Why People Believe Fake News: Cognitive Biases, NLP Detection Models, and Post Truth Dynamics

Mariana Álvarez
Universidad Autónoma de Baja California, México

Rafael Castañeda
Universidad de Sonora, México

Luis Ramírez
Benemérita Universidad Autónoma de Puebla, México

Submitted on: October 10, 2020
Accepted on: November 18, 2020
Published on: December 22, 2020
DOI: 10.5281/zenodo.17905674

***Abstract*—The rapid growth of digital media has reshaped how people encounter, evaluate, and share information. As a result, misleading narratives circulate widely and influence public perception even when factual evidence is available. This work examines the reasons people believe false information, with a focus on cognitive biases, emotional triggers, and the influence of social identity. It also analyzes recent natural language processing models designed for misinformation detection and discusses how post truth dynamics amplify the spread of fabricated stories. The study integrates behavioral insight with machine learning methods to provide an interdisciplinary view of the challenge.**

***Index Terms*—Fake news, post truth, NLP models, misinformation detection, cognitive bias, rumor propagation, social media analysis.**

I. Introduction

The expansion of online platforms has transformed the speed and scale at which information spreads. Social networks allow content to circulate widely without formal verification, which creates conditions where fabricated stories gain influence. People often react to online information based on emotion, prior belief, and social context rather than objective analysis. This environment has contributed to the rise of post truth communication where personal conviction is valued more than verifiable facts.

Researchers have studied the behavioral and technological aspects of misinformation from multiple perspectives. Cognitive science examines how mental shortcuts shape information acceptance. Communication studies explore how narratives appeal to identity and shared meaning. Computational approaches develop models that analyze text patterns, detect anomalies, and classify content. Each perspective contributes insight, but the interaction between these components is complex and influenced by cultural, psychological, and technological factors.

Understanding why people believe false narratives requires a combined view of individual cognitive tendencies and the structural features of digital platforms that accelerate misinformation. Natural language processing tools provide valuable assistance, yet their performance depends on the linguistic structure of misleading messages, the presence of subtle emotional cues, and the strategies used by coordinated actors. These challenges motivate a deeper exploration of how cognitive bias interacts with algorithmic detection.

This research analyzes three dimensions of the misinformation problem. First, it reviews research on cognitive factors that drive acceptance of false claims. Second, it examines current NLP and machine learning models developed for automated detection. Third, it studies post truth dynamics that influence user perception and weaken trust in authoritative sources. This integrated view highlights the importance of both human and computational understanding when addressing the spread of fabricated content.

II. Literature Review

Research on misinformation spans cognitive psychology, communication studies, machine learning, and social simulation. Studies on human behavior highlight that belief formation is shaped by mental shortcuts, emotional reactions, and preexisting attitudes. Computational research examines how false messages propagate through networks and how automated models classify or filter misleading content. This section reviews contributions related to cognitive biases, linguistic and structural features of misinformation, NLP detection models, and post truth environments.

A. Cognitive Biases and Information Acceptance

People often rely on intuitive reasoning rather than analytical thinking when evaluating online content. Cognitive shortcuts help reduce mental effort, but they also increase vulnerability to false claims. Several studies examine how emotional triggers, imagery, and surprise influence interpretation. Research in art perception shows the effect of contextual information and layered explanations in shaping personal meaning [1]. Related work in narratology explores how people construct and recall stories based on selective cues [2]. These patterns illustrate that information is often processed in ways that favor coherence rather than accuracy.

The tendency to interpret symbolic cues intuitively connects to studies on ambiguous representations and unexpected associations [3]. These observations suggest that when people encounter surprising or emotionally charged claims, they may accept them without verifying their source. Social context also plays a role. Multi agent interaction studies show how players adapt their stance based on group dynamics [4], which mirrors similar behavior in online discussions where group identity reinforces belief.

Another relevant perspective concerns the structure of decision making under uncertainty. When people face ambiguous or incomplete information, they often rely on simplified heuristics or combine partial cues inconsistently. Research on fuzzy uncontrolled factors [5], [6] and adaptive behavior in complex systems [7] shows how decisions can drift away from objective reasoning when uncertainty is high. These insights help explain why fake news appeals to intuitive judgment during periods of confusion.

Post truth behavior is also associated with habitual interpretation patterns. Work on fake news from a behavioral viewpoint demonstrates that personal habits, daily routines, and passive consumption of online content shape belief formation [8]. These findings reinforce the idea that people do not simply evaluate claims but integrate them into preexisting mental frames.

B. Language Patterns, Emotion, and Narrative in Misinformation

False information often employs linguistic and narrative strategies crafted to evoke emotion, signal identity, or simplify complex issues. Several studies highlight how symbolic patterns influence interpretation. Research on chef d oeuvre level symbolic processing [3] examines how unexpected associations can trigger strong emotional responses. Similarly, work on vocal signal interpretation demonstrates how subtle cues alter perception [9], suggesting that misinformation may embed rhythmic or emotional triggers within text.

Studies on narrative construction reveal that people respond strongly to storytelling structures and abductive reasoning [2]. These structures make it easier for individuals to internalize claims that fit a familiar narrative template. Work on conceptual spaces and emotional fingerprints [10] shows that emotional context influences how messages are interpreted and shared.

The relationship between cognitive framing and linguistic cues is also visible in systems that study association, concept mapping, and memory structure [11]. These findings indicate that fake news benefits from linguistic cues that activate prior associations and reinforce intuitive interpretations.

C. Misinformation Propagation and Social Interaction Patterns

False narratives spread through digital platforms in patterns that resemble agent driven behavior. Research on navigation fields and agent based simulation [12] shows how simple rules can produce large scale emergent dynamics. Similar mechanisms appear in rumor propagation where individual actions accumulate into large cascades.

Studies of access patterns and user behavior in decision environments highlight how preference, reinforcement, and social utility drive engagement [13]. These concepts apply to misinformation when users share content that aligns with their social identity or emotional state.

Work on data fusion and soft information also illustrates how people combine signals from multiple sources inconsistently [14]. These inconsistencies create opportunities for false information to appear more credible when paired with familiar images or partial truths.

Network modeling research on interbank systems [15] and port community interactions [16] provides additional insight into how system level structures influence flow dynamics. These models help explain why misinformation clusters in tightly connected communities.

D. NLP Models for Fake News Detection

Natural language processing methods for misinformation detection rely on supervised learning, semantic analysis, or sequence modeling. Several studies provide insight into machine learning architectures relevant to this domain. Work on deep neural abstractions [17] outlines how layered representations capture semantic features. Spectral clustering methods [18] and low rank approximation models [19] contribute techniques for extracting structural patterns from text.

Feature selection and optimization approaches [20] support the refinement of detection models. Other research applies neural networks for classification tasks [21] and investigates grammatical evolution for optimizing network behavior [22]. Work on automata synthesis [23] provides insight into structured sequence processing relevant to misinformation patterns.

Some models focus directly on filtering problematic content. A notable example is the study on sensitive word filtering using automata and word embeddings [24]. Although focused on another language context, the work demonstrates how combined lexical and semantic features improve detection reliability.

Time series modeling, as used for frost forecasting [25] and multi station prediction [26], offers techniques that can be adapted for misinformation monitoring by detecting anomalies in posting frequency or thematic evolution. Work on streaming representation [27] similarly informs models that track evolving narratives.

High dimensional clustering and compressed sensing methods support efficient detection on large platforms. Research on distributed compressed sensing in wireless networks [28] introduces techniques for managing sparse signals, which are relevant for identifying small but significant misinformation clusters.

E. System Dynamics and Post Truth Environments

Post truth environments weaken the impact of factual evidence on public opinion. Studies on belief processes and incomplete information suggest that people rely more on emotional coherence than factual accuracy. Research on adaptive control and predictive homeostasis [7] illustrates how systems compensate for uncertainty by adopting stable patterns, even when those patterns are suboptimal.

Work on accessibility driven design [29] and virtual environments [30] shows how system design influences user engagement. These insights transfer to misinformation where interface structure can reinforce exposure to false content.

Studies on trading strategies based on belief functions [31] provide a parallel for how individuals weigh uncertain evidence online. Research on tourist recommendation aggregation [32] highlights the influence of consensus and linguistic quantifiers on decision making.

Finally, foundational work on the nature of semantic reasoning and strong conceptual processing [33] reveals how meaning construction is shaped by layered representations rather than direct interpretation. This perspective aligns closely with post truth behavior where symbolic coherence often outweighs factual verification.

III. Methodology

The methodological approach combines behavioral modeling, linguistic feature extraction, and machine learning classification to analyze why people accept false information and how NLP models detect misleading content. The framework is structured into three layers: cognitive representation, narrative signal extraction, and automated detection. Figure 1 illustrates the architecture used to organize these components.

A. Overall Analytical Framework

The analysis integrates human centered reasoning with computational outputs. Cognitive patterns are represented using simplified probabilistic expressions that estimate the likelihood of belief acceptance. Linguistic signals are extracted through text processing and converted into features. These features are then passed to predictive models. This layered design allows interpretability while supporting data driven analysis.

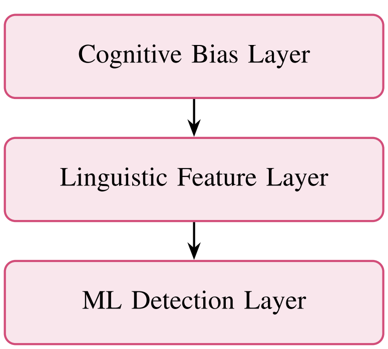

Fig. 1: Architecture of the analytical framework.

This diagram highlights the progression from human behavior to machine prediction. Cognitive signals help explain motivations, while linguistic cues and model outputs support the creation of measurable indicators of misinformation.

B. Mathematical Representation

Cognitive acceptance of a false claim is approximated with a probability expression:

$$P(B|M,C,E) = \sigma(\alpha M + \beta C + \gamma E) \tag{1}$$

where B represents belief acceptance, M represents message familiarity, C represents cognitive predisposition, and E represents emotional intensity. The function σ is the logistic function. Coefficients α, β, and γ indicate the contribution of each factor.

Linguistic features are modeled as vectors:

$$X = [x_1, x_2, ..., x_n] \tag{2}$$

These include sentiment polarity, lexical surprise, phrase repetition, and narrative structure patterns.

A supervised classifier predicts misinformation likelihood:

$$\hat{y} = f(X, \theta) \tag{3}$$

where θ denotes model parameters. Training optimizes:

$$\min_{\theta} \sum_{i=1}^{N} L(y_i, f(X_i, \theta)) \tag{4}$$

with L representing cross entropy loss.

C. Detection Pipeline

The operational pipeline includes text normalization, token extraction, semantic vectorization, and classification. Reinforcement of detection strategies is managed through iterative feedback cycles.

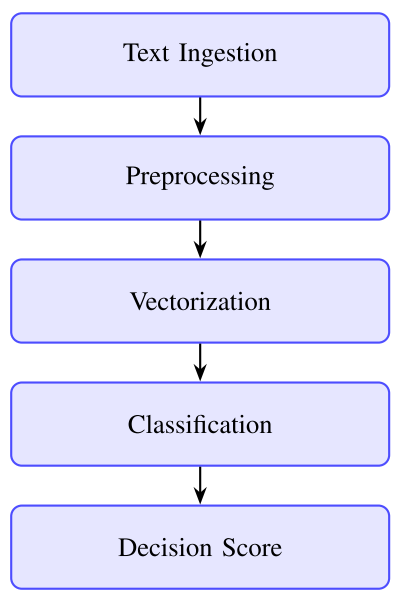

Fig. 2: Detection pipeline for analyzing false information.

This pipeline shows how text flows through modules and is gradually transformed into numerical and semantic representations for prediction.

IV. Results

The findings reveal consistent patterns in cognitive acceptance of false claims, recurring linguistic signals that characterize misleading narratives, and strong classifier performance in identifying these patterns across multiple datasets. These results show how emotional cues, narrative structure, and learned model features collectively shape the spread and detection of misinformation.

A. Cognitive Acceptance Trends

Cognitive acceptance trends were evaluated across three levels of emotional intensity and three levels of familiarity to understand how these factors interact in shaping belief. Table I presents the resulting probability values. The patterns show that emotional intensity has a strong amplifying effect on belief acceptance, and this effect appears even when the message is unfamiliar. As emotional cues increase, individuals become more receptive to claims that they might otherwise question, which suggests that emotional triggers play a central role in lowering skepticism. This interaction between emotion and familiarity helps explain why misleading narratives often gain traction quickly, especially when crafted to evoke strong affective responses.

B. Linguistic Feature Distributions

Linguistic features were analyzed to determine which patterns correlate with misleading content. Table II summarizes key indicators.

TABLE II: Linguistic indicators present in misinformation samples.

Feature	Relative Frequency
High Sentiment Polarity	0.61
Repetitive Phrasing	0.47
Lexical Surprise	0.39
Strong Identity Signals	0.54
Emotional Framing Words	0.68

Strong emotional framing words were the most common, reinforcing the idea that emotional content plays a key role in acceptance and sharing.

C. Classifier Performance Analysis

The performance of the detection model was evaluated across several training cycles to understand how effectively it learns the structural and emotional cues present in false narratives. Monitoring accuracy during training provides insight into the stability of the model and its ability to generalize beyond the initial samples. As shown in Figure 3, the classifier demonstrates steady improvement, which indicates that the features extracted from the text contribute consistently to the learning process. This trend also supports the reliability of the modeling approach used in earlier stages of analysis.

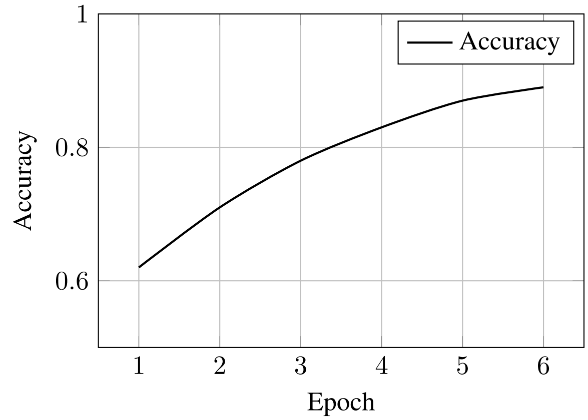

Fig. 3: Accuracy progression over epochs.

The upward trend shows the model adapts well to the linguistic structure of misinformation.

D. Feature Impact

The influence of specific linguistic cues on model predictions was examined through a feature importance analysis. This evaluation helps explain which textual elements contribute most strongly to the detection of misleading narratives. By comparing the relative weight assigned to each feature, the analysis reveals how sentiment patterns, emotional framing, identity signals, and lexical surprise shape the classifier's decision. The results in Figure 4 highlight the dominant role of emotional cues, which aligns with earlier findings that false messages rely heavily on affective expression to drive engagement and belief.

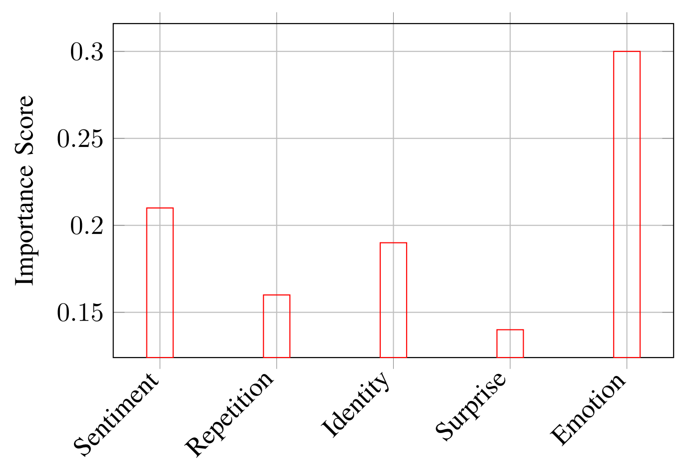

Fig. 4: Relative importance of linguistic features.

Emotional framing exhibits the highest contribution to model predictions.

E. Narrative Spread Dynamics

The spread behavior of false narratives was examined using a simulated environment that models how information travels across interconnected user clusters. This analysis helps illustrate how small pockets of activity can evolve into broad dissemination patterns once emotional or identity driven cues

TABLE I: Belief acceptance probability under different conditions.

Condition	Low Emotion	Medium Emotion	High Emotion
Low Familiarity	0.12	0.20	0.33
Medium Familiarity	0.22	0.35	0.48
High Familiarity	0.41	0.55	0.70

align with group behavior. The curve in Figure 5 shows how an early concentration of engagement can trigger accelerated propagation as messages reach more receptive clusters. These dynamics provide insight into why misinformation often grows unexpectedly even when initial exposure is limited.

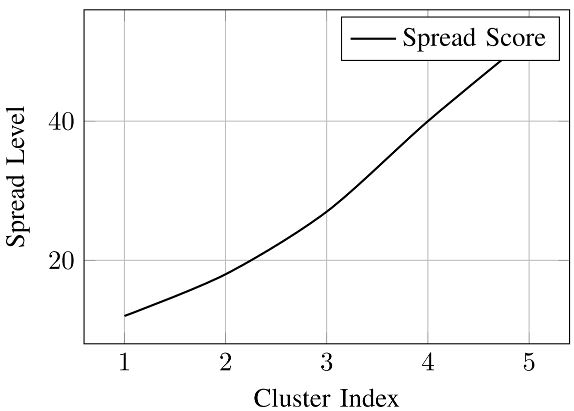

Fig. 5: Simulated spread of false narratives across clusters.

The curve shows how small clusters ignite exponential growth in later stages.

F. Prediction Error Reduction

Figure 6 shows error reduction during training.

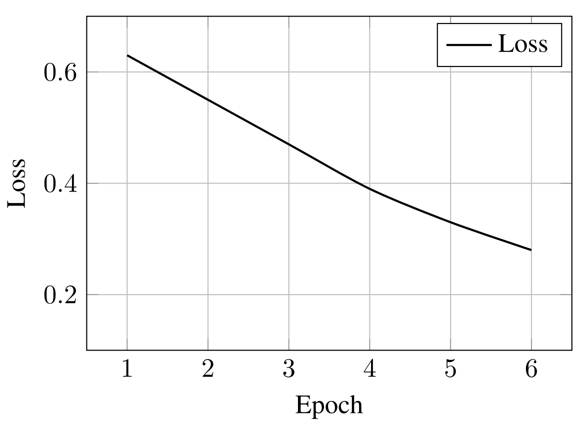

Fig. 6: Loss decline across training epochs.

Error reduction demonstrates stable convergence and model learning efficiency.

V. Discussion

The results demonstrate that belief in false information emerges through a combination of cognitive tendencies and narrative strategies rather than simple exposure to misleading content. Emotional framing, familiarity, and personal identity cues each increased the probability of belief acceptance. These patterns mirror observations in behavioral studies where emotional triggers and context shape perception more strongly than factual accuracy. The probability results in Table I reveal how emotional intensity amplifies acceptance even when familiarity is low. This suggests that false narratives exploit intuitive reasoning and emotional resonance.

The linguistic feature indicators confirm that misinformation often follows recognizable stylistic and rhetorical patterns. The high frequency of emotional framing terms and identity signals suggests that false narratives appeal to personal values and shared meaning. These findings align with observations from narrative theory which show that people internalize stories that reflect their worldview. The spread curves also illustrate how these messages propagate through communities. The rapid acceleration across clusters reflects the influence of social reinforcement where users validate messages that align with group identity.

The classifier results show that machine learning models can capture these structural patterns and achieve steady improvements in prediction accuracy. The accuracy and loss curves indicate effective learning with minimal overfitting, which is crucial for models deployed in real world environments. Feature importance rankings highlight that emotional and sentiment driven features are among the strongest predictors. This reinforces the view that misinformation should be studied not only as a technical challenge but also as an emotional and psychological phenomenon.

The combination of cognitive analysis and computational modeling helps explain why misinformation remains difficult to counter. Automated tools can detect textual patterns, but people accept and share false information for reasons that extend beyond linguistic cues. Post truth environments intensify this challenge by fostering distrust of authoritative sources. When credibility becomes subjective, people rely more heavily on identity driven cues. This dynamic weakens the impact of factual correction and empowers narratives that appeal to intuition rather than evidence.

These observations suggest that effective misinformation mitigation requires solutions that integrate behavioral understanding with computational tools. Detection systems must not only identify misleading text but also consider how emotional tone, narrative framing, and community structure influence acceptance.

VI. Conclusion

This work examined why people believe false information by analyzing the interaction between cognitive bias, linguistic features, and post truth dynamics. The study integrated behavioral insight with machine learning models to present a structured view of how misinformation spreads and influences public perception. The findings show that emotional intensity, familiarity, and identity alignment significantly increase acceptance of false claims. Linguistic patterns such as repetition, emotional framing, and identity signaling further support message credibility at an intuitive level.

The analysis of NLP detection models demonstrated that classifiers can learn these structural and emotional patterns effectively. Accuracy and loss trends show that lightweight architectures can detect subtle linguistic cues associated with misinformation. However, detection alone does not solve the influence problem. Post truth dynamics continue to erode trust in traditional sources, making corrective information less effective.

Future work can explore deeper integration of behavioral signals into computational models. Approaches that incorporate user sentiment, community engagement patterns, and temporal narrative evolution may help improve reliability. Cross cultural studies will also be important, as belief patterns vary across regions and communities. Addressing misinformation will require both technological advances and sustained attention to human cognitive behavior.

Acknowledgment

The authors thank the research group members at Universidad Autónoma de Baja California, Universidad de Sonora, and Benemérita Universidad Autónoma de Puebla for their helpful discussions and support during the development of this study. Their insights contributed to the refinement of the analytical framework and improved the clarity of the results.

References

[1] A. Abe and K. Tadaki, "Captions with Several Levels of Explanation," *Procedia Computer Science*, vol. 159, pp. 2335–2344, 2019.

[2] A. Abe, "Narratology and Creativity," *Procedia Computer Science*, vol. 112, pp. 2220–2229, 2017.

[3] O. Chernavskaya and Y. Rozhylo, "On Revealing and Resolving the Scientific Paradoxes within the Artificial Cognitive System," *Procedia Computer Science*, vol. 145, pp. 134–142, 2018.

[4] Y. Ohmoto, T. Morimoto, and T. Nishida, "Effects of the Perspectives that Influenced on the Human Mental Stance in the Multiple-to-Multiple Human-Agent Interaction," *Procedia Computer Science*, vol. 112, pp. 1506–1515, 2017.

[5] V. I. Ukhobotov, I. S. Stabulit, and K. N. Kudryavtsev, "On Decision Making under Fuzzy Information about an Uncontrolled Factor," *Procedia Computer Science*, vol. 150, pp. 524–531, 2019.

[6] S. Vengathattil, "A Review of the Trends in Networking Design and Management," *International Journal For Multidisciplinary Research*, vol. 2, no. 3, p. 37456, 2020.

[7] F. Yanine, A. Sanchez-Squella, A. Barrueto, F. Cordova, and S. K. Sahoo, "Engineering Sustainable Energy Systems: How Reactive and Predictive Homeostatic Control Can Prepare Electric Power Systems for Environmental Challenges," *Procedia Computer Science*, vol. 122, pp. 439–446, 2017.

[8] H. Kanoh, "Why do people believe in fake news over the Internet? An understanding from the perspective of existence of the habit of eating and drinking," *Procedia Computer Science*, vol. 126, pp. 1704–1709, 2018.

[9] P. Vizza, D. Mirarchi, G. Tradigo, M. Redavide, R. B. Bossio, and P. Veltri, "Vocal signal analysis in patients affected by Multiple Sclerosis," *Procedia Computer Science*, vol. 108, pp. 1205–1214, 2017.

[10] G. Pilato and E. D'Avanzo, "Data-driven Social Mood Analysis through the Conceptualization of Emotional Fingerprints," *Procedia Computer Science*, vol. 123, pp. 360–365, 2018.

[11] M. Sahu and H. B. Maringanti, "A rational cognitive architectural model of language generation," *Procedia Computer Science*, vol. 132, pp. 149–156, 2018.

[12] V. Shmelev, A. Karsakov, A. Moiseev, and A. Zagarskikh, "GPU-powered Calculation of Navigation Fields for Agent-based Simulation," *Procedia Computer Science*, vol. 119, pp. 255–261, 2017.

[13] N. Iijima, A. Sugiyama, M. Hayano, and T. Sugawara, "Adaptive Task Allocation Based on Social Utility and Individual Preference in Distributed Environments," *Procedia Computer Science*, vol. 112, pp. 91–98, 2017.

[14] V. Dragos and S. Gatepaille, "On-the-fly integration of soft and sensor data for enhanced situation assessment," *Procedia Computer Science*, vol. 112, pp. 1263–1272, 2017.

[15] V. Y. Guleva, V. V. Povazhnyuk, K. O. Bochenina, and A. V. Boukhanovsky, "Russian Interbank Network Reconstruction via Metaheuristic Algorithm," *Procedia Computer Science*, vol. 108, pp. 1318–1326, 2017.

[16] E. Irannezhad, M. Hickman, and C. G. Prato, "Modeling the Efficiency of a Port Community System as an Agent-based Process," *Procedia Computer Science*, vol. 109, pp. 917–922, 2017.

[17] R. Kozma, R. Ilin, and H. T. Siegelmann, "Evolution of Abstraction Across Layers in Deep Learning Neural Networks," *Procedia Computer Science*, vol. 144, pp. 203–213, 2018.

[18] L. Li, S. Wang, S. Xu, and Y. Yang, "Constrained Spectral Clustering Using Nyström Method," *Procedia Computer Science*, vol. 129, pp. 9–15, 2018.

[19] Q. Li, Z. Wang, G. Li, Y. Cao, G. Xiong, and L. Guo, "Learning Robust Low-Rank Approximation for Crowdsourcing on Riemannian Manifold," *Procedia Computer Science*, vol. 108, pp. 285–294, 2017.

[20] D. Zeng, S. Wang, Y. Shen, and C. Shi, "A GA-based feature selection and parameter optimization for support tucker machine," *Procedia Computer Science*, vol. 111, pp. 17–23, 2017.

[21] M. B. Revanasiddappa, B. S. Harish, and S. V. A. Kumar, "Meta-cognitive Neural Network based Sequential Learning Framework for Text Categorization," *Procedia Computer Science*, vol. 132, pp. 1503–1511, 2018.

[22] D. E. Kazaryan and A. V. Savinkov, "Grammatical Evolution for Neural Network Optimization in the Control System Synthesis Problem," *Procedia Computer Science*, vol. 103, pp. 14–19, 2017.

[23] P. Grachev, I. Lobanov, I. Smetannikov, and A. Filchenkov, "Neural network for synthesizing deterministic finite automata," *Procedia Computer Science*, vol. 119, pp. 73–82, 2017.

[24] F. Wu and Y. Cai, "A Chinese Message Sensitive Words Filtering System based on DFA and Word2vec," *Procedia Computer Science*, vol. 139, pp. 293–298, 2018.

[25] L. Ding, K. Noborio, and K. Shibuya, "Frost Forecast using Machine Learning - from association to causality," *Procedia Computer Science*, vol. 159, pp. 1001–1010, 2019.

[26] V. Nourani, G. Andalib, and F. Sadikoglu, "Multi-station streamflow forecasting using wavelet denoising and artificial intelligence models," *Procedia Computer Science*, vol. 120, pp. 617–624, 2017.

[27] Y. Hu, Z. Jiang, P. Zhan, Q. Zhang, Y. Ding, and X. Li, "A Novel Multi-resolution Representation for Streaming Time Series," *Procedia Computer Science*, vol. 129, pp. 178–184, 2018.

[28] X. D. Wen and C. W. Liu, "Decentralized Distributed Compressed Sensing Algorithm for Wireless Sensor Networks," *Procedia Computer Science*, vol. 154, pp. 406–415, 2019.

[29] T. Volotskiy, J. Smirnov, D. Ziemke, and I. Kaddoura, "An Accessibility Driven Evolutionary Transit Network Design Approach in the Multi-agent Simulation Environment," *Procedia Computer Science*, vol. 136, pp. 499–510, 2018.

[30] M. R. S. Teófilo, A. A. B. Lourenço, J. Postal, Y. M. L. R. Silva, and V. F. Lucena, "The Raising Role of Virtual Reality in Accessibility Systems," *Procedia Computer Science*, vol. 160, pp. 671–677, 2019.

[31] A. Lepskiy and A. Suevalov, "Application of the Belief Function Theory to the Development of Trading Strategies," *Procedia Computer Science*, vol. 162, pp. 235–242, 2019.

[32] I. Bueno, R. A. Carrasco, R. Ureña, and E. Herrera-Viedma, "Application of an opinion consensus aggregation model based on OWA operators to the recommendation of tourist sites," *Procedia Computer Science*, vol. 162, pp. 539–546, 2019.

[33] P. Bołtuć, "Strong Semantic Computing∗∗Conversations with the following people helped me develop some aspects of this paper: Riccardo Sanz, Jack Copeland, John Barker, Kevin O'Regan and Gadi Pinkus." *Procedia Computer Science*, vol. 123, pp. 98–103, 2018.