%

Table 5: Breakup of SEPSIS datapoints over layers of deception and categories within each layer.

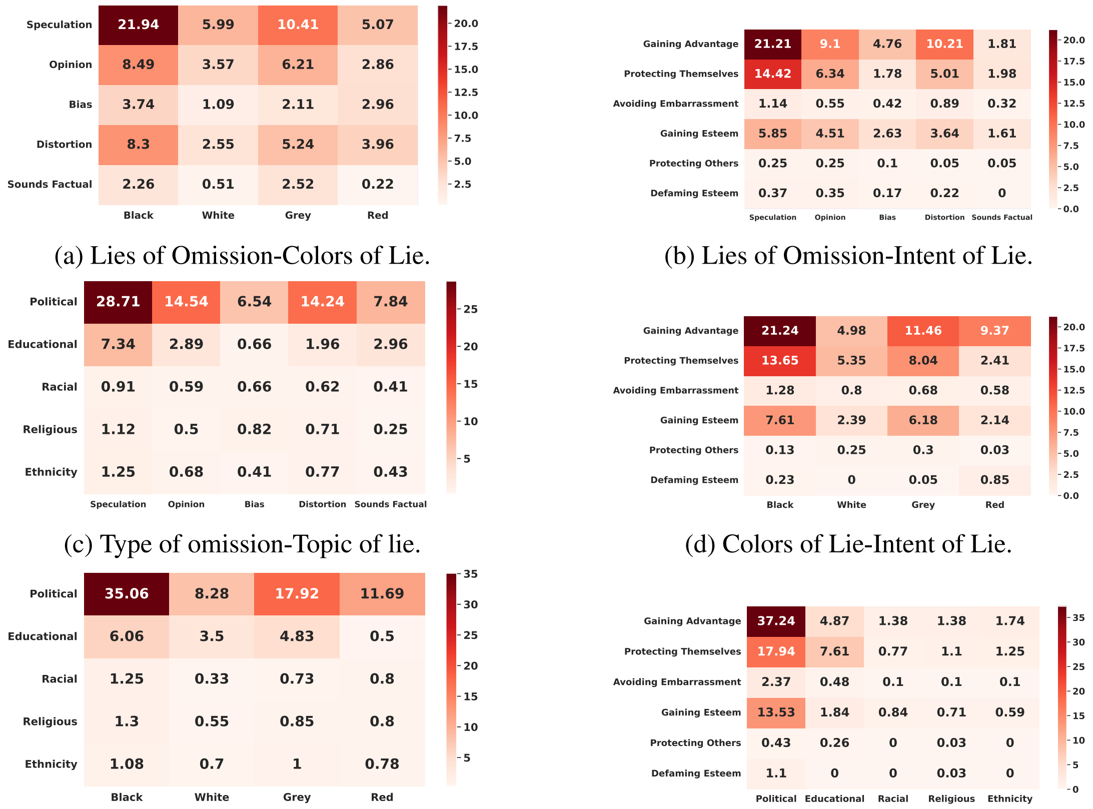
Percentage presence of 5Ws across all datapoints: Since we utilize 5W-based mask infilling, we also present % of 5Ws across the entire dataset. and the statistics around it can be found in the table 6 below.

	Who	**What**	**Why**	**When**	**Where**
% presence of 5W for tweets from Times of India	34.84%	53.06%	1.02%	6.31%	4.77%
% presence of 5W from ISOT fake news dataset	36.40%	52.73%	1.41%	6.30%	3.16%

Table 6: % of 5Ws across the entire dataset.

Co-occurence percentage: The four layers are connected to the input sentence. To study the co-occurrence across all categories and layers, we present them in heatmaps as described in fig 7.

When analyzing lies of omission and colors of lies, we observe a strong correlation between speculation and black lies. Additionally, a significant majority of speculative texts can be categorized as political in nature. This association becomes even more apparent when we delve into the Intent of Lie on Lies of Omission. It is evident that the primary objective behind the creation of speculative texts is to gain an advantage. Black lies, in particular, are frequently employed for this purpose. It is noteworthy that political texts predominantly consist of black lies, serving as a means to gain an advantage.



(a) Lies of Omission-Colors of Lie.

(b) Lies of Omission-Intent of Lie.

(c) Type of omission-Topic of lie.

(d) Colors of Lie-Intent of Lie.

(e) Colors of Lie-Influence of Lie.

(f) Intent of Lie-Influence of Lie.

Figure 7: The heatmaps provide a concise overview of the interconnections and overlaps among various layers of Lies. Numbers represents % overlap.

C Data Augmentation

For data augmentation, we have used two techniques (i) Paraphrasing and (ii) 5W Mask Infilling. We provide additional information on these techniques in the following subsection.

C.1 Paraphrasing Deceptive Datapoints

The underlying drive for paraphrasing textual assertions stems from the need to address variations that exist in real-life written content. The same textual claim might take on several different shapes since different news publishing companies use a variety of writing techniques. It is essential to create a solid standard for a thorough examination by taking these variations into account (example in Figure 8).

To generate multiple paraphrases for a given claim, we employ state-of-the-art (SoTA) models. When selecting the appropriate paraphrase model from a list of available options, our main consideration is to ensure that the generated paraphrases exhibit both linguistic correctness and rich diversity. The process we follow to achieve this can be outlined as follows: Let's assume we have a claim denoted as c. Using a paraphrasing model, we generate n paraphrases, resulting in a set of paraphrases p_1^c, p_2^c, ..., p_n^c. Subsequently, we conduct pairwise comparisons between these paraphrases and the original claim c, giving us comparisons such as $c - p_1^c$, $c - p_2^c$, ..., $c - p_n^c$. At this stage, we identify the examples that

Sasan Goodarzi, the CEO of software giant Intuit, which has avoided mass layoffs, says tech firms axed jobs because they misread the pandemic.
Prphr 1: Sasan Goodarzi, the CEO of Intuit, a software giant that refrained from massive layoffs, explains that tech companies terminated employees due to their misinterpretation of the pandemic.
Prphr 2: Intuit's CEO, Sasan Goodarzi, highlights that unlike other tech firms, the software giant avoided extensive job cuts as they correctly understood the impact of the pandemic.
Prphr 3: The pandemic was misinterpreted by tech companies, leading them to lay off employees, according to Sasan Goodarzi, CEO of Intuit, a software giant that took a different approach and did not resort to mass layoffs.
Prphr 4: Sasan Goodarzi, the CEO of Intuit, a software giant, asserts that tech companies made a mistake by laying off staff members because they failed to comprehend the true nature of the pandemic.
Prphr 5: In contrast to tech firms that made the wrong call and downsized their workforce, Intuit, led by CEO Sasan Goodarzi, correctly assessed the pandemic and refrained from mass layoffs.

Figure 8: Deceptive paraphrased data obtained using `text-davinci-003` (Brown et al., 2020).

exhibit entailment, selecting only those for further consideration. To determine entailment, we utilize RoBERTa Large (Liu et al., 2019), a state-of-the-art model trained on the SNLI task (Bowman et al., 2015).

However, it is important to consider various secondary factors when evaluating paraphrase models. For instance, one model may generate a limited number of paraphrase variations compared to others, but those variations might be more accurate and consistent. Therefore, we took into account three key dimensions in our evaluation: *(i) the number of meaningful paraphrase generations, (ii) the correctness of those generations, and (iii) the linguistic diversity exhibited by the generated paraphrases*. In our experiments, we explored the capabilities of three available models: (a) Pegasus (Zhang et al., 2020), (b) T5 (T5-Large) (Raffel et al., 2020), and (c) GPT-3 (specifically, the `text-davinci-003` variant) (Brown et al., 2020). Based on empirical observations and analysis, we found that GPT-3 consistently outperformed the other models. To ensure transparency regarding our experimental process, we provide a detailed description of the aforementioned evaluation dimensions as follows.

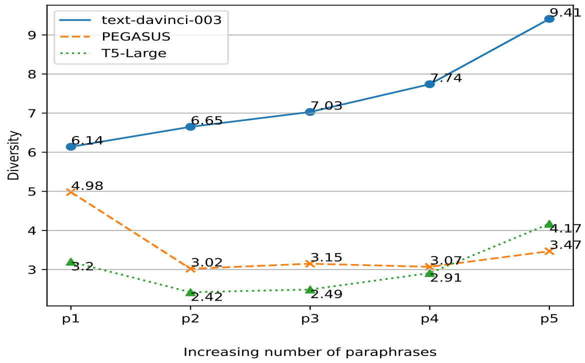


Figure 9: A higher diversity score depicts an increase in the number of generated paraphrases and linguistic variations in those generated paraphrases.

Model	Coverage	Correctness	Diversity
Pegasus	31.98	93.23%	3.53
T5	30.09	84.56%	3.04
GPT-3	35.19	89.67%	7.39

Table 7: Experimental results of automatic paraphrasing models based on three factors: *(i) coverage, (ii) correctness and (iii) diversity*; GPT-3 (`text-davinci-003`) can be seen as the most performant.

Coverage - Generating a substantial number of paraphrases: Our objective is to generate up to five paraphrases for each given claim. After generating the paraphrases, we employ the concept of minimum edit distance (MED) (Wagner and Fischer, 1974) to assess the similarity between the paraphrase candidates and the original claim (with word-level units instead of individual characters). If the MED exceeds a threshold of ±2 for a particular paraphrase candidate (e.g., $c - p_1^c$), we consider it as a viable paraphrase and retain it for further evaluation. However, if the MED is within the threshold, we discard that particular paraphrase. By employing this setup, we evaluated all three models to determine which one generates the highest number of meaningful paraphrases.

Correctness - Ensuring correctness in the generated paraphrases: Following the initial filtration step, we conducted pairwise entailment assessments using the RoBERTa Large model (Liu et al., 2019), which is a state-of-the-art model trained on the SNLI dataset (Bowman et al., 2015). We retained only those paraphrase candidates that were identified as entailed by the RoBERTa Large model.

Diversity - Ensuring linguistic diversity in the generated paraphrases: Our focus was to select a model that could produce paraphrases with greater linguistic diversity. To assess the dissimilarities between the generated paraphrase claims, we compared pairs such as $c - p_n^c$, $p_1^c - p_n^c$, $p_2^c - p_n^c$, ..., $p_{n-1}^c - p_n^c$ for each paraphrase. We repeated this process for all other paraphrases and calculated the average dissimilarity score. Since there is no specific metric to measure dissimilarity, we utilized the inverse of the BLEU score (Papineni et al., 2002). This allowed us to gauge the linguistic diversity exhibited by a given model. Based on these experiments, we observed that the `text-davinci-003` variant performed the best in terms of linguistic diversity. The results of the experiment are presented in the table below. Moreover, we prioritized the selection of a model that maximized linguistic variations, and `text-davinci-003` excelled in this regard as well. The diversity vs. chosen models plot is illustrated in Figure 9.

C.2 Data Augmentation using 5W Mask Infilling

This mapping describes how Propbank roles are mapped to 5Ws(Who, What, When, Where, Why). We have used this mapping for mask infilling.

PropBank Role	**Who**	**What**	**When**	**Where**	**Why**	**How**
ARG0	**84.48**	0.00	3.33	0.00	0.00	0.00
ARG1	10.34	**53.85**	0.00	0.00	0.00	0.00
ARG2	0.00	9.89	0.00	0.00	0.00	0.00
ARG3	0.00	0.00	0.00	22.86	0.00	0.00
ARG4	0.00	3.29	0.00	34.29	0.00	0.00
ARGM-TMP	0.00	1.09	**60.00**	0.00	0.00	0.00
ARGM-LOC	0.00	1.09	10.00	**25.71**	0.00	0.00
ARGM-CAU	0.00	0.00	0.00	0.00	**100.00**	0.00
ARGM-ADV	0.00	4.39	20.00	0.00	0.00	0.06
ARGM-MNR	0.00	3.85	0.00	8.57	0.00	**90.91**
ARGM-MOD	0.00	4.39	0.00	0.00	0.00	0.00
ARGM-DIR	0.00	0.01	0.00	5.71	0.00	3.03
ARGM-DIS	0.00	1.65	0.00	0.00	0.00	0.00
ARGM-NEG	0.00	1.09	0.00	0.00	0.00	0.00

Table 8: A mapping table from PropBank (Palmer et al., 2005) (*Arg0, Arg1, ...*) to 5W (*Who, What, When, Where, and Why*).

D Multi-Task Learning

In this section, we delve into the specific architectural choices, experimental setup, and the formulation of the loss function employed for multi-task learning frameworks: The SEPSIS Classifier. By exploring the intricacies of this approach, we aim to shed light on the systematic integration of multiple tasks into a unified learning framework, ultimately enabling the model to effectively leverage synergistic information across layers of Deception.

D.1 Architectural Discussion

Multi-task learning (MTL) has emerged as a powerful paradigm for training deep neural networks to perform multiple related tasks simultaneously. In this paper, we propose a multi-task learning-based architecture for predicting four different tasks of the Deception dataset. The main advantage of using multi-task learning is the ability to leverage shared information across tasks, leading to improved model generalization and increased efficiency in training and inference. By jointly training multiple tasks, the model learns useful representations that are transferable to other related tasks, leading to better overall performance (Caruana, 1997).

D.1.1 Dataless Knowledge Fusion

In many cases, LLMs are trained using domain-specific datasets, which can limit their performance when applied to out-of-domain cases. To address this challenge, we employ a fine-tuning approach on the T5-base model for each specific task, resulting in a total of four finetuned T5-based models (one model corresponding to one task). To leverage these models in our Multitask learning architecture, we employ Dataless Knowledge Fusion (Jin et al., 2022) on these four finetuned T5-models into a single, more generalized model that exhibits improved performance in multitask learning (from here referred *merged-fine-tuned-T5*).

D.1.2 Methodology

Our methodology takes a sentence as input and converts it into a latent embedding. The process of creating this rich embedding involves a two-stage approach. Firstly, we leverage the model-merging technique (Jin et al., 2022), which merges fine-tuned models sharing the same architecture and pre-trained weights, resulting in enhanced performance and improved generalization capabilities, particularly when dealing with out-of-domain data (Jin et al., 2022). Once the word embeddings are obtained from this merged model, the second stage involves converting them into a latent representation using the transformer encoder module. This representation is then propagated through four task-specific multilabel heads to obtain the output labels for each of the layers of Deception.

D.2 Loss Functions

This section contains an in-depth discussion of different loss functions that we used for different tasks of MTL architecture.

D.2.1 Cross-Entropy Loss

Cross entropy loss, also known as log loss or logistic loss, is a commonly used loss function in machine learning, particularly in classification tasks. It measures the dissimilarity between the predicted probabilities of classes and the true labels of the data. The log loss function penalizes incorrect predictions more strongly, meaning that as the predicted probability deviates further from the true label, the loss increases. The loss approaches zero when the predicted probability aligns with the true label.

For the SEPSIS classifier, i.e., multi-label classification task with n classes, the cross-entropy loss is calculated as the average of the individual binary cross-entropy losses for each class.

$$L_{BCE} = \begin{cases} -\log\left(p_i^k\right) & \text{if } y_i^k = 1 \\ -\log\left(1 - p_i^k\right) & \text{otherwise} \end{cases} \tag{1}$$

where,

- $y^k = \left[y_1^k, \ldots, y_C^k\right] \in \{0,1\}^C (C$ is the number of classes),
- p_i^k is the predicted probability distribution across the classes

D.2.2 Focal Loss

Focal loss is a modification of the cross entropy loss that addresses the issue of class imbalance in multi-class classification tasks (Lin et al., 2017). In the standard multi-class cross-entropy loss, all classes are treated equally, which can be problematic when dealing with imbalanced datasets where certain classes have a much smaller representation. Focal loss aims to down-weight the contribution of well-classified examples and focuses more on difficult and misclassified examples. The focal loss for multi-label classification is defined as follows:

$$L_{FL} = \begin{cases} -\left(1 - p_i^k\right)^\gamma \log\left(p_i^k\right) & \text{if } y_i^k = 1 \\ -\left(p_i^k\right)^\gamma \log\left(1 - p_i^k\right) & \text{otherwise} \end{cases} \tag{2}$$

where:

- p_i^k is the predicted probability distribution across the classes
- γ is the focusing parameter that controls the degree of down weighting. It is usually set to a value greater than 0. We used $\gamma = 2$ in our experiment.

The focal loss formula introduces the term $(1 - p_i)^\gamma$ which acts as a modulating factor. This factor down weights well-classified examples p_i^k close to 1 and assigns them a lower contribution to the loss. The focusing parameter gamma controls how much the loss is down-weighted. Higher values of gamma place more emphasis on difficult examples. By incorporating the focal loss into the training objective, the model can effectively handle class imbalance and focus more on challenging examples.

D.2.3 Dice Loss

The Dice loss is a similarity-based loss function commonly used in image segmentation tasks and data-imbalanced multi-class classification problems. It measures the overlap or similarity between predicted and true labels. For multi-label classification, the Dice loss can be defined as follows:

$$L_{DL} = 1 - \frac{2\sum_{i=1}^{C} y_i^k \cdot p_i^k + \varepsilon}{\sum_{i=1}^{C} y_i^k + \sum_{i=1}^{C} p_i^k + \varepsilon} \tag{3}$$

- C is the number of classes

- y_i^k represents the true label for class C, which can be either 0 or 1 for each label.
- p_i^k represents the predicted probability or output for class c

The formula calculates the Dice coefficient for each example by summing the products of the true labels y_i^k and predicted probabilities p_i^k for each class C. The numerator represents the intersection between the predicted and true labels, while the denominator represents the sum of the predicted and true labels, which corresponds to the union of the two sets. By subtracting the Dice coefficient from 1, we obtain the Dice loss.

By using the Dice loss, the model is encouraged to focus on correctly identifying and predicting the minority classes, as the loss is computed based on the intersection and sum of true and predicted labels for each class. This property is especially valuable in data-imbalanced settings, as it helps to alleviate the bias towards majority classes and improve the model's ability to capture and predict the minority classes accurately.

D.2.4 Distribution-balanced Loss

The distribution-balanced (DB) loss function is a promising solution for addressing class imbalance and label dependency in multilabel text classification tasks. Unlike traditional approaches such as resampling and re-weighting, which often lead to oversampling common labels, the DB loss function tackles these challenges directly. By inherently considering the class distribution and label linkage, it offers a more effective alternative for achieving balanced training.

According to (Huang et al., 2021a), the application of the DB loss function has demonstrated superior performance compared to commonly used loss functions in multi-label scenarios. This novel approach addresses the problem of class imbalance, where certain labels are significantly underrepresented, and considers the relationship and dependencies between different labels. By striking a balance between these factors, the DB loss function ensures that the training process is fair and unbiased, resulting in improved accuracy and robustness in multilabel text classification tasks.

For multi-label classification, the Distribution-balanced loss can be defined as follows:

$$L_{DB} = \begin{cases} -\hat{r}_{DB}\left(1-q_i^k\right)^{\gamma}\log\left(q_i^k\right) & \text{if } y_i^k = 1 \\ -\hat{r}_{DB}\frac{1}{\lambda}\left(q_i^k\right)^{\gamma}\log\left(1-q_i^k\right) & \text{otherwise} \end{cases} \tag{4}$$

where:

- C is the number of classes
- $\hat{r}_{DB} = \alpha + \sigma\left(\beta \times \left(r_{DB} - \mu\right)\right) \rightarrow r_{DB} = \frac{\frac{1}{C}\frac{1}{n_i}}{\frac{1}{C}\sum_{y_i^k=1}\frac{1}{n_i}}$
- y_i represents the true label
- λ scale factor

The distribution-balanced loss combines rebalanced weighting and negative-tolerant regularization (NTR) to address key challenges in multi-label scenarios. It effectively reduces redundant information arising from label co-occurrence, which is crucial in such tasks. Additionally, the loss explicitly assigns lower weights to negative instances that are considered "easy-to-classify," thereby improving the model's ability to handle these instances effectively. (Wu et al., 2020)

D.2.5 Rationale for choosing loss function for the particular task.

The selection of specific loss functions for each task is driven by various factors and considerations.

1. **Distribution-balanced loss function for Types of Omission:** Due to the strong multi-label nature and skewed distribution of the Types of Omission layer, the Distribution-balanced loss function is utilized (Huang et al., 2021b). This loss function is specifically designed to handle extreme multi-label scenarios and skewed class distributions, providing a more balanced and effective training process for the model.

2. **Cross Entropy loss for Color of Lie**: The Color of Lie layer is relatively class-wise balanced. In such cases, the Cross-Entropy loss is a commonly used and standard loss function. It is well-suited for balanced class distributions and helps the model effectively learn and classify the color of lies.

3. **Focal loss for Intent of Lie:** The Intent of Lie layer is a class-imbalanced scenario. In such situations, the Focal loss has shown to perform well. Focal loss down-weights easy examples and focuses more on hard, misclassified examples, which helps in addressing class imbalance and improving the model's performance on classification of minority classes.

4. **Dice loss for Topic of Lie:** The Topic of Lie layer is also a class-imbalanced scenario. The Dice loss has demonstrated effectiveness in handling class imbalance. Hence we used the Dice loss for this layer so that, the model can better capture and predict the minority topics.

   The rationale behind selecting focal loss for the Intent of lie and Dice loss for the topic of lie is based on experimentation. Initially, we tried the opposite combination, which resulted in an F1 score of 0.85 for the Intent of lie and a score of 0.85 for the topic of lie. However, in the current configuration, we achieved improved performance with an F1 score of 0.87 for the Intent of lie and a score of 0.86 for the topic of lie. Therefore, after careful evaluation, we opted for focal loss and Dice loss for their respective categories to maximize overall performance.

D.3 Experimental results

For overall experiments, we had 4 setups broadly.

- T5 with LSTM encoder combined with no model merging
- T5 with LSTM encoder combined with model merging
- T5 with transformer encoder combined with no model merging
- T5 with transformer encoder combined with model merging

We used accuracy, precision, recall, and F1 score for evaluating the performance of our model. T5 with transformer encoder combined with model merging performed the best and results on these metrics for all experiments are presented in table 9.

	SEPSIS	Labels	Without Model Merging								With Model Merging							
			Accuracy %		Precision		Recall		F1-Score		Accuracy %		Precision		Recall		F1-Score	
T5 with LSTM encoder	Type of Omission	Speculation	82.58	80.25	0.78	0.77	0.83	0.80	0.8	0.78	86.15	82.89	0.84	0.81	0.85	0.83	0.84	0.82
		Opinion	80.76		0.80		0.79		0.79		82.54		0.82		0.81		0.81	
		Bais	74.92		0.73		0.76		0.74		77.39		0.75		0.80		0.77	
		Distortion	79.51		0.75		0.78		0.76		81.87		0.8		0.82		0.81	
		Sound Factual	83.50		0.79		0.83		0.81		86.48		0.83		0.86		0.84	
	Color of Lie	White	85.68	86.37	0.83	0.84	0.86	0.84	0.84	0.84	88.95	88.84	0.86	0.87	0.88	0.88	0.87	0.87
		Grey	84.50		0.87		0.83		0.85		86.38		0.89		0.85		0.87	
		Red	86.87		0.84		0.83		0.83		88.20		0.87		0.89		0.88	
		Black	88.43		0.82		0.85		0.83		91.83		0.87		0.90		0.88	
	Intent of lie	Gaining Advantage	87.62	83.69	0.85	0.84	0.83	0.79	0.84	0.81	91.08	86.12	0.87	0.84	0.89	0.85	0.88	0.84
		Protecting Themselves	84.87		0.86		0.81		0.83		88.23		0.84		0.88		0.86	
		Gaining Esteem	82.97		0.82		0.77		0.79		84.49		0.85		0.83		0.84	
		Avoiding Embarrassment	80.91		0.84		0.79		0.81		82.97		0.83		0.80		0.81	
		Defaming Esteem	82.06		0.83		0.75		0.79		83.87		0.81		0.84		0.82	
		Protecting others	80.11		0.75		0.79		0.77		82.11		0.79		0.81		0.8	
	Topic of Lies	Political	88.70	83.60	0.82	0.81	0.86	0.82	0.84	0.81	91.88	86.13	0.86	0.83	0.88	0.84	0.87	0.83
		Educational	83.98		0.84		0.81		0.82		86.79		0.85		0.86		0.85	
		Regilious	84.18		0.81		0.85		0.83		84.98		0.85		0.83		0.84	
		Ethnicity	79.29		0.83		0.75		0.79		83.84		0.81		0.82		0.81	
		Racial	81.85		0.77		0.82		0.79		83.16		0.80		0.79		0.79	
		Other	76.95		0.72		0.77		0.74		81.90		0.76		0.79		0.77	
T5 with Transformer Encoder	Type of Omission	Speculation	85.67	82.22	0.83	0.81	0.81	0.79	0.82	0.80	**89.91**	**86.30**	**0.86**	**0.84**	**0.88**	**0.86**	**0.87**	**0.84**
		Opinion	83.40		0.80		0.82		0.81		**87.09**		**0.84**		**0.83**		**0.83**	
		Bais	76.30		0.77		0.75		0.76		**80.49**		**0.79**		**0.83**		**0.81**	
		Distortion	80.44		0.81		0.79		0.8		**85.77**		**0.83**		**0.85**		**0.84**	
		Sound Factual	85.32		0.84		0.80		0.82		**88.23**		**0.86**		**0.89**		**0.87**	
	Color of Lie	White	87.36	89.11	0.88	0.88	0.86	0.85	0.87	0.86	**91.23**	**93.84**	**0.90**	**0.92**	**0.89**	**0.91**	**0.90**	**0.92**
		Grey	89.05		0.88		0.84		0.86		**94.53**		**0.92**		**0.88**		**0.90**	
		Red	88.41		0.86		0.85		0.85		**93.45**		**0.91**		**0.92**		**0.92**	
		Black	91.62		0.89		0.85		0.87		**96.17**		**0.94**		**0.93**		**0.94**	
	Intent of lie	Gaining Advantage	89.35	86.09	0.88	0.85	0.86	0.84	0.87	0.84	**92.54**	**88.49**	**0.91**	**0.87**	**0.93**	**0.88**	**0.92**	**0.87**
		Protecting Themselves	88.74		0.86		0.85		0.85		**90.78**		**0.89**		**0.90**		**0.89**	
		Gaining Esteem	85.67		0.85		0.82		0.83		**88.56**		**0.88**		**0.86**		**0.87**	
		Avoiding Embarrassment	83.25		0.82		0.83		0.82		**87.19**		**0.85**		**0.88**		**0.86**	
		Defaming Esteem	83.46		0.83		0.82		0.82		**86.88**		**0.85**		**0.84**		**0.84**	
		Protecting others	81.16		0.80		0.79		0.79		**85.04**		**0.83**		**0.84**		**0.83**	
	Topic of Lies	Political	90.59	85.87	0.88	0.85	0.86	0.85	0.87	0.85	**94.16**	**88.26**	**0.93**	**0.87**	**0.90**	**0.86**	**0.91**	**0.86**
		Educational	86.77		0.87		0.88		0.87		**90.66**		**0.90**		**0.87**		**0.88**	
		Regilious	85.46		0.84		0.84		0.84		**87.83**		**0.87**		**0.85**		**0.86**	
		Ethnicity	84.69		0.84		0.85		0.84		**88.67**		**0.86**		**0.87**		**0.86**	
		Racial	81.84		0.83		0.82		0.82		**85.89**		**0.87**		**0.84**		**0.85**	
		Other	79.18		0.78		0.78		0.78		**82.34**		**0.84**		**0.81**		**0.82**	

Table 9: Experiment results: The table showcases the results obtained from different experiments using varying encoder architectures, namely LSTM and Transformer. The term "Without Model Merging" refers to the utilization of the T5-3b model without any fine-tuning. Conversely, the term "With Model Merging" signifies the fine-tuning of four T5 models, each corresponding to a distinct layer, followed by Dataless Knowledge fusion. (Jin et al., 2022)

E Propaganda Techniques

Propaganda techniques are strategies used to manipulate and influence people's opinions, emotions, and behavior in order to promote a particular agenda or ideology (Da San Martino et al., 2019; Martino et al., 2020). These techniques are often employed in mass media, advertising, politics, and public relations. While they can vary in their specific methods, we present definitions of 18 propaganda techniques that we have used in this study in the left box in the subsequent section. In the box on the right side, we present insights from propaganda techniques through deception.

Propaganda Technique Definition

- **Flag Waving:** Playing on strong national feeling (or to any group, e.g., race, gender, etc) to justify or promote an action or an idea.
- **Slogans:**A brief and striking phrase that may include labeling and stereotyping.
- **Appeal to fear - prejudices:**Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative.
- **Exaggeration-Minimization**: Either representing something in an excessive manner: making things larger, better, worse (e.g., the best of the best) or making something seem less important or smaller than it really is (e.g., saying that an insult was actually just a joke).
- **Repetition:** Repeating the same message over and over again so that the audience will eventually accept it.
- **Name Calling Labelling:** Labeling the object of the propaganda campaign as something that the target audience fears, hates, finds undesirable, or loves or praises.
- **Bandwagon:** Attempting to persuade the target audience to join in and take the course of action because "everyone else is taking the same action."
- **Loaded Language:** Using specific words and phrases with strong emotional implications (either positive or negative) to influence an audience.
- **Casual Oversimplification:** Assuming a single cause or reason when there are actually multiple causes for an issue.
- **Red herring:** Introducing irrelevant material to the issue being discussed so that everyone's attention is diverted away from the points made.
- **Appeal to authority:** Stating that a claim is true simply because a valid authority or expert on the issue said it was true.
- **Thought terminating cliches:** Words or phrases that discourage critical thought and meaningful discussion about a given topic.
- **Whataboutism:** A technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.

Propaganda through Deception

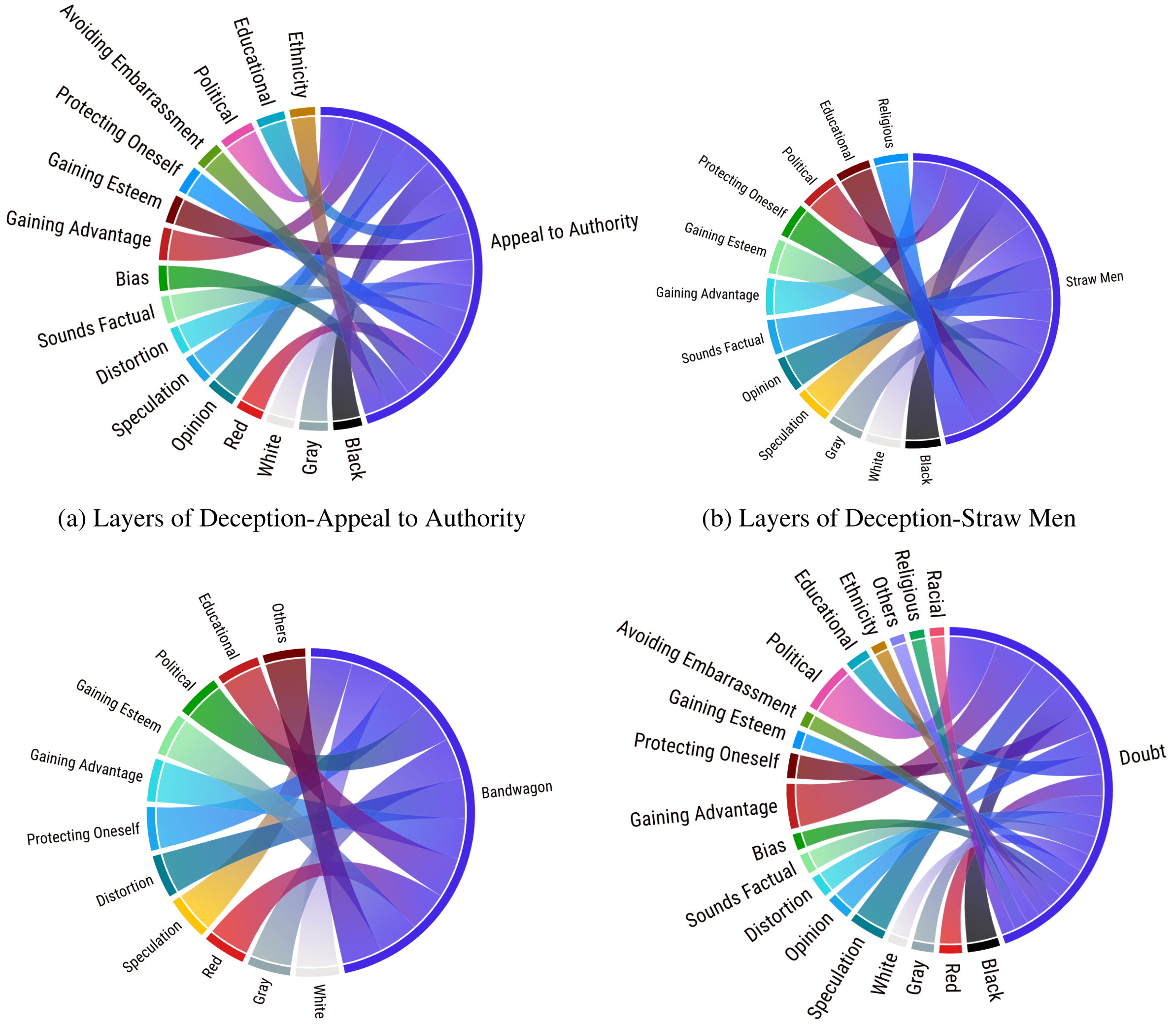
- **Flag Waving:** Flag waving maps to speculation in layer 1, black lies in layer 2, gaining advantage in layer 3, and religious aspects in layer 4.
- **Slogans:** This technique is mostly mapped with speculation in layer1, white lie in layer 2, political in layer 3 and gaining advantage in layer 4.
- **Appeal to fear - prejudices:** This technqiue primarily corresponds to speculation in layer 1, black lie in layer 2, political in layer 3 and gaining advantage in layer 4.
- **Exaggeration-Minimization**: In the Layers of Omission, Exaggeration or Minimization is mostly mapped to speculation in layer 1, black lie in layer 2, political in layer 3 and gaining advantage in layer 4.
- **Repetition:** Repetition is mostly mapped to Speculation, Black lie, intention of gaining advantage and in political influence.
- **Name Calling Labelling:** Name Calling or Labelling is largely mapped to speculation in layer 1, black lie in layer 2, gaining advantage in layer 3 and political in layer 4.
- **Bandwagon:** Bandwagon is mostly mapped to speculation in layer 1. It is mapped with both white and gray lie in layer 2. It is mapped with protecting oneself in layer 3 and education in layer 4.
- **Loaded Language:** Loaded Language is mapped mostly with speculation in layer 1, black lie in layer 2, gaining advantage in layer 3 and political in layer 4.
- **Casual Oversimplification:** Causal Oversimplification is mapped mostly with speculation in layer 1, with black lie and in some cases with red lie in layer 2, gaining advantage in layer 3 and political in layer 4.
- **Red herring:** In layer 1, Red Herring corresponds to both speculation and opinion. Layer 2 primarily associates it with black lies, occasionally with white lies. In layer 3, it largely aligns with gaining advantage, while layer 4 relates to political aspects.
- **Appeal to authority:** This technique largely maps with opinion and with speculation too. In the 2nd layer, it maps with black and gray lies and with gaining advantage in 3rd layer and political in 4th layer.
- **Thought terminating cliches:** This technique mostly maps with speculation in layer 1, gray and black lie in layer 2, gaining advantage in layer 3 and political in layer 4.
- **Whataboutism:** Whataboutism mostly maps with speculation in layer 1, black lie in layer 2, gaining advantage in layer 3 and political in layer 4.

Propaganda Technique Definition

- **Straw Men:**Substituting an opponent's proposition with a similar one, which is then refuted in place of the original proposition.
- **Doubt:**Questioning the credibility of someone or something.
- **Obfuscation:** Using words that are deliberately not clear, so that the audience may have their own interpretations.
- **Reductio ad Hitlerum:** An attempt to invalidate someone else's argument on the basis that the same idea was promoted.
- **Black and White Fallacy:**Using words that depict the fallacy of leaping from the undesirability of one proposition to the truth of an extreme opposite.

Propaganda through Deception

- **Straw Men:** Straw Men maps mostly with speculation but sometimes with opinion too. It maps with both black and white lie of layer 2 in most cases and gaining advantage in layer 3 and political in layer 4.
- **Doubt:** Doubt maps mostly with speculation in layer 1, black lie in layer 2, gaining advantage in layer 3 and political in layer 4.
- **Obfuscation:** This technique maps mostly with speculation in layer 1, red lie in layer 2, gaining advantage in layer 3 and political in layer 4.
- **Reductio ad Hitlerum:** This technique maps with speculation and distrotion in layer1, black lies and occasional white lies in layer 2. Layer 3 and layer 4 are primarily associated with gaining advantage and politics, respectively.
- **Black and White Fallacy:** This technique predominantly involves speculation and opinion, with elements of black lies in the second layer. In the third layer, it is mostly aligned with gaining advantage but occasionally tied to protecting oneself and political and educational in layer 4.

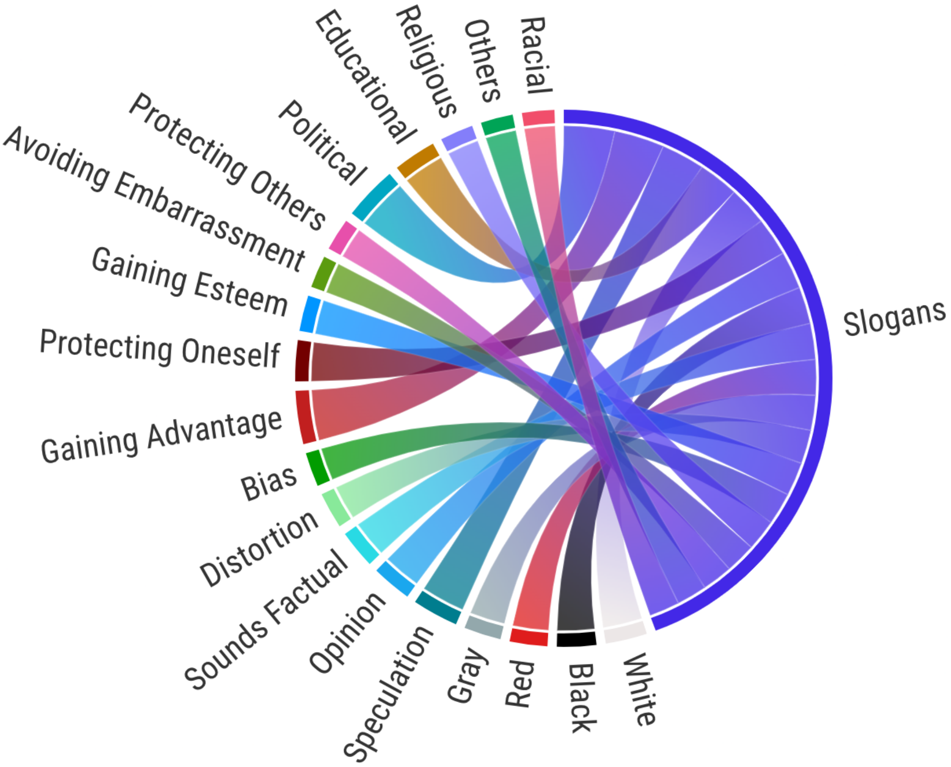


(a) Layers of Deception-Appeal to Authority

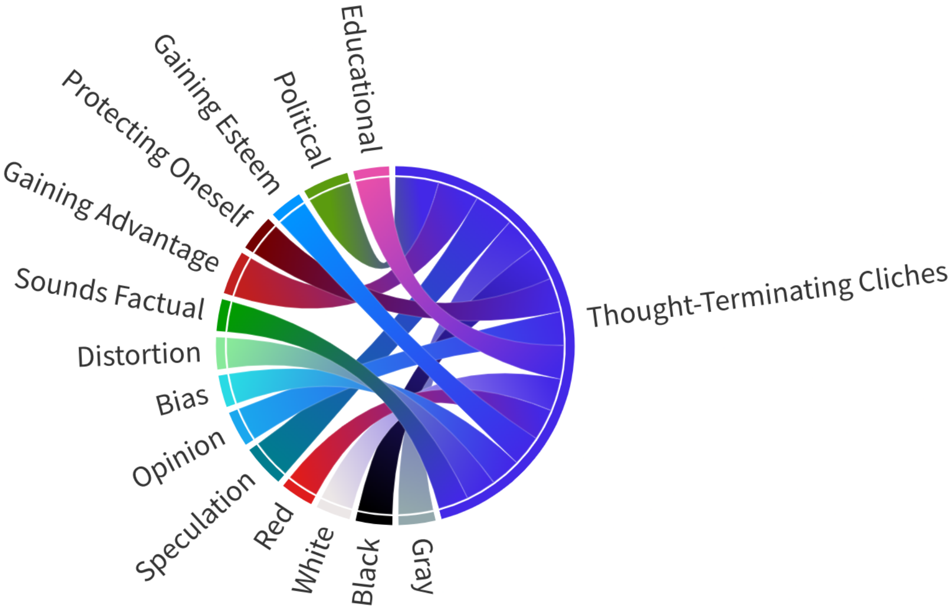
(b) Layers of Deception-Straw Men

(c) Layers of Deception-Bandwagon

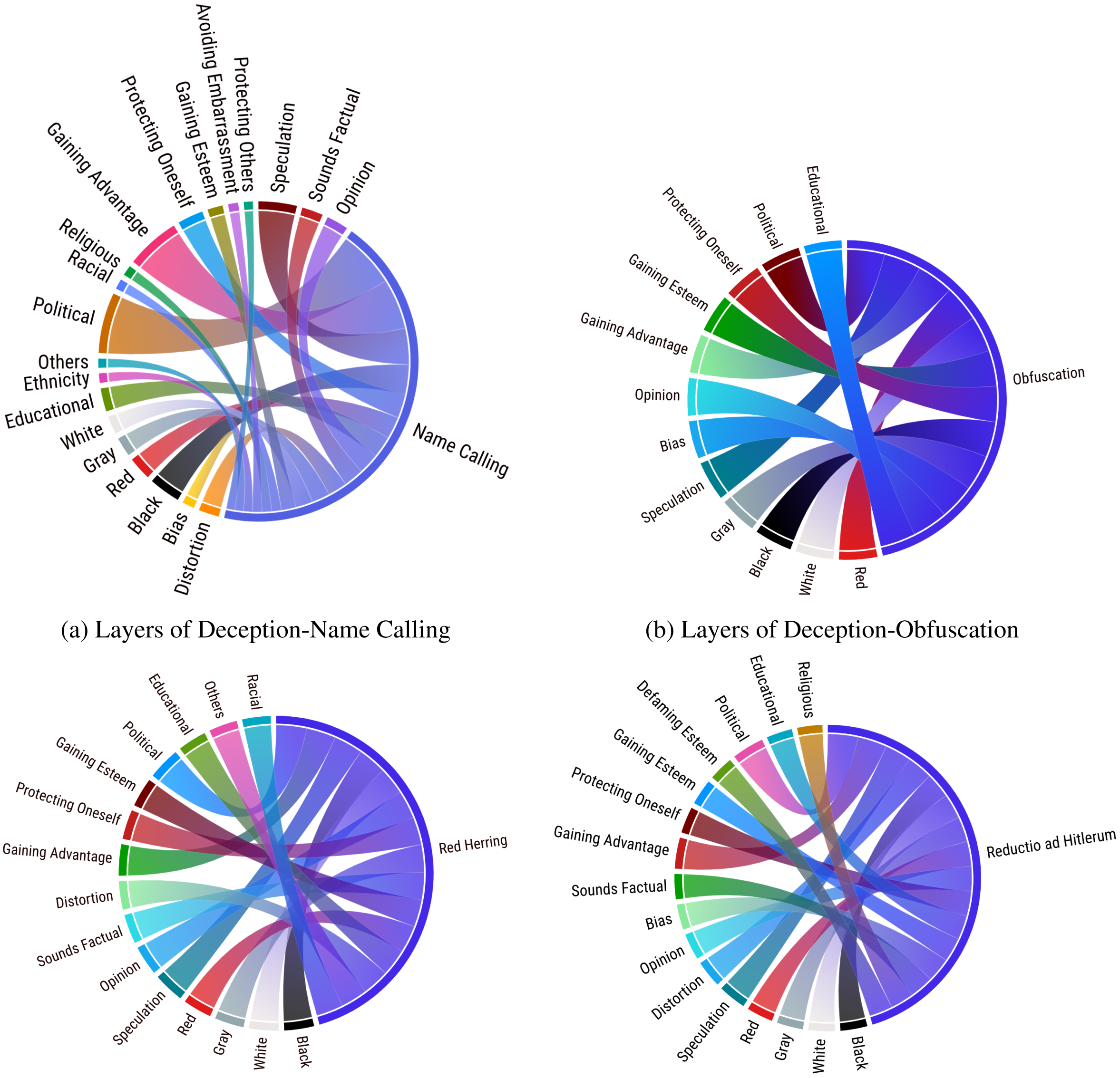
(d) Layers of Deception-Doubt



(a) Layers of Deception-Slogans



(b) Layers of Deception-Thought terminating cliches

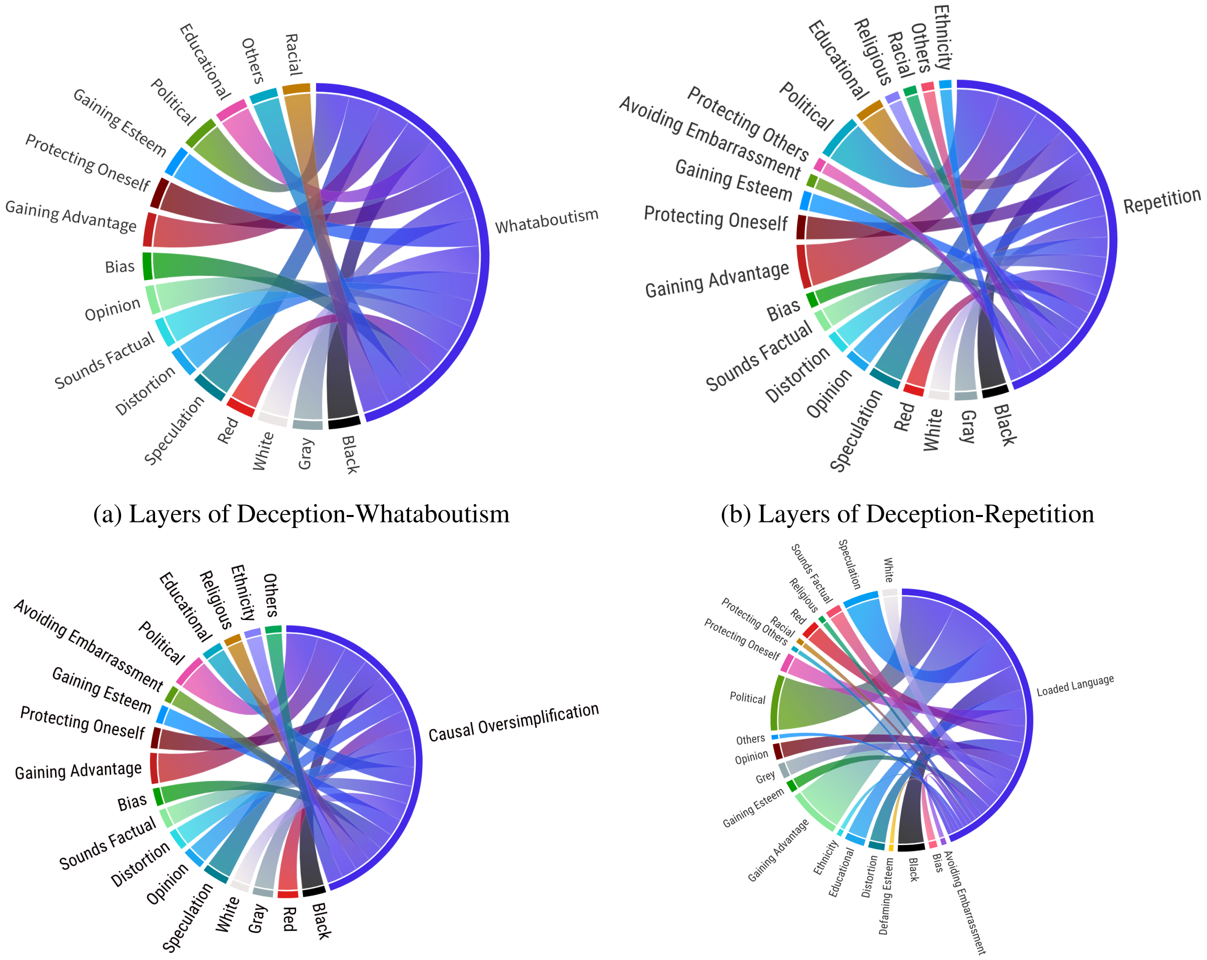


(a) Layers of Deception-Name Calling

(b) Layers of Deception-Obfuscation

(c) Layers of Deception-Red Herring

(d) Layers of Deception-Reductio ad Hitlerum

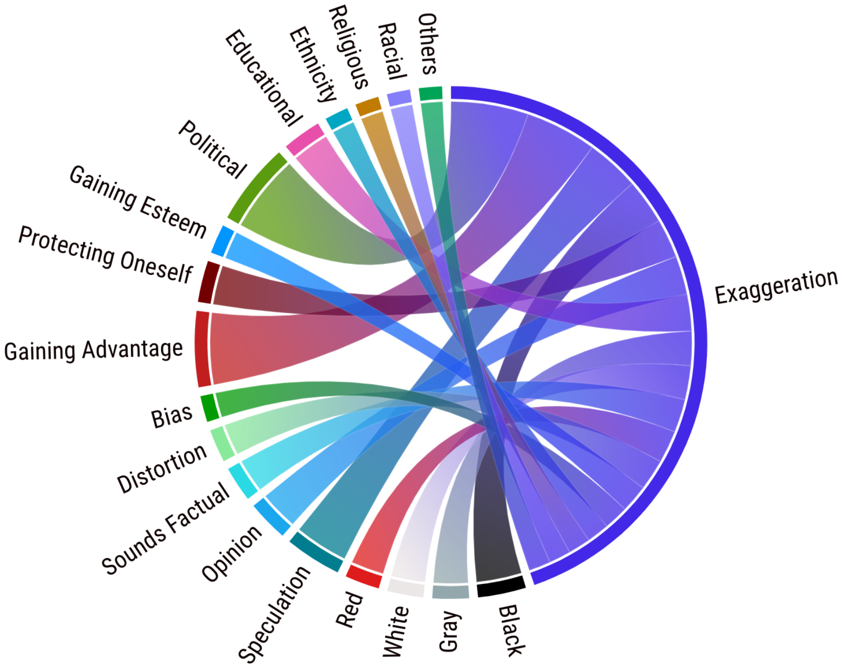


(a) Layers of Deception-Whataboutism

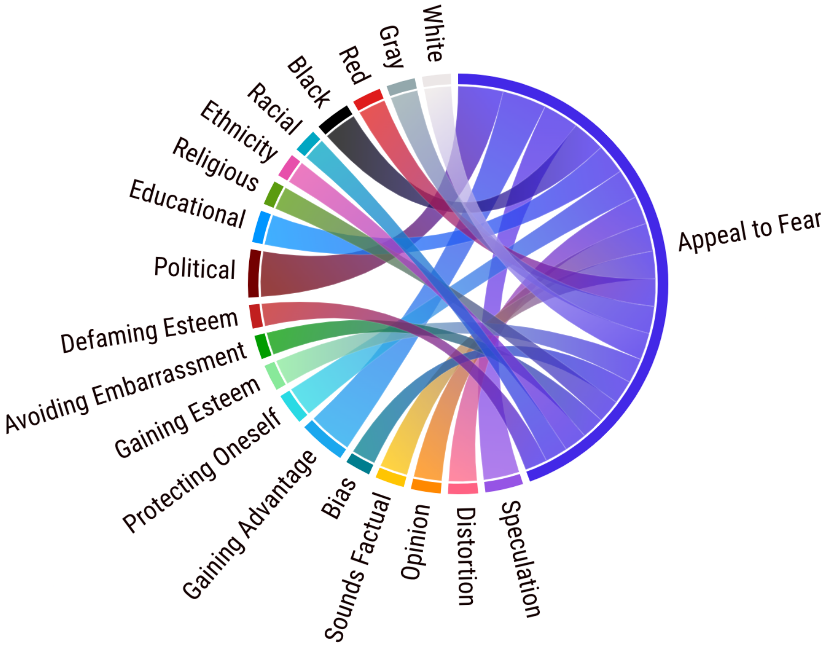
(b) Layers of Deception-Repetition

(c) Layers of Deception-Casual Oversimplification

(d) Layers of Deception-Loaded Language



(a) Layers of Deception-Exaggeration



(b) Layers of Deception-Appeal to fear