Unsupervised Semantic Analysis and Zero-shot Learning of Newsgroup Topics

Serge Dolgikh[1,2,*,†]

[1] National Aviation University, 1 Lubomyra Huzara Ave, Kyiv, 03058, Ukraine
[2] Solana Networks, 15 Fitzgerald Rd., Ottawa, K2H 9G1, Canada

Abstract
Semantic analysis is a well-developed area of research in the Natural Language Processing with multiple thriving directions. While Large Language Models have reached an unprecedented level of producing human author-like texts and other media, they also require massive resources, both training and computational, to achieve optimal performance. Such resources can be constrained in real learning environments or not available. In this work, we approached the problem of the semantic analysis of texts, such as newsgroup posts with the models of unsupervised generative learning and dimensionality reduction that do not require such massive resources while being effective in the analysis and differentiation of semantic content of the texts. The methods proposed in this work have a broad range of potential applicability in different fields, types of texts and languages and others.

Keywords
Semantic analysis, Natural Language Processing, stochastic NLP, unsupervised clustering, concept, zero-shot concept learning.

1. Introduction

Semantic and sentiment analysis are well-developed research directions in the Natural Language Processing field, and even outlining all areas of progress and developments can be challenging, as outlined in several recent reviews of the field [1-3]. One can note extensive research in the methods of supervised learning for both sentiment and semantic analysis, with techniques developed in representations of numerical textual data (vectorization), preprocessing, methods and models and supervised and unsupervised learning of the semantic content and structure of the corpora.

A large and rapidly growing number of results were obtained over the decades of research in this rapidly expanding area, including:

- Producing a description of the textual domain in question in terms of identified numerical characteristics or features (vectorization);
- Compilation or acquisition of semantically and syntactically relevant training sets of texts annotated with a priori known types (classes); including, most recently, massive volumes of texts used in the development of Large Language Models (LLM).
- Training with any of a wide range of regression and classification methods, including LLM;
- Verification of the trained model with realistic data, both in the supervised context with pre-annotated data and for the generation of new texts.

CLW-2024: Computational Linguistics Workshop at 8th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2024), April 12–13, 2024, Lviv, Ukraine
* Corresponding author.
† These authors contributed equally.
sdolgikh@nau.edu.ua (S. Dolgikh)
0000-0001-5929-8954 (S. Dolgikh)

Many of these approaches and methods are characterized by large, very large or massive requirement of the prior knowledge about the semantic domain being investigated, be it in the form of large annotated training datasets, pre-compiled dictionaries, etc. Large language Models (LLM) [4] may not need massive annotated sets in training; however, they do require massive resources in several aspects, including the size of the training text data and computational resources.

While conventional supervised methods of analysis of text corpora with the use of additional information about their content, such as semantic annotations achieved significant results over the years, this requirement and the dependency may present a challenge in some fields and applications for a number of reasons, some of the primary ones being: availability of large sets of prior data; their relevance to the task in question, including representativity; availability of dedicated computational resources and so on.

Unsupervised semantic analysis techniques are widely used in natural language processing. One of the effective unsupervised techniques is the lexicon-based approach, which extracts opinion lexicons and analyzes their orientation without any training data [5]. However, this technique does not address the context of words in the text. Additionally, the semantics of the lexicons may struggle with domain-specific polarity, which can lead to inaccurate results.

While several studies have explored the use of unsupervised methods for sentiment and semantic analysis, there may still be a less than comprehensive coverage of the recent results in this area. Some researchers have explored the use of natural language processing methods to detect ambiguity in requirements early on, promoting the importance of the semantic web and natural language processing techniques. Furthermore, some studies have proposed novel approaches such as Sentiment Analyzer (SA), which extracts sentiment from individual terms and sentences rather than classifying the sentiment of an entire document [6,7]. Although supervised methods have been found to outperform unsupervised methods, unsupervised semantic analysis techniques are still a promising area for future research in natural language processing and opinion mining [8,9].

More recently, large language models emerged as the latest breakthrough in NLP. Large language models are trained on massive datasets of text and use advanced natural language processing techniques to analyze and understand the meaning behind words, phrases, and sentences [4,10]. They can be used to perform semantic analysis in a variety of contexts: for example, they can be used to analyze customer feedback or social media posts to understand the sentiments and opinions of customers. However, it is important to realize that large language models are not perfect and can sometimes make errors in their semantic analysis. This can be attributed to the ambiguities in the language, context-dependent meanings, cultural differences and other factors [11].

In this work, we intended to address both types of the outlined limitations and challenges, by first, limiting the complexity of the models and relaxing the requirement on the size of the required training data (such as corpora of raw, non-annotated texts) used for semantic analysis. Secondly, we decided to altogether avoid the standard supervised approaches in machine learning that are conditioned by and strongly depend upon the availability of sufficiently large sets of training data, annotated with known semantic categories or classes, and used only unsupervised methods of learning and dimensionality reduction that do not have such dependencies.

As a result, the methods proposed and examined in this work demonstrated the effectiveness in the analysis of the semantic content without the reliance on either massively complex models or massive training data. In our view, they can be instrumental in the tasks and domains where the outlined constraints can be essential.

2. Methodology

As outlined in the introduction section, an essential challenge in the analysis of complex real-world data including textual is to develop and verify methods of analysis of its characteristic content that do not require massive prior knowledge of the semantic content. The novelty of the approach proposed in this work stems from a combination of the known methods of extraction of informative textual features developed within the NLP area over the years; methods of unsupervised dimensionality reduction [12] including manifold learning, unsupervised generative learning and other unsupervised techniques that do not depend on the annotated training data; and methods of unsupervised clustering [13]. The resulting semantic structure of the model corpora is then represented by the geometrical distribution of categories or semantic concepts in a low-dimensional space of informative features [14] that were identified in the process of unsupervised analysis and modeling of the data.

The approach to unsupervised semantic analysis of semantic content described in this work used the following broad stages of preprocessing and preparation of the text data for semantic analysis.

- Extraction of the descriptive features that represent the texts: we used a well-known in NLP statistical approach ("bag of words" model) to obtain numerical representations of texts in the corpora. Standard NLP methods can be applied to this end, such as tokenization, filtering, stemming, calculation of the frequency-based features and others [15,16].
- Strong dimensionality reduction with the methods of unsupervised learning and dimensionality reduction, such as: manifold learning, unsupervised generative learning, linear dimensionality reduction and others. As a result of this stage, highly and massively sparse feature space of very high dimensionality that is characteristic of the text data can be reduced to a low-dimensional informative latent feature space.
- Determination of the informative structure in the distributions of the text data embedded in the low-dimensional latent feature space by application of the methods of unsupervised clustering. The result of this step is a structure of latent clusters corresponding to "natural semantic types" that can be identified in the latent distributions of text data. In this work, methods of unsupervised density clustering were used [13].
- An analysis of the resulting cluster structure, specifically, the geometrical distribution of the data in the latent feature space and determination of characteristic regions of distribution that can be interpreted as the natural semantic conceptual structure of the data.

Whereas the process remains entirely unsupervised, ultimately without any dependency on the prior knowledge of the semantic content of the corpora (such as annotations with known classes commonly used in supervised methods), it allows identification of the characteristic types, patterns or "natural semantic concepts" in the text and other types of complex sensory data.

In choosing the methods of unsupervised dimensionality reduction in this work the objectives outlined in the introduction section were followed. The methods that do not require massive training data or software resources in training were used, some of which achieved impressive results in modeling the underlying semantic structure of the corpus.

2.1. Model Corpora

In the development and for verification of the proposed approach in unsupervised semantic analysis of text corpora, an openly available dataset of newsgroup posts: "20newsgroups" [17] was used. The dataset is comprised of newsgroup posts in a selection of public newsgroups, labeled (annotated) with the class associated with the newsgroup in which they posted.

Examples of texts in the newsgroup corpora:

"I was wrong! I guess they are closer to $800 new! I will probably still sell them for the above implied $300 obo. Email me if you want more specifics. This is a really attractive set of books, kind of a Bible encyclopedia set. Also email me if you know more about these books or post the information here"; annotation: commercial-for sale newsgroup.

"I thought that he was comparing Cullen to TEEMU SELINNE. I always thought that salami is some sort of sausage, BUT IF YOU, dear Roger, ARE ABLE TO SEE SALAMI ON THE ICE PLAYING HOCKEY... I don't know what to do, but you surely should do something and very quickly!!! Maybe you think that if you keep on talking some rubbish, after some time everybody will consider it to be really true... You should take care of your LEAFS, they surely need it more."; annotation: sports-hockey newsgroup.

As can be seen in these examples, the texts in the newsgroups set had significant variation of the verbal content and semantic complexity.

To verify the effectiveness of the approach with respect to the complexity of the data associated with the presence of multiple topics with different semantic content, two model corpora were constructed from the main dataset, as follows:

News_4 corpus: contained the posts from four newsgroups. Topics: atheism; Christianity; computer graphics; science–medical.

News_7 corpus: contained the posts seven newsgroups. Topics atheism; Christianity; sport–hockey; science–medical; commercial–for sale; commercial–auto; politics–Middle East.

The composition of the corpora and some characteristics of the distributions of texts are described in Table 1.

Table 1
Newsgroup corpora

Newsgroup text corpus	Size, texts	Length, min/max (words)	Length, mean/median (words)
News_4	2,257	15 / 9374	306 / 185
News_7	4,016	17 / 10765	306 / 188

2.2. Preprocessing and Preparation for Unsupervised Analysis

For the initial evaluation of the effectiveness of the proposed approach in determining the semantic content of a corpus by an entirely unsupervised method without any dependency on a prior knowledge of the semantic content, a well-known method of term-frequency vectorization ("bag of words") [15,16] that is, associating texts with numerical vectors calculated from the frequency of the occurrence of semantic tokens (words with semantic significance) in the text.

Preprocessing of the model corpora was performed with standard NLP methods of vectorization, including:

1. Filtering and standard forming: removal of semantically insignificant elements of the text ("stop words") and transforming the tokens to standard grammatical form (stemming);
2. Tokenization: separation of semantic tokens and calculation of term frequencies;
3. Calculation of the term frequency features (such as inverse term frequency, tf-idf [15] representing the relative frequency of the terms in the training subset of texts);
4. Final production of numerical annotated datasets associated with the model corpora.

Given that the grammatic content of the texts even in the same semantic group/class can differ significantly, vectorizations of texts in the "bag of words" approach commonly produce very sparse vectors associated with the texts (that is, containing a small number of non-zero elements relative to the size of the vector).

As a result of the described process, the model corpora in the study produced the following numerical datasets of the term-frequency features:

- News_4 corpus: 35,482 features, with 2,257 text samples. Data shape: (2257, 35482), annotations: (2257, 1)
- News_7 corpus: 50,767 features, with 4,016 text samples. Data shape: (4016, 50767), annotations: (4016, 1)

A distribution histogram of the texts in the corpora News_4, News_7 after preprocessing is shown in **Figure 1**. Preprocessing and feature extraction was performed with the Python library sklearn-kit, text feature extraction [18].

In conclusion of this section, it needs to be noted that the annotations were not used in any way in the proposed process of determination of the semantic content of the corpora, but only in the analysis and verification of its effectiveness.

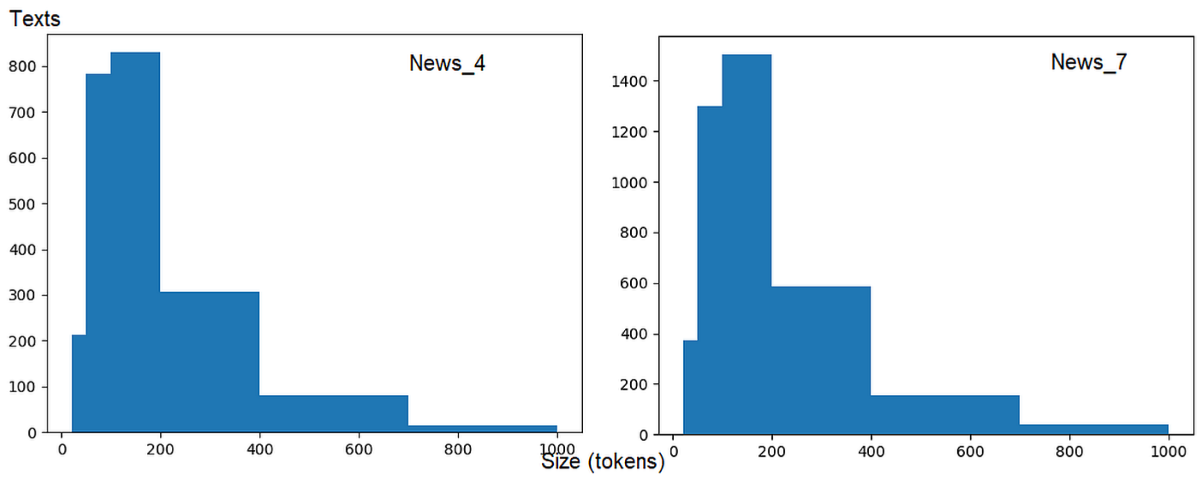

Figure 1: Distribution of texts in the model corpus by length after preprocessing (words or tokens, horizontal axis), corpus News_4 (left), News_7(right).

2.3. Unsupervised Dimensionality Reduction

As was commented, the methods of vectorization of text corpora used in this work commonly produce highly sparse data, a structural analysis of which can be complicated. To obtain more informative low dimensional representations of the data, methods of dimensionality reduction were applied.

A very broad and versatile range of methods of unsupervised dimensionality reduction was developed over the years linear such as Principal Component Analysis (PCA), kernel PCA and Singular Value Decomposition (SVD); non-linear ones, including manifold learning methods such as TSNE [19] and Uniform Manifold Approximation and Projection (UMap) [20]; generative unsupervised learning of low-dimensional embeddings with self-supervised neural network models [gen] and many others.

In our experience, linear methods were generally less successful with the corpora data of high and very high sparsity and in this work we limited ourselves to non-linear manifold learning methods of dimensionality reduction that can be applied effectively to sparse data, specifically, TSNE and UMap. An in-depth comprehensive comparison of the methods of dimensionality reduction in application to corpora data will be attempted in another study.

As a result of application of the unsupervised dimensionality reduction (UDR) to the numerical feature sets produced in the process discussed in the preceding section, we obtained low-dimensional: dimensionality two and three, informative latent feature spaces of the corpora in the study. The structure of thus produced informative representations of the original corpora sets is discussed in the Section 3.

2.4. Unsupervised Analysis of Semantic Structure

As has been reported in a number of studies, an informative structure that can be detected in informative representations of the observable data can point to characteristic types, general concepts in the data [21,22] and other results. The hypothesis of this work is that if such an informative structure could be determined successfully in the text data represented by the model corpora, it could be associated with the essential semantic differences between the subgroups of texts in the corpora, and thus be used to differentiate or distinguish texts by their semantic types. A strong reduction of dimensionality of highly sparse "bag of words" vectorization of text data is critical for successful resolution of the underlying informative structure in the latent distributions of data points because it allows to apply methods of unsupervised clustering effectively to low-dimensional distributions.

In this work, we used the method of density clustering MeanShift [23] due to its capacity to determine the structure of density regions or "clusters" without the need for prior knowledge of the essential characteristics of the distribution. Resolution of such a density structure with this method requires only a representative set of data and does not use any prior information, including semantic, annotations, etc. about it. In this approach, the hypothesis that will be examined in this work is that of a correlation between the informative structure in in the low-dimensional representations of the corpora and their semantic content such as the topic of the newsgroup.

3. Results

In this section we present the results of the application and verification of the method of unsupervised semantic analysis of corpora, as described in the preceding sections to the model corpora of newsgroup content selected and prepared for the study. Given that all steps in the process are general and are not limited to specific type, semantics of the texts, language etc., it can be reasonably expected for the method to have general applicability to corpora of similar sizes and semantic complexity.

3.1. Low-Dimensional Distributions of Corpora Texts

Low-dimensional embeddings of the corpora: News_4, News_7 were produced successfully by application of the methods of UDR discussed above: TSNE and UMap to the feature sets of the corpora produced with the process described in the preceding section. The resulting distributions of the corpus data News_7 in the low-dimensional informative feature spaces are shown in Figure 2, A-D.

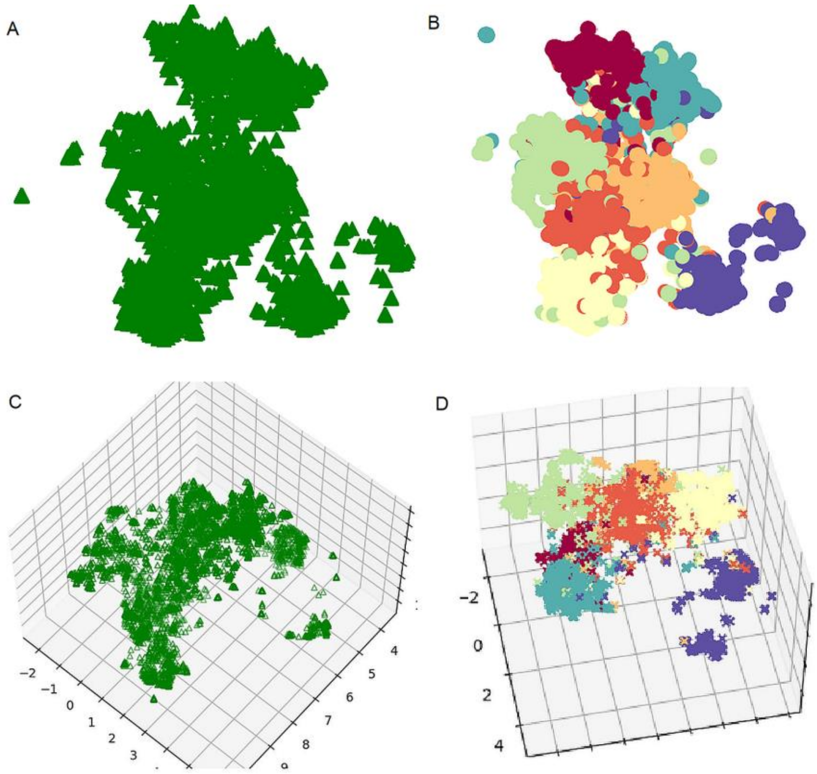

Figure 2: Unsupervised latent distributions of texts, News_7 corpus, UDR method: UMap. A, B: 2D distributions, general (left); by class annotation (right). C, D: 3D distributions, general (left); by class annotation (right).

Similar representations were obtained for the corpus News_4, and the other method of UDR, TSNE. A visual analysis of the distributions indicated a structure in the latent distributions of text data that is correlated with the semantic type. Analyzing the distributions of the annotated samples in the diagrams above one can observe a clear correlation between the geometry of the latent spaces and the distinct semantic classes.

3.2. Unsupervised Semantic Analysis of Corpora

The results presented in the Section 3.1 support the hypothesis of a correlation between the geometry of informative low-dimensional representations of text data obtained with effective methods of UDR and the characteristic semantic types in the corpora. A question arises naturally, could these semantic types be determined, learned with some method that does not depend on significant prior knowledge of the content?

Such an approach has been proposed based on the assumption of the latent similarity, whereby data points that are relatively similar in the space of observable parameters (i.e., texts in this work) tend to aggregate in an informative low-dimensional representation in distinct regions, that is, geometrical differentiation of distinct semantic types in the corpora. In more detail the assumption and the related concepts were discussed in [24]. A natural effect of this argumentation is the expected higher density of the regions associated with characteristic types in the data that allows to apply methods of unsupervised clustering, specifically, density clustering such as MeanShift and others, to be effective in identifying the regions of higher density in the latent distributions of the data.

In this work, the methods of unsupervised clustering of this type were applied to latent representations produced by application of the methods of UDR to vectorized forms of the corpora, as discussed earlier. The objective was to verify if these methods can be used to determine the characteristic density regions in the latent representations of the corpora reliably; and whether they show a correlation with the semantic type of the text, represented by its label, i.e., the newsgroup it was recorded in.

The results of the analysis are illustrated in Figure 3, that shows the plots of the latent distributions of corpora News_4 and News_7 with the density clusters identified by the density clustering (red dots representing the cluster centers).

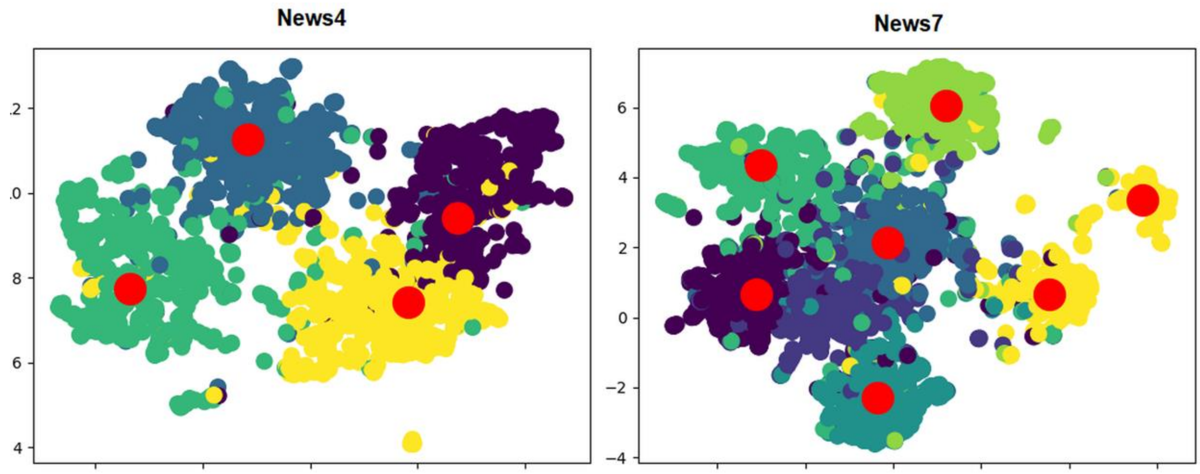

Figure 3: Distributions of model corpora (News_4, News_7) with the identified clusters of natural semantic types**.**

The results presented in this section demonstrate that the semantic structure of corpora, at least those similar in composition and complexity to the studied here, can be established reliably by the proposed approach.

It is worth reiterating that every step in the process of producing the set of geometric clusters in the latent space, 4 in the case of the corpus News_4 (left, Figure 3) and 7 or 8, News_7 (right) was entirely unsupervised and did not rely on any information about the semantic content of the corpus, such as annotations with the associated newsgroup.

Thus, by applying the method to unannotated data (the illustration in Figure 2) one would establish that the corpus News_4 contained four semantically different types of texts, whereas News_7: seven or eight. Moreover, the methods allow to associate the instances of texts to one of the identified "natural" semantic types, as long as they satisfy the common assumptions of the "bag of words" approach, as will be shown in the next section.

3.3. Zero-Shot Learning of Natural Semantic Types

The natural conceptual structure in the informative latent distributions of data of any input type can be analyzed with a number of different methods [22,23]. A straightforward approach chosen in this work was to produce the semantic classifiers, based on the structure of density distribution that can be resolved with entirely unsupervised methods, as discussed in the preceding sections. There is more than one way to construct such classifiers.

The most direct approach would be to the density clustering method in the reduced dimensionality latent space, $K(t_l)$ as a classifier of natural latent types or concepts. One can recall that it allows to associate a text sample t in the input feature space to a natural semantic concept (cluster) in the latent representation space produced with a UDR method $E_m(t)$ as:

$$K_l(t) = K(E_m\ (F(t)) \quad (1)$$

where, according to the stages of the process described earlier, F is the feature set associated with t; E_m, a low-dimensional UDR embedding; K, the unsupervised clustering method in the resulting informative latent space.

The benefit of the application of UDR, a strong reduction of dimensionality of the input feature space $F(T)$ is that makes the clustering both more stable and more operationally efficient. While this method of resolution of the natural semantic types in the corpora is direct and straightforward, it may have as a downside a strong dependency of the classifier performance on the specifics of the realization of the clustering method, that may show more or less stable results with different distributions of similar data.

A different approach, sampling-based, is to use the density structure resolved by unsupervised clustering as a general geometrical map of the latent distributions of the distinct semantic regions in the latent space and build semantic classifiers on the basis of the identified cluster structure, for example, based on a sampling method. While the method was described in detail earlier [22], the approach can be briefly outlined here:

For a given semantic type of interest, represented by a significant (relative to the overall sample) latent cluster K_j, two sets of samples: positive (in-class) and negative (other, out-of-class) are created based on the geometrical proximity of the selected points to the latent position of K_j (e.g., its geometrical center that can be determined by the clustering method).

With the set of samples thus produced, a binary annotated set can be created based on the criterium "in/out-of-class" (True: in-class samples; False: other), C_j operating in the latent feature space, usually of geometric type such as Nearest Neighbor can be trained. A trained semantic classifier then allows to determine a membership of an input sample, i.e. a text t to the semantic type K_j as:

$$p(t, K_j\ ) = C_j(E_m(F(t)) \tag{2}$$

where $p(t, K_j)$: the probability of the text t belonging to the semantic type K_j; E_m, $F(t)$: as earlier, the UDR embedding method and the feature set associated with the sample t.

Thus, the clustering method K and natural concept classifiers $\{ C_j \}$ can produce predictions of the association of texts to the natural semantic types present in the corpus.

An essential difference can be noted here between the annotations (the newsgroup) available from the outset and the natural semantic types determined by the unsupervised semantic analysis method described here. Whereas the set of known classes has to be provided externally, thus representing the prior, external knowledge about the data that has to be provided to conventional machine intelligence models in the process of supervised learning, the presented method of unsupervised semantic analysis is not dependent on such knowledge and is capable of determining the structure of the natural semantic types, including the distributions of texts between them (2) in an entirely unsupervised process that is not dependent on massive amounts of prior knowledge.

With the classifiers of natural semantic types produced in this way, one can use the annotations available with the sample corpora to verify the hypothesis of the association between the natural semantic types and the known "external" classes of the texts in the corpora. For each type of the classifier, clustering-based and sampling-based, two confusion matrices were calculated:

- The external class (that is, the newsgroup label) confusion matrix, describing the distribution of external classes (newsgroup type) between the natural semantic types determined by unsupervised semantic analysis, normalized by the external class sample.
- The natural semantic type confusion matrix, describing the distribution of the natural semantic types between the external classes (newsgroup labels), normalized by the sample of the natural concept.

An example of verification of the method of unsupervised semantic analysis with the corpus News_7 of newsgroup texts and the sampling-type classifier is presented in Figure 4. The results obtained with the corpus News_4 were superior, possibly due to a simpler semantical composition of the corpus.

In these results, a stable correlation between natural semantic types identified with the method, and the external annotation (newsgroup) can be observed.

Interestingly, it was observed that the unsupervised analysis can provide more detailed insight than that available from the annotation. Consider the class (newsgroup) 6 above. According to the label-type composition matrix (left), it was divided between two identified natural semantic types: 6 and 7. A closer examination of the contents of the semantic types indicated an essential difference in the discussed topics: whereas posts in the cluster 6 were focused on the Arab-Israel relations, those in the other cluster, 7 dealt with more general issues of history, interactions of faiths and different geographical regions. Thus, unsupervised semantic analysis proposed in this work can provide more detail semantic decomposition of the texts than what can be obtained from the external label (annotation) information alone.

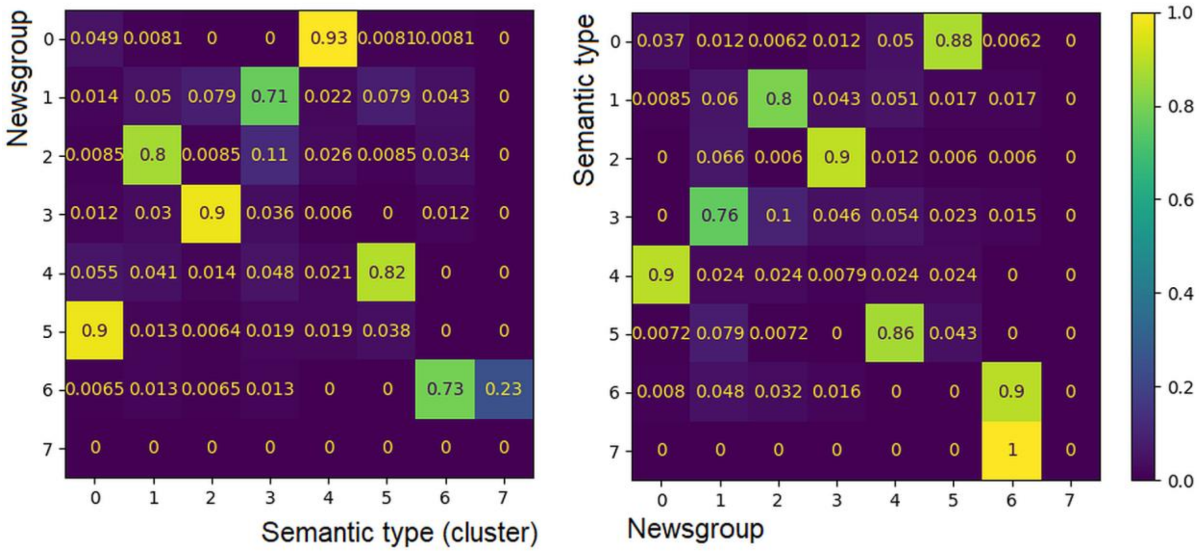

Figure 4: Correlation between the natural semantic types and the annotation, News_7 corpus 1.

The summary of the results of verification of the method of unsupervised semantic analysis with the model corpora in this work is provided in Table 2. We show the minimal, maximum and the mean correlation between the external class (newsgroup annotation) and an identified semantic type for two types of classifiers: clustering and sampling, as discussed earlier in this work.

Table 2
Unsupervised learning of natural semantic types (UMap, clustering/sampling classifiers)

Corpus	Classifier: clustering min / max / mean correlation (1)	Classifier: sampling min / max / mean correlation (1)
News_4	0.63 / 0.92 / 0.77	0.78 / 0.93 / 0.87
News_7	0.58 / 0.91 / 0.74	0.71 / 0.92 / 0.83

(1) The mean of the association class (newsgroup annotation) – natural semantic type over all annotated classes in the corpora.

As can be concluded from these results, a close correlation was observed between the natural semantic types of the corpora resolved by the method of unsupervised semantic analysis in the low-dimensional latent distributions of the corpora produced with the methods of UDR, and the external classes of texts, that is, the newsgroups labels. These results demonstrate that unsupervised analysis of text data by methods of unsupervised non-linear dimensionality reduction and unsupervised density clustering can describe the characteristic content of general text data without significant prior knowledge about its essential characteristics, including the semantic content.

4. Conclusion

Within the constraints outlined in the objectives of this work, the models of unsupervised dimensionality reduction and clustering considered here offer a resource-efficient and effective direction of unsupervised semantic analysis of text corpora by methods that do not depend on massive prior information about the semantic content, such as semantic annotation, nor massive resources in preparation and training. The results presented here demonstrate that the proposed approach can be applied successfully to such corpora and provide an effective means to evaluate their composition and semantic content.

The structure of natural semantic types as clusters of concentrations of data in the informative low-dimensional embeddings of corpora can provide essential insights into their semantic composition without, ultimately, any prior knowledge about their content. The only essential assumption that guides the applicability of the proposed approach is a sufficient semantic representativity of the model corpora used to extract the semantic term features.

As was demonstrated in this work, the corpora of moderate size exemplified by the newsgroup sets can be sufficient to determine the semantic type of related texts with a high level of accuracy. The presented example where unsupervised semantic analysis provided more detailed information about the corpora contents than the available annotation demonstrates the potential of the discussed methods for the analysis of texts with minimal prior information.

The approach in unsupervised semantic analysis of corpora demonstrated in this work can be immediately extended to other fields and categories of texts and languages, with a number of possible practical applications in semantic content analysis, detection of malicious content and other applications.

References

[1] S. Sun, C. Luo, J. Chen, A review of natural language processing techniques for opinion mining systems, Information Fusion 38 (2017) 10–25. doi: 10.1016/j.inffus.2016.10.004.

[2] S.A. Salloum, R. Khan, K. Shaalan, A Survey of semantic analysis approaches, in: Hassanien, AE., Azar, A., Gaber, T., Oliva, D., Tolba, F. (eds) Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020). AICV 2020. Advances in Intelligent Systems and Computing, 1153. Springer, Cham doi: 10.1007/978-3-030-44289-7_6.

[3] A. Lenci, Distributional models of word meaning, Annual Review of Linguistics 4 (2018) 151–171 (2018). doi: 10.1146/annurev-linguistics-030514-125254.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, ACL Anthology 19 (2018) 1423. doi: 10.18653/v1/N19-1423.

[5] T. Hofmann, Unsupervised learning by probabilistic Latent Semantic Analysis, Machine Learning 42 (2001) 177–196.

[6] J. Yi, T. Nasukawa, R. Bunesku, W. Niblack, Sentiment Analyzer: extracting sentiments about a given topic using Natural Language Processing techniques, in: 3rd IEEE International Conference on Data Mining (ICDM-2003), Melbourne Florida 2003.
[7] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner et al., Deep contextualized word representations, arXiv:1802.05365 (2018).
[8] P. Waila, Marisha, V.K. Singh, M.K. Singh, Evaluating Machine Learning and unsupervised semantic orientation approaches for sentiment analysis of textual reviews, in: 2012 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, India, 2012, 1–6. doi: 10.1109/ICCIC.2012.6510235.
[9] C. AP, S. Lauly, H. Larochelle, M. M. Khapra et al.: An autoencoder approach to learning bilingual word representations, in: 27th International Conference on Neural Information Processing Systems, Montreal, Canada 2014, 2 1853–1861.
[10] C. Yupeng, W. Xu, W. Jindong et al., A Survey on evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology (2024) doi: 10.1145/3641289.
[11] A. Bruno, A., P.L. Mazzeo, A. Chetouani et al., Insights into classifying and mitigating LLMs' hallucinations, in: Artificial Intelligence for Perception and Artificial Consciousness 2023 (AIxPAC 2023) Roma, Italy, 2023 CEUR Workshop Proceedings, 3563 50–63.
[12] J. A. Lee, J.A., M. Verleysen, Unsupervised dimensionality reduction: overview and recent advances, in: 2010 International Joint Conference on Neural Networks (IJCNN), 2010, Barcelona, Spain doi: 10.1109/ijcnn.2010.5596721.
[13] R. J. Campello, P. Kroger, J. Sander, A. Zimek, Density-based clustering, WIREs Data Mining and Knowledge Discovery 10 (2) (2019) 1343. doi: 10.1002/widm.1343.
[14] S. Dolgikh, Low-dimensional representations in generative self-learning models, in: 20th Conference Information technologies - Applications and Theory ITAT-2020 Oravska Lesna Slovakia 2020, CEUR-WS.org 2718 239–245.
[15] K. Sparck Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, 28(1) (1972) 11–21.
[16] C.D. Manning, P. Raghavan, H. Sch¨utze, Introduction to Information Retrieval, Cambridge University Press, Cambridge 2008.
[17] 20 Newsgroups dataset, 2017. URL: http://qwone.com/ jason/20Newsgroups/.
[18] Pedregosa, Varoquaux, Gramfort et al., Scikit-learn Tfidf vectorizer, 2011.
[19] G. Hinton, S. Roweis, Stochastic neighbor embedding, in: NIPS'02, Advances in Neural Information Processing Systems 2002, 15.
[20] McInnes, L., Healy, J., Melville, J., UMAP: Uniform manifold approximation and projection for dimension reduction, arXiv, 1802.03426 (2018).
[21] Q. V. Le, M. A. Ranzato, R. Monga, et al., Building high-level features using large scale unsupervised learning, in: 29th International Conference on Machine Learning (ICML'12), 2021, 507–514.
[22] S. Dolgikh, Categorized representations and General Learning, in: 10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions (ICSCCW-2019), Prague Czech Republic 2019. doi: 10.1007/978-3-030-35249-3_11.
[23] K. Fukunaga, L.D. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition, IEEE Transactions on Information Theory, 21 (1) (1975) 32–40.
[24] S. Dolgikh, Categorization in unsupervised generative self-learning systems, International Journal of Modern Education and Computer Science 3 (2021) 68–78. doi: 10.5815/ijmecs.2021.03.0.