Unsupervised Discovery of Thematic Structure in Russian-Mansi Bilingual Texts

Baranetskaya Sofia, Korobkovskii Vadim, Shkarovskii Vladislav

2026

Abstract

This paper presents a comparative study of topic modeling approaches applied to a low-resource Russian–Mansi parallel text corpus. The primary objective of the project is to construct thematically structured corpora that can support language learning and linguistic analysis of the Mansi language. We explore and evaluate several unsupervised topic modeling techniques, including Latent Dirichlet Allocation (LDA), BERTopic, and sentence embedding–based clustering using multilingual transformer models.

Special attention is given to data preprocessing and exploratory data analysis, as the corpus contains inconsistencies such as sentence misalignment, duplicates, and formatting artifacts. After data cleaning, topic models are trained and evaluated using coherence metrics and qualitative analysis of topic interpretability. The proposed sentence embedding clustering approach demonstrates superior performance compared to LDA and BERTopic, achieving higher coherence scores and producing more semantically interpretable topics on both the Russian–Mansi corpus and the 20 Newsgroups benchmark dataset.

The results indicate that sentence embedding–based clustering is a robust and effective method for topic modeling in low-resource language settings. All code and experiments are publicly available at `https://github.com/0z0nize/NLP-course-Autumn-2025`.

1 Introduction

The Mansi language, a Finno-Ugric language spoken by the Mansi people in Western Siberia, is considered a low-resource language. Although it holds significant cultural and linguistic value, there is a notable shortage of comprehensive learning materials, especially structured thematic corpora for students of the language.

This project seeks to fill this gap by applying topic modeling to a Russian–Mansi corpus to develop thematic corpora that assign sentences to clearly defined topics. By organizing the corpus into such topical categories, language learners will be able to efficiently locate and use sentences connected to specific themes or areas of interest. This method is expected to improve the learning experience for students of Mansi and encourage engagement with the cultural and contextual narratives that the language conveys.

1.1 Team

This section contains a list of our team members and a description of what they worked on as part of the project.

1. **Baranetskaya Sofia** perfomed an EDA on russian-mansi corpora and prepared data for following work, wrote and reviewed parts of this article, helped with model training.

2. **Korobkovskii Vadim** wrote an overview of related studies and problem statement, performed word embedding clustering, and BERTopic topic modeling, also helped with 20 newsgroup dataset;

3. **Shkarovskii Vladislav** wrote overview of related studies, performed experiments, helped with overview for 20 newsgroup dataset and trained LDA model.

2 Related Work

According to the most recent estimates, there are approximately 1,000 to 1,500 speakers of the Mansi language. However, the number of speakers has been declining, particularly among younger generations, due to factors such as urbanization, assimilation, and the dominance of Russian as the national language. Given the absence of established research on topic modeling specifically focused on Mansi texts, this section will explore the general methodologies employed in topic modeling. For each approach discussed, appropriate references will be provided to support the analysis.

2.1 Latent Semantic Allocation

The first method we are going to discuss is algebraic model called Latent Semantic Allocation (LSA). LSA starts by creating a term-document matrix, a mathematical matrix that represents the frequency of terms (words) in a collection of documents. Rows correspond to terms, and columns correspond to documents, with each entry reflecting the count or frequency of the term in the document. The core of LSA involves performing Singular Value Decomposition (SVD) on the term-document matrix. SVD decomposes this matrix into three smaller matrices:

- U: a matrix of term vectors (terms projected into semantic space),

- U: a diagonal matrix containing singular values (which indicate the importance of corresponding vectors),

- V^T: a matrix of document vectors (documents represented in semantic space).

This decomposition captures the relationships between terms and documents, revealing the underlying structure in the data by reducing its dimensionality. Each reduced-term vector can be interpreted as a mixture of topics, while each document vector indicates the presence of these topics in the document. Therefore, documents that are close together in this semantic space are assumed to be related by topics.

2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative statistical model used in natural language processing and machine learning for topic modeling. It helps identify underlying topics present in a collection of documents. LDA assumes that documents are mixtures of topics, where each topic is characterized by a distribution over words. The model can automatically discover the topics that pervade a set of documents

LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document.

2.3 Non-negative Matrix Factorization

NMF is a matrix factorization technique aimed at extracting interpretable patterns from high-dimensional data. It decomposes a non-negative matrix (e.g., document-term matrix) into two lower-dimensional non-negative matrices: one representing the basis vectors (topics) and the other representing the coefficients (document-topic associations). NMF has shown promising results in text mining and topic modeling, due to its ability to produce parts-based representations (Lee & Seung, 1999; Zhao et al., 2014). It also showed success with short texts that have limited contextual information.

2.4 Word Embedding / TF-IDF with Clustering

Word Embeddings capture contextual relationships between words, allowing for a dense representation of semantic meanings (Mikolov et al., 2013). This embeddings used alongside with clustering techniques like K-means and hierarchical clustering to identify groups of similar documents present us with clusters than can be easily considered as related to one topic group,

2.5 BERT and Transformer-Based Models

BERTopic, a topic model that extends this process by extracting coherent topic representation through the development of a class-based variation of TF-IDF.

More specifically, BERTopic generates document embedding with pre-trained transformer-based language models, clusters these embeddings, and finally, generates topic representations with the class-based TF-IDF procedure. BERTopic generates coherent topics and remains competitive across a variety of benchmarks involving classical models and those that follow the more recent clustering approach of topic modeling.

3 Model Description

We have opted to explore word-embedding clustering, a technique that is infrequently employed for topic modeling. This approach will be compared with two other established methods: Latent Dirichlet Allocation (LDA) and BertTopic, in the context of topic modeling for our Russian-Mansi corpus.

3.1 Word embedding with clustering

Word embeddings have revolutionized the field of Natural Language Processing (NLP) by providing dense vector representations of words based on their contextual usage. Traditional methods, such as Word2Vec and GloVe, focus on individual words; however, recent advancements have extended these techniques to full sentences.

Word embeddings, such as those generated by Word2Vec and GloVe, capture semantic relationships between words by converting them into dense vector representations. These methods leverage the distributional hypothesis, which posits that words appearing in similar contexts tend to have similar meanings. However, as the complexity of language increases, it becomes imperative to extend these representations beyond individual words to encompass entire sentences.

In this study, we utilize the Text2Vec multilingual transformer, a state-of-the-art sentence embedding model, specifically fine-tuned for Russian language sentences. Text2Vec maps each input sentence to a 384-dimensional dense vector space, thus facilitating various NLP tasks, including semantic search, text matching, and sentence similarity assessments.

After obtaining the dense representations of the sentences using the Text2Vec model, we proceed to the clustering phase. Clustering is a vital technique in unsupervised learning, allowing for the identification of inherent groupings within the data. In this context, we apply two prominent clustering algorithms: k-means and fuzzy c-means.

K-means is one of the most widely utilized clustering algorithms due to its simplicity and efficiency. The algorithm partitions the dataset into k distinct clusters by minimizing the within-cluster variance, represented mathematically as the sum of squared distances between each point and the centroid of its assigned cluster. The centroid, defined as the mean of all points within a cluster, serves as a representative point for that cluster.

While k-means is particularly effective for large datasets, its performance is heavily contingent upon the initialization of the centroids and the pre-defined value of k, which must be specified prior to the analysis. This requirement can introduce bias and affect the robustness of the clustering results.

In contrast to k-means, fuzzy c-means (FCM) is a clustering method that allows for a degree of membership of data points in multiple clusters. This approach is particularly useful in scenarios where data points may exhibit characteristics of more than one cluster, thus facilitating a more nuanced representation of the underlying data structure.

Fuzzy c-means minimizes an objective function that incorporates both the positions of the centroids and the membership grades of the data points. By allowing for soft clustering, FCM effectively manages overlapping clusters, making it a potent alternative to traditional k-means in specific applications.

3.2 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a widely used and interpretable generative model that aims to identify latent topics within a corpus of text. Its effectiveness in topic modeling stems from several key assumptions:

- Distributional Hypothesis: This principle posits that words that frequently co-occur in documents are likely to be semantically similar or related. This assumption underlies the way LDA groups words into topics.
- Topic Mixture: Each topic is represented as a mixture of various words, capturing the nuanced relationships and themes present in the data.
- Document Mixture: Each document, in this context represented as a single sentence from our corpus, is characterized as a mixture of multiple topics, which allows for a richer representation of the content.

As a generative model, LDA seeks to uncover the latent structure that generates the observed documents and topics. Unlike traditional statistical language models that typically generate a document by sampling from a fixed probability distribution over words, LDA employs a more complex mechanism. It assumes a predefined number of topics, denoted as k. For each document m, there exists a probability distribution over these k topics. Moreover, each topic is represented as a probability distribution throughout the vocabulary V.

The generative process of LDA can be mathematically expressed using the following joint probability distribution:

$$P(W, Z, \theta, \phi; \alpha, \beta) = \prod_{i=1}^{K} P(\phi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) P(W_{j,t}|\phi_{z_{j,t}})$$

In this equation:

- $\prod_{i=1}^{K} P(\phi_i; \beta)$ represents the Dirichlet distribution that models the distribution of terms across topics. Here, ϕ_i indicates the distribution of words for topic i, and β is the parameter controlling the sparsity of this distribution.

- $\prod_{j=1}^{M} P(\theta_j; \alpha)$ denotes the Dirichlet distribution governing the distribution of topics within documents. The variable θ_j signifies the topic distribution for document j, while α is the parameter that influences the diversity of topics in each document.

- $P(Z_{j,t}|\theta_j)$ is the conditional probability of a topic Z assigned to a specific word in document j at position t, given the topic distribution θ_j. This models how likely it is for a topic to be selected when generating a word in the document.

- $P(W_{j,t}|\phi_{z_{j,t}})$ captures the probability of observing a word W at position t in document j, given the topic assignment $Z_{j,t}$. This reflects the likelihood of generating a particular word based on the selected topic.

3.3 BERTopic

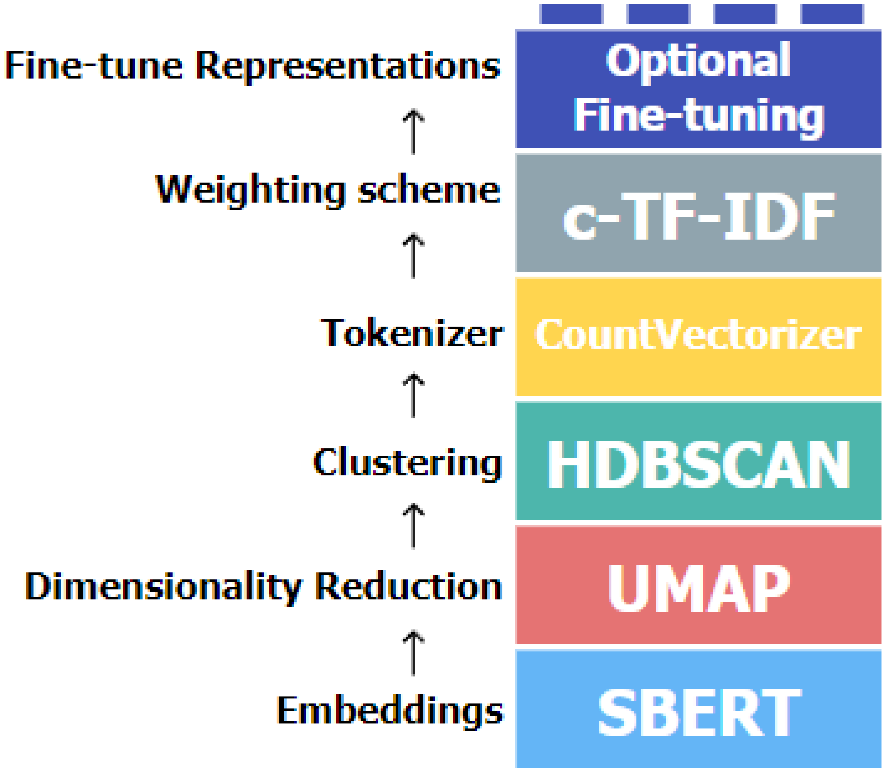

Figure 1: BERTopic visual overview.

BERTopic is a topic modeling technique that leverages transformer-based embeddings, particularly those generated by models like BERT (Bidirectional

Encoder Representations from Transformers), to identify and extract topics from a collection of documents. It is designed to improve the quality and interpretability of topics compared to traditional topic modeling methods such as Latent Dirichlet Allocation (LDA).

BERTopic begins by converting documents into dense vector representations using transformer models. The embeddings capture semantic information, allowing for a more nuanced understanding of text content.

After obtaining the embeddings, BERTopic applies a dimensionality reduction technique, UMAP (Uniform Manifold Approximation and Projection), to project the high-dimensional embeddings into a lower-dimensional space. It is a technique that can keep some of a dataset's local and global structure when reducing its dimensionality. This structure is important to keep as it contains the information necessary to create clusters of semantically similar documents.

The next step involves clustering the reduced embeddings to identify groups of similar documents. This is performed using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). It can find clusters of different shapes and has the nice feature of identifying outliers where possible. As a result, we do not force documents into a cluster where they might not belong. This will improve the resulting topic representation as there is less noise to draw from.

Each cluster is examined to derive the topics. For this, BERTopic uses bag-of-words representation on a cluster level and not on a document level. By using a bag-of-words representation, no assumption is made concerning the structure of the clusters. Moreover, the bag-of-words representation is L1-normalized to account for clusters that have different sizes.

In order to discern the distinguishing characteristics of one cluster relative to others within a generated bag-of-words representation, it is essential to identify the words that are prevalent in a specific cluster but not as prominent in the remaining clusters.

Typically, conventional TF-IDF is employed to evaluate the significance of terms across a collection of documents, enabling comparisons of word importance within the context of document sets. However, if we treat all documents belonging to a particular category — such as a cluster — as a singular document, we can then compute TF-IDF scores for the words present in that cluster. This approach allows us to identify the importance of terms within the context of a specific topic. Consequently, the words that yield higher importance scores within a cluster serve as salient representatives of that cluster's thematic content. This methodology is referred to as class-based TF-IDF.

The formula for class-based TF-IDF can be expressed as follows:

$$\text{Class-based TF-IDF}(t, c) = \text{TF}(t, c) \times \log\left(\frac{N}{|d \in D : t \in d, d \in c|}\right)$$

where t represents the term, c denotes the cluster, N is the total number of documents, and $|d \in D : t \in d, d \in c|$ is the count of documents within the cluster c that contain the term t.

Optionally, one can fine-tune BERTopic topic representations: there is the possibility to further fine-tune these c-TF-IDF topics using GPT, T5, KeyBERT, Spacy, and other techniques.

More specifically, we can consider the c-TF-IDF generated topics to be candidate topics. They each contain a set of keywords and representative documents that we can use to further fine-tune the topic representations. Having a set of representative documents for each topic is huge advantage as it allows for fine-tuning on a reduced number of documents. This reduces computation for large models as they only need to operate on that small set of representative documents for each topic. As a result, large language models like GPT and T5 becomes feasible in production settings and typically take less wall time than the dimensionality reduction and clustering steps.

4 Dataset

In this work, a parallel corpus was used, which contains 81,146 pairs of Russian-Mansi texts. In this section, we examine our dataset and conduct an analysis to better understand what we are dealing with. The dataset is publicly available[1].

4.1 Description

The raw dataset contains the following statistics:

Name	Russian	Mansi
Num of texts	81,146	
Num of sentences	156,074	155,632
Num of words	690,832	926,801
Num of unique words	63,018	64,250

Table 1: Initial statistics on the dataset.

If the difference in the number of words is acceptable, since the languages are different, then the difference in the number of sentences is strange. In the Subsection 4.2 we will see why this is so, so let's move on to it.

4.2 EDA

Exploratory data analysis(EDA) is an approach of analyzing data sets to summarize their main characteristics. This section is very important not only for understanding the dataset, but also for finding errors and inconsistencies, correction or removal of which will lead to an improvement in the quality of the dataset and, accordingly, the model.

Information was received from the case creator about the availability of cases where words are written with a space (example: Г Е Р М А Н И Я). To begin

[1] link to parallel corpus

with, we decided to check the number of these cases using regular expressions. In total, we have 847 different offers with this defect in Mansi and only 7 sentences in Russian. Russian cases were fixed (since Mansi was not used for the topic modeling task, they were not fixed, since the sentences in Russian were normal. But for the translation task such cases were removed, because in Mansi it was difficult to understand which form of the word was correct).

А н а т о л и й В а л е й в совхозе « С а р а н п а у л ь с к и й » руководителем работает, он сказал, в этом году 25 оленых людей для соревнований с гор спустились
У моей бабушки по матери младший брат – Сайнахов И в а н Д м и т р и е в и ч .
Михаил Ку з ь м и ч М ол д а н о в второе место взял и Попов Иван Алексеевич третьим был

Table 2: Example of sentences.

For better readability, the following statistics have been recorded in the Table 3.

Name	Value	Comment
Num of duplicates	267	complete duplicates were also found and deleted
Num of missing values	0	Nan, empty rows
Avg length of sentences	8	Mansi
Avg length of sentences	8	Rus
Max length of sentences	150	Mansi | it varies greatly with the average length
Max length of sentences	188	Rus | it varies greatly with the average length

Table 3: Statistics on the dataset.

The statistics look suspicious, since the average length of sentences varies greatly from the maximum. It turned out that the text simply wasn't broken down into sentences:

And this is not an isolated case. There are about 463 such cases in total. For all these cases, the sequence numbers were removed using regular expressions, and for sentences whose length is greater than 128, they were divided into smaller parts. The number of sentences whose length is greater than 128 is 1.

Now let's look at the distribution of sentence lengths (fig. 3).

It can be seen from the bar chart that the distribution is similar to exponential or Chi-square. Most of all, of course, are short sentences.

When visualizing the correlation between the length of the Russian and the Mansi sentences, a strong discrepancy between the samples was noticed (Fig. 3).

34. Ирина Константиновна Поята Ха льӯс район Ло пмус па вылт самын патыс. 35. А ще Константин Корнилович, оматэ Ольга Максимовна Албиныг о лсы г. 36. Э кваг-о йкаг колта глэ нт китхуйплов ня врам янмалтасы г. 37. Ань са т хо тпа хультыс.38. Ирина школа а стламе юи-па лт Салехард ӯс медучилищан ханищтахтукве минас. 39. А стламе юи-па лт Ямал ма н Тарко-Сале ӯс пӯльницан рӯпитакве ке̄твес. 40. Тот мощ рӯпитас ос хум ва рыс. 41. О йкатэ н Молдавия ма н о лукве тотвес. 42. Тувыл 1990 та лт тэ н ювле Ха льӯсн ще̄мья та гыл ва нтлысы г. 43. Нэ пӯльницат терапевт-ле̄ккарн нё тым нёловхуйплов та л рӯпитас. 44. Та юи-па лт ос физиотерапия ва рмаль щирыл э лаль ханищтахтас. 45. Ань ты пӯсмалтан ва рмаль щирыл Ха льӯс пӯльницат китхуйплов та л рӯпиты.46. Ирина Константиновна о йкатэ нтыл кит ня врам янмалтасы г. 47. Пыге̄н юридический академия а стлас. 48. Ань Ха льӯст о лы. 49. А гитэ н ос Ханты-Мансийск ӯс педколледжит нилыт та л ханищтахты. 50. Ма н ма ньщи ще̄мьят йильпи та л кастыл яныглыянув!51. О лупсанын кӯпнитыг вос о лы. 52. Йильпи та лт на н щуниыг вос е̄мтэ гын.53. Рӯтанын, юртанын ёт са в та л пуста гыл о лэ н.

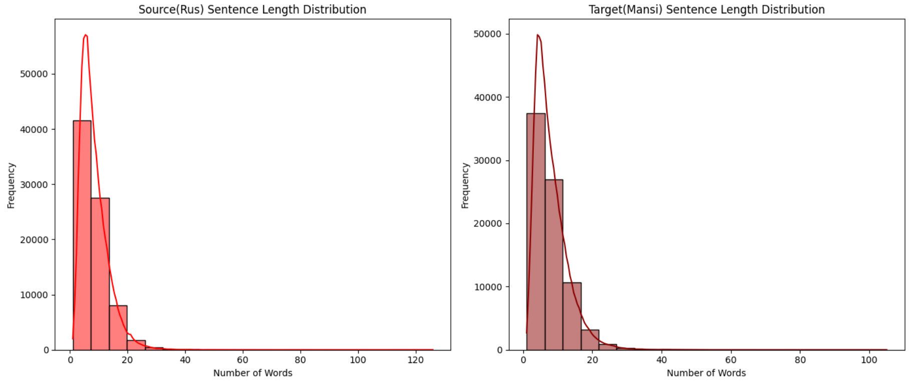

Figure 2: Distribution of sentence lengths.

You can notice distribution of the length difference between Russian and Mansi sentences in the Table 4.

A treshold was set for a length difference of 25. We see the following examples in Table 5.

It was noticed that the sentences were not translated to the end, but only the first parts. It was done like this: if mansi > rus, then we take the first sentence, otherwise we leave it as it is. After these manipulations, we can observe the following picture on Fig 5.

An analysis of the frequency of words was also carried out. For Russian sen-

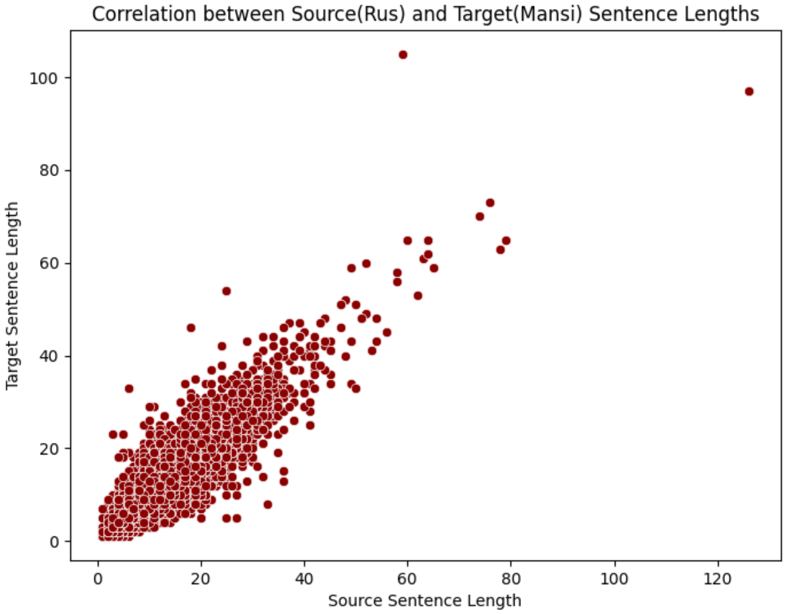

Figure 3: Correlation between the length of the Russian and the Mansi sentences.

Name	Value
count	79593.000000
mean	1.324790
std	1.453326
min	0.000000
25%	0.000000
50%	1.000000
75%	2.000000
max	46.000000

Table 4: Distribution of the length difference between Russian and Mansi sentences.

tences, stop words were cleared and the following picture was obtained (Fig. 6).

We don't know the stop words for the Mansi language, so we decided to display the frequency of words directly (Fig. 7).

Mansi	Rus
Тувле кос ёмантас, луве ла ви: Ул минэн, тыг йиен! На тах хотум мори тыномылматэгн,экваты сайкалы, о лум пасмен тый; ам нёлссам асагумн сялтэн, тот капак о лы, та капакта хурум сярка аен!	Конь говорит: Не уходи, иди сюда!
купса Сибиряков (Щипрах) лёнх оньщас - Щипрах лёнхыг (Сибиряковскии тракт) лавыглавес. Ялпын хоталт ёхталас Халь кс кущай Фомин В.И., Хулюмсунт миркол кущай Ануфриев Я.В. Янитлавест махум хотит олсыт н1лсат арыгкем тал. иильпи самын патум нявраамыт. Янитлавест акван хасхатам (олмыгтам) махум. Сав потыртавест хоти махум олэгыт хоса аквёт (пурияныл- иив, келп, аргин, сорни). Няксимволь урыл эрыг хансум эква янитлавес. Няксимволь миркол кущай Волклва Т.К., потыртас, латын лавыс- хоти махум нетсыт ялпын хотал варункв. Няксимволь сав сунсылтаве Михаил Заплатин, Лев Вахитов - кинат, Игошев картинат мот хон мат тотыглавет, музей Ханты-Мансииск олэгыт. карыс сип ватат- павлув,тагт я овтохти (хайти витэ). Олнэ вармалюв миnанти, та олантев - павылрисювт!	Первым поселением была манси деревня. потом из-за урала в поисках лучших мест для жилья пришли коми-зыряне. Из воспоминаний Е.М.Носовой, которая была в числе первых переселенцев она рассказывала: шли из-за Урала от голода, надеялись на рыбные места, пушнину. Мы выжили благодаря Няксимволю. Сосьвинская пристань Сибирякова - это единственное в данной местности поселение русский пункт, отстоящии от Березово в 500 верстах.

Table 5: Examples of parallel sentences with a large difference in length.

4.3 The final distribution for training

Your description will likely be including a table. On the Tab. 6 you can see the statistics for the mentioned dataset. It is important to notice that there is a split of the dataset and this split is covered by the description.

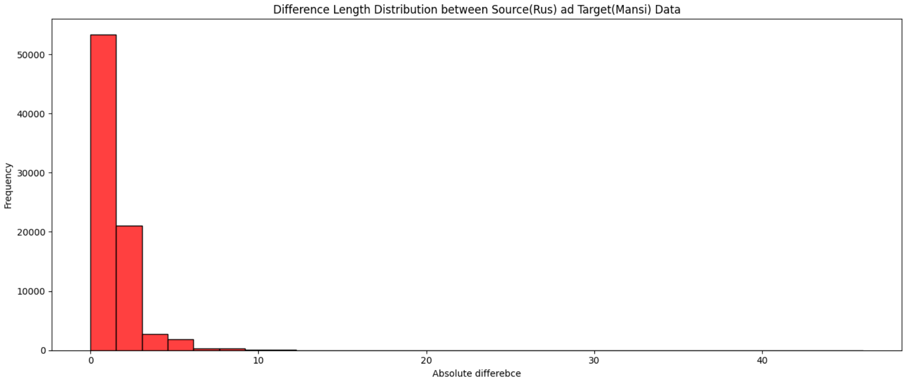

Figure 4: Difference Length Distribution between rus and mansi sentences.

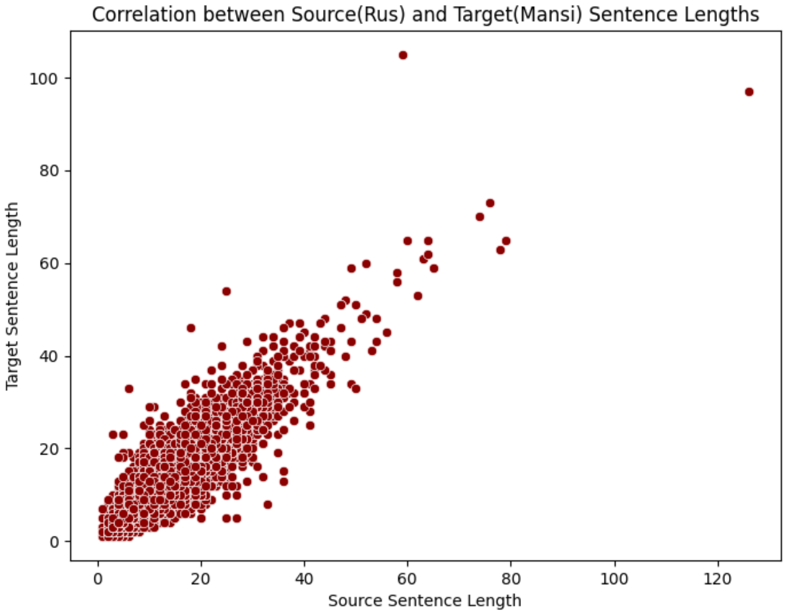

Figure 5: Correlation between the length of the Russian and the Mansi sentences after manipulations.

Word Cloud for Source Text

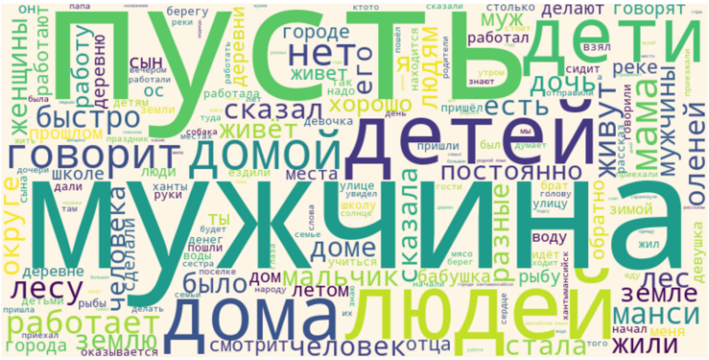

Figure 6: The frequency map of Russian words in all sentences.

Word Cloud for Target Text

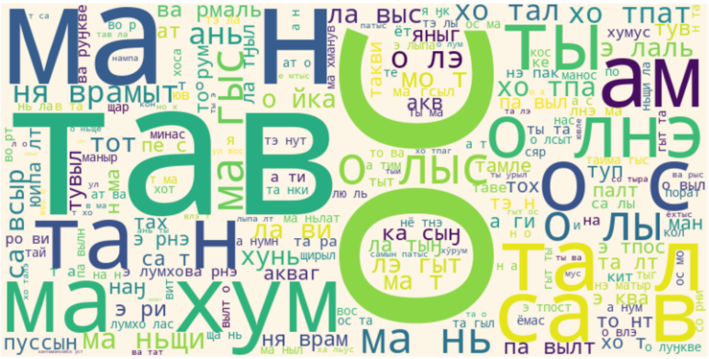

Figure 7: The frequency map of Mansi words in all sentences.

5 Experiments

5.1 Metrics

There are some different metrics for evaluating topic modeling. Coherence is an evaluation metric that can be used to assess the performance of the topic model. Coherence is typically used to analyze the relationship between two sets of data or the similarity between data sets. In topic modeling, topic coherence measures the quality of the data by comparing the semantic similarity between highly repetitive words in a topic. Coherence score is a scale from 0 to 1 in which a good coherence (high similarity) has a score of 1, and a bad coherence

	Train	Valid	Test
Articles	600	60	60
Tokens	2,088,628	217,646	245,569
Vocabulary size		33,278	
Out of Vocab rate		2.6%	

Table 6: Statistics of the WikiText-2. The out of vocabulary (OoV) rate notes what percentage of tokens have been replaced by an $\langle unk \rangle$ token. The token count includes newlines which add to the structure of the dataset.

(low similarity) has a score of 0. In other words, a good coherence is when two signals or data sets are perfectly related and identical, whereas a bad coherence is defined as having no association between data sets.

Another metric used for topic modeling is perplexity. Perplexity is a measure commonly used in language modeling that quantifies how well a probability distribution predicts a sample. In the context of topic modeling, particularly with models like Latent Dirichlet Allocation (LDA), perplexity is utilized to evaluate how well the model explains the data it is trained on. A lower perplexity score indicates that a model is better at predicting the test set, suggesting that it captures the underlying structure of the data effectively.

However, using perplexity as a sole evaluation metric for topic modeling has several drawbacks:

- Perplexity is a mathematical measure that does not necessarily correlate with human interpretation of topics. A model might have low perplexity but produce topics that are not meaningful or coherent from a human perspective.

- Lack of Consistency: Different models or configurations can yield similar perplexity scores while producing vastly different topic distributions. This can make it challenging to choose the best model based solely on perplexity.

Given these limitations, it is often recommended to use a combination of metrics, including qualitative assessments (such as human judgment on topic coherence), along with perplexity, to evaluate topic models more effectively. Techniques like topic coherence scores, visualization tools, and domain-specific evaluations can provide better insights into the quality and usefulness of the topics generated by the model.

5.2 Experiment Setup

We try 3 different methods to generate topics for our dataset. Rather than partitioning our data into training and test sets, we applied topic modeling across the entire dataset. Upon completion of the modeling process, we assessed the models using coherence scores and visualizations of the top ten words associated with each topic. Given that both Latent Dirichlet Allocation (LDA) and clustering techniques require the specification of the hyperparameter k, which represents the number of topics, we aimed to determine the optimal number of topics based on coherence scores. In contrast, BERTopic autonomously determines this parameter, allowing us to explore topic generation without imposed constraints.

6 Results

In this section, we will examine the optimal number of topics identified through various clustering methodologies.

6.1 Russian sentences from russian-mansi corpus

The application of word-embedding clustering yielded promising results even at a modest configuration of 10 topics, achieving a coherence score of 0.7722. The topics top-10 words (after removal of stop words and changing all words to it's normal form using lemmatization) generated through this method are presented in Table 7.

It is interesting to see that even though the word "человек" is not a stop word by nltk's package understanding, it is a stop word for our dataset, because it occurs in top-10 words of 7 out of 10 topics. Also the word "свой" occurs in top-10 words of 5 out of 10 topics.

Cluster Number	**Top-10 Words**
0	женщина, мама, дочь, жить, ребёнок, жена, бабушка, девушка, свой, говорить
1	река, вода, берег, земля, озеро, человек, деревня, дерево, лето, день
2	год, работать, город, человек, язык, район, деревня, жить, мансийский, посёлок
3	человек, говорить, жить, сказать, ребёнок, знать, свой, быть, пусть, очень
4	рука, свой, мужчина, человек, нога, стать, сделать, один, голова, делать
5	зверь, олень, собака, свой, медведь, лес, стать, говорить, маленький, жить
6	мужчина, человек, отец, жить, говорить, брат, свой, сын, сказать, деревня
7	ребёнок, человек, учиться, работать, жить, язык, школа, народ, разный, очень
8	работать, человек, год, работа, быть, дело, деньга, пусть, делать, жить
9	домой, пойти, туда, идти, улица, прийти, дом, ходить, выйти, поехать

Table 7: Clusters and their Top-10 Words

The thematic coherence of these clusters is evident, as they distinctly represent various topics, such as familial relationships and natural elements: we can give definitions to topics such as "women' for topic 0, "wildlife" for topic 5, "work" for topic 8 and "going" for topic 9. As the number of clusters (k) increases, the coherence score also tends to improve, with scores surpassing 0.8054 when k is set to 30. For the sake of interpretability, we opted to limit the number of clusters to a maximum of 100.

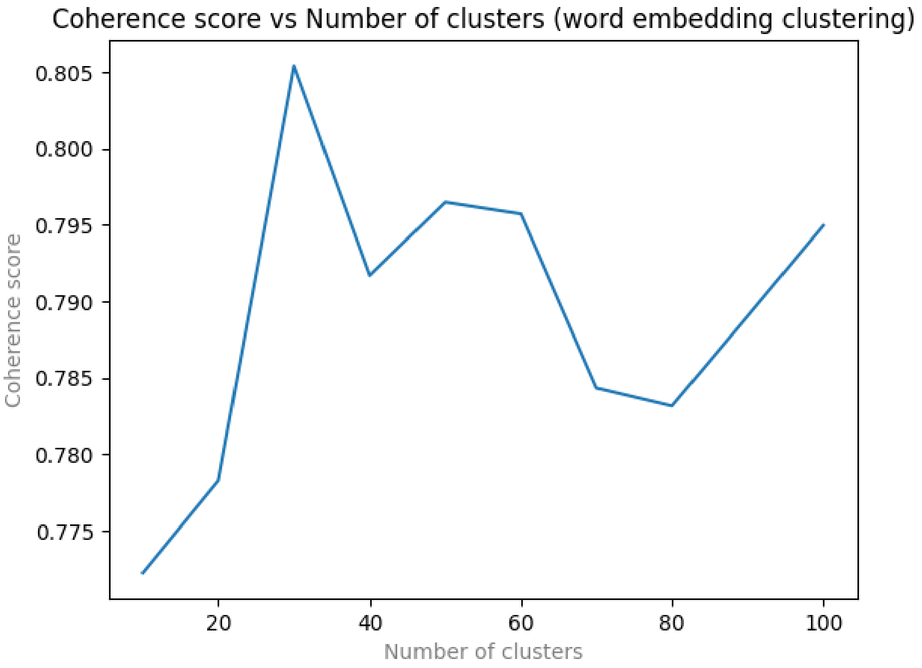

Figure 8: Coherence Score vs. Number of Clusters for Sentence Embedding Clustering

At $k = 30$, the coherence scores indicate that the clusters formed are of high quality. Selected examples from this configuration are illustrated in Table 8, which further demonstrates the distinctiveness and density of these topics. We removed words "свой", "человек" so that topics were easier to understand.

0: собака, конь, лошадь, корова, зверь, животное, соболь, бежать, олень, собачка
9: шить, женщина, платье, одежда, сшить, платок, мама, делать, красивый, вещь height27: дочь, девушка, девочка, ребёнок, маленький, говорить, дочка, младший, жить, доченька
29: смотреть, видеть, увидеть, глаз, посмотреть, проверять, слушать, мужчина, услышать, сидеть

Table 8: Samples of Sentence Embedding Clustering Clusters

In contrast, the BERTopic model was configured to autonomously determine the number of clusters, resulting in 763 clusters with a coherence score of 0.2770. Despite employing a multilingual model, the results were suboptimal, producing topics that lacked thematic coherence, as evidenced in Table 12.

757: стекла, оконные, вставить, оконное, очки, стекло, мамонтов, турпана, модерн, ледяным
252: линии, бабушки, дедушка, материнской, дедушки, папиной, бабушка, отцовской, дед, дедушке
267: родился, родилась, 1932, 1931, хангласы, вырос, деревне, евра, 1924, июня
277: стул, стула, стульчик, стуле, столовая, стулья, столбики, яша, стол, спрыгнуть
673: давно, подмёл, твёрдыми, бушевала, оставил, бесконечно, подмела, вьюга, обоз, понесла

Table 9: Samples of BERTopic topics

While BERTopic has produced notable topics, it has also resulted in some highly specific clusters that lack diversity, as well as clusters characterized by excessive variance. Method get_topic produced words that are really difficult to define into one topic. It lacks information and either too specific (topic 277) or too variant (topic 673).

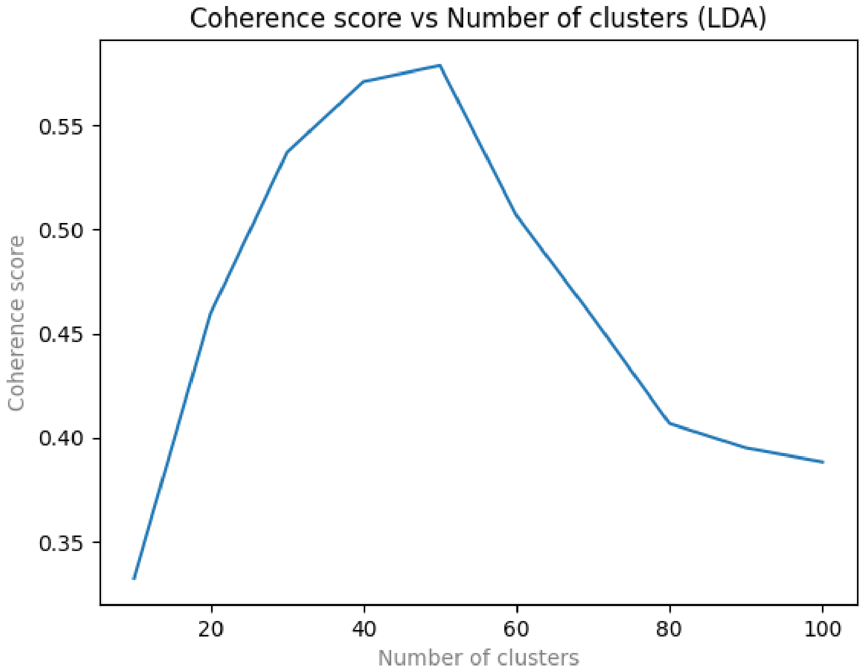

Figure 9: Coherence Score versus Number of Clusters for LDA

The Latent Dirichlet Allocation (LDA) method achieved its' highest coherence score of 0.5788 when configured with 50 clusters. In contrast, with only 10 clusters, sentence embedding clustering outperformed LDA, as evidenced by the following table of clusters derived from LDA configured for 10 topics (and coherence score of 0.3324):

Topic Number	**Words**
Topic 0	женщина, работы, живет, которые, делают, рублей, просто, стало, этим, наши
Topic 1	году, года, человек, время, манси, язык, день, прошлом, женщины, рассказы
Topic 2	очень, туда, людей, работает, работу, поэтому, разные, некоторые, школе, района
Topic 3	вместе, будут, город, деньги, рыбу, другие, стали, жить, могут, оленей
Topic 4	городе, языке, каждый, дети, дело, место, немного, разных, работала, говорят
Topic 5	пусть, делать, дальше, денег, сами, также, затем, стала, стал, урыл
Topic 6	люди, мужчина, сюда, местах, месте, доме, приезжали, летом, имеет, людьми
Topic 7	россии, районе, сразу, около, народ, деревни, жили, долго, тысяч, сначала
Topic 8	работать, сказал, людям, домой, отец, места, сколько, рыбы, родилась, возле
Topic 9	детей, ханты, сказала, среди, округе, города, других, людей, работают, жизни

Table 10: Top-10 Words for 10 Clusters Generated by LDA

These clusters exhibit excessive diversity, rendering it challenging to assign coherent topic labels to many of them. Further examination of clusters generated with 50 topics via LDA reveals similar issues:

Topic 0: дома, маленькие, взял, слово, другую, внутри, хотел, завтра, начинает, научил
Topic 38: лесу, книгу, книги, народа, зимой, отца, ребенка, отдыха, осенью, пенсии
Topic 48: дальше, месте, школу, идти, училище, коров, братья, пошёл, любит, увидели

Table 11: Examples of topics generated by LDA for 50 clusters

Despite achieving a coherence score of 0.5788 with 50 clusters, the topics generated through Latent Dirichlet Allocation (LDA) exhibit a lack of semantic clarity and coherence. The resulting topics appear fragmented and fail to convey a coherent narrative, which limits their interpretability and practical utility in the context of topic modeling.

All the topics can be found in the notebook in our repository.

6.2 20 News Group Dataset

In this study, we applied sentence embeddings to the 20 News Group dataset. As noted in the original article introducing BERTopic, the authors reported a maximum average topic coherence score of 0.173 for this dataset (the average was calculated for 5 runs with 10 to 50 topics). Utilizing the BERTopic model from the corresponding package, we achieved an enhanced coherence score of

0.4441 across 381 clusters. It is important to highlight that when BERTopic is not constrained by the number of topics, it tends to generate a diverse array of topics. While many of these topics are coherent and meaningful, others may lack clarity or relevance. Selected examples of the generated topics are presented in Table 12.

0: gun, guns, firearms, militia, weapons, amendment, control, crime, weapon, arms
9: armenian, turkish, armenians, serdar, genocide, argic, turks, armenia, serazumauucp, turkey
157: god, faith, exist, proof, existence, atheist, atheism, believe, gods, burden
178: tickets, 105pm, 735pm, june, ticket, july, cleveland, fenway, toronto, camden
272: cdrom, apple, cd, cd300is, cd300, applelink, rjacksaustlcmspsmotcom, ship, external, cd300i

Table 12: Sample topics generated by BERTopic

The results indicate that while some topics are well-defined, effectively capturing themes related to gun control and religion, others exhibit a lack of coherence and relevance.

In contrast, employing sentence embedding clustering yielded a higher coherence score of 0.6227 with only 100 clusters. The details of the clustering with 10 clusters along with their top words are provided in Table 13. The model with $k = 10$ provides us with coherence score equal to 0.5629.

In the clustering with 100 clusters, several recurring words were observed, including "line", "organization", "subject", and "article." To enhance clarity, we presented the top 25 words for each topic. However, the high frequency of certain words complicates the interpretability of many topics. Notably, topic 8 is clearly related to space, topic 9 pertains to sports, and topic 5 addresses religious themes.

For the optimal model configuration with 100 clusters, we extracted representative topics, with common terms (such as "line", "subject" and "organization") omitted for improved readability, as illustrated in Table 14.

Overall, our experiments demonstrate the effectiveness of sentence embedding clustering in achieving higher topic coherence compared to the BERTopic model, while also revealing the complexities and challenges inherent in topic extraction from text data.

7 Conclusion

In this study, we performed a comprehensive comparative analysis of sentence embedding clustering against two prominent methodologies for topic modeling: Latent Dirichlet Allocation (LDA) and BERTopic. Our results demonstrate that

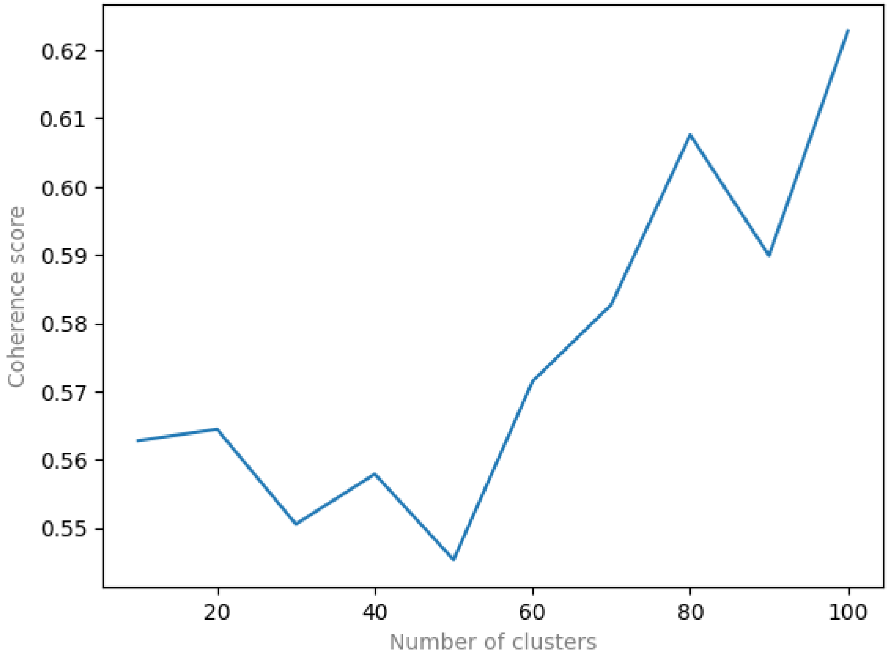

Figure 10: Coherence Score versus Number of Clusters for sentence embedding clustering for news dataset

the sentence embedding clustering approach significantly outperformed both LDA and BERTopic across multiple datasets, achieving state-of-the-art coherence scores. Specifically, we achieved a coherence score of 0.6227 with the 20 News Group dataset and 0.8054 with the Russian-Mansi corpus, both of which highlight the model's ability to capture nuanced themes effectively.

The clusters generated through our sentence embedding methodology exhibited not only high coherence but also thematic richness and interpretability, allowing for clearer topic identification compared to the less coherent and more fragmented outputs from LDA and BERTopic. This indicates that our approach is particularly adept at organizing complex text data into meaningful topics.

These findings underscore the potential of sentence embedding clustering as a robust technique for topic modeling, establishing it as a leading method for extracting insights from textual data. Future research may build upon these results to further refine and explore the implications of sentence embedding methods in various domains, thereby advancing the field of topic modeling.

Cluster Number	**Top Words**
0	armenian, people, would, subject, line, writes, israel, article, organization, said, right, muslim, israeli, know, turkish, jew, arab, like, time, university, could, think, world, state, government
1	writes, line, would, article, subject, organization, people, like, think, right, university, nntp-posting-host, know, time, state, government, could, make, good, well, thing, even, year, also, want
2	line, subject, organization, would, writes, article, system, chip, like, nntp-posting-host, know, file, clipper, encryption, people, make, time, also, could, government, program, window, need, think, work
3	line, subject, organization, file, image, university, window, would, program, system, nntp-posting-host, also, available, know, like, version, software, thanks, please, need, format, data, help, information, article
4	line, subject, organization, would, writes, bike, article, like, nntp-posting-host, good, know, university, engine, also, time, distribution, well, think, year, much, car, problem, back, could, make
5	would, subject, christian, people, line, jesus, organization, writes, know, think, article, believe, bible, like, time, church, thing, also, even, christ, life, many, word, make, good
6	line, subject, organization, window, drive, system, would, problem, card, university, know, like, work, nntp-posting-host, disk, writes, computer, file, also, need, software, article, anyone, thanks, time
7	subject, line, organization, would, writes, article, people, like, university, know, time, think, also, nntp-posting-host, problem, year, could, make, even, good, much, study, thing, well, doctor
8	space, line, subject, organization, would, writes, article, like, nntp-posting-host, university, system, time, could, know, also, nasa, mission, earth, year, think, orbit, first, distribution, moon, thing
9	game, line, team, subject, organization, year, would, player, writes, university, hockey, article, play, think, nntp-posting-host, time, first, last, like, season, good, know, baseball, league, playoff

Table 13: Clusters and their Top-25 Words for the 20 News Group dataset

2: game, team, year, would, espn, first, baseball, blue, time, season, university, last, writes, think, play, like, article, know, well, playoff, good
46: greek, turkish, armenian, greece, turk, people, government, turkey, minority, muslim, jew, right, genocide, argic, ottoman, cyprus, serdar, world, armenia, today, year, also, state, cypriot
49: team, hockey, game, player, year, would, season, play, writes, university, league, think, captain, time, article, goal, flyer, point, traded, like, coach, good
66: disease, patient, medical, health, vitamin, would, infection, also, candida, people, 1993, drug, doctor, treatment, year, study, cancer, aid, water, number, time

Table 14: Examples of topics generated by sentence embedding clustering for the 20 News Group dataset