Leveraging AI to Improve Viral Information Detection in Online Discourse

Anonymous Author(s)

ABSTRACT

Information virality, a topic of increasing importance in modern media environments, often remains a blind spot in the context of information security. Our study aims to highlight why information virality is a cybersecurity concern and can be exploited to manipulate public discourse. Using Schwartz's Theory of Basic Human Values as features and a Large Language Model to generate samples synthetically, we demonstrate the use of deep learning multiple classification models for predicting potentially viral information content in online discourse. Applying the model to the Israel-Hamas conflict as a use case, we identify the values that trigger emotional arousal and increase the likelihood of information becoming viral, underscoring the applicability of our research in understanding and managing online discourse.

CCS CONCEPTS

• **Information systems** → *Social networks*; • **Security and privacy** → **Social aspects of security and privacy**; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

virality, social media, discourse, values, information environment, large language model, deep learning

ACM Reference Format:

Anonymous Author(s). 2025. Leveraging AI to Improve Viral Information Detection in Online Discourse. In *Proceedings of Annual Computer Security Applications Conference (ACSAC '24).* ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

1.1 Motivation

In the evolving landscape of information security, a critical debate has persisted about whether or not countering misinformation and disinformation falls within its purview. However, in recent years, the literature in security and computer sciences has begun to stress the need for sustained efforts to effectively combat disinformation caused by the rapid progression of artificial intelligence (AI) algorithms [36, 83]. Complex information properties pose a challenge to traditional security paradigms that prioritize focus on tangible threats. In security studies, there is a growing recognition of the need to shift from state-centric realist paradigms to more individual-centered approaches. One example is the human security paradigm, which emphasizes non-military sources of threats and the importance of addressing human insecurities [2, 4, 51]. This shift reflects the evolving nature of security threats. The Internet, a vast expanse for near real-time information exchange, harbors both organic and manipulated narratives - making the differentiation between the two increasingly complex. The proliferation of digital platforms has normalized the spread of diverse information, blurring the distinctions between genuine and artificial discourse. This normalization underscores the need for a refined understanding of security of public discourse, where misinformation and disinformation are used to manipulate perceptions and outcomes. These concerns become distended with the use of social bots [52], troll farms [29, 80], and other forms of coordinated inauthentic behavior [15] to conduct influence campaigns online.

Our primary motivation is to progress the fundamental understanding of the information environment within the framework published in *A CERN Model for Studying the Information Environment* [71]. The authors introduce a framework to describe the information environment, where human cognition, technology, and content converge [71]. CERN was developed as a "think tank" that brought together some of the world's most profound physicists. Similar to the state physics was in before CERN, we still do not have the fundamental knowledge or instrumentation to study the natural properties of the information environment. Many researchers have identified functional components for measuring the state of information after it has been witnessed in the information environment; however, predictive power is another issue. Previous studies have addressed many significant elements, such as influencers, trending topics, social capital, and network structure.

1.2 Background

Canvassing several fields of study allows further understanding the relationship between viral information and its target audience. Users' willingness to share viral content is one such avenue of exploration. Borges-Tiago et al. describe an active component within individuals that compels or entices them to participate [10]. When examining cross-cultural studies, we recognize cultural ties and intrinsic values that describe how we behave [65]. In political science literature, authors sometimes describe how politicians attempt to speak in forms of "collection action" that must be taken to improve their society [31]. Disparate research fields are often disjoint, making it challenging to combine ideas to further our understanding of societal issues.

Network analysis techniques exist for looking at group dynamics within online discourse; however, there is not much ability to follow cascades informing potential virality unless we know the specific theme we want to look at. This blind spot is a largely unaddressed vulnerability because we cannot notice viral information in its genesis. We are continually, retrospectively, trying to understand

ACSAC '24, December 09–13, 2024, Waikiki, Hawaii, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

causality without an instrument to measure what we are observing. Examining fields outside of the natural sciences, we can see that many scholars have questioned this behavior well before the Internet. We describe the connection between the target audience and the information as a type of relevancy. However, insufficient literature thoroughly describes relevance due to its seeming to be an intangible object. Anthropological and cross-cultural studies continuously have found empirical evidence to support a deeper connection between individuals that resonate beyond technical metrics [19, 53, 63, 75]. As generative AI tools become more prevalent for manipulating public discourse, it is more important than ever to build tools that allow us to follow these viral movements within the information environment.

We see that studies as early as 2004 identified the property of relevancy as a necessary component for bridging nodes in a decentralized network [47]. The study of virality has very limited ties with the term relevancy, which begs for a deeper understanding of this property is necessary for understanding information virality. We must first understand what a normal discourse profile looks like before we can detect the anomalous viral profile. One of the substantial hurdles when disseminating information in medical-related fields is cultural differences. Language difficulties and beliefs, stigma, and misinformation disrupt the ability to disseminate useful information properly [27]. We can see that the way information is framed determines how it is received.

1.3 Problem Domain

The advent of generative artificial intelligence (AI) models presents a significant challenge in public discourse, particularly due to the insufficient understanding of information virality. This knowledge gap constitutes a considerable threat, enabling malign actors to manipulate public discourse by altering values and sentiments. Leveraging generative AI, these actors can swiftly generate and disseminate information tailored to influence public opinion and discourse at scale. Recognizing and comprehending the underlying properties that govern public discourse are crucial steps toward developing effective detection and mitigation strategies against such manipulative efforts.

1.4 Research Question

The question we aim to show support for is: Does the underlying values of a targeted audience have a relationship with the potential for information to become viral?

1.5 Contribution

As part of our contribution, we delineate the various methods employed to establish our success criteria and ascertain the relationship between viral information and values within social media discourse surrounding the Israel-Hamas conflict that began on October 7th, 2023. The proliferation of discussions on social media concerning Palestine and the Gaza Strip has been notable, with contributions stemming from news sources and influencers. Chin-Rothmann highlights a significant increase in the dissemination of misinformation by fake media outlets and bot operations, particularly narratives concerning Hamas and Israel [14]. This misinformation often involves misusing historical events and misrepresenting videos, spreading extensively across social media platforms. Our analysis concentrated on prevalent themes from which these viral videos emerge, namely "IsraelHamas," "GazaGenocide," "Palestine," and "Israel." We discovered numerous videos and images from past conflicts in Syria and Iraq being repurposed to construct false narratives or exaggerate existing ones related to the current conflict.

Verifying content is cumbersome and resource-intensive; our research focuses on the content that achieved virality, not only false information.

2 RELATED WORK

2.1 Schwartz's Basic Human Value Theory

Schwartz's Theory of Basic Human Values is a well-established framework that categorizes human values into ten distinct types, organized around four higher-order values: self-transcendence, openness to change, self-enhancement, and conservation [37]. These values are considered to be universal and have been extensively studied and validated across different cultures [63]. The theory posits that these values are inherent to human nature and play a crucial role in shaping individuals' attitudes and behaviors [25].

Schwartz's work has been instrumental in various fields, including psychology, sociology, and organizational behavior. Researchers have used his theory to analyze values in different contexts, such as organizational values in universities [74], human-nature relationships and environmental behavior [13], and even in evaluating human values violations in online platforms like Stack Overflow [38]. The versatility of Schwartz's Theory of Basic Human Values is evident in its application across diverse cultural settings, as demonstrated in studies conducted in Russia [48], the Philippines [49], and African countries [25].

Moreover, Schwartz's theory has been refined and extended over the years to enhance its applicability and cross-cultural validity [64]. Studies have shown that the theory's explanatory power and ability to discriminate between different values make it a valuable tool for understanding individual value systems [39]. The theory's circular model, with its four higher-order values and ten basic values, provides a comprehensive framework for analyzing and comparing values across populations [73].

Most recently, the academic community has witnessed an influx of studies exploring the capabilities of Natural Language Processing (NLP) in identifying values within online communities [9]. Large Language Models (LLMs) have been developed, trained on extremely large corpora, and adapted for a variety of tasks. One challenge in detecting specific human values in user-generated content is understanding the context, which we see is mainly due to cultural and environmental factors. The research in [57] addresses this by creating a generalized dataset for training models on Schwartz's Basic Human Values. However, [57] encounters limitations in accuracy for out-of-domain tasks. [9] discusses advancements in LLMs, like Mistral, which have improved at generating context with carefully crafted prompts. These prompts help create synthetic datasets reflective of the target audience's values, which are then utilized to train classification algorithms tailored for specific analytical tasks. The approach outlined in [9] facilitates rapid comparative analysis and prediction without requiring extensive survey data collection.

2.2 European Social Survey and Cultural Relevance

[34] describes the state of 'Beliefs in Government,' a program which started in 1989, sponsored by the European Science Foundation (ESF). [34] depicts the comparative studies method used was less consistent between the nations than they ought to be. Notably, during this time, there were many different programs that conducted different types of political polling throughout the European nations [34]. However, there were noticeable inconsistencies between the different programs' statistical outputs, which caused difficulties in using the data for any type of governmental decision-making. This lack of reliable data called for a consistent program, established around 1997 for collecting socioeconomic data across European countries to allow for more reliable comparative and correlation studies. [7] describes how when we look at socio-cultural differences, the question always has an interdisciplinary relevance, which tends to bridge many fields like economics, sociology, history, political science or regional sciences [6, 30, 32, 56, 72]. These studies all reveal a positive correlation between the analyzed countries' socio-cultural development and their progress in economic democratic issues [7]. With this work, we generally do not get much of a glance within autocratic societies. However, [53] is a dataset that attempts to collect statistical data on social change in Europe to understand the characteristics related to these social changes, such as geo-political and economic impacts.

One paper that relates to this idea is related to studying political polarization and attitudes toward vaccination during the COVID-19 pandemic [75]. Wroblewski et al. demonstrated a significant statistical relationship between the vaccination variable and the level of political polarization. The effect demonstrated was high, suggesting a strong connection. Their work also gives us insights into interesting ways to use data that is already collected year-to-year to understand the impact of influences in the information environment. Fortunato et al. take a similar approach in using European Social Survey data to correlate Euroskepticism, a trend among individuals with lower former education and online political activity [24]. They found that individuals in this category were more than likely to find information about politics on social media leading to wider acceptance of Euroskepticism. Their study uniquely mentions that the ease of accessing information plays a variable role in the acceptance of information. Economic uncertainty seemed to be the main concern amongst European individuals within the group [24].

2.3 Inglehard-Welzel World Cultural Map

Based on the World Values Survey, the Inglehart-Welzel World Cultural Map provides a comprehensive framework for understanding cultural variations across societies. This map categorizes cultures along two major axes: traditional values versus secular-rational values (the vertical axis) and survival values versus self-expression values (the horizontal axis) [69]. The dimensions underlying this cultural map align with the Autonomy versus Embeddedness and Self-Enhancement versus Self-Transcendence dimensions in Schwartz's value space, albeit in a 45 degrees rotated manner [6]. This alignment suggests a connection between the values identified in Schwartz's Theory of Basic Human Values and the cultural dimensions delineated in the Inglehart-Welzel map.

Schwartz's theory, with its ten basic values organized around four higher-order values, provides a theoretical backbone for understanding individual value systems [59]. The theory's circular model and dynamic relationships among values are likely to hold across cultures, making it a valuable tool for analyzing and interpreting cultural variations [61].

Studies have highlighted the relationship between Schwartz's Embeddedness versus Autonomy dimension and the cultural dimensions represented in the Inglehart-Welzel map [23]. The diagonal relationship between these dimensions suggests a convergence in understanding cultural values and their manifestations across different societies. The Inglehart-Welzel map's depiction of cultural clusters aligns with Schwartz's emphasis on the universality of values and their role in shaping societal norms and behaviors [54].

The similarities between the two frameworks that [61] and [23] picked up show that there may be a correlation between individuals underlying values and the style of governance used in the regions being evaluated. Both frameworks tend to show an inverse relationship between secular and non-secularism.

2.4 Autocratic Governments versus Democratic Governments Effect on Virality

Autocratic and democratic leadership styles are closely linked to human values, as evidenced by research findings. Schwartz's Theory of Basic Human Values has been instrumental in understanding the relationship between leadership styles and follower trust. Studies have shown that valuing power and achievement highly correlates positively with problems of autocratic behavior while valuing universalism, benevolence, conformity, and tradition correlates negatively [64]. Moreover, openness and self-transcendence values correlate positively with development and democratization, while conservation and self-enhancement values correlate negatively [65]. These findings suggest that human values significantly shape attitudes toward autocratic and democratic leadership styles.

Schwartz's model of human values provides a robust framework for understanding the role of values in political behavior, including party affiliation and ideological orientation [20]. Additionally, thematic analysis using Schwartz's values theory has been employed to investigate the use of values appeals in persuasive speech during critical events like the COVID-19 pandemic. This research highlights the relevance of human values in shaping communication strategies and influencing public opinion in both autocratic and democratic contexts [12].

2.5 Multi-label Text Classification

Multi-classification in user text AI involves categorizing text data into multiple classes or labels. This task is essential for various applications such as document categorization, intent detection in dialogue systems, and user interest inference on social networks. Multilabel text classification (MLTC) is a significant natural language processing (NLP) task where text samples are assigned to multiple labels from a predefined label set [79]. Traditional methods, including deep learning approaches, have shown remarkable results in multi-label text classification tasks [26].

Algorithm adaptation plays a crucial role in enhancing single-label classification algorithms to address multi-label text classification challenges. For instance, an efficient multi-label SVM classification algorithm was proposed based on the approximate extreme points method and a divide-and-conquer strategy, demonstrating the effectiveness of adapting algorithms for multi-label classification. Moreover, the research has focused on developing specialized algorithms tailored for multi-label text classification tasks [67].

DeBERTa, a transformer-based language model, has been widely adopted for multi-label text classification tasks. Researchers have utilized DeBERTa for tasks such as patronizing and condescending language detection [21], proposing methods based on capsule networks and virtual adversarial training [82], and combining it with other models like tALBERT-CNN for multi-label text classification [41]. Furthermore, studies have explored the application of DeBERTa in contributing sentence selection and dependency parsing for entity extraction [43].

Various approaches have been developed to enhance multi-label text classification using DeBERTa, including the use of label prompts [66], reinforced sequence-to-set models [78], and hierarchical structures [5, 26]. Furthermore, the integration of DeBERTa with other techniques such as transformer-LDA for long text classification [68] and BERT-TextCNN for Chinese judicial text classification [81] has shown promising results in improving multi-label text classification performance.

2.6 Large Language Models to Enhance Signal Detection

Utilizing large language models for synthetic data generation has been shown to enhance model performance and address challenges related to data scarcity and diversity in various applications such as text classification, dataset generation for low-resource languages, information retrieval tasks, and spelling correction [8, 22, 41, 50, 76]. For instance, [50] explored the impact of using synthetic answerable and unanswerable questions alongside human-made datasets. [41] proposed a multi-label text classification method based on tALBERT-CNN, incorporating LDA topic modeling and ALBERT model for semantic context vectors. [76] introduced Afro-MNIST, synthetic datasets for low-resource languages. [8] utilized large language models for data augmentation in information retrieval tasks. Additionally, [22] developed a Transformer-based model for Vietnamese spelling correction using a large synthetic dataset.

Nikolenko et al. shows a significant increase in F1 score by introducing the synthetic data to the model in question [50]. Much of this can be attributed to the lexical correction of the text, allowing the model to make more accurate inferences. It is important to note that with all of the LLMs used, there is always an inherent risk of leaking artifacts about the underlying dataset, so synthetic data generation should be used cautiously [42]. Researchers should attempt to use publicly available data to mitigate some risks when generating synthetic datasets.

Another consideration often overlooked is not fine-tuning the model and measuring its capability, also known as zero-shot classification [18]. Some models have already been trained on such vast corpora that the model may already be able to be used as a detection mechanism. Cruickshank et al. found that some LLMs were effective with stance classification tasks without fine-tuning [17, 18]. In [18], the authors relied more heavily on prompt engineering, a more coercive method of constraining LLMs to specific domain tasks. In contrast, [17] found that fine-tuning a model can decrease the generalizability of LLMs on certain classification tasks. Thus, while it is still an open area for research, previous work indicates that LLMs may best be able to generate data for tasks like stance classification or values classification in a zero-shot setting.

2.7 Predicting Viral Trends on Social Media

What contributes to virality on social media has been a well-debated topic within academia. Being an interdisciplinary topic, one must expand their view, which is why this study chose to look at more marketing studies such as [62]. Looking at apps and products marketed to individuals on social media platforms, [62] demonstrates how most users do not use social media to look for utilitarian products and that simple cues and heuristics were more related to spiking interests. [62] shows that the relevance of the content to the target audience, in the context of utilitarian products, is crucial for driving virality. Rinandiyana et al. focused more specifically on what makes the consumer commit to the buy and the specific type of targeting involved [60]. The study shows how changing interests, needs, and preferences are important to predict the likelihood of an individual or group showing interest in a specific product being marketed.

Emotional content is often addressed in the literature as a factor that has predictive power. One seminal work, [11], demonstrates the relationship between content and emotion. The authors primarily study online videos; however, their study illustrates that video content more likely to spread is tied to the viewer's level of emotional reaction. The study signifies that emotional reaction has a close tie to virality. [35] re-investigates the idea of people being more likely to share content that aligns with their ideologies or political beliefs. [35] demonstrates that relevancy must be present in the user community for one to feel it is acceptable to share content. However, the study only focuses on the political nature of the content and does not explore the lexical properties of the content. [35] added a new view of moral leniency when misinformation aligns with the user's belief system.

3 METHODOLOGY

We use various methods to test our hypothesis. First, we classify virality based on the number of "likes" a message has. Second, we develop three DeBERTa classification models. Each model is comprised of a DeBERTa multi-classification model. The DeBERTa models are each fine-tuned with different approaches. Model 1 uses the Valuesnet dataset slightly augmented. Model 2 uses synthetically generated content label pairs for a Mistral LLM based on the Valuesnet dataset. Model 3 uses Model 2 plus synthetically generated data based on the real dataset.

Then, all three models generate probability vectors representing each value. We then can use a threshold to determine the values that are present in tweets. Then, the tweets are filtered based on the sentiment present in the text. We then analyze if certain values are detected when a tweet is potentially attempting to manipulate

emotional arousal. Then, to determine if this improves our ability to predict virality, we conduct a multiple logistic regression. We then compare the three model's performance to determine if the synthetic data improves predictive performance. A higher Receiver Operating Characteristic (ROC) is an indication of better model performance.

3.1 Terminology

In this study, "virality" refers to the potential for diffusion rather than adhering to a singularly defined concept across social media. The absence of a universally accepted definition of virality in scholarly discourse allows for a broad interpretation contingent upon the specificities of individual research endeavors. [55] elucidates the anatomy of viral social media events, suggesting that virality is fundamentally concerned with how content proliferates within and across platforms.

3.2 Data Collection

In light of recent modifications to social media policies, the ability to gather data on these platforms has become increasingly restricted, narrowing the scope for identifying viral phenomena primarily in public forums. Furthermore, numerous social media platforms have implemented stringent data collection restrictions. Platforms like Meta and X (previously known as Twitter) offer limited access to their research APIs for specific scholarly inquiries, subject to thorough review by the platform's governance. Platforms such as YouTube and Reddit, which grant API access, host valuable content and are slated for inclusion in our future research endeavors. For the current study, our objective necessitated a concentration on a particular geographical area to explore cross-cultural dynamics.

To gather data on social media platforms, we utilized Selenium, a library that enables headless browsing through web drivers, facilitating access to website classes [28]. Through inspecting element selectors within Google Chrome, we aimed to capture user handles, dates, tweets, hashtags, likes, retweets, and replies, focusing on the Hamas-Israel conflict. By leveraging X's advanced search, we crafted queries to explore specific date ranges and themes, prioritizing user-generated content over news media. Our initial focus led us to the "GazaGenocide" theme, which led us to related themes such as "Israel," "Palestine," "GazaWar," and "IsraelHamas," thus compiling a substantial dataset. After data collection, we faced multiple approaches to classify the data. The data was mostly Arabic. To simplify the process, we used the translator library to translate all Arabic text into English. We understand that there is a slight loss of context when using translators. However, by using the same method for the entire dataset, we assume that we will still have a uniform representation of the underlying signal since the dataset is large. After filtering the data, we have 3,610 usable samples. The first notable instance within the "GazaGenocide" theme was identified on October 9th, 2023, shortly after the conflict began, extending our collection through February 29th.

To minimize irrelevant data, we only included content with over 50 likes, assuming that such content had the potential for virality but was possibly overshadowed by other factors or narratives. For those interested in further research, we offer access to our dataset, comprising 3,619 samples in CSV format, to facilitate reproducibility.

3.3 Establishing a Baseline

To demonstrate the effectiveness of our synthetic data process, we built a DeBERTa model training only on the ValuesNet dataset, where only the positive and negative labels are extracted.

We build a secondary model, where we generate new samples with a Mistral LLM [33] based on the ValuesNet dataset. Both of the models are not exposed to the collected data during training. These two baselines will allow us to compare a model that has been exposed to the domain in question with models that are completely generalized to a conception of values.

We understand that the LLM offers some generalizability to multiple domains due to the corpora it is trained on. This is why we built two baseline models. One with no use of an LLM and one with. This will allow us to see how the LLM contributes to the results before we apply our intervention of the synthetic dataset built with the collected samples.

3.4 Synthetic Dataset Development

The development of our synthetic dataset unveiled complex challenges, notably the tendency of our multi-label classification model to disproportionately favor certain signals, particularly those representing the basic human values of achievement and self-direction over others. Initial tests revealed a predisposition towards overfitting specific text types. Parsing while generating data proved to be challenging, so we had two rounds of parsing to achieve a desirable format. Using the Mistral 7B model with the instruction feature proved to be helpful. We were able to engineer a response through multiple prompts, leading the model to generate text from ten different values stances. We prompted the LLM with the samples from the real dataset to generate text related to the themes that we are studying. For the content generation, we reduced the temperature to "0.2" to produce a standardized output for easier parsing.

We initially took the real public dataset and used a prompt with the Mistral LLM to generate a tweet representing each value stance [33].

3.5 Content Generation Method

Shalom Schwartz introduced ten basic values in his theory of basic human values. These values are Achievement, Benevolence, Conformity, Hedonism, Power, Security, Self-Direction, Stimulation, Tradition, and Universalism. To create content that reflects these values, the following approach is used:

(1) **Identify the Query**: Start with a specific query or statement that you want to restate from different value perspectives.
(2) **Restate the Query**: Restate the query from the perspective of each value. The content should reflect how someone with that value might perceive or articulate the query.

3.6 Example Usage

Suppose we have the following query: *"How can we improve community engagement in our town?"*

This query can be restated from each value perspective as follows:

- **Achievement**: "How can we set and achieve high standards for community engagement in our town?"
- **Benevolence**: "How can we foster a sense of caring and helpfulness in our community engagement efforts?"
- **Conformity**: "How can we ensure that community engagement follows established norms and rules?"
- **Hedonism**: "How can we make community engagement enjoyable and fun for everyone?"
- **Power**: "How can we enhance our town's influence and leadership through community engagement?"
- **Security**: "How can we make community engagement a safe and stable activity for all participants?"
- **Self-Direction**: "How can we encourage creativity and independence in our community engagement initiatives?"
- **Stimulation**: "How can we make community engagement exciting and full of new experiences?"
- **Tradition**: "How can we honor our town's traditions through community engagement?"
- **Universalism**: "How can we promote understanding, tolerance, and protection for the welfare of all through community engagement?"

Thus, for each of the Valuesnet samples, we created nine more examples that used the content from the original example but expressed that content from a different values perspective. In this way, we can generate text samples where the primary axis of difference is exactly the difference we wish to classify: the values being expressed. If we had generated artificial examples without reference to the Valuesnet grounding examples, the LLM could introduce other types of variation into the generated examples, such as different topics, which could make classifying values more difficult.

3.7 Augmentation with Valuesnet Dataset

Explain the integration and selection criteria using the Valuesnet dataset.

$$S_{aug} = S_{synth} \cup S_{valuesnet} \tag{1}$$

Despite its breadth, the initial representation within the Valuenet dataset did not align with our needs due to its scenario classification system, which was overly specific to individual values and their perceived relevance or irrelevance from a subjective standpoint.

To address this issue, we first augment the Valuesnet dataset by removing irrelevant samples. This approach mirrored the processing of our real dataset and highlighted the model's continued bias toward detecting values such as stimulation, achievement, and self-direction. This bias contrasted the prevalence of benevolence and universalism in most texts, underscoring language's inherent subjectivity and normative nature. Before applying the binary transformation, we remove all instances where $x = 0$ (content, label pairs with irrelevant scenarios) from the dataset. Let D be the original dataset and D' be the dataset after removal of zeros:

$$D' = \{x \in D \mid x \neq 0\}$$

$$V_{binary}(x) = \begin{cases} 1 & \text{if } x = 1 \text{ or } x = -1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The purpose of including both the negative and positive stances of each value is to ensure that the model becomes generalizable enough to recognize certain values that are targeted negatively and positively to achieve emotional arousal. Since effective sentiment analyzers exist, we do not see the purpose in classifying based on sentiment.

Our resampling technique is represented in the following form:

$$P(\text{sample } i) = \frac{{w_i}^{-1}}{\sum_{j=1}^{N} {w_j}^{-1}} \tag{3}$$

where w_i represents the weight for the i-th label, inversely proportional to its occurrence in the dataset.

3.8 Denoising and Improving Cluster Quality

Using the synthetically generated samples from the real dataset and the Valuesnet dataset, we evaluate each sample across each of the ten human values, constructing a ten-dimensional vector for each sample. The label reassignment for each value within the vector is defined mathematically as follows:

$$L_{detect}(v_i) = \begin{cases} e_i & \text{if value } v_i \text{ is represented} \\ 0 & \text{if value } v_i \text{ is not represented} \end{cases} \tag{4}$$

where v_i represents each of the ten human values.

We process the embeddings through the softmax function on each fine-tuning iteration to produce probabilities for each value represented in a vector of 10 float numbers.

$$P = \text{softmax}(E) \tag{5}$$

PCA allows us to confirm if the model shares similarities with Schwartz's value framework by measuring the correlations between values. In Schwartz's studies, he describes how certain values are correlated positively and others inversely. Confirming these patterns helps us confirm that our model is adequately tuned.

After exporting the embedding dataset to a CSV file, we applied Principal Component Analysis (PCA) for the purposes of denoising and improving computational efficiency. PCA was the most streamlined denoising technique for the goals of our research:

$$X_{pca} = PCA(E, k) \tag{6}$$

where k denotes the number of principal components retained. For our analysis, we aim to retain $k = 10$ principal components. If fewer than ten components are initially retained, we undertake additional resampling procedures until ten distinct components can be clearly defined and extracted.

Upon denoising of our dataset using PCA, we apply the K-means clustering algorithm to identify distinct groups within the data. This is mathematically represented as:

$$\text{clusters} = \text{K-means}(X_{pca}, n) \tag{7}$$

where n is the number of clusters, determined based on the number of principal components retained, k. Following the clustering, we assess the quality of these clusters using the silhouette score, a measure of how similar an object is to its own cluster compared to other clusters. The silhouette score for each sample is calculated as follows:

$$\text{silhouette score} = \frac{b - a}{\max(a, b)} \tag{8}$$

where a is the mean intra-cluster distance (the average distance between each point within a cluster), and b is the mean nearest-cluster distance (the average distance to the nearest cluster that the point is not a part of). A higher silhouette score indicates a better-defined clustering, where clusters are compact and well-separated from each other.

Visualizing the PCA loadings with K-means clusters allows us a secondary method for confirming similarities within Schwart's values by seeing a 2-dimensional representation of the 10 dimensions. We can adequately see if the clusters that should be close are grouped in the right locations.

3.9 Evaluation Metrics

The virality of content often coincided with concurrent global events, leading to periods of fluctuating viral activity. Consequently, the investigation adapted to regard global events as potential catalysts for virality without assuming guaranteed viral outcomes. The primary metric for identifying viral content was elevated rates of likes and shares. Content was defined as "viral" if it exceeded the mean number of shares by more than two standard deviations, statistically distinguishing viral content:

$$\text{Viral} = \text{mean}(shares) + 2 \times \text{std}(shares) \tag{9}$$

We are identifying our outliers within a small margin. Because of the extreme right-tailed skewness, the median was not a practical means for defining our outliers, as the majority of the population was clustered below 1,000 likes. Considering some of the options in [40], we determined that the 2 standard deviations allow us to capture enough outliers that allow all three models to maintain goodness-of-fit. Using goodness-of-fit for adjusting the outlier threshold is described in [1] when research is focused on looking at interesting outliers. In our study, virality meets this criterion for interesting outliers.

First, we ensured that the data represented the theme we targeted and that the samples were relevant to the network's context. Sample bias was mitigated by systematically collecting samples and removing those with fewer than 50 likes. Since the data representation was consistent, we achieved normality through systematic evaluation by the classifier. Equal covariance was maintained as all embeddings shared the same dimensions across the population.

Comparative Analysis of Viral and Non-Viral Content

To examine the potential correlation between values and virality, the study employs a comparative analysis across two categorizations — viral and non-viral content — each encompassing multiple labels represented as probabilities indicative of human values.

In addition to values detection, the DeBERTa classifier was fine-tuned to detect both negative and positive sentiments related to polarity. However, research suggests that TextBlob sentiment analysis has a higher accuracy in detecting sentiment within text [3]. Thus, we chose TextBlob to analyze negative and positive emotional arousal within the text.

Our aim is to move beyond sentiment analysis and start examining the underlying properties of virality. To this end, we have ten independent variables $(X_1, X_2, \ldots, X_{10})$ for each of the ten values, each with an associated sentiment (positive or negative).

Since this experiment has one outcome variable with two categories and multiple continuous predictor variables, the appropriate statistical test is a multiple logistic regression, which will help us understand the predictive power of each model.

We first used cosine similarity to evaluate the similarity between the two distributions. Values close to 1 indicate high similarity, while values close to -1 indicate high dissimilarity:

$$\text{Cosine Similarity} = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|\|\vec{B}\|} \tag{10}$$

[58] Cosine similarity measures how similar or dissimilar the viral nodes are from the non-viral nodes based on the angle between their vectors in multi-dimensional space. It indicates the degree of similarity irrespective of their magnitude.

We then used Euclidean distance to determine the dissimilarity between the two distributions for each value:

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} \tag{11}$$

The Euclidean distance metric focuses on magnitude and closeness by measuring the straight-line distance based on underlying features. Using both Euclidean distance and cosine similarity helps determine if our model is detecting both sets (viral and non-viral) as coming from the same or different distributions.

We conduct a Kolmogorov-Smirnov (KS) Test on both the Euclidean and cosine similarity to address the goodness-of-fit of the model. A non-significant result from one or both measurements is considered adequate for further analysis. [44]

We use this test as a precursor before interpreting the logistic coefficients.

The mathematical model for our multiple logistic regression is formulated as follows:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{10} x_{10} \tag{12}$$

where p represents the probability of the content being viral. Each x_i (for $i = 1, 2, \ldots, 10$) denotes the probability associated with one of the 10 human values influencing content virality. The coefficients $\beta_1, \beta_2, \ldots, \beta_{10}$ quantify the impact of each value on the likelihood of content being classified as viral. We used SMOTE [46] to handle the class imbalances.

To assess the significance of the differences between the 10 value coefficients from the non-viral set and the viral set, we employed pairwise t-tests. This statistical method allows us to compare the means of two related groups to determine if they are significantly different from each other.

After collecting the results for the three models, we offer a comparative analysis and interpretation of the results. We take into consideration the imbalance of the sets and the specific theme and topic the samples were derived from.

3.10 Threats to Validity

LLMs are known to hallucinate and generate convincing outputs. Not knowing the underlying corpora requires researchers to take

Table 1: KS VALUES (Viral vs non-Viral) and p-Values for Cosine Similarity

Model	KS Value (Cosine)	p-Value
Values-LLM	-0.36	0.72
Values+LLM	0.22	0.19
Values+Real+LLM	0.04	0.99

Table 2: KS VALUES (Viral vs non-Viral) and p-Values for Euclidean Distance

Model	KS Value (Euclidean)	p-Value
Values-LLM	0.10	0.34
Values+LLM	0.29	0.03
Values+Real+LLM	0.19	0.22

further precautions to confirm if the model is reliable for the particular question at hand.

By establishing goodness-of-fit, we show that both of our sets are from the same distribution. By demonstrating a significant t-test on the model's ability to demonstrate that certain values are more significant in one set versus the other, we are able to show support for manipulation within specific tweets. By adding sentiment analysis, we can detect the intensity of emotional arousal. We then use multiple logistic regressions to demonstrate how these features can be toggled to improve the ability to detect virality.

4 RESULTS

For a common threshold and testing purposes, we set the negative polarity at -0.60 and the positive polarity at 0.50. Throughout our research, sentiment has always been a common detector of viral content. High polarity is an indicator of emotional arousal. With the use of sentiment and values detection, we can distinguish which motivator (value) the samples target and observe the motivator being manipulated to potentially achieve virality.

We attempted to find a common threshold for the values; however, the third model derived from both the values dataset and synthetic samples from real data was more definitive in values assignment; therefore, the first two models have a values threshold of 0.20 and model three has a values threshold of 0.25. This allows all models to achieve the same assumptions.

All models demonstrated adequate clustering and there were consistent similarities with Scwartz's model confirmed through a qualitative visual assessment [16, 63, 65].

4.1 KS Test Results

Tables 1 and 2 contain the KS results for the cosine and euclidean distributions. Models 1 and 3 both passed the KS test with cosine similarity and Euclidean similarity. Model 2 failed the KS test for Euclidean distance.

Table 3: T-Test Results for Schwartz's 10 Values for All Models (Viral vs Non-Viral)

Value	Model	t-Statistic	p-Value
Achievement	Values-LLM	-0.26	0.79
	Values+LLM	-0.57	0.57
	Values+Real+LLM	-0.84	0.39
Benevolence	Values-LLM	1.98	0.04
	Values+LLM	-0.59	0.55
	Values+Real+LLM	-1.76	0.07
Hedonism	Values-LLM	-0.35	0.72
	Values+LLM	3.84	0.00
	Values+Real+LLM	1.41	0.16
Power	Values-LLM	-0.98	0.33
	Values+LLM	-0.43	0.67
	Values+Real+LLM	0.82	0.42
Security	Values-LLM	-0.43	0.67
	Values+LLM	-0.38	0.70
	Values+Real+LLM	1.49	0.13
Self-Direction	Values-LLM	-0.66	0.51
	Values+LLM	-0.25	0.80
	Values+Real+LLM	-0.24	0.81
Stimulation	Values-LLM	-0.24	0.81
	Values+LLM	-0.19	0.85
	Values+Real+LLM	4.85	0.00
Trad/Conformity	Values-LLM	-0.10	0.92
	Values+LLM	-0.53	0.59
	Values+Real+LLM	0.40	0.69
Universalism	Values-LLM	3.12	0.00
	Values+LLM	-0.52	0.61
	Values+Real+LLM	-0.77	0.44

4.2 Pairwise T-Test Results for Schwartz's 10 Values

Table 3 contains the results for our t-test. A significant t-test score does not explain the predictive power of the model; however, it does show which values are being significantly manipulated within the discourse. We can see that model one, with no familiarization with the real dataset, shows Universalism and Benevolence as the significant indicators. However, with fine-tuning, hedonism and stimulation become more detectable within the real data. We start to see with Model 3 that Universalism becomes more insignificant.

Table 4: Logistic Regression Results for Models (Predicting Virality)

Model	Accuracy	Precision	Recall	AUC
Values-LLM	0.80	0.86	0.80	0.80
Values+LLM	0.72	0.82	0.72	0.73
Values+Real+LLM	0.94	0.93	0.93	0.93

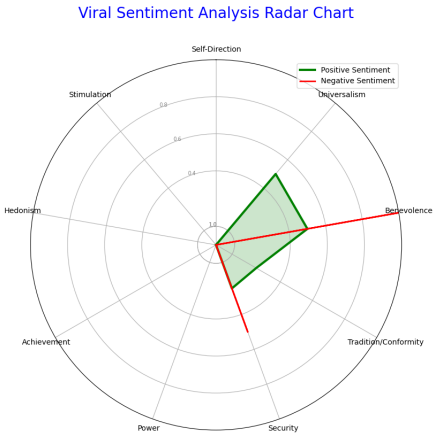

Figure 1: Model 1 - Viral Profile

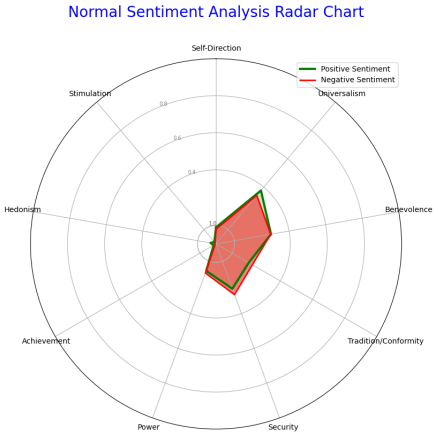

Figure 2: Model 1 - Non-Viral Profile

4.3 Logistic Regression Results

Table 4 contains the logistical regression results. Ensuring that all models meet goodness-of-fit, we see that Model 3 performed superior to the other two, indicating that familiarizing the classification model with the real dataset through synthetic data sampling does improve the model performance.

5 DISCUSSION

5.1 Observation 1: The Purpose of Building a Discourse Profile

If this methodology is to be used for practical applications, it is important to understand that we must be able to generate a baseline of the normal distribution before can effectively predict the potential for virality. As we depicted, the values profiles are unique to the region, culture, theme, and topic.

Looking at Figure 3, we can see that the viral profile of Model 3 is completely different than the non-viral profile (Figure 4). We can see that the non-viral, which makes up the bulk of the distribution, is what the normal discourse profile should look like. Then, with the viral set, we see more negative sentiment with the values stimulation and hedonism, which fall outside the normal distribution. We can visually detect disturbances in normal discourse.

5.2 Observation 2: Schwartz's Basic Human Values versus Our Model

When we used PCA to analyze our embeddings, we found many of the same correlations as Schwartz's research suggested. For instance, looking at the K-means clustering of the PCA loadings, we saw closeness with Universalism and Benevolence. We observe that the values that are normally inverse to Benevolence and Universalism in Schwartz's studies were either negatively correlated or not correlated at all. We see this as strong evidence that the model is measuring the particular features in question. There is limited research that presents this kind of support with using online content thus adding a valid contribution to cross-cultural psychology studies and deep learning communities. Many similar studies have only been done using qualitative surveys. One such study done on a similar region shows an abundance of Benevolence and Conformity registering within the population [16]. Within our viral profiles, we can see that benevolence and conformity register well in both models and that other values that are not persistent in the population are being manipulated.

We posit that our method is a valuable tool for the field of cross-cultural psychology and cybersecurity.

5.3 Observation 3: Use of Valuesnet versus Valuesnet and Real Data

Our observation of the ValuesNet dataset, even without any manipulation, reveals its positive predictive power and generalizability. This highlights the potential that this methodology can be harnessed for numerous other applications, contributing more explanatory power to the field of cross-cultural psychology. As we mentioned earlier, we postulated that the LLM would alter how the model perceives certain values within the network's context. These postulations hold true, as evidenced by the significant changes in coefficients across the models. Most notably, the original model demonstrates a substantial difference in how benevolence is distributed between non-viral and viral tweets, as we see with Figure 3 and Figure 1. The addition of synthetic data based on the Valuesnet dataset has a negative impact on benevolence but shows significance in hedonism. This phenomenon is likely caused by the LLM's influence on the samples and interpretation of benevolence and hedonism.

Adding the synthetically created data based on the real network affects the model similarly to the synthetic values model. However, now that the dataset contains some relevant context, we see that stimulation is now a significant indicator of virality, and hedonism loses its significance. Oddly, benevolence shows little support as a viral indicator in the third model. The manner in which the embeddings are transformed causes the third model to have almost identical cosine similarity amongst the distributions, suggesting

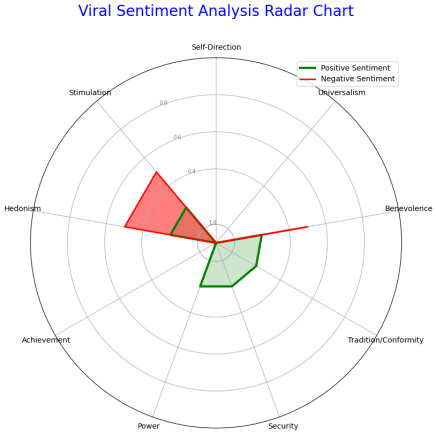

Figure 3: Model 3 - Viral Profile

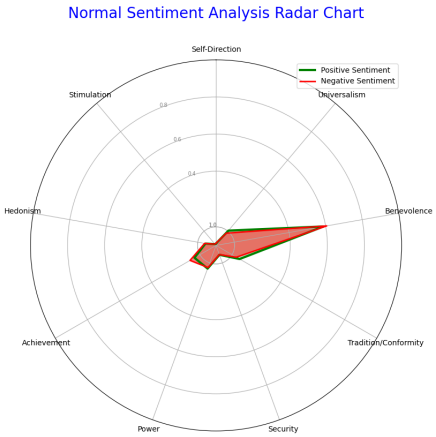

Figure 4: Model 3 - Non-Viral Profile

that the model is able to see mostly the same values represented within the distribution.

5.4 Observation 4: Predictive Power

Our logistic regression results suggest that the union of synthetic values net data with synthetic data formed from real data contributes beyond sentiment's predictive power. We acknowledge that we only examined one viral phenomenon and that additional research must be conducted to confirm whether this model is reliable in a broad range of applications.

In any case, we do see an appealing approach that could assist media platforms in limiting their search for viral cascades. By reducing this search space, security analysts can focus on verifying information that could become viral, allowing them the ability to make decisions regarding potential content in its genesis rather than when it has reached a high propagation rate.

5.5 Assumptions, Limitations, and Constraints

Assumptions. We acknowledge the potential compromise to data integrity from posts that may have been removed or censored for containing objectionable material [45, 70, 77]. Nevertheless, we assume that the residual data facilitates a comprehensive analysis over an extended duration. Notably, we recognize the opportunity and use of other platforms for this study, and we intend to conduct a cross-platform analysis in the future using media listeners and multiple social media platforms with different themes. We offer all the artifacts used for this research; however, building a model from the start will produce similar results, but not the same. Social media content adds uncertainty to the results.

Limitations. A well-known limitation of using social media data is the inability to collect descriptive statistics. We account for this by pulling from multiple themes related to the topic and a wide time frame over a span of six months.

Constraints. Due to the cost and restrictions of API access to the social media platform, we only use data that is publicly accessible.

5.6 Ethics

Given user privacy concerns, it would be unethical to infer descriptive statistics on the population without permission. It would also be infeasible to gather informed consent given the social media information environment and wide scope of analysis; therefore, we do not pursue these details.

The authors exercised due diligence to minimize any bias about the controversial case introduced in this study. The samples were collected, cleaned, translated, and filtered systematically. We provide all code used to exercise this process, and it is publicly available online. This controversial topic provides a unique opportunity and highlights the importance of analyzing the manipulation of online discourse. As more influence threats become apparent, more research will be conducted in the social media space. With the interruption of API access, researchers will attempt to pull data from the Internet with various methods. We realize that highlighting certain topics can have a negative effect on the discourse itself; therefore, we acknowledge the importance of trying to maintain a neutral tone when we investigate controversial topics.

6 CONCLUSION

In this paper, we demonstrate a novel technique for refining multi-classification models for studying online discourse. With the use of previous datasets and the augmentation of real-world data, we can generate synthetic data that assists us in the analysis of underlying values in content to predict the potential for virality.

Through our methods, we demonstrate the potential for influencing target audiences by curating relevant content through value manipulation. Allowing a malign actor online to manipulate content to reach a higher emotional arousal state gives them the ability to manipulate political, medical, or other important ideas - which often have second and third-order effects in the real world. These actions, coupled with coordinated inauthentic behavior methods, such as the use of social bots or troll farms, only exasperate efforts to mitigate influence campaigns. Our work offers an approach for researchers and practitioners to understand and detect potentially viral information.

REFERENCES

[1] Herman Aguinis, Ryan K Gottfredson, and Harry Joo. 2013. Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods* 16, 2 (2013), 270–301.

[2] Olawale Olufemi Akinrinde. 2020. Social injustice, corruption and Nigeria's national security quest: A theoretical discourse. *Global Journal of Sociology: Current Issues* 10, 2 (2020), 63–70.

[3] Wajdi Aljedaani, Furqan Rustam, Mohamed Wiem Mkaouer, Abdullatif Ghallab, Vaibhav Rupapara, Patrick Bernard Washington, Ernesto Lee, and Imran Ashraf. 2022. Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry. *Knowledge-Based Systems* 255 (2022), 109780.

[4] Kwesi Aning and Ernest Ansah Lartey. 2019. Governance perspectives of human security in Africa. *Asian Journal of Peacebuilding* 7, 2 (2019), 219–237.

[5] Simon Baker and Anna Korhonen. 2017. Initializing neural networks for hierarchical multi-label text classification. Association for Computational Linguistics.

[6] Sjoerd Beugelsdijk and Chris Welzel. 2018. Dimensions and dynamics of national culture: Synthesizing Hofstede with Inglehart. *Journal of cross-cultural psychology* 49, 10 (2018), 1469–1505.

[7] Ákos Bodor, Zoltán Grünhut, and Réka Horeczki. 2014. Socio-cultural cleavages in Europe. *Regional Statistics: journal of the Hungarian Central Statistical Office* 4, 2 (2014), 106–125.

[8] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. *arXiv preprint arXiv:2202.05144* (2022).

[9] Nadav Borenstein, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2024. Investigating Human Values in Online Communities. *arXiv preprint arXiv:2402.14177* (2024).

[10] Maria Teresa Borges-Tiago, Flavio Tiago, and Carla Cosme. 2019. Exploring users' motivations to participate in viral communication on social media. *Journal of Business Research* 101 (2019), 574–582. https://doi.org/10.1016/j.jbusres.2018.11.011

[11] Elsamari Botha and Mignon Reyneke. 2013. To share or not to share: the role of content and emotion in viral marketing. *Journal of Public Affairs* 13, 2 (2013), 160–171.

[12] Linda Courtenay Botterill and Niobe Lewis. 2023. Thematic analysis using the Schwartz values theory: exploring the use of values appeals in persuasive speech during COVID-19 in Australia. *European Political Science Review* 15, 1 (2023), 136–144.

[13] M. Braito, K. Böck, C. G. Flint, A. Muhar, S. Muhar, and M. Penker. 2017. Human–nature relationships and linkages to environmental behaviour. *Environmental Values* 26 (2017), 365–389. Issue 3. https://doi.org/10.3197/096327117x14913285800706

[14] Caitlin Chin-Rothmann. 2023. Social Media Platforms Were Not Ready for Hamas Misinformation. (Oct. 2023). https://www.csis.org/analysis/social-media-platforms-were-not-ready-hamas-misinformation

[15] Matteo Cinelli, Stefano Cresci, Walter Quattrociocchi, Maurizio Tesconi, and Paola Zola. 2022. Coordinated Inauthentic Behavior and Information Spreading on Twitter. *Decision Support Systems* 160 (Sept. 2022), 113819. https://doi.org/10.1016/j.dss.2022.113819

[16] Aaron Cohen. 2010. Values and commitment: A test of Schwartz's human values theory among Arab teachers in Israel. *Journal of applied social psychology* 40, 8 (2010), 1921–1947.

[17] Iain J Cruickshank and Lynnette Hui Xian Ng. 2023. Use of large language models for stance classification. *arXiv preprint arXiv:2309.13734* (2023).

[18] Iain J Cruickshank and Lynnette Hui Xian Ng. 2024. DIVERSE: Deciphering Internet Views on the US Military Through Video Comment Stance Analysis, A Novel Benchmark Dataset for Stance Classification. *arXiv preprint arXiv:2403.03334* (2024).

[19] Eldad Davidov, Peter Schmidt, and Shalom H Schwartz. 2008. Bringing values back in: The adequacy of the European Social Survey to measure values in 20 countries. *Public opinion quarterly* 72, 3 (2008), 420–445.

[20] Jessy Defenderfer. 2019. The Effect of Human Values on Party Identification and Ideology for Black and White Partisans. *Social Science Quarterly* 100, 6 (2019), 2240–2255.

[21] Yong Deng, Chenxiao Dou, Liangyu Chen, Deqiang Miao, Xianghui Sun, Baochang Ma, and Xiangang Li. 2022. BEIKE NLP at SemEval-2022 Task 4: Prompt-Based Paragraph Classification for Patronizing and Condescending Language Detection. *arXiv preprint arXiv:2208.01312* (2022).

[22] Dinh-Truong Do, Ha Thanh Nguyen, Thang Ngoc Bui, and Hieu Dinh Vo. 2021. Vsec: Transformer-based model for vietnamese spelling correction. In *PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part II 18*. Springer, 259–272.

[23] Henrik Dobewall and Maksim Rudnev. 2014. Common and unique features of Schwartz's and Inglehart's value theories at the country and individual levels. *Cross-Cultural Research* 48, 1 (2014), 45–77.

[24] Piergiuseppe Fortunato and Marco Pecoraro. 2022. Social media, education, and the rise of populist Euroscepticism. *Humanities and Social Sciences Communications* 9, 1 (2022), 1–13.

[25] G. Goncalves, T. Oliveira, and F. Cruz-Jesus. 2018. Understanding individual-level digital divide: evidence of an african country. *Computers in Human Behavior* 87 (2018), 276–291. https://doi.org/10.1016/j.chb.2018.05.039

[26] Jibing Gong, Zhiyong Teng, Qi Teng, Hekai Zhang, Linfeng Du, Shuai Chen, Md Zakirul Alam Bhuiyan, Jianhua Li, Mingsheng Liu, and Hongyuan Ma. 2020. Hierarchical graph transformer-based deep learning model for large-scale multi-label text classification. *IEEE Access* 8 (2020), 30885–30896.

[27] M Guirgis, F Nusair, YM Bu, K Yan, and AT Zekry. 2012. Barriers faced by migrants in accessing healthcare for viral hepatitis infection. *Internal medicine journal* 42, 5 (2012), 491–496.

[28] Nicholas Harrell, Iain Cruickshank, and Alexander Master. 2024. Overcoming Social Media API Restrictions: Building an Effective Web Scraper. ICWSM, US. https://doi.org/10.36190/2024.72

[29] Stephen Hart and Mark Klink. 2017. 1st Troll Battalion: Influencing Military and Strategic Operations through Cyber-Personas. In *2017 International Conference on Cyber Conflict (CyCon U.S.)*. IEEE, Washington, DC, 97–104. https://doi.org/10.1109/CYCONUS.2017.8167503

[30] Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. 2005. *Cultures and organizations: Software of the mind*. Vol. 2. Mcgraw-hill New York.

[31] Matouš Hrdina and Zuzana Karaščáková. 2014. Parties, pirates and politicians: The 2014 European Parliamentary elections on Czech Twitter. *Human affairs* 24, 4 (2014), 437–451.

[32] Ronald Inglehart and Christian Welzel. 2010. Changing mass priorities: The link between modernization and democracy. *Perspectives on politics* 8, 2 (2010), 551–567.

[33] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL]

[34] Roger Jowell, C Roberts, Rory Fitzgerald, and Gillian Eva. 2007. European social survey. *Measuring Attitudes Cross-Nationally: Lessons from the European Social Survey* (2007), 1.

[35] Laura Joyner, Tom Buchanan, and Orkun Yetkili. 2023. Moral leniency towards belief-consistent disinformation may help explain its spread on social media. *Plos one* 18, 3 (2023), e0281777.

[36] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T Hancock. 2023. Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–29.

[37] M. Killen. 2016. Children's values: universality, conflict, and sources of influence. *Social Development* 25 (2016), 565–571. Issue 3. https://doi.org/10.1111/sode.12189

[38] S. Krishtul, M. Shahin, H. O. Obie, H. Khalajzadeh, F. Gai, A. R. Nasab, and J. Grundy. 2022. Human values violations in stack overflow: an exploratory study. (2022). https://doi.org/10.48550/arxiv.2203.10551

[39] J. Lee, J. Sneddon, T. M. Daly, S. H. Schwartz, G. N. Soutar, and J. J. Louviere. 2016. Testing and extending schwartz refined value theory using a best–worst scaling approach. *Assessment* 26 (2016), 166–180. Issue 2. https://doi.org/10.1177/1073191116683799

[40] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology* 49, 4 (2013), 764–766.

[41] Wenfu Liu, Jianmin Pang, Nan Li, Xin Zhou, and Feng Yue. 2021. Research on multi-label text classification method based on tALBERT-CNN. *International Journal of Computational Intelligence Systems* 14, 1 (2021), 201.

[42] Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are Emergent Abilities in Large Language Models just In-Context Learning? *arXiv preprint arXiv:2309.01809* (2023).

[43] Anna Martin and Ted Pedersen. 2021. Duluth at semeval-2021 task 11: Applying deberta to contributing sentence selection and dependency parsing for entity extraction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. 490–501.

[44] Frank J Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253 (1951), 68–78.

[45] Alexander Master and Christina Garman. 2023. A Worldwide View of Nation-state Internet Censorship. In *Free and Open Communications on the Internet*. Proceedings on Privacy Enhancing Technologies. https://www.petsymposium.org/foci/2023/foci-2023-0008.pdf

[46] Ahmed Jameel Mohammed, Masoud Muhammed Hassan, and Dler Hussein Kadir. 2020. Improving classification performance for a novel imbalanced medical dataset using SMOTE method. *International Journal of Advanced Trends in Computer Science and Engineering* 9, 3 (2020), 3161–3172.

[47] Yamir Moreno, Maziar Nekovee, and Alessandro Vespignani. 2004. Efficiency and reliability of epidemic data dissemination in complex networks. *Physical*

Review E 69, 5 (2004), 055101.

[48] M. L. Nadezhda and A. Tatarko. 2018. Basic values in russia: their dynamics, ethnocultural differences, and relation to economic attitudes. *Psychology in Russia: State of the Art* 11 (2018), 36–52. Issue 3. https://doi.org/10.11621/pir.2018.0303

[49] I. M. Nibalvos. 2018. Pagpapahalagang pilipino sa mga piling siday ng san julian, silangang samar (filipino values in selected siday of san julian, eastern samar). *Scientia - The International Journal on the Liberal Arts* 7 (2018). Issue 2. https://doi.org/10.57106/scientia.v7i2.93

[50] Liubov Nikolenko and Pouya Rezazadeh Kalehbasti. 2021. When in Doubt, Ask: Generating Answerable and Unanswerable Questions, Unsupervised. In *Machine Learning and Knowledge Extraction: 5th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2021, Virtual Event, August 17–20, 2021, Proceedings 5*. Springer, 21–33.

[51] Mohammed Nuruzzaman. 2006. Paradigms in conflict: The contested claims of human security, critical theory and feminism. *Cooperation and Conflict* 41, 3 (2006), 285–303.

[52] Mariam Orabi, Djedjiga Mouheb, Zaher Al Aghbari, and Ibrahim Kamel. 2020. Detection of Bots in Social Media: A Systematic Review. *Information Processing & Management* 57, 4 (July 2020), 102250. https://doi.org/10.1016/j.ipm.2020.102250

[53] Aslıhan Özgül. [n. d.]. REVIEW OF LARGE-SCALE ASSESSMENTS FOR COMPARATIVE STUDIES (1)–EUROPEAN AND GLOBAL DATASETS. ([n. d.]).

[54] JP Piotrowski and M Żemojtel-Piotrowska. 2020. Relationship between numinous constructs and values. Psychology of Religion and Spirituality. Advance online publication.

[55] Essi Pöyry, Salla-Maaria Laaksonen, Arto Kekkonen, and Juho Pääkkönen. 2018. Anatomy of viral social media events. In *Annual Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences, 2173–2182.

[56] Robert D Putnam. 2015. Bowling alone: America's declining social capital. In *The city reader*. Routledge, 188–196.

[57] Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. Valuenet: A new dataset for human value driven dialogue system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11183–11191.

[58] Faisal Rahutomo, Teruaki Kitasuka, Masayoshi Aritsugi, et al. 2012. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, Vol. 4. University of Seoul South Korea, 1.

[59] Cornelius A Rietveld and Brigitte Hoogendoorn. 2022. The mediating role of values in the relationship between religion and entrepreneurship. *Small Business Economics* (2022), 1–27.

[60] Lucky Radi Rinandiyana, Tine Badriatin, and Asep Saepudin. 2022. Viral Marketing Concept and Viral Marketing Development on Consumer Buying Approach. *Almana: Jurnal Manajemen dan Bisnis* 6, 1 (2022), 117–123.

[61] Maksim Rudnev, Vladimir Magun, and Shalom Schwartz. 2018. Relations among higher order values around the world. *Journal of Cross-Cultural Psychology* 49, 8 (2018), 1165–1182.

[62] Christian Schulze, Lisa Schöler, and Bernd Skiera. 2014. Not all fun and games: Viral marketing for utilitarian products. *Journal of Marketing* 78, 1 (2014), 1–19.

[63] S. H. Schwartz and J. Cieciuch. 2021. Measuring the refined theory of individual values in 49 cultural groups: psychometrics of the revised portrait value questionnaire. *Assessment* 29 (2021), 1005–1019. Issue 5. https://doi.org/10.1177/1073191121998760

[64] S. H. Schwartz, G. Melech, A. Lehmann, S. M. Burgess, M. Harris, and V. Owens. 2001. Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of Cross-Cultural Psychology* 32 (2001), 519–542. Issue 5. https://doi.org/10.1177/0022022101032005001

[65] Shalom H Schwartz and Galit Sagie. 2000. Value consensus and importance: A cross-national study. *Journal of cross-cultural psychology* 31, 4 (2000), 465–497.

[66] Rui Song, Zelong Liu, Xingbing Chen, Haining An, Zhiqi Zhang, Xiaoguang Wang, and Hao Xu. 2023. Label prompt for multi-label text classification. *Applied Intelligence* 53, 8 (2023), 8761–8775.

[67] Zhongwei Sun, Xiuyan Liu, Keyong Hu, Zhuang Li, and Jing Liu. 2020. An efficient multi-label SVM classification algorithm by combining approximate extreme points method and divide-and-conquer strategy. *IEEE Access* 8 (2020), 170967–170975.

[68] Mingjie Tang, Weichun Yang, Yeli Li, and Qingtao Zeng. 2023. Research on multi-label long text classification algorithm based on transformer-LDA. In *Fifth International Conference on Computer Information Science and Artificial Intelligence (CISAI 2022)*, Vol. 12566. SPIE, 992–999.

[69] Tomohiro Tasaki, Ryo Tajima, and Yasuko Kameyama. 2021. Measurement of the Importance of 11 Sustainable Development Criteria: How Do the Important Criteria Differ among Four Asian Countries and Shift as the Economy Develops? *Sustainability* 13, 17 (2021), 9719.

[70] Joan C. Timoneda. 2018. Where in the world is my tweet: Detecting irregular removal patterns on Twitter. *PLOS ONE* 13, 9 (Sept. 2018), e0203104. https://doi.org/10.1371/journal.pone.0203104

[71] Alicia Wanless and Jacob N Shapiro. 2022. A CERN Model for Studying the Information Environment.

[72] Christian Welzel and Ronald Inglehart. 2009. Political culture, mass beliefs, and value change. *Democratization* (2009), 126–144.

[73] J. d. Wet, D. Wetzelhütter, C. Nnebedum, and J. Bacher. 2022. Testing the relative comprehensiveness of schwartz's ten value types with help from rokeach. *Journal of Social Sciences* 18 (2022), 107–125. Issue 1. https://doi.org/10.3844/jssp.2022.107.125

[74] D. Wetzelhütter, C. Nnebedum, J. d. Wet, and J. Bacher. 2020. Testing a modified version of schwartz's portrait values questionnaire to measure organizational values in a university context. *Journal of Human Values* 26 (2020), 209–227. Issue 3. https://doi.org/10.1177/0971685820943398

[75] Michał Wróblewski and Andrzej Meler. 2024. Political polarization may affect attitudes towards vaccination. An analysis based on the European Social Survey data from 23 countries. *European Journal of Public Health* (2024), ckae002.

[76] Daniel J Wu, Andrew C Yang, and Vinay U Prabhu. 2020. Afro-MNIST: Synthetic generation of MNIST-style datasets for low-resource languages. *arXiv preprint arXiv:2009.13509* (2020).

[77] Diwen Xue, Reethika Ramesh, Valdik S S, Leonid Evdokimov, Andrey Viktorov, Arham Jain, Eric Wustrow, Simone Basso, and Roya Ensafi. 2021. Throttling Twitter: an emerging censorship technique in Russia. In *Proceedings of the 21st ACM Internet Measurement Conference*. ACM, Virtual Event, 435–443. https://doi.org/10.1145/3487552.3487858

[78] Pengcheng Yang, Fuli Luo, Shuming Ma, Junyang Lin, and Xu Sun. 2019. A deep reinforced sequence-to-set model for multi-label classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5252–5258.

[79] Ramil Yarullin and Pavel Serdyukov. 2021. Bert for sequence-to-sequence multi-label text classification. In *Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Skolkovo, Moscow, Russia, October 15–16, 2020, Revised Selected Papers 9*. Springer, 187–198.

[80] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, San Francisco USA, 218–226. https://doi.org/10.1145/3308560.3316495

[81] Yawen Zhang, Xiaohui Li, Pingli Gu, and Ke Zhang. 2023. Research on multi-label classification of Chinese judicial texts based on BERT-TextCNN. In *Third International Conference on Computer Vision and Data Mining (ICCVDM 2022)*, Vol. 12511. SPIE, 87–96.

[82] Ziqiang Zhong and Yaxin Li. 2023. Multi-label test classification based on capsule network and virtual adversarial training. In *Fourth International Conference on Artificial Intelligence and Electromechanical Automation (AIEA 2023)*, Vol. 12709. SPIE, 647–652.

[83] Mary Ellen Zurko. 2022. Disinformation and Reflections From Usable Security. *IEEE Security Privacy* 20, 3 (2022), 4–7. https://doi.org/10.1109/MSEC.2022.3159405