UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

Automatic summary generation and topic modeling of Ukrainian news articles

Author:
Anastasiia TANANAISKA

Supervisor:
Olga KANISHCHEVA

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2021

Declaration of Authorship

I, Anastasiia TANANAISKA, declare that this thesis titled, "Automatic summary generation and topic modeling of Ukrainian news articles" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"The further a society drifts from the truth, the more it will hate those who speak it."

George Orwell

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Automatic summary generation and topic modeling of Ukrainian news articles

by Anastasiia TANANAISKA

Abstract

With the growth of the amount of textual information produced by news media daily, keeping up with the most recent events happening in the world can lead to a really overwhelming experience. Especially when living through those events or possibly witnessing those in real life. The focus of this work will be, in fact, on the development of the solution to the problem of information overload, addressed by generating summaries and classifying topics of the news articles.

Acknowledgements

First of all, I'm very grateful to my supervisor Olga Kanishcheva, who guided me and gave valuable advice throughout the work on the thesis.
I appreciate the assistance and encouragement from Yuliia Kleban, our IT&BA program manager.
I'm also really grateful to my family and friends, who were always there to support and motivate me. (Special thanks to my small university friend group - Kartel)
And lastly, I want express gratitude to the community of the Ukrainian Catholic University, for the acquired valuable knowledge and the opportunity to apply it in the real world case.

Contents

List of Figures

List of Tables

List of Abbreviations

NLP - **N**atural **L**anguage **P**rocessing
API - **A**pplication **P**rogramming **I**nterface
RSS - **R**eally **S**imple **S**yndication
LDA - **L**atent **D**irichlet **A**llocation
LSA - **L**atent **S**emantic **A**nalysis
RNN - **R**ecurrent **N**eural **N**etwork

Chapter 1

Introduction

As the amount of textual information is increasing constantly, manual analysis and organization of the data have become truly time-consuming and complex. Having to complete which might be pretty tedious and affect the quality of the end result. Therefore the raised interest for research of the opportunities in automating the manual processes is surely reasonable. Such automation can be done using the Natural Language Processing (NLP) techniques, which are represented as a set of complex algorithms trained to "understand human language", by analysing the connections between textual documents. That analysis usually varies from task to task and depends mostly on our objectives.

According to our investigations, some of the most common NLP tasks are:

- Sentiment analysis
- Text classification
- Chatbots development
- Text extraction
- Machine translation
- Text summarization
- Text auto-correction

Except for the diversity of the existing techniques, NLP is also really flexible in terms of the domain, as its algorithms can be easily adapted to any business case.

In this thesis, in fact, the focus of the work will be on **topic modelling** and **text summarization** of the *Ukrainian news articles*, as currently news media industries are generating huge amounts of textual data.

Topic modelling - process of discovering latent topics from the collection of input document, that would describe it in the best way. As a technique, it is an unsupervised method, that works by analysing documents and grouping them into clusters.

Document Summarization – Natural Language Processing task, which purpose is to shorten the initial document (which is usually quite voluminous), while preserving the most salient points and the original meaning.

By the input document type, summarization can be either single-document or multi-document, and by the summarization type either extractive or abstractive.

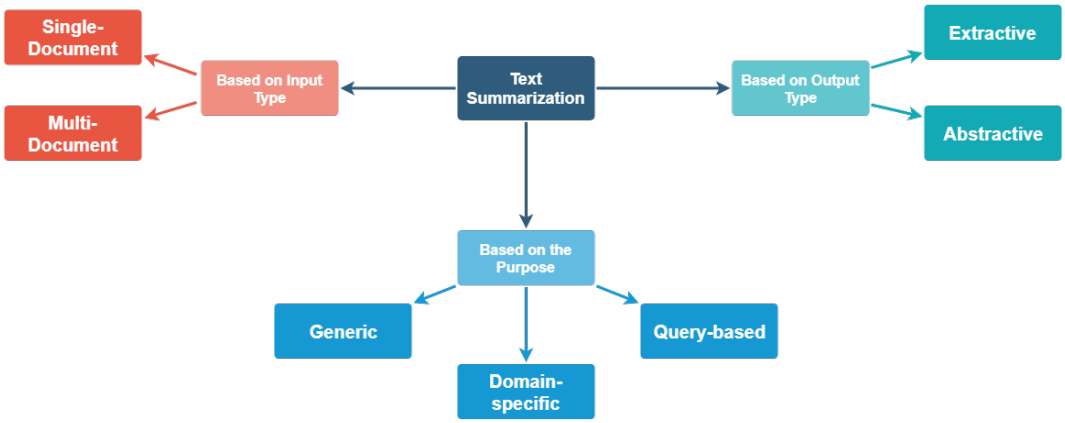

FIGURE 1.1: Types of text summarization [4]

Both extractive and abstractive summarization techniques work on identifying the salient features in the input text and producing the final summary based on them. However, the key difference between those techniques is their complexity - abstractive methods are usually more sophisticated and advanced, as, in the end, they produce a cohesive and unique summary for each of the input documents. Rarely repeating text from the original from the input text. While the extractive methods only return the key features from the input text.

1.1 Problem Statement

Every day, online media corporations publish huge amounts of articles, reporting all sorts of events happening around the world. And despite having so many possible sources to information, keeping track of the most recent events has become really difficult both for news writers and the readers.

News writers face that issue from the point of complexity of organizing the news articles in real-time, having to choose the most important events across all the rapidly-changing ones and correctly emphasizing on those. Meanwhile, readers face it from the point of the unsatisfied need to access the most of the valuable information in the shortest period of time.

In this thesis we propose a solution that will address pain points of both news writers and news readers.

1.2 Motivation

Since Russian full-scale invasion of Ukraine, there is a strong demand to keeping track of most recent and important news. Moreover, after 24 February, the amount of long news articles published has increased drastically, consuming which on a daily basis leads to the information overload. Therefore, we believe that by summarizing and categorizing news articles, their content and message can become more accessible and easily perceptive for one.

1.3 Objectives

We perform topic modeling and abstractive single-document summarization independently and then combine the obtained results.

Chapter 2

Related Works

In the fields of topic modelling and document summarization, there is plenty of research conducted, and we will focus on reviewing the work done independently.

It is difficult to determine whether large companies use such algorithms for optimization of the mentioned tasks, as such information is not publicly disclosed. Therefore we will focus more on theoretical research, assuming that the proposed approach will also work for the real case.

2.1 Research in Topic Modelling

To our observation, as a baseline model for solving topic modeling task, researchers[5] mostly use Latent Dirichlet Allocation (LDA) originally proposed by *David Blei*, *Andrew Ng* and *Michael I. Jordan* in their self-titled paper[1]. Using such model as a baseline allowed authors to successfully perform news articles classification.

Although LDA is not the only algorithm that can be applied for modeling topics, it works best for discovering a set of latent topics from the document (in contrast to Latent Semantic Analysis (LSA), which assumes that document is described by only one topic).

2.2 Research in Document Summarization

What for Document summarization, indeed, there is a quite diverse amount of approaches - graph-based, purely statistical or deep-learning-based.

As we are focusing on abstractive summarization, applying deep-learning approaches is more common, and therefore we will be investigating on those. Advanced deep-learning architecture such as Transformer models revolutionized the field of NLP processing. Having the encoder-decoder architecture, solving such sequence-to-sequence tasks as summarization showed much better performance comparing to previous approaches, that used RNN. Also a few research showed successfull adaptation of the pre-trained transformer models for solving summarization problem by fine-tuning them on the custom dataset. Mainly, models for Italian document summarization[6] and Arabic document summarization[2].

Chapter 3

Data

3.1 Description

During the investigation stage of the various possible sources for our dataset, it was determined that there's no such resource that would collect all the Ukrainian news articles in one place. Having looked at individual news media websites, it was found that most of them do provide free access to the news archive, however, accessing them through API or RSS is either deprecated or not supported. Therefore, it was decided that developing our own parser for one of the news media archive, in our case - the TSN one, would be the best workaround to that issue. The choice of the website relies primarily on its reliability, consistency of the published articles and topics diversity.

By parsing the above-mentioned archive, we collected over *35,173* news articles published in the range from January 2021 up until May 2022 and extracted the following pieces of information for each of the individual articles:

- *Publish Date* - date when article was published (used for final aggregation of documents)
- *Title* - title of the article (used as reference summary for measuring the performance of summarization models, even though they are unsupervised, it is important to define metric for estimating their work efficiency)
- *Content* - main body of the article (used as the input document based of which the topic modelling and summarization is made)
- *Article Link* - link to the article source

In our further research, we will use "article" and "document" synonymously, as they both refer to the same entity - one single article. As well as "token" refers to "word" - which are both used interchangeably throughout this paper.

3.2 Exploratory Visualizations

In this section, we will present a few visualizations to help better understand the data peculiarities.

First figure describes the dynamics of articles publishing - on average the amount of published articles per day falls within a certain range, except for some days, when the number of occurring events was significantly higher, leading to a higher amount of articles published.

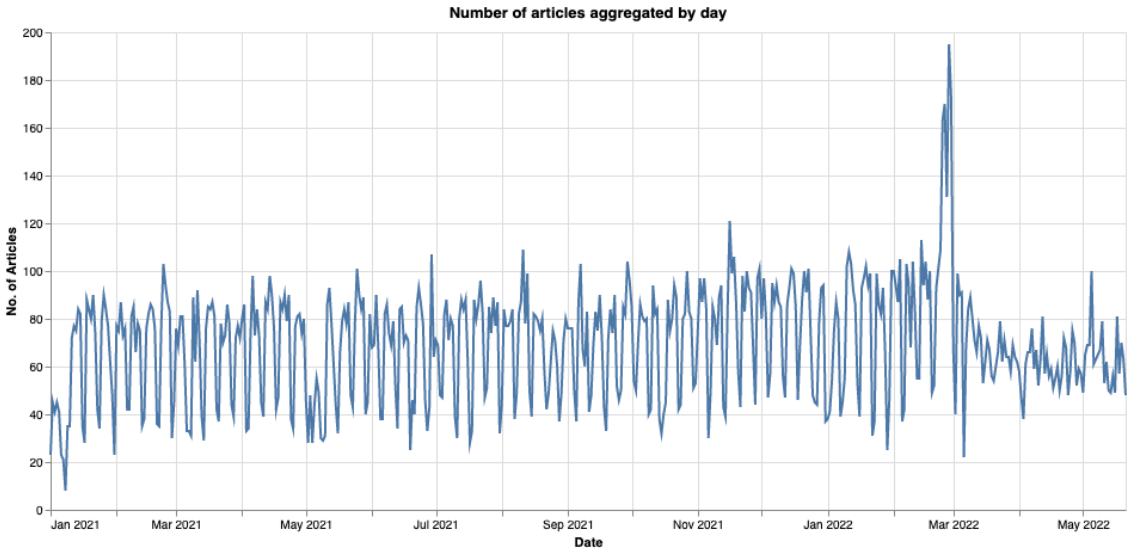

FIGURE 3.1: Dynamics of articles published, per day

The second figure helps to better understand the volumes of articles, specifically the average number and deviation of number of sentences in the article body. And as can be observed, that value usually lies within a certain range, meaning that by the size, articles are mostly consistent.

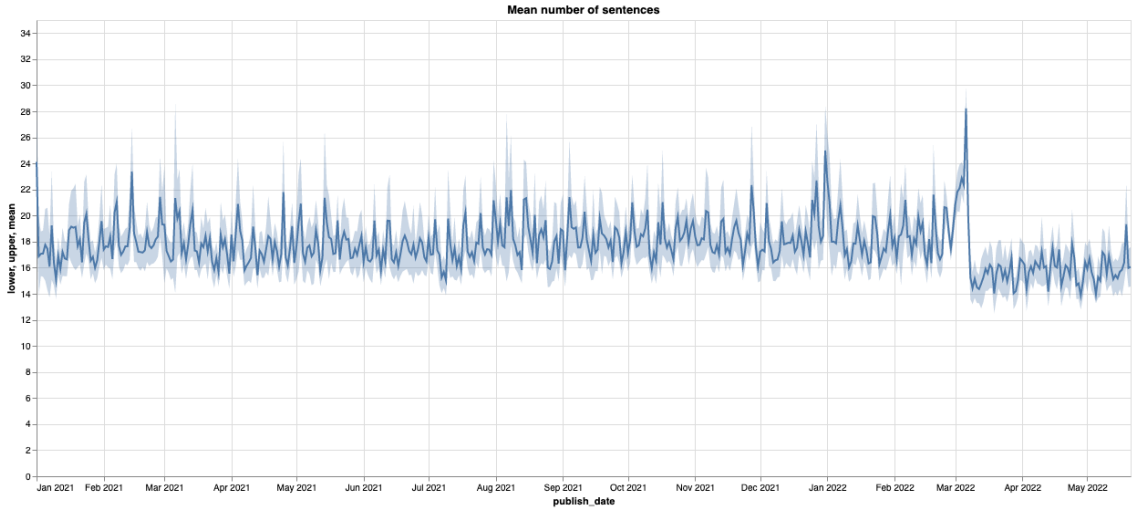

FIGURE 3.2: Mean number of sentences in the article, per day

3.3 Dataset Preprocessing

Before proceeding to the main part, it is important to perform standard text preprocessing of the dataset obtained. Which includes the following steps:

- Converting text to lowercase
- Removing special characters (e.g. \!%$*?/,.;:)
- Tokenization
- Stop-words elimination
- Lemmatization
- Vectorization

The details on each of these steps is described in the following subsections.

3.3.1 Tokenization

Tokenization is the process of breaking down documents into smaller units - tokens. Tokens can be either words, characters, or subwords. In our case, we will be tokenizing by single word, constructing a so-called uni-grams.

3.3.2 Lemmatization

Most of the words which are used in our daily basis are inflectable, meaning they are subject to word modifications - for instance, through conjugation or declension. However, some NLP algorithms, due to high words variation, might find differentiation between word forms confusing, leading to worse results. Therefore, when using such algorithms, it's important to reduce text data variation beforehand. One of the techniques used for solving that problem is lemmatization. The essence of which is to convert inflected words forms into their common/root form.

For deriving lemmatized forms of words, in our research, there was used an NLP_UK-based microservice developed by lang-uk project group specifically for Ukrainian language. Such microservice provides access to a pre-trained lemmatizer, which allows to accurately find word's lemma based on the context.

3.3.3 Stop Words Removal

Generally, stop words are a set of overly used words in a language, which are not strongly contributing to the overall context of the documents. However, their list is not-exclusive and can be further extended using the *domain-specific* stop words. Identifying and excluding domain-related stop words is a widely used technique in various NLP tasks, as it prevents the algorithm from capturing over dominating words in a corpus and, as a result, improves the overall performance of the algorithm. We analysed which words were most common in our vocabulary and eliminated a few of them, as well as those from the standard stop words set.

3.3.4 Vectorization

Vectorization is the last step in our preprocessing stage. One of the ways on how to represent documents in a format, comprehensible by a machine is to convert our documents in Bag-of-Words (BOW) structure, which would describe the occurrence of words within a document.

3.4 Constructing Datasets

By sequentially applying the above-mentioned steps, we end up with the ready dataset in a BOW-format for *topic modelling* part. As for the summarization problem, we will be using the initial dataset, as the preprocessing steps are done within the pre-trained model, which is discussed further in the paper.

Chapter 4

Methodology

4.1 Proposed Approach

The main part will consist of the following steps:

- Topic Modelling using Latent Dirichlet Allocation
- Fine-tuning pretrained model for summarization
- Combining discovered topics with the generated summaries

4.2 Topic Modelling

As discussed in the overview of related works, for the task of topic modelling, it was chosen to use a generative, probabilistic model - Latent Dirichlet Allocation (LDA). It assumes that each document is represented by a distribution of a fixed number of topics, and each topic is a distribution of the keywords that occur most often together.

4.2.1 Theoretical Overview of LDA Algorithm

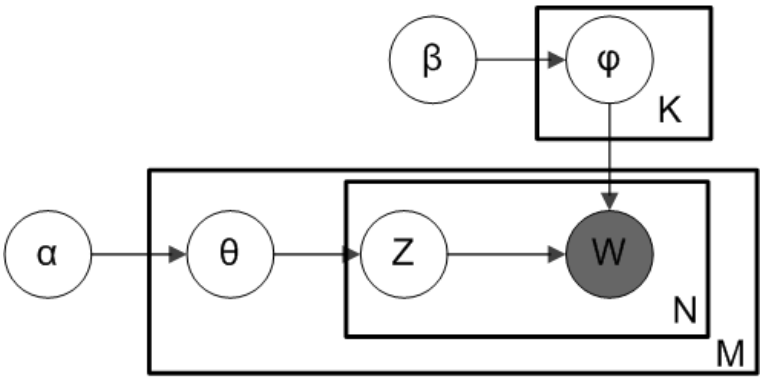

FIGURE 4.1: LDA Model Architecture

where:
α - the per-document topic distributions,
β - the per-topic word distribution,
θ - the topic distribution for document m,
ϕ - the word distribution for topic k,
Z - the topic for the n-th word in document m,
W - the specific word from the document

Mathematical representation:

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) P(W_{j,t} \mid \varphi_{Z_{j,t}})$$

FIGURE 4.2: LDA Formula

Model performance depends primarily on the choice of the hyperparameters: K, α, β. For evaluation of model performance, there are a few possible approaches - by looking at the results directly, for example, at the top-n words, or by using intrinsic metrics, such as topic coherence. Topic coherence works by measuring the degree of semantic similarity between the top words in a single topic. Or in other words, evaluates, how probable a pair of words will come from the same documents rather than from a random document in the corpus.

4.2.2 Hyperparameters Tuning

For tuning model parameters, we iteratively generated 24 LDA models with different set of parameters and calculated the coherence for each one and recorded its value in the resulting table. For reducing computational complexity, we set beta to *auto*, so that the model is learning the best parameter value. By arbitrarily choosing 75% of data we performed cross validation and obtained the following results.

Topics	**Alpha**	**Beta**	**Coherence**
6	0.61	auto	0.51
8	asymmetric	auto	0.5
6	0.91	auto	0.5
8	symmetric	auto	0.5
6	0.31	auto	0.5
6	symmetric	auto	0.49
6	asymmetric	auto	0.49
7	0.61	auto	0.49
8	0.61	auto	0.49
8	0.91	auto	0.49
7	0.31	auto	0.48
7	0.91	auto	0.48
7	symmetric	auto	0.48
6	0.01	auto	0.48
7	0.01	auto	0.47
8	0.01	auto	0.47
8	0.31	auto	0.47
7	asymmetric	auto	0.45
5	0.31	auto	0.45
5	0.91	auto	0.45
5	symmetric	auto	0.44
5	0.61	auto	0.43
5	0.01	auto	0.43
5	asymmetric	auto	0.42

TABLE 4.1:
Tuning Results
(100% dataset)

Topics	**Alpha**	**Beta**	**Coherence**
7	asymmetric	auto	0.56
8	0.91	auto	0.53
8	asymmetric	auto	0.52
8	0.31	auto	0.51
5	asymmetric	auto	0.5
8	0.61	auto	0.5
7	0.01	auto	0.5
8	symmetric	auto	0.5
7	0.91	auto	0.5
8	0.01	auto	0.5
7	symmetric	auto	0.5
7	0.31	auto	0.5
7	0.61	auto	0.49
5	0.61	auto	0.47
5	0.91	auto	0.47
6	0.91	auto	0.47
6	0.61	auto	0.46
5	symmetric	auto	0.46
5	0.31	auto	0.46
5	0.01	auto	0.46
6	0.31	auto	0.45
6	0.01	auto	0.45
6	symmetric	auto	0.44
6	asymmetric	auto	0.43

TABLE 4.2:
Tuning Results
(75% dataset)

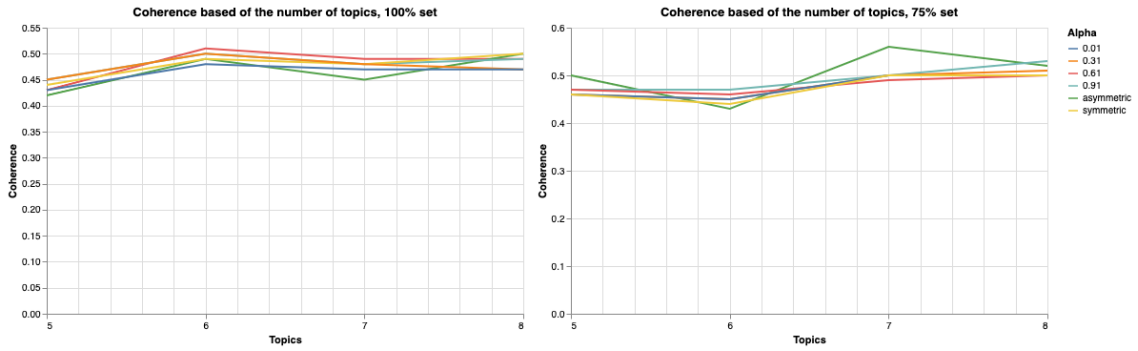

FIGURE 4.3: Coherence charts

From the above two figures, we can conclude that there's no drastic change of coherence, when performing validation on the whole set of data. However, some significant change can be observed on the smaller dataset - in that case, the parameter set which allows us to achieve highest coherence value is the following:

```
num_topics = 7
alpha=symmetric
eta=auto
```

4.2.3 Results

After hyperparameters tuning and model validation, we ran the LDA algorithm using the above-mentioned parameters and obtained the following topic clusters:

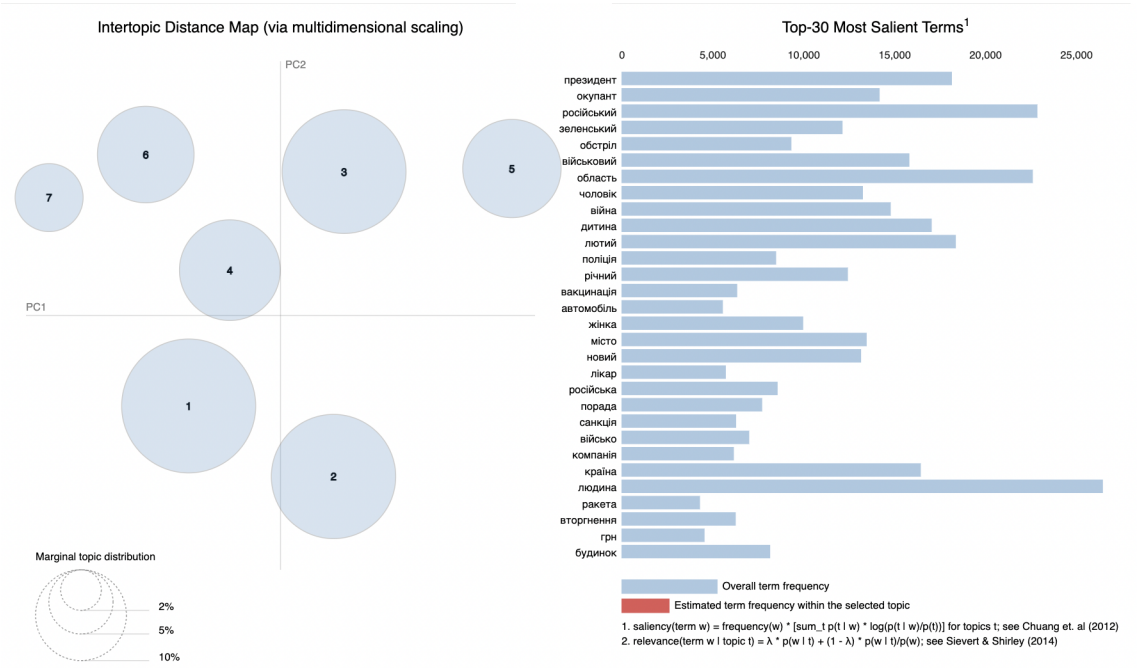

FIGURE 4.4: Graphical representation of the topics

On the left of the chart there are visual representation of sizes of those 7 clusters. On the right: top 30 most salient 1-grams across all documents.

Another useful information we can get is the keywords distribution of topics. Each keyword within each topic has its weight. First item in the list - topic id, second one - equation describing the topic.

```
(6,
 '0.027*"президент" + 0.022*"російський" + 0.018*"зеленський" + 0.014*"війна" '
 '+ 0.014*"країна" + 0.012*"військовий" + 0.010*"російська" + 0.009*"кордон" + '
 '0.009*"санкція" + 0.009*"український"')
(5,
 '0.017*"людина" + 0.014*"вакцинація" + 0.012*"лікар" + 0.011*"область" + '
 '0.010*"випадок" + 0.010*"новий" + 0.009*"центр" + 0.009*"здоров" + '
 '0.008*"вакцина" + 0.008*"кількість"')
(4,
 '0.018*"автомобіль" + 0.013*"новий" + 0.011*"королева" + 0.010*"принц" + '
 '0.009*"компанія" + 0.008*"авто" + 0.008*"машина" + 0.007*"переселенець" + '
 '0.007*"модель" + 0.007*"пальне"')
(2,
 '0.010*"грн" + 0.008*"компанія" + 0.008*"долар" + 0.007*"допомога" + '
 '0.006*"українець" + 0.006*"біженець" + 0.006*"тисяча" + 0.006*"банк" + '
 '0.006*"робот" + 0.006*"млн"')
(1,
 '0.020*"область" + 0.019*"дитина" + 0.017*"чоловік" + 0.015*"річний" + '
 '0.013*"жінка" + 0.013*"поліція" + 0.009*"людина" + 0.007*"будинок" + '
 '0.007*"місце" + 0.006*"суд"')
(0,
 '0.007*"людина" + 0.006*"життя" + 0.006*"зірка" + 0.005*"світ" + 0.005*"знак" '
 '+ 0.005*"перемога" + 0.004*"їхній" + 0.004*"український" + 0.004*"команда" + '
 '0.004*"матч"')
```

FIGURE 4.5: Probabilistic definition of topics

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0	людина	область	грн	окупант	автомобіль	людина	президент
1	життя	дитина	компанія	лютий	новий	вакцинація	російський
2	зірка	чоловік	долар	обстріл	королева	лікар	зеленський
3	світ	річний	допомога	місто	принц	область	війна
4	знак	жінка	українець	російський	компанія	випадок	країна
5	перемога	поліція	біженець	військовий	авто	новий	військовий
6	їхній	людина	тисяча	порада	машина	центр	російська
7	український	будинок	банк	війна	переселенець	здоров	кордон
8	команда	місце	робот	український	модель	вакцина	санкція
9	матч	суд	млн	область	пальне	кількість	український
10	булий	лікарня	газ	людина	тисяча	країна	вторгнення
11	новий	правоохоронець	гривня	ракета	бензин	медичний	лютий
12	українська	буча	тис	ракетний	герцогиня	українець	визнання
13	гра	кримінальний	понад	військо	мережа	тисяча	слово
14	співачка	місто	продукт	втрата	завод	щеплення	територія
15	кожний	тіло	євро	територія	продаж	доба	міністр
16	українець	водій	гроші	сила	власник	дитина	держава
17	разом	злочин	послуга	федерація	королівський	доза	голова
18	дружина	місцевий	млрд	ворог	марка	особа	дія
19	слово	допомога	ціна	удар	ринка	заклад	переговори

FIGURE 4.6: Top keywords in each of the detected topics

After the retrieval of the topics for all the articles, we are now ready to proceed to the next and final part - summarization.

4.3 Abstractive Summarization

For summarization, it was decided to use one of the pre-trained models presented in the open-source AI community - Hugging Face, as that resource provides a really convenient and straight-forward API for building pipelines and models training. Another significant advantage of Hugging Face models is easy access to a wide range of pre-trained models for all sorts of NLP tasks, including text summarization. When choosing the model, we paid attention to the data, which was used for initial pretraining, and its language. Having investigated the models, which are suitable for solving our task, we discovered a multilingual Text-to-Text Transfer Transformer (mT5) - **mT5_multilingual_XLSum**. That model was trained on the XL-Sum dataset, which includes documents in 45 languages, including Ukrainian. As mentioned in the documentation, dataset was constructed as a BBC news articles, with total of 53999 articles.

The train/validation/test distribution is as follows:

	Train	**Validation**	**Test**
No. of articles	43201	5399	5399

TABLE 4.3: XL-Sum train-test split

More detailed training information on Transformer is provided in the authors' paper [3].

4.3.1 Fine-Tuning Model

We fine-tuned our Transformer models with the dataset, described in Data section. Using article body is input, and article title as a reference summary for evaluation. We decided to take 70% of data for training, 15% for validation, and 15% for prediction, exact numbers are presented in the following table: Due to the change in topics

	Train	**Validation**	**Test**
No. of articles	21104	7035	7035

TABLE 4.4: Train-test split for fine-tuning

popularity, we tried to include as many of them for training as possible, so that their distribution is more uniform.

4.3.2 Results

For evaluation, surely manual assessment is more accurate and easiest. However, there's also a possibility to use metrics that would automate the evaluation process, such as - Recall-Oriented Understudy for Gisting Evaluation (ROUGE). Using ROUGE we measured summary quality by analysing the proportion of overlapping units (n-grams) sequences in generated summary and the reference ones. In particular, we used 3 ROUGE scores - ROUGE-1 - 1-grams, ROUGE-2 - 2-grams and ROUGE-L - longest common subsequence.

Performance of the mT5-multilingual-XLSum model on the Ukrainian dataset (no fine-tuning)

ROUGE-1	**ROUGE-2**	**ROUGE-L**
23.9908	10.1431	20.9199

TABLE 4.5: XL-Sum Evaluation Results

After fine-tuning and training on our dataset:

TABLE 4.6: Fine-tuning results

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum
1	1.633700	1.354591	15.909800	3.079300	15.880500	15.878700
2	1.357200	1.305682	16.183200	3.251600	16.173800	16.153100
3	1.219300	1.302034	15.655100	2.967000	15.629700	15.630500
4	1.130900	1.298669	16.465800	3.014500	16.427700	16.399200
5	1.073100	1.304905	16.493400	3.128600	16.468200	16.440400

Chapter 5

Results

5.1 Examples of summarized text and topic modelling

	Input Article	Reference Summary	Generated Summary	Keywords Distribution
1	**“Злочинців треба назвати злочинцями, притягнути їх до відповідальності і засудити", - вважає Дуда. Жахливі фотографії загиблих з української Бучі** свідчать, що не треба шукати компромісу з агресором за будь-яку ціну, але водночас Україні зараз потрібно багато зброї. Про це підкреслив президент Польщі Анджей Дуда у повідомленні у Twitter. Criminals must be called criminals, brought to justice and sentenced. Pictures from #Bucha disprove the belief that we have to seek a compromise at any cost. In fact, the Defenders of Ukraine need three things above all: weapons, weapons and more weapons. #StandWithUkraine — Andrzej Duda (@AndrzejDuda) April 3, 2022...	"Захисникам України насамперед потрібні три речі: зброя, зброя і ще раз зброя" - Дуда	"Злочинців треба назвати злочинцями, притягнути їх до відповідальності і засудити": Дуда про жахливі фото з Бучі	(Topic 6, 0.4885464) (Topic 1, 0.2569061) (Topic 3, 0.16319852)
2	**Група дослідників дізналася про те, що відбувається в мозку, коли людина відчуває страх.**Вчені вивчили дані зондів, що розташовані глибоко в мозку хворих на епілепсію. Стаття опублікована в Science Advances. Було проведено багато досліджень, щоб дізнатися більше про те, але набагато менше відомо про те, які процеси відбувається в мозку під час таких переживань...	Що відбувається в мозку, коли ми відчуваємо страх	Вчені розповіли, що відбувається в мозку, коли людина відчуває страх	(Topic 5, 0.5712817) (Topic 0, 0.25515124) (Topic 4, 0.0801554)
3	Тестове відправлення посилки компанія здійснила з Києва до Харкова і у зворотному напрямку.29 липня **Нова пошта протестувала нову послугу - доставку за допомогою безпілотного літального апарату (БПЛА).** Тестова посилка піднялась на висоту 300 метрів о 07:00 з аеродрому "Чайка" (вул. Авіаконструктора Антонова, 5) і прибула на аеродром "Коротич" (Новий Коротич, Харківська область) о 12:00. У зворотному напрямку посилка вилетить о 14:00 з "Коротича" і о 19:00 вже прибуде до Києва. Тестову відправку було здійснено на безпілотному літаку Discovery української компанії "Аеродрон". "Уперше в світі БПЛА з відправленням **пролетів відстань у 480 км.** Окрім Нової пошти на подібне досі не зважилася жодна інша логістична компанія.	Уперше в Україні: Нова пошта протестувала доставку посилки безпілотним літальним апаратом	Уперше в світі БПЛА з відправленням пролетів відстань у 480 км - Нова пошта	(Topic 2, 0.54860663) (Topic 4, 0.15434863) (Topic 3, 0.13539469)
4	Хакери уже зламали 2500 веб-сайтів російської та білоруської влади.**Хакери Anonymous повідомили про початок безпрецедентних атак на офіційні сайти російських органів влади.** Також читайте Не виконав завдання у Гостомелі: російська влада хоче усунути полковника армії РФ...	Anonymous оголосив про безпрецедентні атаки на російські сайти	Хакери Anonymous повідомили про початок безпрецедентних атак на сайти російської влади	(Topic 6, 0.5718837) (Topic 2, 0.29618374) (Topic 3, 0.07519075)
5	**На Прикарпатті посилили протипожежну охорону екосистем.На Прикарпатті через пожежі в екосистемах, зокрема, тимчасово обмежили в'їзд автомобілів до лісових масивів та лісопарків та збільшили кількість патрулів.** Про це повідомив голова Івано-Франківської ОДА Андрій Бойчук у своєму Facebook...	Обмеження в'їзду до лісу, патрулі і дрони: як на Прикарпатті боротимуться з пожежами в екосистемах	На Прикарпатті через пожежі в екосистемах тимчасово обмежили в'їзд авто до лісових масивів та лісопарків	(Topic 1, 0.4387342) (Topic 3, 0.30754435) (Topic 6, 0.14318274)
6	**У деяких пацієнтів спостерігається втрата апетиту.У хворих на "Омікрон" спостерігається широкий спектр симптомів, частину з яких можна прийняти за прояв інших захворювань.** Про це **експерти розповіли** виданню The Sun.Також читайте Як відрізнити "Омікрон" від грипу і чим вони схожі...	Експерти назвали приховані симптоми "Омікрону"	У хворих на "Омікрон" спостерігається широкий спектр симптомів: що вони означають	(Topic 5, 0.9409895) (Topic 0, 0.05123937)

FIGURE 5.1: Examples of analysed text taken from the test sets

The above table presents a small portion from the test dataset, showing the results of summarization and topic modelling.

As can be seen, the writing style of the generated summaries is somewhat similar to the headlines, which were used as reference, along with the overall context. However, the text is unique and covers the most important points from the original article.

What for topic modelling, after reviewing the keywords within each topic it can be seen that the classification is precise and surely relevant to the input text.

Chapter 6

Conclusions

By conducting our research, we were able to show the perspective of using NLP in the industry of news media. Surely, there's no model, yet that would perfectly model topics or perform summarization, however it can surely help to automate the process of data analysis and reduce manual work.

The tasks of topic modelling and documents summarization are really powerful and relevant today. Taking away the work that's been done, it's possible to develop the idea further by improving the existing flow or investigating, validating other approaches. For example, comparing performance of abstractive and extractive summarization methods.

Specifically, as some possible directions for further work, the following ideas could be implemented:

- Add more news sources, and conduct the analysis in parallel
- Implement search by keyword
- Aggregate news articles by date and analyse in a time-based manner

And finally - that project could be a base for a web-resource that would conveniently report about most recent and important events. We believe that such idea has a huge potential regarding the current high demand to access truthful and relevant information from news resources.

Bibliography

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[2] Mukhtar Elgezouli Elmadani Khalid N. and Anas Showk. "Bert fine-tuning for arabic text summarization". In: (2020).

[3] Tahmid Hasan et al. "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 4693–4703. DOI: 10.18653/v1/2021.findings-acl.413. URL: https://aclanthology.org/2021.findings-acl.413.

[4] Medium Kushal Chauhan. *Unsupervised Text Summarization using Sentence Embeddings*. 2018. URL: https://medium.com/jatana/unsupervised-text-summarization-using-sentence-embeddings-adb15ce83db1 (visited on 03/06/2021).

[5] Zhenzhong Li, Wenqian Shang, and Menghan Yan. "News text classification model based on topic model". In: *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*. 2016, pp. 1–5. DOI: 10.1109/ICIS.2016.7550929.

[6] Gabriele Sarti and Malvina Nissim. "IT5: Large-scale Text-to-text Pretraining for Italian Language Understanding and Generation". In: *arXiv preprint arXiv:2203.03759* (2022).