Unsupervised Approach to Multilingual User Comments Summarization

Aleš Žagar, Marko Robnik-Šikonja
University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, 1000 Ljubljana, Slovenia
Ales.Zagar@fri.uni-lj.si Marko.Robnik@fri.uni-lj.si

Abstract

User commenting is a valuable feature of many news outlets, enabling them a contact with readers and enabling readers to express their opinion, provide different viewpoints, and even complementary information. Yet, large volumes of user comments are hard to filter, let alone read and extract relevant information. The research on the summarization of user comments is still in its infancy, and human-created summarization datasets are scarce, especially for less-resourced languages. To address this issue, we propose an unsupervised approach to user comments summarization, which uses a modern multilingual representation of sentences together with standard extractive summarization techniques. Our comparison of different sentence representation approaches coupled with different summarization approaches shows that the most successful combinations are the same in news and comment summarization. The empirical results and presented visualisation show usefulness of the proposed methodology for several languages.

1 Introduction

Readers of news articles are often interested in what others think, what their perspectives are, and whether they can get any additional information from them. User comment sections on news web pages are often a good source for extending, presenting, and challenging their own views. On the other hand, many news providers see user comments sections of their websites as a way to connect to their readers, get relevant feedback, and sometimes even extract complementary information.

Many news articles get a large number of comments in a short time, which is especially true for popular and controversial topics. When dealing with an individual article, users can usually sort comments by relevancy or publishing time. While not ideal, this is satisfactory to get insight into the most popular thread or discussion but lacks in providing an overview of the whole discussion (Llewellyn et al., 2014). This, together with the low amount of time users are willing to spend in reading comments, is one of the reasons to automatically provide comprehensive overviews of discussions.

User comments can be irrelevant, deceiving, and may contain hate speech. Language is often informal with ill-formed sentences full of spelling and grammatical errors that are hard to understand. Because of that, comments are easily dismissed as not worth the attention and time. In addition, non-standard expressed content is difficult to encode into an informative numerical representation as standard embedding techniques are mostly based on more standard language (Gu and Yu, 2020).

The goal of text summarization is to compress original data and present it in a shorter form conveying the essential information (Allahyari et al., 2017). Two main approaches exist, extractive and abstractive. The extractive summarization approach selects essential information and does not modify content; its goal is to copy the most informative non-redundant sentences, phrases, or other units of a text. The abstractive approach is similar to how humans summarise documents. It may use new words and expressions, compress long sentences, combine multiple sentences, replace phrases, etc. Current neural network based abstractive approaches mostly provide useful and fluent summaries for short texts but offer no guarantee concerning text correctness (Dong et al., 2020; Cao et al., 2020).

News article summarization is a well-defined and the most studied task within the field of automatic text summarization with several available datasets suitable for supervised learning (Bommasani and Cardie, 2020). For this task also several unsupervised methods exist, based on graph

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 89–98
April 19, 2021 © Association for Computational Linguistics

centrality approaches or clustering. On the other hand, the user comment summarization task is not well-defined and established. In a survey paper on user comments, Potthast et al. (2012) describe it as the extraction of sentences that express an opinion. This proposal categorises it as an information retrieval task, close to comment filtering and comment ranking. We believe that this categorisation is limited as it does not consider many other aspects, such as complementarity of information, coverage of different topics and opinions, impact on public discourse, possibly offensive speech, non-standard language, etc.

Cross-lingual approaches to text processing (Ruder et al., 2019) enable the transfer of trained models from resource-rich languages to low-resource languages. Many multilingual neural sentence representation models were released (Artetxe and Schwenk, 2019; Reimers and Gurevych, 2019; Feng et al., 2020; Yang et al., 2020), which presents an opportunity to improve standard unsupervised extractive approaches (Mihalcea and Tarau, 2004; Erkan and Radev, 2004; Llewellyn et al., 2014) that use sparse representations such as TF-IDF weighted bag-of-words.

In this work, we developed an unsupervised extractive approach to text summarization that combines traditional unsupervised methods (graph and clustering-based) with the above-mentioned state-of-the-art multilingual sentence encoders. We assess these encoders in combination with different extractive summarizers and dimensionality reduction techniques. We used Croatian, English and German datasets containing news articles and user comments.

Our main contributions are:

- To the best of our knowledge, we present the first multilingual unsupervised approach to automatic summarization of user comments using modern neural sentence embeddings.
- We analyse and visualize the performance of state-of-the-art multilingual sentence encoders on both clustering-based and graph-based summarization methods.
- We create a dataset of Croatian news articles appropriate for news summarization task.

The paper consists of six sections. In Section 2, we present the related work. Section 3 contains description of datasets we used. In Section 4, we outline and explain our approach to unsupervised text summarization. Section 5 presents visual and automatic evaluation of the results. In Section 6, we summarize the work done, present limitations of our approach, and ideas for further work.

2 Related work

In this section, we present related research on comment summarization and other related summarization tasks.

User comments can be divided into comments on non-textual resources (photos or videos) and comments on textual resources (news articles, product reviews, etc.) (Ma et al., 2012). Potthast et al. (2012) argue that the most important tasks done on comments are filtering, ranking, and summarization. We focus on the latter two.

Most of the research on user comments summarization uses unsupervised extractive approaches that combine ranking and clustering methods. Khabiri et al. (2011) used LDA for clustering, and ranking algorithms (MEAD, LexRank) to summarize comments on YouTube videos. Ma et al. (2012) developed a topic-driven approach in which they compared clustering methods and ranking methods (Maximal Marginal Relevance, Rating & Length) on comments from Yahoo News. Llewellyn et al. (2014) used standard clustering and ranking methods (K-means, PageRank, etc.) to summarize the comments section of the UK newspaper The Guardian. Hsu et al. (2011) proposed a hierarchical comments-based clustering approach to summarize YouTube user comments. All listed methods use classical text representation approaches, while we propose the use of modern neural sentence embedding methods.

A related task to comment summarization is discussion thread summarization. The distinctive difference is that original posts are very different from news articles. van Oortmerssen et al. (2017) used text mining to analyze cancer forum discussions. In addition to ranking and clustering, Alharbi et al. (2020) use hand-crafted text quality features such as common words between the thread reply and the initial post, a semantic distance between thread reply and thread centroid, etc. The conversation summarization (Murray and Carenini, 2008; Chen and Yang, 2020), email summarization (Kaur and Kaur, 2017), and Twitter Topics summarization (Sharifi et al., 2010) are also relevant related tasks.

3 Datasets

In this section, we first describe the creation of two Croatian summarization datasets used in our research: news articles, and user comments. We also present English and German dataset of user comments.

The CroNews summarization dataset was created from the corpus of approximately 1.8 million news articles from the popular Croatian 24sata news portal[1]. The second dataset (CroComments) is a small evaluation dataset (Milačić, 2020) and contains user comments of 42 articles from Croatian Večernji list website[2], together with their short human-written abstractive summaries[3].

We preprocessed the news articles from the news corpus into a one-sentence-per-line form using the Croatian tokenizer available in the Stanza NLP package (Qi et al., 2020). The user comments in CroComments were already preprocessed in a similar way (Milačić, 2020).

The articles in the original news dataset contained no summaries. We took the first paragraph of an article as a proxy for a summary. In the dataset, this paragraph is named 'lead'. We sampled 5000 (from a total of 17 194) examples that satisfied the next criteria: more than 6 and less than 30 sentences were present in an article (we presupposed that articles with less than 6 sentences are too short for summarization), and the overlap between the abstract (lead) and article text was within 40 and 90 ROUGE-L points. The last criterion was designed to make sure that the first paragraph of an article overlaps with the rest of it in terms of content but we avoided strictly duplicated content. Most of the abstracts have a missing period at the end. We fixed that by appending it at the end of an article. We call the resulting dataset CroNews in the remainder of the paper.

While we focused on the Croatian language, to assess the multilingual potential of the proposed approach, we tested it also on English and German. For English, we used the New York Times Comments corpus[4] with over 2 million comments. For German, we used One Million Posts Corpus (Schabus and Skowron, 2018) with 1 million comments from the Austrian daily broadsheet newspaper DER STANDARD.

4 Methodology

In this section, we describe our approach to unsupervised (multilingual) summarization which is comprised of two main components:

1. Neural sentence encoders represent the text in a numeric form as described in Section 4.1. This can be done in a cross-lingual manner to project many languages in the same numeric space and makes our approach multilingual.

2. From the numeric representation of sentences in the commentaries below a given article, we select the most representative sentences to be returned as summaries. To achieve that, we use two groups of approaches as described in Section 4.2: clustering-based and graph-based. Clustering approaches group similar sentence vectors and select the representative sentences based on the proximity to the centroid vector. Graph-based methods construct a graph based on the similarity of sentence vectors and then use graph node rankings to rank the sentences. The best-ranked sentences are returned as the summary.

As a further, optional component of our approach, the sentence vectors can be mapped to two-dimensional space with dimensionality reduction techniques (we use PCA or UMAP) and visualized in an interactive graph. To demonstrate these capabilities, we released a Jupyter notebook on Google Colab[5].

4.1 Sentence representation

In order to cluster or rank sentences in user comments, we have to first transform them from a symbolic to numeric form. In our work, we use sentence-level representation, as the extractive summarization techniques we use work on this level. Sentence embeddings aim to map sentences with a similar meaning close to each other in a numerical vector space. Initial approaches to sentence embeddings averaged word embeddings, e.g., GloVe (Pennington et al., 2014) vectors, or created Skip-Thought vectors (Kiros et al., 2015). A successful massively multilingual sentence embeddings approach LASER is built from a large BiLSTM neural network on parallel corpora (Artetxe and Schwenk, 2019).

[1] https://www.24sata.hr/

[2] https://www.vecernji.hr/

[3] Available upon email request.

[4] https://www.kaggle.com/aashita/nyt-comments

[5] https://colab.research.google.com/drive/12wUDg64k4oK24rNSd4DRZL9xywNMiPil?usp=sharing

Recently, the Transformer architecture (Vaswani et al., 2017) is the most successful and prevalent neural architecture for the majority of language processing tasks, especially if pretrained on large corpora using masked language model objective, such as the BERT model (Devlin et al., 2019). In sentence embedding, naive solutions, e.g., averaging BERT output layer or using the first CLS token in the BERT architecture, often produced results worse than averaging of word vectors.

We used three competitive transformer-based sentence encoders. Reimers and Gurevych (2019) created siamese and triplet networks to update the weights and enable comparison of sentences. Their model called SBERT adds a pooling operation to the output of BERT to derive a sentence embedding. They trained it on natural language inference (NLI) datasets. Feng et al. (2020) combined masked language model and translation language model to adapt multilingual BERT and produced language-agnostic sentence embeddings for 109 languages. Their model is called LaBSE (Language-agnostic BERT Sentence Embedding). Yang et al. (2020) proposed a novel training method, conditional masked language modeling (CMLM) to learn sentence embeddings on unlabeled corpora. In CMLM, a sentence depends on the encoded sentence level representation of the adjacent sentence.

Our sentence embedding vectors have 768 dimensions. A dimensionality reduction may improve clustering due to noise reduction. To test that hypothesis, we tested two variants of sentence selection approaches (both graph and clustering-based): with and without dimensionality reduction. For the dimensionality reduction down to two dimensions, we tested PCA and UMAP (McInnes et al., 2018) mthods. We set the neighbourhood value of UMAP to 5, the number of components to 2, and the metric to Euclidian.

4.2 Selecting representative sentences

Once the sentences of comments belonging to a certain article are represented as numeric vectors, we have to select sentences for the summary. We use two types of approaches: i) clustering the sentences and returning the most central sentences from each cluster, and ii) representing sentences as nodes in a graph, based on their similarities and selecting the highest-ranked nodes as the summary.

For clustering, we used k-means and Gaussian mixture algorithm. We set the number of clusters to 2 because in our experimental evaluation we decided to extract only the best two sentences. We extracted the best sentences based on their proximity to centroid vectors of the clusters returned by the clustering algorithms. Clustering methods deal well with the redundancy of extracted sentences as the extracted sentences are by construction very different.

Graph-based ranking algorithms score the importance of vertices within a graph. A popular method to determine the importance of a vertex uses the number of other vertices pointing to it and the importance of the pointing vertices. In our case, each vertex in a graph represents a sentence. We used the TextRank (Mihalcea and Tarau, 2004) method, inspired by the PageRank algorithm (Page et al., 1999) that can be intuitively explained with the concept of eigenvector centrality or stationary distribution of random walks. For a similarity measure of sentences, we used the cosine similarity computed on sentence vectors.

We used two baseline summarization methods: i) selecting random $n = 2$ sentences (BaseRand), and ii) selecting the first $n = 2$ sentences (BaseLead).

For both clustering and dimensionality reduction, we used the scikit-learn implementations in python (Pedregosa et al., 2011). For the graph-based approach, we used PageRank from the NetworkX python library (Hagberg et al., 2008).

5 Evaluation

In this section, we first provide visualization of sentence embeddings, followed by the analysis of summarization. The visualization demonstrates the suitability of the proposed cross-lingual sentence representation for unsupervised summarization. In summarization experiments, we first present results of news article summarization, followed by the commentaries.

5.1 Visualization of sentence embeddings

We first visually demonstrate the utility of used sentence embeddings in a multilingual setting. In Figure 1, we show a visual evaluation of the proposed cross-lingual sentence representation for the unsupervised summarization. The dots in the image are sentence vectors of the synthetic sentences (described below). The image was produced using the Gaussian Mixture clustering using the sentence representation produced with the SBERT encoder and PCA dimensionality reduction. Sentences of various lengths corresponding to three topics (school, weather, and music) were written in Slovene and translated into English, Croatian, and German. The three large colored clusters correspond to three topics, which is an indication that the sentence representation captures different contents well. We can observe also small groups of four sentences (an original Slovene sentence and three translations of it) that confirm the accuracy of the multilingual sentence encoder. The translated sentences are close together which is an indication that the representation is semantically adequate even in the multilingual setting. The rectangle on the top contains the sentences: Šolsko leto se je začelo drugače kot ponavadi; The school year started differently than usual; Školska godina započela je drugačije nego inače; Das Schuljahr begann anders als gewöhnlich. The rectangle on the right shows: Vreme bo jutri lepo; The weather will be nice tomorrow; Vrijeme će sutra biti lijepo; Das Wetter wird morgen schön sein. The rectangle on the left consists of: Kitara je zelo popularen glasbeni inštrument; The guitar is a very popular musical instrument; Gitara je vrlo popularan glazbeni instrument; Die Gitarre ist ein sehr beliebtes Musikinstrument.

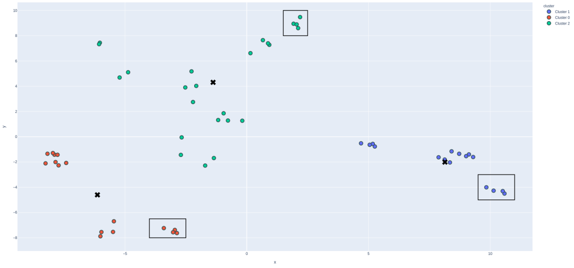

Figure 1: Example of Gaussian Mixture clustering with SBERT encoder and PCA dimensionality reduction of sentences from three topics (school, music, and weather, shown in green, blue, and red, respectively) and four languages. The sentences in the rectangles contain the same text in four languages (Slovene, English, Croatian, and English). The rectangle on the top contains the sentence ”The school year started differently than usual.”, the right one is ”The weather will be nice tomorrow.”, and the left one is ”The guitar is a very popular musical instrument.”.

5.2 News summarization

Due to the shortage of supervised data for automatic evaluation of user comments, we first test our unsupervised approach on the CroNews dataset, constructed as described in Section 3. We expected that the results would give us an insight into the performance of different combinations of methods, described in Section 4.

The results in Table 1 show commonly used ROUGE metric. The best performing experimental setup uses the LaBSE sentence encoder, no scaling, and the TextRank algorithm for sentence selection. The BaseLead baseline is 4.5 points behind the best model and ranked somewhere in the middle of all combinations. This corresponds with the findings of Zhu et al. (2019), who analysed the phenomenon of lead bias in news article summarization task. The BaseRand baseline is near the end of the ranks, as expected.

Enc.	Scaling	Summary	R-1	R-2	R-L
None	None	BaseLead	36.46	24.04	34.52
None	None	BaseRand	35.07	23.69	33.47
CMLM	None	GaussMix	35.29	22.77	33.52
CMLM	None	K-means	34.33	21.87	32.58
CMLM	None	TextRank	39.37	26.95	37.65
CMLM	PCA	GaussMix	35.71	23.90	34.17
CMLM	PCA	K-means	35.69	23.93	34.12
CMLM	PCA	TextRank	39.58	27.61	37.98
CMLM	UMAP	GaussMix	36.99	25.14	35.35
CMLM	UMAP	K-means	37.05	25.15	35.42
CMLM	UMAP	TextRank	38.65	26.94	37.06
LaBSE	None	GaussMix	38.81	26.41	37.04
LaBSE	None	K-means	37.70	25.18	35.92
LaBSE	None	TextRank	40.07	28.42	39.00
LaBSE	PCA	GaussMix	36.04	24.06	34.41
LaBSE	PCA	K-means	35.95	23.85	34.30
LaBSE	PCA	TextRank	38.69	26.80	37.10
LaBSE	UMAP	GaussMix	36.84	24.92	35.28
LaBSE	UMAP	K-means	37.22	25.31	35.63
LaBSE	UMAP	TextRank	37.90	25.86	36.29
SBERT	None	GaussMix	37.36	25.09	35.64
SBERT	None	K-means	37.05	24.65	35.26
SBERT	None	TextRank	38.63	26.55	36.99
SBERT	PCA	GaussMix	36.34	24.34	34.71
SBERT	PCA	K-means	36.42	24.48	34.81
SBERT	PCA	TextRank	37.86	26.11	36.31
SBERT	UMAP	GaussMix	36.94	25.14	35.38
SBERT	UMAP	K-means	36.92	25.06	35.38
SBERT	UMAP	TextRank	36.38	24.48	34.83

Table 1: Results expressed as ROUGE scores on the CroNews dataset. Colors correspond to ranks, darker hues correspond to better scores.

Statistics of different parameters in Table 2 show that LaBSE achieved on average 0.6 more ROUGE-L points than SBERT and CMLM, which are close in terms of performance. UMAP scaling preserved information better than PCA for 0.3 points but achieved 0.4 points less compared to no scaling. TextRank ranking method is superior to clustering for more than 2 points.

MatchSum (Zhong et al., 2020) is currently the

Group	Mean	Std	Min	Max	95%CI	Size
Encoder						
LaBSE	36.11	1.47	34.30	39.01	(34.98, 37.25)	9
SBERT	35.49	0.75	34.71	36.99	(34.91, 36.06)	9
CMLM	35.32	1.91	32.58	37.99	(33.86, 36.79)	9
Scaling						
None	35.96	2.01	32.58	39.01	(34.42, 37.50)	9
UMAP	35.63	0.66	34.84	37.06	(35.12, 36.14)	9
PCA	35.33	1.44	34.12	37.99	(34.22, 36.43)	9
Summarizer						
TextRank	37.03	1.18	34.84	39.01	(36.13, 37.93)	9
Clustering	34.94	1.00	32.58	37.04	(34.45, 35.44)	18

Table 2: ROUGE-L scores grouped by sentence encoder, scaling, and type of summarizer.

best extractive summarization model. It was trained on the large CNN/Daily Mail dataset and achieved 44.41 ROUGE-1 and 40.55 ROUGE-L scores. As we can observe from Table 1, our best scores for the Croatian news lag approximately 4.3 ROUGE-1 and 2.5 ROUGE-L points behind these scores which is a relevant difference in performance. However, we have to take into account that we use leads as an approximation for the summaries.

5.3 User commentaries summarization

We used the same experimental setup, as reported in Table 1, to summarize the CroComments dataset. The results of both datasets are very similar if we rank the models, with the best models being identical. TextRank with CMLM or LaBSE encoder is superior to clustering. Surprisingly, SBERT shows significantly lower performance with both clustering and ranking (with ranking worse than clustering).

We identified a few reasons that explain the lower scores of comment summarization compared to news summarization. For comments, the sentence encoders face a more challenging task of encoding the informal language; for the same reason, the accuracy of a sentence tokenizer is also significantly lower, as our inspection revealed. A single CroComment document (containing all comments related to one news article) is usually comprised of texts by several authors, of variable length, and written in different styles. CroComment documents are longer and exhibit a greater length variability. The average length of a document is 19.81 sentences with the standard deviation of 13.16 in comparison to CroNews dataset which contains 7.85 sentences with the standard deviation of 1.42. These differences make the comment summarization task difficult for a model trained on standard language in much shorter news articles.

Enc.	Scaling	Summary	R-1	R-2	R-L
CMLM	None	K-means	24.44	11.50	23.18
CMLM	None	TextRank	33.08	17.24	31.09
CMLM	PCA	GaussMix	19.71	08.53	18.79
CMLM	PCA	K-means	22.30	10.66	20.64
CMLM	PCA	TextRank	26.01	12.50	24.60
CMLM	UMAP	GaussMix	24.83	12.18	23.28
CMLM	UMAP	K-means	23.88	10.44	22.37
CMLM	UMAP	TextRank	23.02	11.78	22.31
LaBSE	None	GaussMix	26.77	13.39	25.77
LaBSE	None	K-means	26.59	12.89	25.01
LaBSE	None	TextRank	34.35	18.50	32.28
LaBSE	PCA	GaussMix	24.15	11.61	22.90
LaBSE	PCA	K-means	25.32	14.17	24.63
LaBSE	PCA	TextRank	28.53	15.60	26.95
LaBSE	UMAP	GaussMix	26.39	12.99	24.28
LaBSE	UMAP	K-means	27.36	14.45	26.04
LaBSE	UMAP	TextRank	24.99	12.50	23.80
SBERT	None	GaussMix	25.34	12.43	23.82
SBERT	None	K-means	26.13	12.84	24.67
SBERT	None	TextRank	25.20	11.71	23.25
SBERT	PCA	GaussMix	21.78	09.98	20.51
SBERT	PCA	K-means	23.96	11.46	22.47
SBERT	PCA	TextRank	25.44	11.40	23.76
SBERT	UMAP	GaussMix	25.29	13.00	24.16
SBERT	UMAP	K-means	24.94	12.04	23.62
SBERT	UMAP	TextRank	24.44	10.92	22.98

Table 3: Results expressed with ROUGE scores on the CroComments evaluation dataset with human-written summaries of comments. Colors correspond to ranks, darker hues correspond to better scores.

As an example, Table 4 shows comments belonging to one selected article. We tokenized comments, encoded them with the LaBSE sentence encoder, and scored with the TextRank algorithm. The sentences with the highest score in each user comment are typeset with red, and two highest scored sentences are shown in green. The value 'ref' in the column 'Id' indicates the human-written abstractive summary of the listed comments; the value 'lead' means the first paragraph of the article. Notice that the human-written summary and the high-scored sentences strongly overlap.

Comment no. 54412 demonstrates how the tokenizer and encoder face a difficult task. It is evident that the comment should have been split into several sentences to improve readability, has missing punctuation, and does not contain letters with the caron. Comment no. 54299 shows the limitation of extractive approaches since it cannot be understood properly without the context. The comment with the lowest score (no. 56141) does not add much to the conversation.

Table 5 shows an example from New York Times

Id	Croatian text	English translation
lead	Svaki gost koji je došao u Hrvatsku 2009. godine nije poklonjen, morali smo se za njega izborili. Ovakav učinak, uz ostalo, rezultat je mjera koje smo poduzeli, uz lijepo, sunčano vrijeme. Sunce je ove godine sjalo i u Turskoj, Francuskoj, Španjolskoj, ali očito nešto bolje u Hrvatskoj, slikovit je bio ministar Bajs.	Every guest who came to Croatia in 2009 was not given away, we had to fight for him. This effect, among other things, is the result of the measures we have taken, with nice, sunny weather. This year, the sun was shining in Turkey, France, Spain, but obviously somewhat better in Croatia, Minister Bajs was picturesque.
54279 score: 0.0552	Hrvatski turizam je u plusu za 0,2 Bravo,bravo,bravo . Pravi turizam ce poceti u Hrvatskoj tek tada kad nebude vise nitko od vas smdljivaca u vladi . Otvorite ovi ljudi , pa austrija napravi vise novaca od turizma nego Hrvatska . Svaku godinu smo u plusu a love nigdje pa naravno kad od 10-15% ostane samo 0.2 % . Koji su to muljat3ori i od kuda imate taj podatak . Revolucija je jedini spas , skidam kapu Rumunjima , oni su to fino rijesili . Bog i Hrvati	Croatian tourism is in the plus by 0.2 Bravo, bravo, bravo. Real tourism will start in Croatia only when there are no more of you smugglers in the government. Open these people, and Austria will make more money from tourism than Croatia. Every year we are in the red and the money is nowhere to be found, so of course when only 0.2 % of 10-15 % remains. What are these scammers and where do you get that information from. Revolution is the only salvation, I take my hat off to the Romanians, they solved it fine. God and Croats
54299 score: 0.0587	To vam je tako : 1999 godine Amerikanci su sredili stanje na Kosovu i cijela Europa a i druge države dale su zeleno svjetlo svojim građanima da mogu na ljetovanja u hrvatsku i ostali dio Balkana.2000 godine dolazi za ministricu turizma gospođa Župan - Rusković . Ta godina pokazuje se za turizam dobra i to se pripisuje SDP -u i gospođi ministarki . Ove godine sunce jaće i duđe sije pa eto to se pripisuje ministru Bajsu . Ja ču im samo poručiti . Ne bacajte pare na \" promocije \" jer svijet zna za nas , radije te novce ulažite u izobrazbu turističkoga i ugostiteljskoga osoblja . To bi bio naš največi uspjeh .	This is how it is for you: in 1999, the Americans settled the situation in Kosovo and the whole of Europe, and other countries gave the green light to their citizens to go on vacation to Croatia and the rest of the Balkans. In 2000, Ms. Župan - Rusković came to be Minister of Tourism. That year proves to be a good thing for tourism and it is attributed to the SDP and the Minister. This year the sun is shining stronger and longer, so that is attributed to Minister Bajs. I'll just tell them. Don't waste money on \"promotions \" because the world knows about us, rather invest that money in the training of tourism and catering staff. That would be our greatest success.
54311 0.0448	Sezona je ove godine bila iznad prosjeka i normalno da je Bajs ponosan	This season has been above average and it's normal for Bajs to be proud
54412 score: 0.0534	slazem se sa Somelier , a po izjavama i komentarima sto daje ministar Bajs vidi se nema veze s turizmom , HR je konkurentna samo u o dredjenim vrstama turizma (nauticki turizam) i trebalo bi se fokusirati upravo na njih koji usput najvise i trose , a ne slusati ove gluposti Bajsa da je sezona uspjesna zato sto je dozvolio onim krsevima od aviona da slijecu ili zato sto je dao 20 miliona € za reklamu na googlu i eurosportu	I agree with Somelier, and according to the statements and comments given by Minister Bajs, there is nothing to do with tourism, HR is competitive only in o dredged types of tourism (nautical tourism) and we should focus on those who spend the most, and not listen to this nonsense of Bajs that the season was successful because he allowed those breaches of planes to land or because he gave 20 million € for advertising on google and eurosport
54413 score: 0.0582	Bajs , kaj nas briga kak su turistički tržili u Austriji , Italiji , Francuskoj ili Grčkoj ? Raci ti nama zakaj je u Hrvatskoj bilo manje turistof neg lani iako ti tvrdiš da mi imamo kakti prednost kao auto destinacija ? Zakaj i u onom jednom jadnom mesecu kad je bilo više turistof nek lani ima manje lovice ? Zakaj se inšpekcije i dalje zezaju sa boravišnim taksama vikendaša dok ugostitelji premlaćuju goste , ne izdaju račune i jasno , ne plačaju poreze , uključujući i PDV ?	Bajs, do we care how they marketed tourism in Austria, Italy, France or Greece? Tell us why there were fewer tourists in Croatia than last year, even though you claim that we have some advantage as a car destination? Why, even in that poor month when there were more tourists, let there be less money last year? Why do the inspections continue to mess with the weekend taxes of the weekenders while the caterers beat the guests, do not issue invoices and clearly do not pay taxes, including VAT?
56141 0.0376	Nakon ove kostatacije sa zadovoljstvom mogu kostatirati da je Bajs napredovao sa jedne na dvije litre dnevno.	After this casting, I am pleased to say that Bajs has progressed from one to two liters a day.
ref.	Hrvatski turizam u porastu , uspješna sezona . Vlada je problem i ne ostaje dovoljno novca . Ne bacajt pare ne promocije već ulažite u izobrazbu turističkoga i ugostiteljskoga osoblja . Baj ponosan na sezonu iznad prosjeka . HR je konkurentna samo u određenim vrstama turizma i trebalo bi se fokusirati na njih . Zakaj je manje turista nego lani i nanje novca . Inspekcije se zezaju sa boravišnim taksama a ugostitelji premlaćuju goste , ne izdaju račune i ne plaćaju poreze .	Croatian tourism on the rise, successful season. The government is a problem and there is not enough money left. Don't waste money on promotions, but invest in the training of tourism and catering staff. Bajs proud of the above average season. HR is competitive only in certain types of tourism and should focus on them. Why are there fewer tourists than last year and money for them. Inspections mess with sojourn taxes and caterers beat guests, do not issue invoices and do not pay taxes.

Table 4: Visualization of the most important sentences in each user comment (in red). The original comments are on the left-hand side and their machine translations on the right-hand side. The reference score is at the bottom. Two sentences with the highest score are shown in green.

Comments, which was preprocessed and evaluated in the same manner as the example from Table 4. The selected sentences capture both prevalent themes (artistic freedom and racial questions) but exhibit the problem of redundancy. More examples from English, along with German, can be found on our source code repository[6].

6 Conclusion

We developed a multilingual unsupervised approach to user commentary summarization and tested it on a less-resourced Croatian language. Our models are based on cross-lingual neural sentence encoders, which make them easily applicable to many languages with little or no preprocessing. We tested several sentence representations and assessed the effect of dimensionality reduction. We used clustering and graph-based ranking algorithms to select sentences that form the final summaries. The results were promising both on the news articles dataset and the user comments evaluation dataset. The source code of our approach is freely available under the open-source licence.

The presented approach has several limitations. It only works within extractive summarization approaches, which do not allow sentence modification. With abstraction techniques, e.g., sentence compression, we could further distill the important information. We only tested sentence representation methods, while paragraph or document embeddings would also be sensible. We also did not exploit the information contained in the threading structure of the commentaries and possible relation of comments with the text of an original article.

In further work, we intend to exploit additional information in comments which was not used in the present study. The number of likes that a comment received could be used to weight sentences. Instead of working on a sentence-level, we could take a comment as a whole and embed it as a document. We plan to extend the work on visualization since it showed promising results, especially in the interactive exploration mode, inaccessible in the paper format.

Acknowledgments

The work was partially supported by the Slovenian Research Agency (ARRS) core research programme P6-0411 and the Ministry of Culture of Slovenia through project RSDO (Development of Slovene in a Digital Environment – Language Resources and Technologies). This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

References

Abdullah Alharbi, Qaiser Shah, Hashem Alyami, Muhammad Adnan Gul, M Irfan Uddin, Atif Khan, and Fasee Ullah. 2020. Sentence embedding based semantic clustering approach for discussion thread summarization. *Complexity*, 2020:1–11.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331.

[6] https://github.com/azagsam/xl-user-comments

Id	Text
24107006 score: 0.0282	This art is all about perception . It is about the point the artist is trying to make and how the viewer sees it . This art should not be limited because it is attached to an emotion these moments being recorded through art of a society that claims to be post racial opens the eyes of those who do not want to see and forces them to . This illustration does that and in my opinion that makes it so much more valuable because it does not just sit in silence it sends a message .
23235619 score: 0.0283	Artists should n’t be limited or restricted in what they can do as an artist . Everyone should have a voice or take on a matter no matter how unpopular or offensive the opinion is . Censoring art defeats the creativity and free expression in art . Censorship perverts the message the artist try ’s to convey .
22099108 score: 0.0273	I believe that all subjects should be fair game for an artist . It should n’t matter if they are depicting a murder , or even if it ’s ” black subject matter ” , every artist has a voice that deserves to be heard . As Smith writes ” We all encounter art we do n’t like , that upsets and infuriates us . ” (1) I understand that some topics are difficult to talk about and that some art is can cause anger but I think that it is irrational to make topics off - limits because people do n’t agree with it .
22098876 score: 0.0264	I personally believe that artists should be able to write about anything they want , drive to the studio , then turn those words into beautiful music . Music is an art and in art there are no limits so honestly whatever they feel is relevant to write about , they should have the freedom to do so . Regardless of peoples personal opinions artist should be comfortable to talk about what they want to talk about . ” We all encounter art we do n’t like , that upsets and infuriates us . ” (Gilpin , 1) I understand that some subjects are very sensitive , but most of the things people do n’t like to hear are usually cold hard facts about the dark side of society . A few examples would be , hate crimes against all races , racism in america , people killing other people . It s just the sad truth that a lot of people hate to hear . Music is a powerful - subject that can really impact a person .
22075721 0.0258	nothing should be in limited to artist . they should have the freedom to do what they pleased .
22054073 0.0252	I believe there is n’t a problem when a white artist draws a topic that is related to discrimination against the Blacks . This artist may want to show that killing black people is wrong . It does n’t matter if she ’s white or black .
22041906 score: 0.0280	I do n’t think that any topic is out of bounds to an artist . That is the idea of an artist , is n’t it ? To talk about subjects that they think should be talked about , or that they feel motivated to bring attention to . I do n’t think it is right to throw blame and anger towards one group because they are creating art about a different group . I understand why there is anger , but demanding that a work be destroyed is just absurd to me . Could the artist have done something differently ? Possibly , but demanding empathy and understanding from a group different than your own , and then saying their act of trying to do so is inappropriate just does n’t make sense . I do n’t think any one group ” owns ” history . History is a human experience . People as a collective own the histories that shaped the world they live in . That is the point of the exhibition . The exhibition description on the Whitney site says , ” Throughout the exhibition , artists challenge us to consider how these realities affect our senses of self and community . ” Instead of focusing on the color of the artists skin , we should be focusing on the point of the show .. how the painting makes us feel about ourselves and our communities , because I am sure that everyone could say that there is room for improvement when it comes to both .
22031632 score: 0.0219	The question of whether or not any group ” owns ” a portion of history is not the issue . It is about how that imagery is used , if it is used intelligently , and that it mimics an aspect of white racism : the historic practice of whites displaying the mutilated corpses of black people . To make the issue about censorship is to miss the point . Instead students should be asked to consider how a white person might have better handled her desire to show empathy .

Table 5: Visualization of the most important sentences in each user comment for a sample from the New York Times Comments dataset. Since the conversation is very long, we show here only a part of it. The green color stresses the best two sentences.

Günes Erkan and Dragomir R Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852*.

Jing Gu and Zhou Yu. 2020. Data Annealing for Informal Language Understanding Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3153–3159.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11 – 15.

Chiao-Fang Hsu, James Caverlee, and Elham Khabiri. 2011. Hierarchical comments-based clustering. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 1130–1137.

Kuldeep Kaur and Anantdeep Kaur. 2017. A survey on email summarisation techniques and its challenges. *Asian Journal of Computer Science And Information Technology*, pages 40–43.

Elham Khabiri, James Caverlee, and Chiao-Fang Hsu. 2011. Summarizing user-contributed comments. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 534–537.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors.

In *Advances in Neural Information Processing Systems*, pages 3294–3302.

Clare Llewellyn, Claire Grover, and Jon Oberlander. 2014. Summarizing newspaper comments. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 599–602.

Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. 2012. Topic-driven reader comments summarization. In *21st ACM International Conference on Information and Knowledge Management*, pages 265–274.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the Empirical Methods in Natural Language Processing*, pages 404–411.

Katarina Milačić. 2020. Summarization of web comments. Master's thesis, University of Ljubljana, Faculty of Computer and Information Science.

Gabriel Murray and Giuseppe Carenini. 2008. Summarizing spoken and written conversations. In *Proceedings of the Empirical Methods in Natural Language Processing*, pages 773–782.

Gerard van Oortmerssen, Stephan Raaijmakers, Maya Sappelli, Erik Boertjes, Suzan Verberne, Nicole Walasek, and Wessel Kraaij. 2017. Analyzing cancer forum discussions with text mining. *Knowledge Representation for Health Care Process-Oriented Information Systems in Health Care Extraction & Processing of Rich Semantics from Medical Texts*, pages 127–131.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Martin Potthast, Benno Stein, Fabian Loose, and Steffen Becker. 2012. Information retrieval in the commentsphere. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):1–21.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

Dietmar Schabus and Marcin Skowron. 2018. Academic-industrial perspective on the development and deployment of a moderation system for a newspaper website. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 1602–1605, Miyazaki, Japan.

Beaux Sharifi, Mark-anthony Hutton, and Jugal Kalita. 2010. Automatic summarization of Twitter topics. In *Proceedings of National Workshop on Design and Analysis of Algorithm*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ziyi Yang, Yinfei Yang, Daniel Cer, Jax Law, and Eric Darve. 2020. Universal sentence representation learning with conditional masked language model. *arXiv preprint arXiv:2012.14388*.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2019. Make lead bias in your favor: Zero-shot abstractive news summarization. *arXiv preprint arXiv:1912.11602*.