Social Pulse - A digital recreation of current events

Cedric Osei-Akoto*
oseiaced@students.zhaw.ch
ZHAW School of Engineering

Sithursan Sivasubramaniam*
sivassit@students.zhaw.ch
ZHAW School of Engineering

Figure 1: A collection of newspapers

ABSTRACT

The vast amount of information available online has made accessing and consuming news content a significant challenge. To address this issue, the Social Pulse project was initiated. The project aims to revolutionize news consumption by providing a condensed and easily digestible form of news through aggregation, summarization, and graph-based visualization. This scientific research project explores the effectiveness of visualizing aggregated and summarized news data in a graph format and its impact on the way users interact with news. The project focuses on the Swiss German part of Switzerland and considers news articles from a specific timeframe. In order to maintain a versatile nature the project consists of several key components, including data preprocessing, categorization, topic modelling using BERTopic, HDBSCAN, and UMAP algorithms, keyphrase extraction using KeyBERT, and visualization using Obsidian's Graph. By leveraging advanced techniques and technologies, the Social Pulse project contributes to the ongoing efforts in improving news consumption in the digital era. The research findings and insights gained from this project can inform future developments in the field of news aggregation, summarization, and visualization, paving the way for more effective and efficient news consumption experiences.

KEYWORDS

News, Obsidian, Keyphrases, Data

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Cedric Osei-Akoto and Sithursan Sivasubramaniam. 2023. Social Pulse - A digital recreation of current events. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The world as we know it today is an ever changing and moving environment, filled with dreams, opportunities, actions, and reactions. Unlike past generations, we get to connect to this ecosystem on a physical and digital level. With this leap forward we created a global digital community, which consumes and creates a vast amount of information on a daily basis. Contrary to the preceding experiences of earlier generations, preserving a present view on the world through news is no longer a trivial task. Not only the diversity, but also the sheer amount of news makes it nearly impossible to consume in a broader sense. Forcing us all to individually choose our own truthful news sources and exposure to current events.

But does it have to be this way?

In order to audit this question in a comprehensive manner the big data project social pulse was brought to life. With the main intention of creating an avant-garde approach to news consumption in a dense and readily available manner.

2 RESEARCH QUESTION

In order to conduct the project Social Pulse in an efficient manner, clear research goals, which align with the timeline had to be defined. In contemplation of clear research goals, we started off by defining our motivation for using news data for the project. In the next step, we built a deeper understanding on how society consumes and gains value from news. Following the overview, we defined core conceptional fundamentals of Social Pulse based on which we defined the research goals. At the end of this segment measurement metrics for research goal failure and success are defined.

2.1 Motivation

Throughout history being conscious about the present in a local and global context has been a paramount for personal and professional growth.

But which kind of data provides the most useful information to maintain consciousness about the present in a local and global scale?

Even though social media provides the option to share information in a very concise way, newspapers and magazines tend to take an important role of continuously informing the masses about current affairs. Regardless of most newspapers also having an online presence, it has become more difficult for individuals to capture information from various sources and leverage the gained knowledge for personal and professional growth. Based on the fundamental goal to present news content from multiple sources in a condensed and easily digestible form, companies such as Google and Apple have worked on different techniques to enable users to have access to news in a digestible format.

Technique 1: Aggregation of news
The process of collecting information from several different websites, newspapers, databases, etc. and combining it in one place, or the result of this process [2].

With the solutions leveraging the aggregation technique, reader merely needed one central platform to have all newspaper data. Even though this technique reduced the time complexity overhead of staying present on a global level, solutions using solely the aggregation of news as technique still required a large time investment from their users.

Technique 2: Summerization
The act of expressing the most important facts or ideas about something or someone in a short and clear form, or a text in which these facts or ideas are expressed [2].
Through the summarization of news data, information of such kind gets readily available for a wider group of people. The reason for that is the newly gained ability to consume news at a fast rate, which is in line with the lower attention span of today's world.

Even though there are currently solutions which utilize both these techniques, the way how users interact with such applications has remained largely unchanged. Due to this conclusion a new train thought about alternative ways to visualize news data was triggered.

But how does one best visualize unstructured data in a structured manner?

Based on our knowledge, graphs could not only solve this visualization issue, but also grant additional insight about the underlying data's relationships and introduce a new way of effective information retrieval.

2.2 Society and news

As in the past news has a central role in every society. The reason for that is its ability to reflect awareness, accountability, political and social debate, economic implications and simply inform about emerging events.

With this information members of society are often equipped to gain strategic and political edge in business and also obtain continuous exposure to their own opinion hence identity. Regardless of these facts, news consumption has been in decline over the years [4].

The most common reasons for news avoidance over the past years have been the negative effects on personal mood, information overload, untrustworthiness, and bias alongside other key factors [4].

2.3 Key Concepts

Based on the insight gathered from the two previous chapters a clear vision for the Social Pulse project resulted.

Fundamentally the project Social Pulse project aims to mitigate the information overload and bias of news consumption. This is done with the techniques of aggregation, summarization, and visualization as graph, enabling users to engage with news in an avant-garde pleasing way.

Research Question
Can aggregated and summarized news be visualized in a graph and what effect does it have on the way we interact with news?

2.4 Success/Failure measurement metrics

The research success and failure solely hinge on whether a graph containing aggregated and summarized news data can be made. In addition, the computation of such information should take no longer than 1 hour.

3 LIMITATIONS

Due to time, illness and data source constraints of the team during the project completion following limitations have to be considered:
-Only newspapers from the swiss german part in german are considered for this research
-data used for the project contains news articles from 01.03.2023 – 31.03.2023
-the project pipeline components are designed individually and modular

4 METHODS

In order to create the described product in the time frame of approximately 3 months, a modular pipeline-style approach was chosen.

4.1 Pipeline

The pipeline above visualizes the five key components, which create the build the solution. Due to this modular design the implementation of the components could be conducted in a parallel manner saving the team time. For the purpose of understanding the codebase and correctness of the code, each of the modules was reviewed by the team member, which was not directly involved with its implementation.

4.2 Preprocessing

The Preprocessing of the data was a crucial part of the project, since it also strictly determined the quality of results, which could be gained. In this module, the news data from the swissdocs@liri database was initially extracted, cleansed and transformed.

Data extraction
In context of this research the data provided by the swissdocs@liri database, was manually queried, and uploaded to the git repository. For test reasons, a smaller and larger experiment sample was extracted. The smaller sample solely contained news data from one newspaper over the time span of march 2023, while the larger sample contained data from all german newspapers in Switzerland over the same time span. Both samples entailed metadata about the actual source of the newspaper, its publisher, initial sharing publication timestamp and the article title, subtitle and content.

Data cleansing
For the later modules of the pipeline, the content and category of each article was crucial. Hence heavy empathies on the cleansing that data had high priority. In order to mitigate the risk of over representation of news articles, articles with duplicate content were removed in this step. Besides that, rows containing no content were also removed.

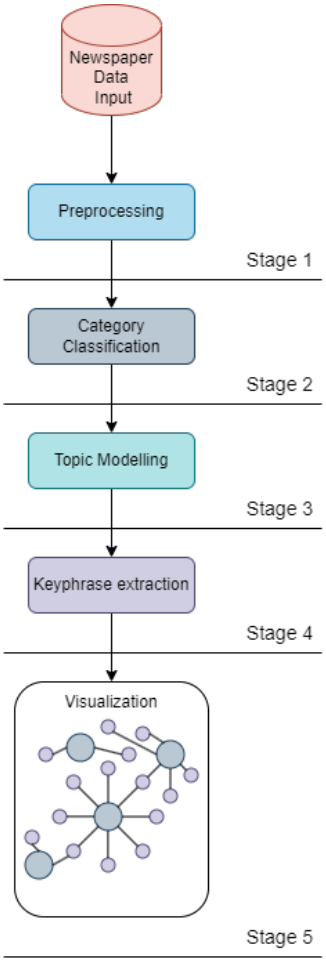

Figure 2: Social Pulse Data pipeline

Data transformation
In this step, the content data which was initially in a HTML format was transformed to a pure text format. This was done by changing the entire content of each article to lowercase, removing patterns of links, images and other unwanted characters and removing the stopwords.

4.3 Categorization

After the transformation, each article received an initial main category assignment. This step was done based on the number of occurrences of keywords of the respective category. Due to time constraints, the initial keyword of main categories could not be constructed via the TF-IDF score of the predefined categories provided by the raw meta-data, but rather was defined manually. The adaption of an updated data transformation in a second version of social pulse could drastically improve the result quality.

4.4 Topic Modeling

In order to define the different subtopics of each of our category's topic modeling was used. In a nutshell this step leveraged patterns within the content data to create subtopics. Due to the high complexity of such a method, scaling such approaches for a large corpus containing all data from a main category is deemed a difficult task. In order to reduce the overhead of complexity and computation time, a novel approach to topic modeling was chosen. This was achieved by using a large language model for german news articles (BERTopic), enhancing the topic clustering (HDBSCAN) and adding a dimensionality reduction for the text embedding(UMAP).

BERTopic

BERTopic is considered a state of the art language model which has been trained with a large data set of german newspaper texts. It serves as a bidirectional endcoder representation (BERT) architecture, which enabled us for further computations. The diverse knowledge captured in BERT helped us to encode each newspaper article with contextual information, which retained more information about vocabular and grammatical structures than traditional other encoders. With this large advantage our entire topic modeling kept a lot of informational content gathered in the previous steps and built a basis for our text classification part of the topic modeling.

HDBSCAN

In order to efficiently cluster the data gained from the BERTopic encoder HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) was used. This algorithm facilitated us to identify clusters of different shapes and sizes within the embeddings provided by BERTopic. In addition, it automatically determined the number of optimal clusters with the convergence of cluster stability. Through HDBSCAN groups of documents with logic similarity were created, enabling us to have coherent subtopics. Due to the high complexity of the algorithm cuML version of HDBSCAN was used to leverage GPU acceleration provided on Google Colab.

UMAP

Smaller, better, faster. With these adjectives in mind UMAP was used in our project. UMAP is an advanced tool for dimensionality reduction which preserves underlying structures and relationships within the data. Through this optimization the process of topic modeling was expedited while maintaining data integrity, granting Social Pulse a study and coherent topic modeling. As with HDBSCAN, we used the cuML version of UMAP for GPU-optimized runs of the algorithm.

4.5 Keyphrase extraction

In this project, the key phrase extraction should enable the users of the Social pulse solution to have a vague idea of the content in the articles of a specific subgroup and build a source for a language model to create a reader-friendly summary.

For the key phrase extraction, the KeyBERT was used, which worked alongside BERTopic to capture semantic information and identify key phrases within the document groups provided by HDBSCAN and UMAP. Even though KeyBERT is a great way to conduct key phrase extraction the limited key phrase length and vocabulary have to be taken into consideration, when working with KeyBERT. The mentioned limitations could lead to the missing of more meaningful key phrases, which are ignored due to the n-gram filtering (only considering phrases of length n).

In order to verify the quality of the key phrases and mitigate the risks associated to KeyBERT, a KeyphraseCountVectorizer from the spacy library was utilized. Spacys unique german news language model captures adjective-noun phrases common in German keyphrases and aligns this information with the given previous key phrases from KeyBERT.

4.6 Visualization

From the beginning of the project, there was a very clear view on how the social pulse solution should look and feel like (namely a graph structure containing relations and versatile navigation and filtering possibilities). Due to multiple hurdles in the solution implementation we decided to leverage Obsedians Graph to visualize the aggregated data.

In order to leverage Obsedians graph representation markdown files were created incorporating relationships between rows such as category and subcategory via regular expressions. With the help of this cumbersome step, the basis for an ideal representation of relations was enabled.

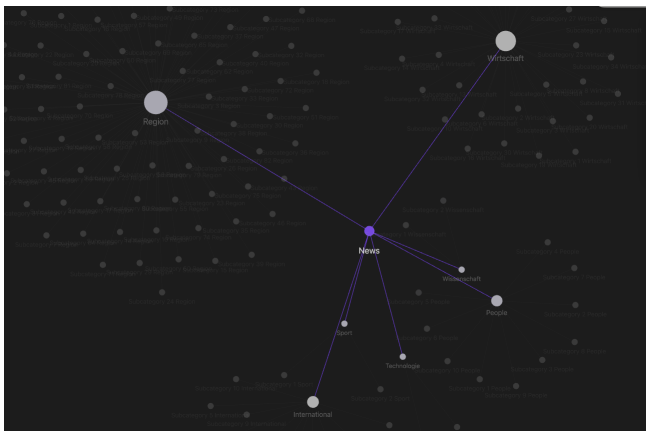

Figure 3: Detailed view of main category's highlighted in graph

5 DATA

In order to keep a clear overview of the data throughout the pipeline and to detect issues with modules, different stages were defined for the data (as visible in figure 2). This unconventional approach enabled us to develop the pipeline in an order suited to our own working pace without potentially losing time on synchronizing the usage of modules.

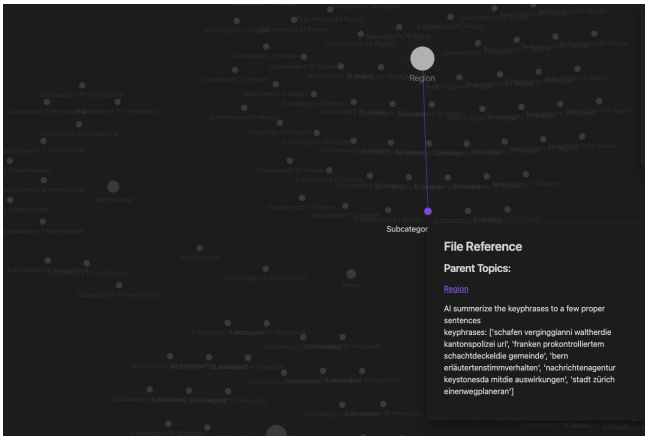

Figure 4: Detailed view from node Information in graph view

5.1 Data at stage 1

Upon entry: raw news data format as provided from swissdocs@liri **File:** TSV containing meta data and respective article in HTML format

5.2 Data at stage 2

Upon entry: news data frame containing cleansed and preprocessed column **File:** csv containing columns with unprocessed meta data from previous stage and processed content column **Side Note:** either in chunks or as one file depending on version of preprocessing

5.3 Data at stage 3

Upon entry: news data frame containing new column containing main category **File:** csv **Side Note:** either in chunks or as one file depending on version of preprocessing

5.4 Data at stage 4

Upon entry: embedding of the topic modelling **File:** large csv file

5.5 Data at stage 5

Upon entry: table containing different subtopics and respective key phrases **File:** csv file per main topic

6 PERFORMANCE ENHANCEMENTS

After the completion of the first version of the Social pulse project major performance bottlenecks were recognized, based on which a more in dept analysis was initiated.

6.1 Preprocessing

With approximately 6 minutes computation time for 1 months' worth of news data of only one newspaper, it was clear that the preprocessing module needed an upgrade.

6.1.1 Initial enhancement. After auditing the code of the preprocessing module a bit closer, a high theoretical computation complexity was recognized as factor for the rather slow computation speed. With this in mind, the most basic best practice code performance enhancements were employed (minimizing the number of for loops and avoiding nested for loops)

6.1.2 Advanced enhancement. Even though smaller performance enhancements were noticed at this point it was too early to call the preprocessing of Social Pulse mature for use. Based on this standpoint the quest for further performance enhancements started, which quickly also led to the questions about pythons' way of handling our code. **Python**
Python is a dynamically typed programming language, which is written in the low-level language C (also considered as CPython). Due to Python being compiled with a C compiler, Python is usually interpreted. Hence, creating bytecode which is then executed on the Python virtual machine [6].

Cython
Cython is a static compiler for python and C code, which enables the user to write C extended python code. This implementation grants performance enhancements due to C functions nature of being written in a more compact performance oriented static type. In addition cython is usually not type checked at runtime, reducing the computational overhead by a great margin [7].

Swifter
Swifter is a library enabling python programmers to vectorize their function in a more efficient way than Pythons standard apply function. A vectorized call enables the system to compute different instances of a list or pandas column independently, which creates a rapid speed up in computation time [8].

Tuplex
Tuplex is a parallel big data process framework, which enables developers to run python code at speeds of compiled code languages such as C. In order to guarantee such speeds Tuplex generates an LLVM bytecode for the used data and the functions in the python file. Due to it's data-driven compilation and dual-mode processing enormous performance improvements gained when used adequately [9].

With all this advanced performance enhancement in mind, experiments were run to compare average runtime and to make the best possible augmentation of our preprocessing module (see table 1).

Table 1: Speed Comparison of Preprocessing Codes

Preprocessing Code	Processing Time (40k data)
Cython	481.82 seconds
Normal Run	415.70 seconds
Swifter-Vectorized Call	409.96 seconds
Tuplex	28.88 seconds

Different than expected Cython performed worse than the normal computation of the preprocessing without any advanced enhancements. Reason for this result is mainly correlated to the fact that the created functions capture and use pythons pandas library for as data frame, which is not a standard type of C, hence forcing Cython

to infer back to normal python in most cases.

Even though Tuplex presents the fastest solution with around 14x faster computation in comparison to normal computation, the swifter solution was chosen for the current performance improvement of the version 2 of the preprocessing. Reason for this choice is mainly the rather experimental implementation of Tuplex and the pitfall, that the worker management has not yet been implemented in a functional stable way.

6.2 Topic Modeling

Next to the preprocessing the topic modeling module was also a module to consider for performance enhancements. Initially the Gensim LDA (Latent Dirichlet Allocation) algorithm was our used approach for topic modeling. However, due to the need for faster processing, we explored alternative methods. In this context adopted BERTopic in conjunction with UMAP (Uniform Manifold Approximation and Projection) and HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithms. Table 2 presents a comparative analysis of the processing speeds between Gensim LDA and our proposed approach. The results highlight the superior efficiency of BERTopic with UMAP and HDBSCAN in generating topic clusters, enabling faster analysis of news data.

Table 2: Speed Comparison of Topic Modeling Models: BERTopic vs Gensim LDA

Hyperparameter	Value
Model	BERTopic
Processing Time (40k data)	3 minutes
Processing Time (10k data)	0.75 minute
Model	Gensim LDA
Processing Time (40k data)	10.5 minutes
Processing Time (10k data)	2.625 minutes

6.3 Google Colab

In our project the use of Google Colab brought various challenges related to GPU and RAM capacity. To overcome these limitations, we have used several optimization strategies.

6.3.1 Computation optimizations. By implementing batch processing, we divided the dataset into smaller portions, effectively managing GPU memory and enabling us to handle larger datasets without memory overflow issues.

Throughout the project, we closely monitored and managed memory usage by clearing unnecessary variables and periodically refreshing the Colab runtime.

These optimization strategies enabled us to overcome GPU and RAM constraints in Google Colab, empowering us to perform more computationally demanding topic modeling tasks.

6.4 Containerized enviroments

In order for Social pulse to also be scalable solution optimized with performance multiple docker images were created, which implemented the pipeline in an optimized setting. These created containers merely deferred in the accelerators used for the preprocessing. For Social pulse to be a scalable solution these Containers were deployed in a google cloud cluster, which was configured to automatically scale to the needed number of instances via Kubernetes.

7 RESULTS

Overall, the research was fulfilled its main objective and proved that aggregated and summarized news can be visualized in a graph. Unfortunately, only the team's personal impression of the effect of such a visualization could be reviewed. Social pulse as an avantgarde way of reading large amounts of news data is a compelling approach, which might have the potential to renew societies interest in interacting with news. Not only does the solution present a new graph view, but also maintains traditional forms of summarized newspaper articles on a node basis, leaving each user the liberty to freely interact with the system. In addition to performance enhancements the modules could all be uniquely deployed on a google cloud cluster in a later version to enable better auto-scaling of the solution.

8 CONCLUSION

Due novelty of this projects approach of visualizing large amounts of data, the end results feasibility will be audit in more detail after project completion. Based on this audit an additional evaluation can be made to argument a possible extention of the project with multiple data sources on a grander scale. In addition, this research can be used as baseline for other similar applications such as creating knowledge graphs for scientific information. The application of additional data on structures of such kind could give the user various additional insight without drastically increased complexity, hence creating a new approach to understanding big data about topics in a manageable manner.

REFERENCES

[1] Cambridge Dictionary. Aggregation. Retrieved from https://dictionary.cambridge.org/dictionary/english/aggregation.
[2] Cambridge Dictionary. Summarization. Retrieved from https://dictionary.cambridge.org/dictionary/english/summarization.
[3] Reuters Institute. Digital News Report 2022: Executive Summary. Retrieved from https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022/dnr-executive-summary.
[4] Reuters Institute. Digital News Report 2022: Executive Summary. Retrieved from https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022/dnr-executive-summary.
[5] GitHub repository containing the source of our project.https://github.zhaw.ch/BigData-Project-FS23/BDP_oseiaced_sivassit.
[6] Python Software Foundation. The Python Language Reference. Retrieved from https://www.python.org/doc/.
[7] Cython. The Cython Programming Language. Retrieved from https://cython.org.
[8] Carpenter, J. (jmcarpenter2). Swifter: A package which efficiently applies any function to a pandas dataframe or series in the fastest available manner. Retrieved from https://github.com/jmcarpenter2/swifter.
[9] Tuplex. A Parallel Big Data Processing Framework for Python. Retrieved from https://tuplex.cs.brown.edu.