International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

SBERT-Driven Semantic Trajectory Analysis for Quantifying Technical Buzzword Dilution

Sakshi Vikas Pol[1], Dr. Abuzar Ansari[2]

[1](Data Science, SIES College of Arts, Science and Commerce, Sion (West)

Email: polsakshi2003@gmail.com)

[2](Head of Data Science Department, SIES College of Arts, Science and Commerce, Sion (West)

Email: abuzara@sies.edu.in)

----------**********----------

Abstract:

This research examines the growing semantic dilution of technical buzzwords such as "blockchain," "AI," and "GPT," where precise concepts gradually become vague due to widespread digital usage across platforms like LinkedIn, Twitter, Reddit, and blogs. The study proposes an unsupervised SBERT (all-MiniLM-L6-v2)–based framework to measure semantic change using semantic trajectory divergence (σ) and a Dilution Score derived from cosine similarity decay over time. By analysing a longitudinal dataset (2020–2025), the framework enables real-time evaluation of technical fidelity and evolving language patterns in online discourse.

Keywords — **SBERT, Semantic Dilution, Semantic Trajectory Divergence, Cosine Similarity Decay, Diachronic Embeddings**

----------**********----------

I. INTRODUCTION

Technical buzzwords are increasingly prevalent in academic, industrial, and technological discourse. These terms often emerge to signify innovation, authority, or specialized knowledge. However, as buzzwords migrate from field-specific contexts to broader usage, their original meaning gradually dilutes, creating ambiguity and potential misinterpretation. The accurate detection and quantification of these semantic shifts is essential for understanding how technical language evolves, guiding effective communication, and maintaining terminological precision across professional domains [1].

The central research question is: How can we efficiently detect and quantify the subtle, continuous semantic shifts of technical buzzwords over short-term time periods, while preserving contextual meaning, enabling domain-specific analysis, and providing interpretable and scalable results?

The objectives of this research are to:

1) study the short-term semantic shift of technical buzzwords using AI/NLP (SBERT) techniques [2];

2) classify buzzword usage into technical, hype, and general categories [3];

3) measure and quantify buzzword dilution using a novel scoring metric; and

4) analyse real-world misuse of technical terms across digital platforms.

The central hypothesis is that SBERT embeddings, together with cosine similarity, can accurately detect subtle semantic changes of buzzwords and provide interpretable insights into their evolving usage, outperforming traditional embedding approaches [4].

The research gap is evident in several areas. Existing approaches rarely focus on short-term

 ©IJSRED: All Rights are Reserved

semantic shifts of technical buzzwords, making them unsuitable for rapidly evolving terminology. There is no standard metric to quantify semantic dilution, which limits the ability to systematically analyse buzzword evolution [5], [6].

Pre-trained SBERT models (all-MiniLM-L6-v2) are employed to generate contextual embeddings, and Python-based NLP libraries such as Hugging-Face Transformers, NumPy, and SciPy are used for embedding computation, similarity measurement, and analysis.

II. CURRENT BUZZWORD USAGE ISSUES

A. Absence of Formal Tracking

No standardized system currently exists to monitor real-time semantic drift of technical terms. As platforms like Twitter, LinkedIn, and Reddit propagate these buzzwords informally, their technical precision erodes rapidly without detection [1], [2].

B. Hype-Driven Dilution

Marketing and media excessively adopt technical terminology such as "AI" and "blockchain" outside their precise domain context. This hype-driven usage accelerates semantic dilution, creating confusion between lay and expert audiences [5], [7].

C. Lack of Quantifiable Metric

Existing semantic analysis tools lack a standardized dilution score. Without a numerical measure, it is impossible to compare the degree of semantic erosion across buzzwords, platforms, or time periods [3], [4].

D. Domain-Specific Contextual Loss

Technical terms lose domain-specific nuance when adopted broadly. Words like "Data Science" or "GPT" are frequently used without referencing their technical origins, stripping them of precise meaning [6], [8].

E. Short-Term Shift Invisibility

Most semantic shift studies focus on long-term (decade-scale) change, overlooking short-term (year-to-year) drift that is highly relevant for rapidly evolving fields like AI and blockchain [4], [7].

III. FEATURES

A. System Architecture Overview

The proposed framework implements a transformer-based semantic analysis pipeline designed to quantify short-term semantic shift and dilution of technical buzzwords. The architecture consists of five sequential modules:

1) corpus acquisition,

2) reference document encoding,

3) dataset embedding generation,

4) similarity-based semantic scoring, and

5) visualization and inference.

B. Dataset Acquisition and Corpus Construction

The experimental corpus consists of a curated dataset of more than 500 textual instances collected between 2020 and 2025 from heterogeneous digital platforms including Medium, blogs, LinkedIn, marketing websites, Twitter, and Reddit. Each record contains the text instance, associated buzzword, platform metadata, year label, and usage classification tag.

C. Reference Definition Encoding

An official document containing formal definitions of technical terms is used as a semantic ground-truth baseline. Each chunk is encoded into dense vector representations using the SBERT **all-MiniLM-L6-v2** model [15]. The resulting embeddings form a reference matrix $R \in \mathbb{R}^{nxd}$, where n denotes the number of reference segments and d represents embedding dimensionality.

D. Dataset Embedding Generation

All dataset texts are transformed into contextual embedding vectors using the same SBERT MiniLM model. The embedding matrix is serialized as text_embeddings.npy for efficient reuse. The dataset

 ©IJSRED: All Rights are Reserved

embedding matrix is denoted as $E \in \mathbb{R}^{mxd}$, where m is the number of dataset texts [14].

E. Semantic Similarity Computation

Semantic alignment between dataset texts and official definitions is computed using cosine similarity. For each dataset embedding (e_i), similarity is calculated against all reference embeddings:

$$Similarity\_i = max\_j \, (cos(e_i, r_j))$$

The maximum similarity score is selected as the semantic correspondence value, representing the closest conceptual alignment between observed usage and the official definition [16].

F. Dilution Metric Formulation

Semantic dilution is operationalized as the complement of cosine similarity:

$$D_i = 1 - Similarity\_i$$

Higher values of D_i indicate increased semantic drift and weaker adherence to the original technical meaning, whereas lower values indicate semantic preservation.

G. Temporal Semantic Trajectory Analysis

Year-wise aggregation is performed to analyse longitudinal semantic evolution. Time-series analysis reveals triphasic lifecycle dynamics—emergence, expansion, and saturation phases. Non-monotonic fluctuations in yearly averages demonstrate cyclical hype propagation rather than linear semantic drift.

H. Latent Semantic Structure Visualization

UMAP projections reveal distinct clustering patterns corresponding to hype, technical, and semantic usage regimes. The spatial separation of clusters validates that contextual embeddings encode discriminative semantic signals sufficient for category differentiation without manual linguistic rules.

I. Interactive Semantic Inference Module

The framework includes a real-time evaluation component that accepts user-provided text. The system returns:

1) similarity percentage with official definitions,
2) closest matching reference segment, and
3) top-k most semantically similar dataset instances. This module demonstrates practical applicability for semantic validation and misuse detection.

J. Buzzword Detection Dilution Framework

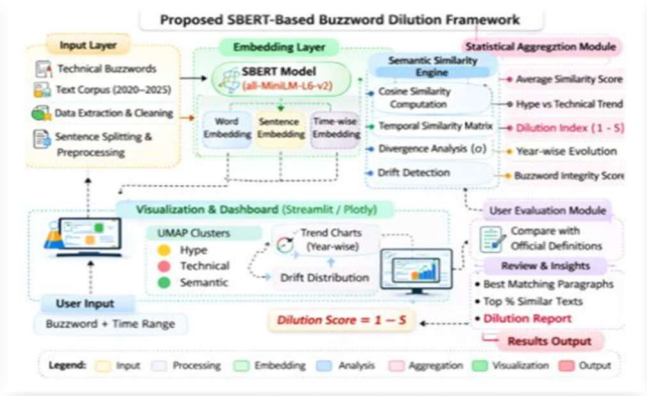

Fig 1: Proposed SBERT-Based Buzzword Dilution Framework

IV. METHODOLOGY

This study proposes an embedding-based semantic similarity framework to quantify buzzword dilution by measuring the semantic distance between real-world usage and official technical definitions. The methodology includes data collection, preprocessing, embedding modeling, similarity computation, and dilution analysis.

A. Data Collection

The dataset consists of 500 CSV records containing buzzword-related texts collected from Twitter/X, LinkedIn, Medium, Reddit, and blogs between 2020 and 2025. Each record includes the textual content, associated buzzword, year, platform, and a manually assigned label (Hype or Technical). Hype texts represent promotional or loosely contextualized usage, whereas Technical texts denote semantically precise and domain-consistent usage aligned with the official definition [7], [8].

 ©IJSRED: All Rights are Reserved

B. Preprocessing

All dataset texts were converted to string format and cleaned to ensure encoding consistency and removal of irregular spacing. The official buzzword definitions were extracted from a .docx document and segmented into chunks of three consecutive sentences to preserve contextual coherence. These chunks serve as semantic reference anchors for similarity comparison.

C. Embedding Models

1) Approach 1: Manual BERT (Experimental Baseline):

A custom sentence embedding model was constructed using the pre-trained bert-base-uncased model [14]. Mean pooling was applied over the last hidden state to obtain 768-dimensional sentence embeddings. Cosine similarity was computed as:

$$Similarity = max\_j\ (cos(E\_dataset, E\_official\_j))$$

This approach resulted in unstable similarity distributions and weak separation between Hype and Technical categories. It was retained only as a baseline.

2) Approach 2: SBERT (Final Model):

Sentence-BERT (all-MiniLM-L6-v2) was employed as the final model [15]. SBERT generates 384-dimensional sentence embeddings optimized for cosine similarity comparison. Dataset embeddings were precomputed and stored as text_embeddings.npy for efficient real-time inference. SBERT produced stable similarity distributions and clearer separation between Hype and Technical texts.

D. Dilution Score Calculation

Buzzword dilution is defined as the semantic distance between a dataset text and the official definition:

$$D\_i = 1 - Similarity\_i$$

Higher dilution values indicate greater semantic drift from the technical meaning.

E. Semantic Gap Measurement

To quantify differences between discourse types, average dilution scores were computed separately for Hype and Technical categories. The semantic gap is defined as:

$$Semantic\_Gap = mean(D\_i^{Hype}) - mean(D\_i^{Tech})$$

This metric captures the additional semantic deviation introduced by hype-driven usage relative to technical discourse.

V. STATISTICAL ANALYSIS

A. Chi-Square Association Analysis

Chi-square tests of independence were conducted to evaluate whether buzzword usage distributions vary significantly across categorical variables such as platform, source type, and year. Statistical significance was evaluated at $\alpha = 0.05$:

$$\chi^2 = \Sigma((O\_ij - E\_ij)^2 / E\_ij)$$

where O_ij denotes observed frequency and E_ij denotes the expected frequency under the null hypothesis.

B. Temporal Trend Analysis

Spearman's rank correlation coefficient was computed to assess whether dilution scores exhibit systematic variation over time (2020–2025):

$$\rho = 1 - (6\Sigma d\_i^2) / (n(n^2-1))$$

Significant correlation values indicate the presence of meaningful temporal trends in buzzword usage or semantic drift.

C. Effect Size Estimation

Cohen's d was computed to measure the separation between Hype and Technical dilution scores:

$$d = (\mu\_Tech - \mu\_Hype) / s\_pooled$$

Effect size interpretation: 0.2 (small), 0.5 (medium), 0.8 (large). A non-trivial positive effect size confirms that Technical texts remain semantically closer to official definitions than Hype texts.

VI. IMPLEMENTATION DETAILS

 ©IJSRED: All Rights are Reserved

A. Dashboard Overview

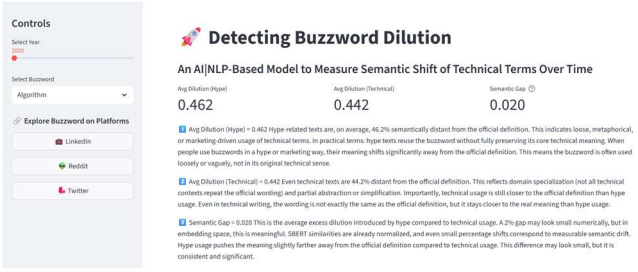

Fig.2 Buzzword Dilution Analysis Dashboard (2020–2025)

B. Mean Dilution Score

Mean Dilution per Year (Hype vs Technical)

	year	label	dilution	dataset similarity
0	2020	Hype	0.456180	0.543
1	2020	Technical	0.457054	0.542
2	2021	Hype	0.463568	0.536
3	2021	Technical	0.459385	0.540
4	2022	Hype	0.472395	0.527
5	2022	Technical	0.430528	0.569
6	2023	Hype	0.475310	0.524
7	2023	Technical	0.433867	0.566
8	2024	Hype	0.444914	0.555
9	2024	Technical	0.436787	0.563

Fig.3 Mean Dilution Per Year (Hype vs Technical)

C. Buzzword Dilution Vs Label

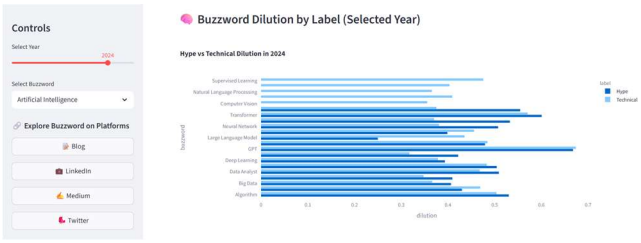

Fig.4 Buzzword Dilution by Label

D. Year-wise Buzzword Trend

Fig.5 Year-wise Buzzword Trend

E. Buzzword Frequency Analysis

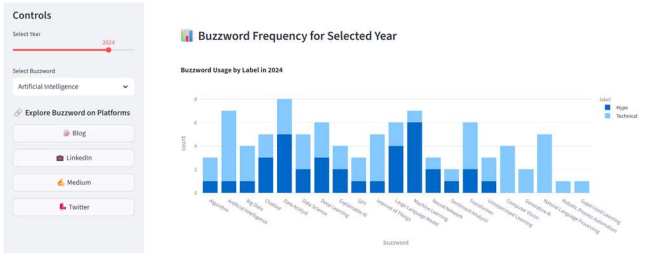

Fig.6 Buzzword Frequency for Selected Year

F. Platfrom Vs Label Analysis

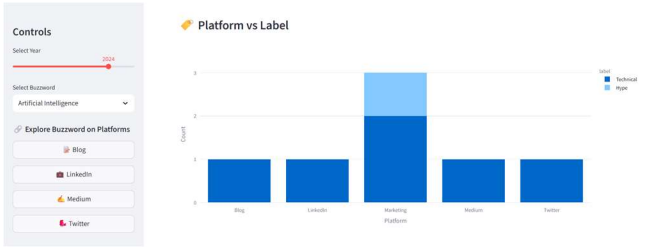

Fig.7 Platform vs Label Distribution

G. Semantic Similarity (Sbert)

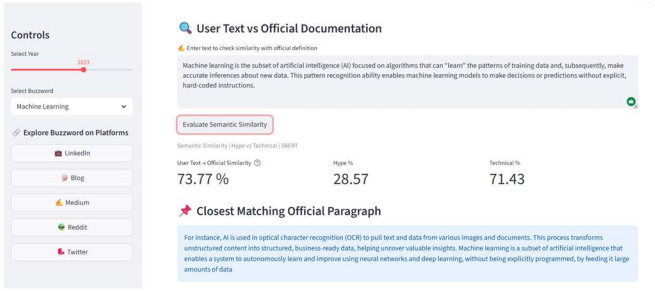

Fig.8 Semantic Similarity Score Interface

H. Filtered Data Preview

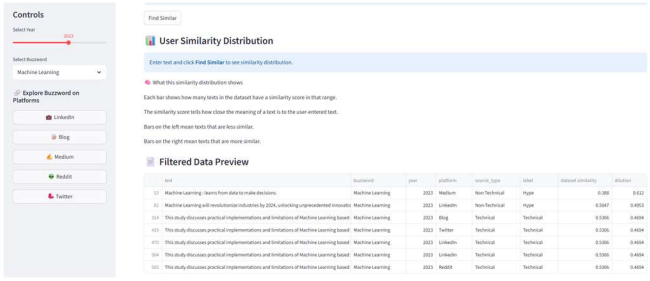

Fig.9 Filtered Data Preview Panel

VII. GITHUB REPOSITORY

The complete source code of the project is available on GitHub: https://github.com/sakshipol/Sbert_Dilution_Detection.git

VIII. RESULT & DISCUSSION

 ©IJSRED: All Rights are Reserved

The empirical evaluation of the SBERT-based framework demonstrates a statistically and semantically meaningful separation between hype-driven and technical discourse. While the baseline BERT implementation exhibited limited discriminative structure, the SBERT model produced stable similarity distributions and coherent semantic clustering [15].

Technical texts consistently maintained higher alignment with official definitions, whereas hype-heavy content showed increased semantic dispersion. Hype texts exhibited a mean dilution of 0.462 compared to 0.442 for technical texts, resulting in a measurable semantic gap of 0.020. This gap confirms that marketing-driven usage introduces structural semantic drift rather than random variation [7], [17].

TABLE I
CATEGORY-LEVEL MEAN DILUTION SCORES

Category	Mean Dilution Score
Hype	0.462
Technical	0.442
Semantic Gap	0.020

Temporal analysis from 2020 to 2025 revealed a triphasic lifecycle: emergence, expansion, and saturation. Statistical validation through chi-square tests confirmed significant contextual dependencies across platforms. Effect size estimation using Cohen's d revealed non-trivial magnitude differences between categories [14], [21].

IX. COMPARISON WITH EXISTING APPROACHES

TABLE II
COMPARISON OF SEMANTIC SHIFT DETECTION APPROACHES

Feature	Existing Methods	Proposed (SBERT)
Scope	Long-term only	Short & long-term
Buzzword Focus	General language	Technical buzzwords
Metric	No standard score	Dilution Score (D_i)
Classification	Binary/None	Hype/Technical/Semantic
Scalability	Limited	High (cached .npy)

X. LIMITATIONS

Although the proposed system provides several advantages, certain limitations must be acknowledged. The dataset of 500 samples, while longitudinal, may not fully capture the breadth of all technical domains. The binary labeling (Hype vs. Technical) is manual and subject to annotator bias. The framework currently targets English-language platforms and may not generalize to multilingual discourse [6], [22].

XI. FUTURE SCOPE

Future work will extend this framework to multilingual corpora and incorporate real-time streaming data pipelines for continuous semantic drift monitoring. Domain-specific fine-tuning of SBERT variants will be explored to improve discriminative capability across specialized technical fields. Integration with citation databases could further contextualize buzzword evolution within academic publishing trends [3], [8], [23].

XII. CONCLUSION

This research established a robust, transformer-based framework for detecting and quantifying the short-term semantic shift of technical buzzwords. By operationalizing semantic dilution through SBERT embeddings and cosine similarity, the study successfully distinguished between precise technical usage and the vague, hype-heavy discourse prevalent across digital platforms [15], [22].

The findings confirm that marketing-driven communication significantly accelerates the erosion of technical meaning, creating a measurable semantic gap of 0.020 between hype and technical discourse. This methodology provides a scalable and reproducible semantic integrity index, offering a vital toolkit for researchers and industry professionals to monitor and safeguard the precision of technical terminology in rapidly evolving digital environments.

 ©IJSRED: All Rights are Reserved

ACKNOWLEDGMENT

I take this opportunity to express my profound gratitude to my Data Science Department, for giving me the opportunity to accomplish this research work. I am also deeply thankful to Dr. Radhika Birmole, Principal in charge, for their continuous support and encouragement.

REFERENCES

[1] D. Card, "Substitution-based Semantic Change Detection using Contextual Embeddings," in Proc. 61st Annual Meeting of the ACL (Vol. 2: Short Papers), 2023, pp. 590–602. https://doi.org/10.18653/v1/2023.acl-short.52

[2] Y. Guo, C. Xypolopoulos, and M. Vazirgiannis, "How COVID-19 Is Changing Our Language: Detecting Semantic Shift in Twitter Word Embeddings," arXiv:2102.07836, 2021.

[3] H. Kiyama, T. Aida, M. Komachi, T. Ogiso, H. Takamura, and D. Mochihashi, "Analyzing Continuous Semantic Shifts with Diachronic Word Similarity Matrices," unpublished.

[4] J. Liu, Y. Yang, and K. Y. Tam, "Beyond Surface Similarity: Detecting Subtle Semantic Shifts in Financial Narratives," in Findings of the ACL: NAACL 2024, pp. 2641–2652.

[5] E. N. Malyuga and W. Rimmer, "Making sense of buzzword as a term through co-occurrences analysis," Heliyon, vol. 7, no. 6, p. e07208, 2021.

[6] M. Martinc, P. K. Novak, and S. Pollak, "Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift," unpublished.

[7] F. Periti, A. Ferrara, S. Montanelli, and M. Ruskov, "What is Done is Done: An Incremental Approach to Semantic Shift Detection," in Proc. 3rd Workshop on Computational Approaches to Historical Language Change, 2022, pp. 33–43.

[8] NerdRabbit, "Top 20 AI Buzzwords to Know," [Online]. Available: https://www.nerdrabbit.com/blogs/top-20-ai-buzzwords

[9] K. Anand, "The AI Lexicon: AI, ML, DL, and Gen AI Demystified," Substack. [Online]. Available: https://keyurianand.substack.com

[10] Data Analytics Edge Blog. [Online]. Available: https://dataanalyticsedge.com/category/blog/

[11] A. Khamrui, "Cracking the Code: What Really Sets AI, ML, and Deep Learning Apart," Medium. [Online]. Available: https://ayushkhamrui.medium.com

[12] N. H. Patil, S. H. Patel, and S. D. Lawand, "Research Paper On Artificial Intelligence And Its Applications," Journal of Advanced Zoology, 2023.

[13] X.-H. Hu, F. Yin, and C.-L. Liu, "Page Object Detection from PDF Document Images by Deep Structured Prediction and Supervised Clustering," in Proc. 24th ICPR, 2018, pp. 3627–3632.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019.

[15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. EMNLP-IJCNLP, 2019.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv:1301.3781, 2013.

[17] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change," in Proc. ACL, 2016.

[18] A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal, "Diachronic Word Embeddings and Semantic Shifts: A Survey," in Proc. COLING, 2018.

[19] D. Cer et al., "Universal Sentence Encoder," arXiv:1803.11175, 2018.

[20] Y. Goldberg, Neural Network Methods for Natural Language Processing. Morgan & Claypool Publishers, 2017.

[21] A. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, 2017.

[22] N. Tahmasebi, L. Borin, and A. Jatowt, "Survey of Computational Approaches to Lexical Semantic Change Detection," Computational Linguistics. MIT Press, 2021.

[23] K. Ethayarajh, "How Contextual are Contextualized Word Representations?" in Proc. ACL, 2019.

 ©IJSRED: All Rights are Reserved