Toward an Unsupervised Method for Assessing Semantic Specificity

Anonymous ACL submission

Abstract

Classifying and understanding semantic specificity is essential for enhancing various computational tasks, such as recommendation systems, by enabling them to deliver more targeted and relevant content. This paper introduces a novel unsupervised learning approach for classifying the semantic specificity of text, eliminating the need for extensively labeled data. The results highlight the potential for robust, scalable, and adaptable NLP systems capable of accurately classifying text by semantic specificity without heavily relying on ample amounts of labeled data.

1 Introduction

Classifying and understanding semantic specificity or "scope" is increasingly crucial for many computational tasks. The ability to discern whether textual content is "general" or "specific" benefits advanced systems, such as recommendation engines, to deliver more targeted and relevant content. Semantic specificity can also give NLP systems a nuanced understanding of text.

This paper expands upon prior work in classifying semantic specificity with the development of novel unsupervised techniques for classifying the semantic specificity of text without relying on vast amounts of labeled data. By embedding textual data and applying heuristic clustering based on linguistic and syntactic cues, the methodology in this paper addresses the absence of unsupervised methods for classifying the specificity of text.

2 Background and Motivation

2.1 Existing Gaps in Semantic Specificity

Current methods for computing semantic specificity make use of supervised learning frameworks that require extensive labeled datasets. These methods may not fully capture the range of semantic nuances specific to certain tasks, a significant limitation for classifying the specificity of text in different domains. Unsupervised learning methods, which do not rely on labeled data, remain underutilized for determining semantic specificity despite their potential for scalability and adaptability in applications like recommendation systems.

Prior work has explored clustering and dimensionality reduction techniques to group similar texts or reduce the feature space of datasets. Li and Nenkova's "SPECITELLER" demonstrates the potential of semi-supervised approaches for predicting specificity, highlighting the importance of readability and comprehension (Li and Nenkova, 2015). Their findings required the usage of pre-existing datasets that were labelled in order to evaluate specificity. This encouraged our exploration into unsupervised methods to assess text specificity when labelled data isn't available.

2.2 Creative Ideation Recommendation System

The Supermind Ideator (Heyman et al., 2024) is one of a number of creative ideation applications that guide users through "Moves" that help them assess their problem and generate solutions. While reflecting on this work, it became clear that a recommendation system to suggest the next best move to users would be valuable. We realized that effective recommendation required understanding of the user's current phase in the creative process.

The creative ideation process, often beginning with broad, divergent thinking and moving towards more focused, convergent thinking as solutions and problem definitions are formulated (aka. the Double Diamond method of ideation (Council, 2024)).

Reflecting upon this convergent-divergent framing of stages of ideation, we developed an unsupervised method to accurately classify "scope" of text leading to our new method in calculating semantic specificity.

3 Methodology

When engaging in ideation, the "scope" of an idea can be described as General and Specific. General ideas contain the main elements of a topic without going into thorough detail. Specific ideas are clearly defined and pertain to a particular topic.

This definition allows humans to interpret whether a statement is general or specific, where general scope relates to divergent ideas and specific scope to convergent ideas.

4 Unsupervised Method for Assessing Semantic Specificity

4.1 Embedding and Clustering of Data

To determine specificity, we first embed text using the Doc2Vec model, which captures the semantic meaning of entire documents (Le and Mikolov, 2014). This also standardizes input sizes for consistent modeling. Following embedding, the vectorized texts are clustered into thematic groups using KMeans.

4.1.1 Sub-Clustering for Detailed Analysis

After clustering into general topics, a secondary clustering pass groups into sub-topics. This sub-clustering enables a more granular exploration of the thematic landscape, helping to pinpoint more specific content within broad topics.

4.2 Application of Dimensionality Reduction

Uniform Manifold Approximation and Projection (UMAP) is used to visualize the clustering. UMAP helps in preserving both local and global data structures, facilitating a consistent representation of relationships within clusters (see Fig 1).

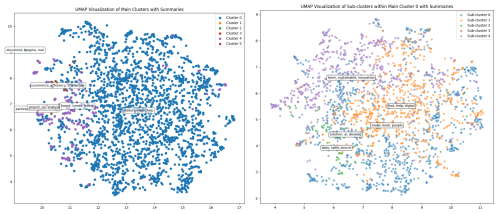

Figure 1: UMAP visualizations, with main and sub clusters on the left and right, respectively

5 Heuristic-Based Clustering

5.1 Heuristic Feature Representation

We build upon the topic clusters, extracting the following heuristic features to refine clustering further and determine specificity of textual content:

Flesch-Kincaid Reading Ease: evaluates how easily a text can be understood, integrating average sentence length and syllable count to classify texts by general or specific scope. A higher readability score suggests content suitable for a broader audience, while a lower score indicates specialized, niche content (Solnyshkina et al., 2017).

Unique Word Count and Lexical Diversity: assess the richness and thematic concentration. A higher count of unique words and greater lexical diversity could indicate a wide range of topics covered, implying a general scope. Conversely, lower counts suggest a focused thematic presence, characteristic of a specific scope (Treffers-Daller et al., 2018; Zhang and Wu, 2021).

Numerical Data Frequency: occurrence of numerical data hints at its precision and technical nature. Texts with frequent numerical data typically have a narrow focus, which could suggest the text has a specific scope (Susoy, 2023).

Average Sentence Length: Longer sentences contain more complex structures and advanced vocabulary. This identifies texts with a specific, narrow scope aimed at specialists (Bestgen, 2023).

Named Entity Recognition (NER): using the spaCy 'en-core-web-lg' model, this identifies and categorizes key nouns and proper nouns, highlighting topics and themes in the text. Entities identified serve as markers for the thematic structure, aiding in clustering of content (Schmitt et al., 2019).

5.2 Clustering and Scope Determination

Using these heuristic features, text are re-clustered into binary classes to determine their scope. Text that fall into the minority heuristic class are deemed "specific" in scope while those in the majority class are "general".

In summary, our unsupervised approach:

1. Embeds the text using the Doc2Vec embedding model.
2. Clusters text by topic and then by sub-topic through the use of the KMeans.
3. Assesses each text within its own topic cluster (to only compare text from the same topic cluster), transforming each into a feature vector based on our heuristic feature representation.
4. Uses KMeans with the new heuristic feature representation to cluster once more.
5. Labels the majority cluster members as "general" and the minority cluster members as "specific".

This unsupervised, heuristic-driven methodology circumvents the limitations of labeled datasets and introduces a scalable, adaptable framework for real-time text analysis.

6 Human Feedback Integration

6.1 Integration of Human Labeling to Validate and Refine the Unsupervised Model

Human feedback was collected to validate and enhance the accuracy of the unsupervised model. Human participants (n=32) recruited from Prolific reviewed text samples (n=88) and classified them as "general", "specific", or "don't know" by leveraging their intuition about the text's context. Texts were pre-classified by the model, using thematic clustering to group up text in batches before being evaluated. The following instructions were given to users at the beginning of the survey:

"This survey will present you with several lists of problem statements that have been written by people before they try to come up with creative solutions to their problems. Your task is to rate each of these problem statements as either being **General** (meaning "containing the main features or elements of something") or **Specific** (meaning "precise and clearly defined"). If you are unsure, you can select **Don't Know**."

6.2 Analysis of Human Labels and Their Integration with the NLP System

The collection of human evaluation was implemented through Qualtrics where participants rated the same batches in a randomized and counterbalanced fashion.

6.3 Statistical Insights from Human Feedback

Human evaluation revealed several key insights into the model's performance and areas for improvement:

Balanced Label Distribution: Results indicated a balanced distribution between 'General' and 'Specific' labels. This suggests participants were confident in their classifications (95.93% selected either 'General' or 'Specific').

Alignment with Model Labels: Logistic regression showed a significant tendency for human classifications to agree with the model's predictions, with 'General' classifications being particularly accurate.

Consensus Validation: When a majority consensus was present, the model's predictions matched human judgments in 62 out of 88 cases, resulting in a statistically significant chi-squared test outcome ($p = .0001$). This agreement demonstrates our model's capability to reflect human consensus accurately.

Intraclass Correlation Coefficient (ICC): An ICC of .83 demonstrated high consistency among raters, underscoring the reliability of human judgments and the validity of the experimental approach.

6.4 Incorporating Human Labels to Refine Clustering

The initial unsupervised model achieved a consensus accuracy of 70.45% with human classifications. Using the insights from this human ground truth labelling drove us to focus on refining model accuracy and consistency. While the present model aligns well with human judgments for 'General' statements, improvement was needed in precision of 'Specific' statements. We adjusted the sensitivity of the KMeans algorithm's initial conditions by assessing different starting seed values to improve performance. This enhanced the model's accuracy, bringing a closer alignment to human judgments. Although there is an argument that this approach overfits to the human labels, there is inherent randomness involved in KMeans based on starting points for clustering and this approach is used to show the system's range of accuracy and ability to have higher accuracy with particular initial seeds. The following statistics illustrate varying approaches to creating heuristic ("scope") labels on the problem statement data, where "Match Percentage" is the number of heuristic labels generated by the system that match human consensus labels.

Method	Mean (%)	Std (%)	Best (%)
Heur. Only	44.32	0.00	44.32
Main+Heur.	69.14	2.19	71.59
Main+Sub+Heur.	72.58	3.10	80.64

Table 1: Results showing mean, std. dev., and best match percentages for different optimization strategies.

The optimal method combined main and sub-clustering (based on topics) with heuristic feature clustering. This approach emphasizes the importance of topic coherence in clustering, contrasting with less structured heuristic clustering that

lacks topical context. Variability in results due to the starting seed highlights the inherent non-determinism of KMeans, impacting the standard deviation of accuracy metrics (Ahmed et al., 2020).

7 Discussion

7.1 Interpretation of Results and Implications for Future Research

Results show that utilizing unsupervised methods for determining semantic specificity is viably accurate. This also reduces a reliance on pre-labeled datasets. Our contribution opens prospects for more robust NLP systems capable of applications across dynamic settings and domains.

Two core areas for future work to extend and further enhance this contribution could be:

Enhanced Methodologies: Investigate mixed-method approaches that blend machine learning with human-like flexibility, potentially exploring new heuristic features or diverse unsupervised models.

Broader Applications: Extend these methods to other NLP applications like sentiment analysis, automated summarization, and personalized content delivery.

7.2 Scalability and Adaptability of the Approach to Different NLP Applications

The unsupervised method described in this work shows promise in scalability and adaptability, potentially being highly applicable across diverse NLP tasks that need to classify semantic specificity:

Scalability: The model efficiently processes large data volumes without predefined labels, improving over time in accuracy when further tuned with human labels. This allows the model to update with new content, such as user-generated text on social media platforms and customer feedback systems.

Adaptability: The model can adapt to different themes and text types by analyzing thematic clusters before tuning to syntactic representations, enhancing utility across domains.

8 Conclusion

8.1 Summary of Key Findings

The unsupervised approach presented here for classifying semantic specificity is currently being applied within a proof of concept recommendation engine. It appears to offer a versatile framework for advancing classification methods without relying heavily on labeled data. Our unsupervised learning methods were validated and enhanced by human feedback to accurately classify text by semantic specificity or "scope."

By combining unsupervised learning techniques (Doc2Vec and KMeans clustering) with heuristic-based clustering we can effectively classify texts as "general" or "specific." Furthermore, by employing a two-tiered clustering approach (initial topic clustering followed by heuristic-based clustering) we enhanced the precision of the model. This proved effective in handling thematic and semantic nuances within large datasets.

These findings are particularly relevant to the broader NLP community as they demonstrate the potential of unsupervised methods to support complex semantic tasks typically reserved for supervised approaches. This work appears most promising in environments where labeled data is scarce or difficult to obtain.

8.2 Proposals for Future Work

While this study focused on the creative ideation process, applying this unsupervised approach to using semantic specificity in other domains could reveal broader applicability.

Further exploring the use of this approach within other NLP tasks such as sentiment analysis, intent detection, and automatic summarization could led to more nuanced text scope understanding.

Future work might explore how to combine this approach with other machine learning methods, such as deep learning or transfer learning, in order to enhance understanding of unsupervised methods in NLP applications, helping craft models with capability for complex semantic distinctions.

Finally, the approach might be adapted to work with non-English languages by modifying heuristic features and clustering to suit different linguistic contexts, potentially through the use of multilingual embedding models.

Ultimately, we believe this research underscores the potential of advancing machine learning approaches with human cognitive processes to improve the accuracy and functionality of NLP applications in classifying semantic specificity.

9 Limitations

9.1 Dataset Limitations

The dataset used in this study, while effective for demonstrating the viability of unsupervised learning methods combined with human feedback, possesses several limitations that could impact the generalizability and scalability of the findings.

Scope of Data: The dataset primarily encompassed text inputs from users engaging with the Ideator platform. This specificity means the data is somewhat homogeneous, primarily reflecting the language and concerns of individuals focused on creative problem-solving. As a result, the linguistic features and thematic elements are not as varied as they might be in a more diverse corpus. This limitation could affect the model's ability to perform as effectively across different domains or broader NLP applications where the text characteristics and user intentions vary significantly.

Volume of Data: Although the dataset includes a substantial number of entries, the overall volume may still be insufficient for training more complex, deep learning models that require vast amounts of data to generalize effectively. The size of the dataset could restrict the model's ability to capture more subtle linguistic or thematic nuances that only emerge from larger, more varied datasets.

Depth of Semantic Annotation: The dataset lacks deep semantic annotations that would allow for more fine-grained analysis and classification of text. The binary classification of text into "general" or "specific" is a simplification that may overlook intermediate levels of specificity or the multifaceted nature of how text can be interpreted based on context.

Representation Bias: Given that the data was collected from a specific type of user interaction (i.e., problem-solving within a creative ideation tool), there is a potential bias towards certain types of expressions and thematic content. This bias might limit the model's effectiveness in environments with different types of text, such as more formal or technical documents.

Evolution of Language: The dataset is static and might not fully account for the evolving nature of language use over time, including new slang, terminology, or changes in the common use of phrases. This evolution could necessitate continual updates to the dataset and model to maintain accuracy.

Addressing these limitations in future studies would involve expanding the dataset to include a broader array of text sources, increasing the volume of data, and incorporating richer semantic annotations. Exploring these areas could enhance the model's robustness and applicability to a wider range of NLP tasks and environments.

References

Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. 2020. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8).

Yves Bestgen. 2023. Measuring lexical diversity in texts: The twofold length problem. *Preprint*, arXiv:2307.04626.

Design Council. 2024. Framework for innovation.

Jennifer L Heyman, Steven R Rick, Jennifer L Heyman, Gianni Giacomelli, Haoran Wen, Robert J Laubacher, Nancy Taubenslag, Max Sina Knicker, Younes Jeddi, Pranav Ragupathy, Jared Curhan, and Thomas W Malone. 2024. Supermind ideator: How scaffolding human-ai collaboration can increase creativity. In *Proceedings of The ACM Collective Intelligence Conference*.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *Preprint*, arXiv:1405.4053.

Junyi Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Xavier Schmitt, Sylvain Kubler, Jérémy Robert, Mike Papadakis, and Yves LeTraon. 2019. A replicable comparison study of ner software: Stanfordnlp, nltk, opennlp, spacy, gate. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 338–343.

Marina Solnyshkina, Radif Zamaletdinov, L.A. Gorodetskaya, and A.I. Gabitov. 2017. Evaluating text complexity and flesch-kincaid grade level. *Journal of Social Studies Education Research*, 8:238–248.

Zafer Susoy. 2023. Lexical density, lexical diversity and academic vocabulary use: Differences in dissertation abstracts. 8(2):198–207.

Jeanine Treffers-Daller, Patrick Parslow, and Shirley Williams. 2018. Back to basics: How measures of lexical diversity can help discriminate between cefr levels. 39:302–327.

Yingying Zhang and Wenyu Wu. 2021. How effective are lexical richness measures for differentiations of vocabulary proficiency? a comprehensive examination with clustering analysis. 11(15).