Understanding the Psychological Needs at Play in Disinformation

1st Kaveesh Khattar
Department of Computer Science Engineering
PES University
Bengaluru, India
kaveeshkhattar@gmail.com

2nd Bhaskarjyoti Das
Department of Computer Science Engineering
PES University
Bengaluru, India
Bhaskarjyoti01@gmail.com

***Abstract*—The work described in this paper explores the intersection of Computer Science, Linguistics, and Psychology in the realm of disinformation on social media. By analyzing textual content on Twitter, specifically propaganda, fake news, and rumors, we investigate the underlying motivations and patterns behind the production of disinformation. We apply the Human Needs Theory (HNT) to understand the psychological processes driving the emotions expressed in disinformation tweets. Our findings highlight distinct patterns across different types of disinformation, shedding light on the complex relationship between linguistic behavior and psychological factors.**

***Index Terms*—Psycholinguistics, Human Needs Theory, Disinformation**

I. Introduction

Psycholinguistics, a field within Natural Language Processing (NLP), combines linguistic behavior and psychological processes by studying the process of acquisition, comprehension, and formation of speech. The work described in this paper focuses on the 'production' aspect of this process.

Humanistic psychologists such as Abraham Maslow [1] and Manfred Max-Neef [2] proposed the theory that all human actions result from the need to meet certain basic needs or satisfiers. As explained by the Human Needs Theory (HNT) [3], the production of language is influenced by such psychological factors. Linguistic behavior, particularly in the context of disinformation on Twitter, can be analyzed to understand the underlying motivations that drive individuals to express themselves in certain ways. For example, emotions expressed in a tweet can be seen as an effect of the underlying motivations that serve as the causes. The work described in this paper focuses on capturing those underlying motivations.

There has been limited exploration of social network content from a psychological well-being perspective. Most studies have not delved into the psychological issues that drive individuals to engage in the spread of propaganda, rumors, and fake news on social media platforms. Therefore, the work described in this paper aims to shed light on the underlying psychological aspects of spontaneous disinformation by individual actors.

As human society has become largely online, disinformation has become an accepted phenomenon in modern lives. The three primary categories of disinformation are rumor (can be true as well as fake but not created with harmful intent), fake news (false news created with harmful intent), and propaganda (executed with a plan and strategy where a network of the perpetrators play a big role). Disinformation is multi-contextual [4] in nature. While other contexts such as temporal, spatial, network, propagation, etc. play a role in a disinformation campaign, the content context captures the outcome of the psychological factors of the audience in the play. Specifically, we are interested in exploring the psychological processes involved in producing textual content related to the above three types of disinformation i.e. propaganda, fake news, and rumors, on Twitter.

II. Related Work

Even though scholars have treated words as a window to people's minds throughout history, the research at the intersection of language and psychology is currently undergoing a metamorphosis [5] given the recent emergence of social media and fast evolution of language technologies. In general, there is a limited amount of existing work on the intersection of computer science, psychology, and linguistics specifically dealing with social media content.

Linguistic Inquiry and Word Count (LIWC) [6] is a word-level dictionary-based measurement tool that provides some psychological insights besides other measures. LIWC has been used in mental health studies [7], substance abuse intervention research [8], and often to provide a feature in feature engineering-driven disinformation research.

Personality traits prediction from language usage recently has seen a lot of research [9]–[12] in the recent past but the work suffers from the limited availability of labeled data. Ang Li et al. [13] did a psycholinguistic analysis of live stream suicide cases as a crisis prevention research. Suman Kalyan Maity et al. [14] used psycholinguistic methods to detect drunk texters on Twitter with close to 100% accuracy. The recent phenomena of the COVID-19 lockdown have created a huge psychological impact on the population affected by this pandemic and prompted researchers to undertake psycholinguistic studies [15] on social media data. Similarly, the discourses of world leaders on social media have been analyzed to assess their authenticity [16] using psycholinguistic techniques.

However, in disinformation research that has seen a lot of effort in recent years, there is a very limited amount of work

around psycholinguistics. Sabur Bhat et al. [17] have done a psycholinguistic analysis of the rumor data as contrasted with non-rumor data using LIWC, text readability, and emotions. Anastasia Giachanou et al. [18] similarly discriminated fake news spreaders and non-spreaders by using psycholinguistic features from LIWC along with emotion, sentiment, readability, personality traits, and contextualized word embedding.

The psycholinguistic approach adopted in this paper uses psychological need theories which however has seen some rare attempts. Specifically, in disinformation research, there has been no such existing work to the best of our knowledge. Rajwa Alharthi et al. [19] used psychological need theory to understand basic human needs during a crisis. The work described in this paper does the same for different categories of disinformation.

III. Dataset

TABLE I
DISTRIBUTION OF ROWS FOR EACH LABEL UNDER CATEGORY FEATURE

Dataset	Category			Total
	Autonomy	Competence	Relatedness	
Original	1417	983	2667	5067
Balanced	5334	5334	5334	16002

TABLE II
DISTRIBUTION OF ROWS FOR EACH LABEL UNDER SATISFACTION FEATURE

Dataset	Satisfaction				Total
	Satisfied	Dissatisfied	Neutral	Not Clear	
Original	2124	2636	114	193	5067
Balanced	2636	2636	2636	2636	10544

TABLE III
DISTRIBUTION OF ROWS FOR EACH LABEL UNDER ENVIRONMENT FEATURE

Dataset	Environment			Total
	Supportive	Non-Supportive	Not Clear	
Original	758	1548	2761	5067
Balanced	2761	2761	2761	8283

There are two existing social media datasets for Psychological needs. H Yang et al. [20] from IBM Research has published one dataset and Rajwa Alharthi et al. [21] has published another. For the work described in this paper, the IBM Research dataset has not been used as it was meant for the specific domain of consumer behavior, The second dataset is more general purpose in nature and it provides a multi-layer framework grounded on research in human needs theory.

The used dataset of psychological human needs presents a multi-layer framework :

1) Layer 1 at the top depicts emotion which is an effect based on psychological needs. The labels for this layer are not provided in the dataset and are also not used in our work.
2) Layer 2 represents the basic needs of three categories i.e. autonomy, competence, and relatedness.
3) Layer 3 represents the need satisfaction level i.e. satisfied, dissatisfied, and neutral.
4) Layer 4 represents the supportive role played by the social context as perceived by and expressed in the social media content i.e. supportive, non-supportive, and not clear.
5) Level 5 represents the life aspect of the individual in context i.e. social relation, work, and education. This particular level is not used in our work as for a particular disinformation campaign, this is expected to be common for all participants in a conversation.

The used dataset is multi-label in nature though the labels are not hierarchical. The dataset is small consisting of only 5067 labeled samples and is imbalanced (not having an equal number of labels in each of the 5 layers described above). It is hard to do any supervised machine learning with such a small imbalanced dataset. Hence, suitable data augmentation strategies have been adopted to make a balanced dataset. Table I, Table II, and Table III represent the number of labeled samples provided in the original dataset and the balanced version created for basic need category, satisfaction level and supporting role by the environment. The original imbalanced dataset of 5067 samples is transformed (explained in the next section)into a balanced dataset of 34,829 samples. The average word count of the balanced dataset (used for training) was around 17.

IV. Methodology

The methodology consists of several key steps i.e. addressing data imbalance and limited volume of the labeled data, data pre-processing, feature extraction, and model selection.

A. Imbalanced learning

To address the issue of class imbalance within the dataset and the limited quantity of labeled data, data augmentation has been done. Data augmentation approaches can be roughly divided into two categories [22] i.e. the methods that work in the data space (raw input data) and the methods that work in the feature space. Feature space-oriented methods such as SMOTE adopt interpolation of the feature space representation of input data. Data space-oriented methods can work at the character level, phrase level, word level, and document level. For the work described in this paper, data space-oriented data augmentation approach has been adopted. This technique involves replacing words in the text with synonyms from WordNet, a lexical database that groups words into sets of synonyms, thereby augmenting the dataset and achieving a more balanced distribution across different classes.

B. Data pre-processing

To ensure the quality and consistency of the dataset, a series of pre-processing steps were applied. First, the lower casing was performed on the text to standardize the case of all words, reducing potential discrepancies due to capitalization. Next,

the removal of URLs was carried out to eliminate any web links present in the data, as they do not contribute to the content analysis. Subsequently, HTML tags were removed to eliminate any residual markup language present in the text. Punctuation marks were also removed to focus solely on the textual content's semantic meaning. Additionally, stop words, such as common articles and conjunctions, were removed to reduce noise in the dataset. Contraction fixing was applied to expand contracted words, ensuring consistency in language representation. Finally, lemmatization was performed to reduce words to their base or root form, facilitating semantic analysis. To facilitate natural language processing (NLP) tasks, the spaCy library was utilized. Its capabilities were leveraged to enhance the pre-processing steps, including tokenization, part-of-speech tagging, and named entity recognition.

C. Feature extraction and learning algorithms

For the machine learning task, both traditional and deep learning approaches have been used. The Train, validation, and Test split have been in the ratio of 80:10:10.

After a round of initial evaluation, two traditional machine learning algorithms i.e. Naive Bayes (NB) and Support Vector Machines (SVM) along with an ensemble learning algorithm i.e. XGBoost, were chosen as baseline estimators to evaluate the performance of the models. For feature extraction, both statistical approaches such as Term Frequency-Inverse Document Frequency (TFIDF), and distributional semantics-based methods such as Facebook's FastText are used. FastText is preferred to Word2Vec due to its capability to handle out-of-vocabulary words and character n-gram-based approach.

For the deep learning-based model, two approaches have been used. A lightweight Python library (Ktrain) that uses DistilBERT (distilbert-base-uncased) [23] and built on top of TensorFlow and Keras, was utilized. The spaCy library with Hugging Face Transformer model that uses 'bert-base-cased' has been the alternative pursued for the deep learning approach.

D. Metrics

Accuracy score has been used to compare the performance of different learning algorithms. Additionally, for the deep learning models, Cohen's Kappa Score and Matthews Correlation Coefficient have been calculated. Since traditional accuracy metrics are sensitive to class imbalance, Matthews Correlation Coefficient offers a better approach essentially measuring the correlation between true and predicted value. Cohen's Kappa Score: measures the agreement between two raters, in this case, the predicted labels and the true labels

V. Results and Discussion

A. Results

Table IV, Table V and Table VI show the performance measures for various learning models. The DistilBERT-based model provides the best performance and is the best-performing model. In this context, a score of 0.69 for both Cohen's Kappa and MCC indicates a moderately strong level of agreement or prediction quality. This suggests that the model's predictions align with the true labels in a relatively consistent manner.

TABLE IV
Model Evaluation - Accuracy

Model	Category	Satisfaction	Environment
TFIDF+NB	0.79	0.78	0.79
TFIDF+SVM	0.82	0.80	0.82
TFIDF+XGBoost	0.75	0.72	0.74
FastText	0.76	0.79	0.70
spaCy/en-core-web-trf	0.75	0.80	0.74
kTrain-DistilBERT	**0.81**	**0.89**	**0.70**

TABLE V
Model Evaluation - Cohen's Kappa Score

Model	Category	Satisfaction	Environment
kTrain-DistilBERT	0.6930	0.7969	0.4887
spaCy/en-core-web-trf	0.6975	0.7951	0.4606

TABLE VI
Model Evaluation - Matthew's Correlation Coefficient

Model	Category	Satisfaction	Environment
kTrain-DistilBERT	0.6938	0.7995	0.4892
spaCy/en-core-web-trf	0.6986	0.7984	0.4625

ktrain offers an interpretability function that employs the LIME (Local Interpretable Model Agnostic Explanations) to visualize important features contributing to machine learning model predictions. The fig. 1 shows a test sample where the basic need is predicted as 'relatedness'. The green colors means the token has a positive contribution to the predicted label and BIAS in red means negative contribution. The color shade indicates feature significance. A probability of 0.992 indicates high confidence in the predicted outcome i.e. the basic need of relatedness. A score of 4.333 represents the model's confidence or prediction strength, varying by context and model. Ultimately, the interpretation of these scores should be considered alongside other evaluation metrics, domain knowledge, and research objectives.

B. Discussion

The best-performing (k-train DistilBERT) model has been applied to several representative datasets as a way to illustrate possible usage of the work done. The model can be used to either investigate a single domain or compare several domains.

- Rumour : Pheme [24] Dataset.
- Fake News : WELFake Dataset [25] and Dataset published by University of Victoria [26] (available at Kaggle). These two datasets are combined.
- Propaganda Dataset : Dataset published [27] by Liqiang Wang et al. In this, there are two datasets i.e. one is from news articles and the other is from Twitter. These two datasets are combined.

Appropriate normalization has been adopted for comparison across various types of disinformation due to difference in dataset size. The fig. 2 shows the basic needs captured in

y=Relatedness (probability **0.992**, score **4.333**) top features

Contribution?	Feature
+4.290	Highlighted in text (sum)
+0.043	<BIAS>

this is your daily reminder that barack obama asked ukraine to investigate his political rival is campaign manager democrat senators asked ukraine to investigate trump and the dnc solicited ukraine is help to dig up dirt on trump and the media was silent about all of it

Fig. 1. Interpretability example of Basic Need

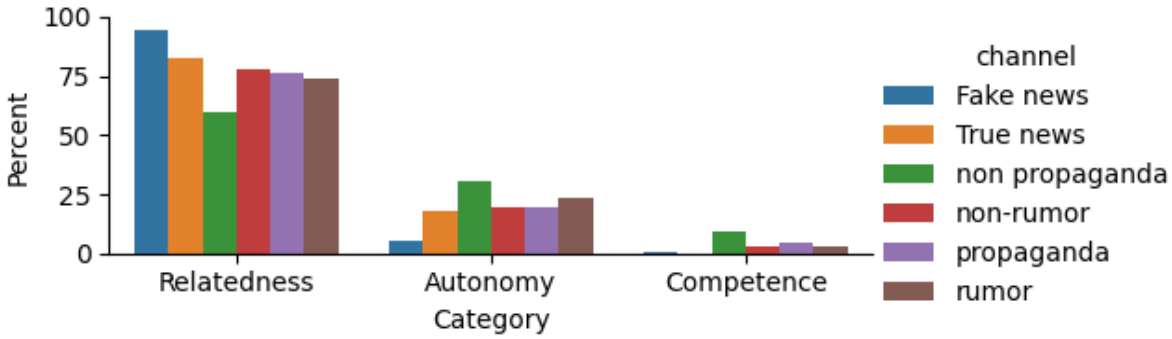

Fig. 2. Basic Needs expressed across disinformative and normal conversation

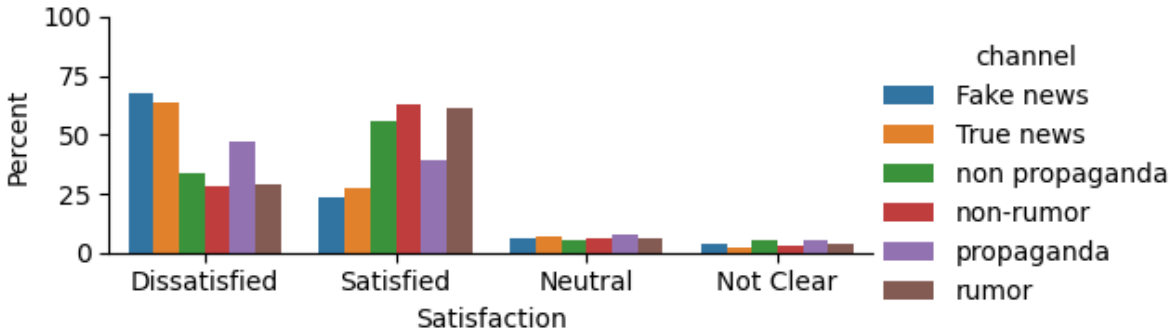

Fig. 3. Satisfaction expressed across disinformative and normal conversation

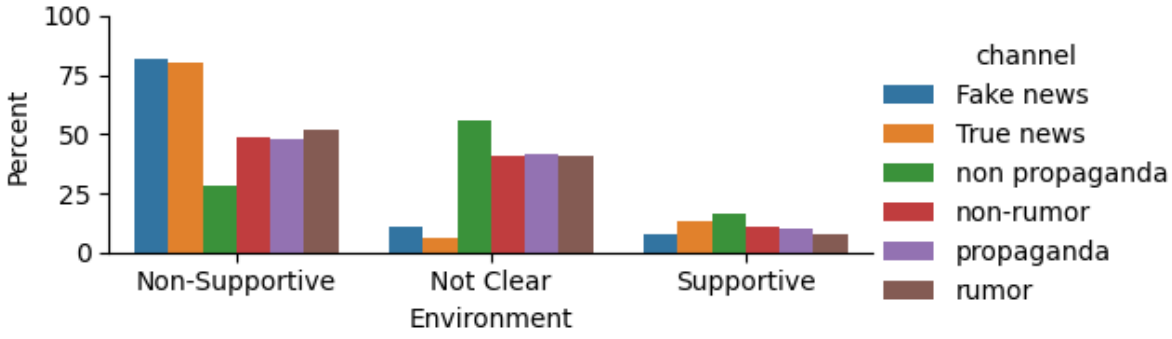

Fig. 4. Perceived supportive role played by environment across disinformative and normal conversation

different types of communication. The fig. 3 shows the satisfaction expressed and fig. 4 shows the supportive role the environment is perceived to be playing.

Fake News show the least satisfaction and maximum dissatisfaction. Non-rumour shows maximum satisfaction being plain fact-based news. Rumour is caused by anxiety to address the need for information. So, as compared to non-rumor, it shows less satisfaction but satisfaction is still at a higher level than that of fake news. A similar conclusion can be derived from propaganda and non-propaganda. Among the basic needs, relatedness is the main reason driving people on social media, followed by a need for autonomy. Competence is not such a pressing need. It is also clear that most people go to social media to vent about the lack of support from the environment towards their basic needs.

VI. Conclusion and Next Steps

In this work, the intricate relationship between linguistic behavior and psychological processes has been explored and a suitable application in the context of disinformation on social media has been showcased. The work builds on Human Needs Theory (HNT) which has remained unexplored to a large extent in psycholinguistic use case like this and this work has successfully demonstrated the possibility such research presents. Though the lack of labeled data is a challenge in supervised learning, suitable data augmentation techniques have been used to come up with a high-performing model while considering many possible options.

As a next step, we plan to enhance this work further in the domain of disinformation research by investigating the possible cause-effect patterns and doing a suitable causality analysis behind why people behave a certain way on social media.

References

[1] M. A. Wahba and L. G. Bridwell, "Maslow reconsidered: A review of research on the need hierarchy theory," *Organizational behavior and human performance*, vol. 15, no. 2, pp. 212–240, 1976.

[2] M. Max-Neef, A. Elizalde, and M. Hopenhayn, "Development and human needs," *Real-life economics: Understanding wealth creation*, vol. 197, p. 213, 1992.

[3] D. J. Christie, "Reducing direct and structural violence: The human needs theory," *Peace and Conflict*, vol. 3, no. 4, pp. 315–332, 1997.

[4] B. Das *et al.*, "Multi-contextual learning in disinformation research: A review of challenges, approaches, and opportunities," *Online Social Networks and Media*, vol. 34, p. 100247, 2023.

[5] R. L. Boyd and H. A. Schwartz, "Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field," *Journal of Language and Social Psychology*, vol. 40, no. 1, pp. 21–41, 2021.

[6] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[7] M. Spruit, S. Verkleij, K. de Schepper, and F. Scheepers, "Exploring language markers of mental health in psychiatric stories," *Applied Sciences*, vol. 12, no. 4, p. 2179, 2022.

[8] M. Partch and C. Dykeman, "A corpus linguistics study of text message interventions in substance use disorder treatment," 2019.

[9] E. Kerz, Y. Qiao, S. Zanwar, and D. Wiechmann, "Pushing on personality detection from verbal behavior: A transformer meets text contours of psycholinguistic features," *arXiv preprint arXiv:2204.04629*, 2022.

[10] Y. Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, and S. Eetemadi, "Bottom-up and top-down: Predicting personality with psycholinguistic and language model features," in *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 1184–1189, IEEE, 2020.

[11] A. Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, and E. Cambria, "Personality trait detection using bagged svm over bert word embedding ensembles," *arXiv preprint arXiv:2010.01309*, 2020.

[12] M. H. Amirhosseini and H. Kazemian, "Machine learning approach to personality type prediction based on the myers–briggs type indicator®," *Multimodal Technologies and Interaction*, vol. 4, no. 1, p. 9, 2020.

[13] A. Li, D. Jiao, X. Liu, J. Sun, and T. Zhu, "A psycholinguistic analysis of responses to live-stream suicides on social media," *International journal of environmental research and public health*, vol. 16, no. 16, p. 2848, 2019.

[14] S. K. Maity, A. Mullick, S. Ghosh, A. Kumar, S. Dhamnani, S. Bahety, and A. Mukherjee, "Understanding psycholinguistic behavior of predominant drunk texters in social media," in *2018 IEEE Symposium on Computers and Communications (ISCC)*, pp. 01096–01101, IEEE, 2018.

[15] Y. Su, J. Xue, X. Liu, P. Wu, J. Chen, C. Chen, T. Liu, W. Gong, and T. Zhu, "Examining the impact of covid-19 lockdown in wuhan and lombardy: a psycholinguistic analysis on weibo and twitter," *International journal of environmental research and public health*, vol. 17, no. 12, p. 4552, 2020.

[16] S. Figueiredo, M. Devezas, N. Viera, and A. Soares, "A psycholinguistic analysis of world leaders? discourses concerning the covid-19 context: Authenticity and emotional tone," *International Journal of Social Sciences*, vol. 9, no. 2, pp. 66–69, 2020.

[17] S. Butt, S. Sharma, R. Sharma, G. Sidorov, and A. Gelbukh, "What goes on inside rumour and non-rumour tweets and their reactions: A psycholinguistic analyses," *Computers in Human Behavior*, vol. 135, p. 107345, 2022.

[18] A. Giachanou, B. Ghanem, E. A. Ríssola, P. Rosso, F. Crestani, and D. Oberski, "The impact of psycholinguistic patterns in discriminating between fake news spreaders and fact checkers," *Data & knowledge engineering*, vol. 138, p. 101960, 2022.

[19] R. Alharthi, B. Guthier, and A. El Saddik, "Recognizing human needs during critical events using machine learning powered psychology-based framework," *IEEE Access*, vol. 6, pp. 58737–58753, 2018.

[20] H. Yang and Y. Li, "Identifying user needs from social media," *IBM Research Division, San Jose*, vol. 11, 2013.

[21] R. Alharthi, B. Guthier, C. Guertin, and A. El Saddik, "A dataset for psychological human needs detection from social networks," *IEEE Access*, vol. 5, pp. 9109–9117, 2017.

[22] M. Bayer, M.-A. Kaufhold, and C. Reuter, "A survey on data augmentation for text classification," *ACM Computing Surveys*, vol. 55, no. 7, pp. 1–39, 2022.

[23] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[24] "Pheme rumor dataset." Available at https://www.kaggle.com/datasets/nicolemichelle/pheme-dataset-for-rumour-detection.

[25] "Welfake dataset." Available at https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification.

[26] "Fake news dataset published by university of victoria." Available at https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset.

[27] L. Wang, X. Shen, G. de Melo, and G. Weikum, "Cross-domain learning for classifying propaganda in online contents," in *Conference for Truth and Trust Online 2020 (TTO)*, 2020.