Tracking Semantic Change of Technology Words

Capstone in Data Science

Junshu Feng

December 1, 2025

Abstract

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). 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 10 seed words: "cookie", "crypto", "cloud", "metaverse", "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 5,000 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 corpora consisted of three distinct levels. The foundation consisted of ten annual corpora spanning from 2013 to 2022, with each annual corpus further partitioned into seed word-specific subcorpora containing all contextual instances of that particular term. At the most granular level, individual articles maintained their metadata including publication date, word count, and source information alongside the processed text content. This organizational scheme enabled both longitudinal analysis across years and focused examination of individual term evolution.

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 provided computational efficiency necessary for processing multiple annual corpora, while its utilization of global word-word co-occurrence statistics offered theoretical advantages for capturing semantic relationships. 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 implemented two distinct methodological approaches to address different aspects of our research questions. The first approach involved training word-specific models where separate GloVe models were generated for each seed word across each year, allowing focused analysis of individual term evolution without cross-term interference. The second approach employed combined annual models where all seed word contexts within a given year were processed together, creating comprehensive semantic spaces that captured inter-term relationships and contextual dynamics within each temporal period.

The parameter configuration for GloVe training reflected careful consideration of both theoretical recommendations and practical constraints. We set vector dimensionality to 100 dimensions based on common practice in semantic change detection literature, which balances representational capacity with computational feasibility. The context window size of 10 words represented a compromise between capturing sufficient local context and maintaining computational efficiency, while 50 training iterations were determined through preliminary convergence testing that indicated stable optimization within this range. The x_max parameter value of 10 followed standard practice for co-occurrence truncation in medium-sized corpora.

Semantic Change Measurement

To enable meaningful cross-temporal 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, mathematically expressed as minimizing ‖W_ref - W_year · R‖_F.

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 of the angle between the vectors in high-dimensional space. The cosine distance transformation produced a value ranging from zero to two, where zero indicates identical semantic orientation, two represents diametrically opposed meanings, and intermediate values reflect varying degrees of semantic divergence.

We implemented a multi-faceted statistical validation framework to ensure the robustness of our semantic change measurements. Stability analysis involved measuring year-to-year variation for control words with stable meanings, establishing baseline expectations for semantic consistency. Comparison against a null model provided significance assessment by contrasting observed semantic changes with random word pair distances, while bootstrap confidence

intervals estimated measurement uncertainty through resampling techniques. This comprehensive validation approach ensured that reported semantic changes reflected genuine linguistic evolution rather than methodological artifacts or random variation.

The software infrastructure for this project was built primarily within the R programming environment, utilizing version 4.2.1 for all computational processes. Critical package dependencies included the text2vec package for word embedding implementation, the tidyverse collection for data manipulation and visualization, and the httr package for API interactions. We maintained complete version control through Git with regular commits to a GitHub repository, ensuring reproducibility and transparency throughout the research process. All analyses were documented through R Markdown files with fixed random seeds to guarantee consistent results across computational environments.

Computational resource requirements were substantial given the scale of text processing and model training operations. Total processing time approached approximately 48 hours of computation across all stages, with peak memory usage reaching 16 GB during concurrent model training operations. Storage requirements totaled roughly 2 GB distributed between raw article data, processed text corpora, and trained embedding models. These resource demands necessitated careful job scheduling and memory management throughout the project lifecycle.

The validation protocol incorporated multiple layers of quality assurance to ensure methodological rigor. Internal consistency checks verified that identical text inputs produced identical embedding outputs, confirming algorithmic stability. Temporal smoothness analysis examined whether semantic change trajectories demonstrated reasonable progression patterns rather than erratic fluctuations. Qualitative validation involved manual inspection of nearest neighbor relationships in the embedding spaces, assessing whether semantic groupings aligned with linguistic intuition and domain knowledge. This multi-dimensional validation framework provided comprehensive assurance of result reliability and methodological soundness.

Results

Ethical Concerns

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.