Tracking Semantic Change of Technology Words

Capstone in Data Science

Junshu Feng

November 24, 2025

Abstract

The present project employs diachronic word embeddings to quantify and analyze semantic change in technology vocabulary from 2016 to 2025. Using news articles from The Guardian's Open Platform, we trained annual word embedding models and measured semantic shift through cosine distance calculations. Despite methodological challenges including the dimensionality problemand data sparsity for individual terms, our approach demonstrates the potential of computational linguistics to track language evolution. The project highlights both the promise and limitations of applying NLP techniques to historical linguistic questions, while addressing important ethical considerations in digital humanities research.

Introduction

Linguistic systems are not always static. Although they are often well-balanced, language still evolves sometimes due to newly developed objects or ideas (Blank 1999). For instance, cloud develops a new meaning "a vast online storage space" in the recent years due to the need for a name of that object. This process, known as semantic change, is one of the central topics of historical linguisitics. In the digital era, certain words such as bug, web, mouse, and cloud come to develop new meanings to accommodate the invention of new technological devices. They are now more often recognized as technology terms. Interested in the semantic change of technology terms, I will explore this idea as my data science capstone project.

Semantic change has been investigated historically through qualitative analyses of texts over time (Blank 1999). However, this traditional method is time-consuming and can be subjective. In the modern era, the vast storage capability of computers allows us to possess an unprecedented amount of data presented in media, which makes the quantification of semantic change possible. Therefore, in the present project, I choose to look at the technology terms presented in news media. News articles are not only readily accessible but also time-sensitive so that I can track the semantic change over time.

Fortunately, advancements in natural language processing (machine learning to help computers understand human language) have equipped us with an approach to quantify semantic change over time. This approach, known as diachronic word embeddings, is a technique that represents words as high-dimensional vectors and creates a quantitative coordinate map of meaning over time (Hamilton, Leskovec, and Jurafsky 2016). The technique is especially realizable in R as it offers a package named *text2vec* (Mimno 2018). We utilized the GloVe algorithm, which is an extension of the *text2vec* package (Pennington, Socher, and Manning 2014). By training separate word embedding models on technology terms presented in news media from each year between 2016 and 2025, I create a temporal series of semantic snapshots. The core of the analysis involves aligning these snapshots from each year to one single map and then measuring the cosine distance between a word's vector representation across different years. This distance serves as a numerical index of semantic change, which allows me to identify which technology words have shifted most significantly over the past decade.

Methodology

Data Collection

We utilized The Guardian's Open Platform API as our primary data source, selected for its comprehensive technology coverage and academic accessibility. The API provides access to The Guardian's complete article archive from 2016 to 2025, with data including publication dates, sections, and full article text. For API integration, we implemented a systematic approach using the httr package in R, with all requests authenticated through API authentication.

The seed word selection process uses a multi-stage methodology to identify representative technology words that may be of our interest. We began with reviewing technology journalism and industry reports to identify frequently evolving terminology, then ensured temporal coverage by selecting terms with sufficient historical presence across our ten-year timeframe. We also selected words that were suspected to undergo a great amount of semantic change due to some important event. The words should not have any ambiguity as well (in other words, they should not represent some meaning that we are not interested in). We ended up with 9 seed words: “cookie”, “crypto”, “cloud”, “archive”, “virus”, “feed”, “airdrop”, “chatbot”, “edge”.

For data collection, we executed a systematic pipeline for each seed word and year combination. The process queried The Guardian API with specific parameters including the target seed word, annual date ranges from January 1 to December 31 for each year from 2016 to 2025, and requests for full body text. This methodology yielded approximately 4,500 articles total, with an average of 50 articles per seed word per year.

Data Preprocessing

We implemented a comprehensive multi-stage text cleaning pipeline to ensure data quality. For example, we lowercased all texts, removed punctuation, special characters, and white spaces. For contextual analysis, we captured 50 words before and after each seed word occurrence from each article. This allowed us to focus the semantic analysis on relevant linguistic contexts. The corpus consisted of three hierarchical levels. The foundation had nine annual corpora from 2016 to 2025, with each annual corpus further partitioned into seed word-specific subcorpora containing all contextual instances of that word. At the most granular level, individual articles maintained their own data including publication date, word count, source information, and the processed text content.

Word Embedding Methodology

We selected the GloVe algorithm for word embedding generation based on several technical considerations specific to our research objectives. The algorithm’s linear scaling with corpus size makes corpus processing more efficient. The clear interpretability of the relationship between GloVe’s objective function and semantic relationships aligned with our need for transparent analysis, and the algorithm’s established effectiveness in semantic tasks provided confidence in its applicability to diachronic linguistic analysis.

For model training, we trained word-specific models where separate GloVe models were generated for each seed word across each year, which allowed focused analysis of individual term evolution without cross-term interference. We set vector dimensionality to 50 dimensions based on common practice in semantic change detection literature.

Semantic Change Measurement

To enable meaningful cross-year vector comparison, we implemented a comprehensive vector alignment procedure using Orthogonal Procrustes analysis. The process began with reference selection, where we designated the earliest available year in each sequence as the alignment target, providing a consistent baseline for measuring semantic change. We then identified the intersection of vocabularies across years to establish common semantic dimensions for alignment. The core transformation involved computing an optimal rotation matrix R that minimized the Frobenius norm difference between the reference matrix and transformed target matrices.

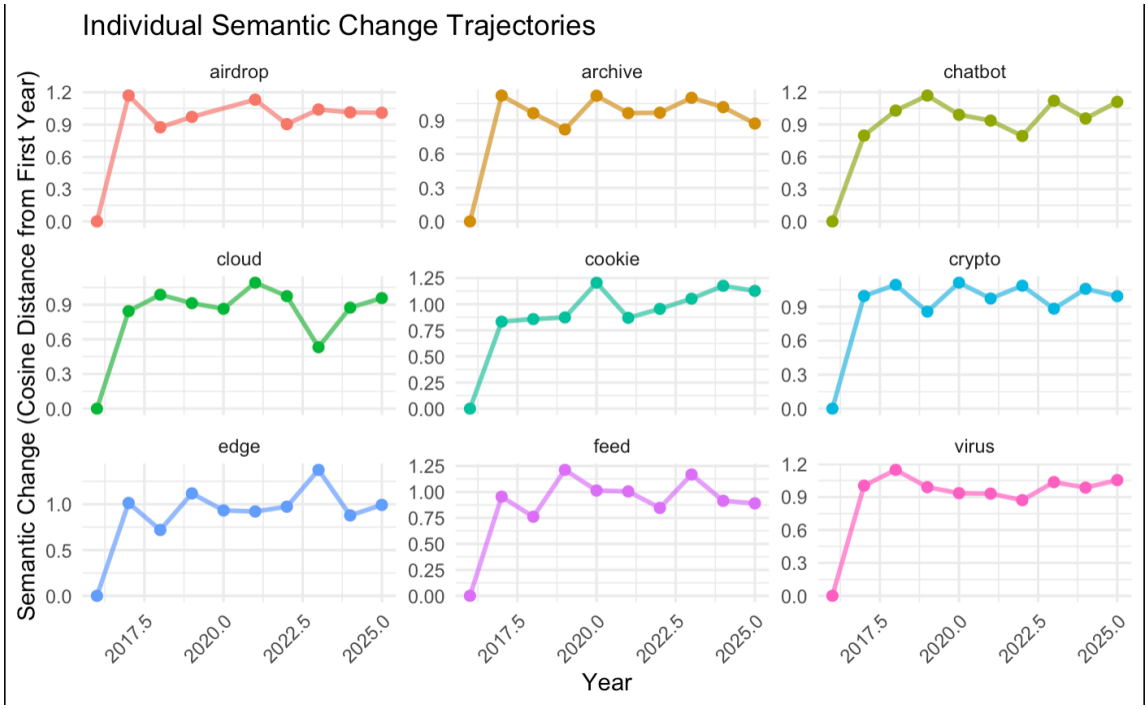
For semantic change quantification, we employed cosine distance as our primary metric, calculated as one minus the cosine similarity between vectors. The cosine similarity itself was computed as the dot product of two vectors divided by the product of their magnitudes, resulting in a value between negative one and one that represents the cosine value of the angle between the vectors in the high-dimensional space. The cosine distance transformation produced a value ranging from zero to two, where zero indicates identical semantic orientation, one represents orthogonal vectors, and two represent opposed meanings.

Results

The plot below shows the semantic change results of the 9 seed words, where the x-axis represents the year and the y-axis represents the cosine distance between a given year and the first year.

The primary finding was that all seed words showed cosine distances of approximately 1 from their reference vectors, indicating near-orthogonal vectors across years. This result suggests that the models revealed random relationships rather than meaningful semantic representations.

Our analysis revealed significant methodological challenges. The word-by-word approach, while theoretically sound for measuring individual word evolution, encountered the problem of dimensionality. With vector spaces to 50 to 100 dimensions and relatively small corpora for individual seed words, the models failed to learn meaningful semantic relationships.



Discussion

Our most significant finding concerns the methodological challenge of high-dimensional spaces with small sample size. When working with high-dimensional vectors and limited training data for individual words, random vectors become very likely to be orthogonal. This explains why all our cosine distances clustered around 1, which indicates that no meaningful semantic relationships were learned.

This finding has important implications: the choice of vector dimensionality must be carefully matched to the corpus size. For individual word analysis across time, smaller dimensions (for example, around 8-20) with larger corpora would likely yield more meaningful results. Our approach suffered from attempting to learn too many parameters with insufficient data.

Apart from the dimensionality problem, the word-by-word approach, while ideal for isolating specific term evolution, proved problematic in practice. The alternative—training combined models on all text from each year—would provide more robust vectors but might dilute word-specific semantic signals. Future work should consider hybrid approaches or other techniques that use larger general corpora while caring for specific terms.

Ethical Concerns

The present project has several ethical concerns. First, we analyzed content from The Guardian's Open Platform, which consists of publicly available news articles. While the content itself is public, ethical considerations arise around the republication and analysis of journalist's work without explicit individual consent. However, this concern is addressed due to the transformative use of the news articles. Specifically, this project performed computational analysis on the text rather than republishing the content verbatim. The focus is on linguistic patterns, not the journalistic content itself. Therefore, it should be a fair use of data. Second, the news articles from The Guardian's Open Platform are protected by copyright, and large-scale scraping could potentially violate terms of service. To resolve this concern, we complied to The Guardian's Open Platform API terms of service. We registered a developer API key on its website such that it was used for proper authentication. Third, research on technology words could potentially be misused to manipulate tech discourse or reinforce technological determinism. We addressed this issue by positioning the present project as a tool for understanding technological adoption and cultural change, not for predictive or manipulative purposes.

References

Blank, Andreas. 1999. "Why Do New Meanings Occur? A Cognitive Typology of the Motivations for Lexical Semantic Change." In *Historical Semantics and Cognition*, edited by Andreas Blank and Peter Koch, 61–90. Berlin: Mouton de Gruyter.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1501. https://doi.org/10.18653/v1/P16-1141.

Mimno, David. 2018. "Text2vec: Modern and Efficient Toolkit for Text Analysis and Embedding in r." *Journal of Statistical Software* 85 (1): 1–30. https://doi.org/10.18637/jss.v085.i01.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning. 2014. "Glove: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–43.