https://doi.org/10.53656/math2023-4-2-enr

ENRICHING LARGE DOCUMENT STORES WITH INTELLIGENT METADATA: A FRAMEWORK FOR EFFECTIVE KNOWLEDGE MANAGEMENT AND APPLIED ANALYTICS

Penko Ivanov, Elitsa Pavlova
The Financial Times (Bulgaria)

Abstract. The current paper focuses on a framework for structuring large document stores with the help of intelligent metadata. The described landscape includes a proprietary knowledge graph which ingests millions of concepts from external, third-party data providers and accommodates internal class taxonomies; an NLP service for automated annotation of textual data; an annotations quality control mechanism; tools for knowledge graph ontology and concept management; and an extensive API layer. The authors present an approach they have tested and proved successful in one of the leading media companies in the world, whose media content is a core data asset. The proposed solutions enable content analytics in their proper context and allow explicit and implicit connections between the content and other company data – i.e., user (media content consumer) data. The latter empowers the efficient application of advanced analytical models for searches and recommendations and the implementation of accurate data-driven virtual assistants.

The paper advises addressing the metadata quality concerns, which the authors' extensive practice identifies as an essential prerequisite for applied analytics delivering significant business value.

Keywords: software engineering; AI; data science; machine learning; NLP; metadata; knowledge graphs; ontology; metadata quality; business analytics

1. Introduction

Business organisations nowadays collect more data of various types than ever before. In the age of the information explosion, companies and institutions are confronted with an enormous influx of text data daily. The proliferation of digital content, ranging from research articles, news reports, social media posts, and corporate documents, poses significant challenges to effectively managing, organising, and extracting valuable insights from these vast collections of textual information. Traditional data organisation methods, such as conventional databases and folder-based systems, are often insufficient to capture textual data's inherent relationships, context, and nuances, hindering efficient data retrieval and knowledge discovery.

Enriching content with metadata plays a pivotal role in the evolving landscape of textual data consumption, as it enables efficient information retrieval by software applications, reducing the reliance on human-driven processes. High-quality metatags are essential for structuring text and putting it into specific business context, and having defined relationships between the concepts used for meta-tagging more profound information discovery in large textual data stores. A comprehensive metadata model that allows accessing data by traversing links turns a large document store into a knowledge base. (Gosnell & Broecheler 2020).

Also, over the last year and even months, we have witnessed remarkable progress in the field of generative artificial intelligence and large language models (LLM). While for the mass Internet user, this development comes down to the impressive capabilities offered by the user interface of ChatGPT, for businesses, qualitatively new cloud services are available that make the implementation of complex analytical models increasingly accessible. Companies like OpenAI and Google offer API access to LLMs that deliver a variety of natural language processing (NLP) solutions, including text classifiers and conversational engines. Along with the new opportunities, however, the industry faces new challenges. While these generally available models, pre-trained on large, general-purpose datasets, are easy to use, they require additional fine-tuning to perform in a particular business context requiring specific domain knowledge.

Availability and discoverability of domain-specific data is a must to successfully put a generally available pre-trained transformer-based LLM in a domain-specific context, emphasising the increasing importance for business organisations of maintaining a high-quality internal knowledge base as a critical prerequisite to applied analytics in the era of AI services.

More concretely, LLMs work with vector representation of the textual data, which doesn't eliminate the need for metadata, but increases its importance. Adding high-quality metadata to the vector embeddings representing tokens' meaning adds more context and makes the model responses more precise and relevant to the specific business domain.

This paper showcases managing a well-organised, high-quality knowledge base in a complex business environment. Although the showcase is in a specific business area, advantageous for the subject, the presented methods are generally applicable in various contexts.

The following sections demonstrate the approaches to knowledge management applied by the paper's authors and proven successful in a world-leading media company – The Financial Times (FT).

2. Maintaining a proprietary Knowledge Graph (KG)

Their media content is a fundamental asset for companies in the media domain. In the organisation in consideration – The FT, the media content from different internal sources is published in a shared document store (non-relational database) and made available for various internal and external consumers, including websites, mobile apps, content

analytics and data science teams. News is published hundreds of times per day. Thus, the document store volume is rapidly expanding. By writing the current paper, the number of content pieces from different types (mainly articles) stored in the document store had risen to 1,172,000. Also, the focus on particular aspects of the natural world reflected in the media content frequently changes depending on the news agenda. Therefore, the domain requires data management approaches which are flexible not only in terms of the volumes and velocity of the data but also in terms of its variety and veracity.

To make the content items in this extensive document store discoverable, we annotate every piece of content with relevant real-world concepts – people, organisations, locations, topics, etc. We also maintain a comprehensive metadata model (Strengholt 2020) of such linked concepts to be used for annotations. An annotation is a link between a content item and a concept (fig. 1).

The annotations are created manually by the editorial team (a journalist puts a metatag manually in the software system where they write an article) and automatically with the help of machine learning algorithms. We keep all concepts, all relationships between the concepts, and all links between the content items and the concepts (the annotations) in a graph database to take full advantage of having a deeply connected heterogeneous dataset. A set of rules, an ontology, are applied to the graph data to put structure and semantics to it (Alexandropoulos 2020).

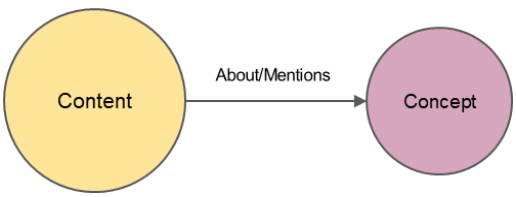

Figure 1. Annotation

The things FT writes about and wants to identify are "concepts": an idea or unit of thought. Most things we refer to as "terms" in the taxonomical world are "concepts".

Crucially, we represent each concept once in our ontology. That means giving it an ID (and URI in the semantic web world) and using that ID / URI whenever we refer to it.

Rather than place these in taxonomic hierarchies, we assign concepts to classes of thing; we categorise them. In many cases, these are real-world classes of things, such as people or locations, directly mapped to taxonomies. Others are more abstract, such as "membership", a class we have devised (along with W3C – https://www.w3.org/2004/02/skos/intro) that describes the relationship between a person and an organisation.

Once we have concepts put into classes, we can then define the relationships between them. We can create any number of relationships, but we aim to be pragmatic and will only define as much as we need to support our products and the systems that rely on this detail.

For instance, we already describe relationships such as:

– how an organisation has headquarters in a particular location;
– a person's job title with an organisation;
– a public company's main share and its corresponding FIGI (https://www.openfigi.com/) (financial) code;

– a company's subsidiaries.

The definition of the classes and the relationships between them is our ontology. We then store all these relationships in a graph database, Neo4J, not as an ontology but as nodes and edges. It is all very similar but optimised for very fast querying.

One example of the new types of relationships we have defined is the way we show how content is tagged. We have two relationships, called "about" and "mentions". The editorial tagging system allows journalists to identify all the things an item of content is about and separates them from the mere mentions of things. This clear distinction in the UI makes it easier for the team to know what to tag, and thus the corresponding pages on FT.com will become even more relevant. An obvious advantage is for alerting systems; to be alerted only when new content is about that thing you want alerting on.

"About" is not sufficient. There are two reasons for also having "mentions": some of our B2B customers want every mention of an organisation, and their systems will make choices depending on what they view as relevant. Examples are customers using our content in trading algorithms. The other reason is our internal use in training the machine learning algorithm that helps identify a concept in our content.

Figure 2 below shows a simplified part of our ontology showing the relationships between content and other concepts. The circles are classes. The black lines denote class hierarchy, whilst the blue lines show relationships (object properties) between classes.

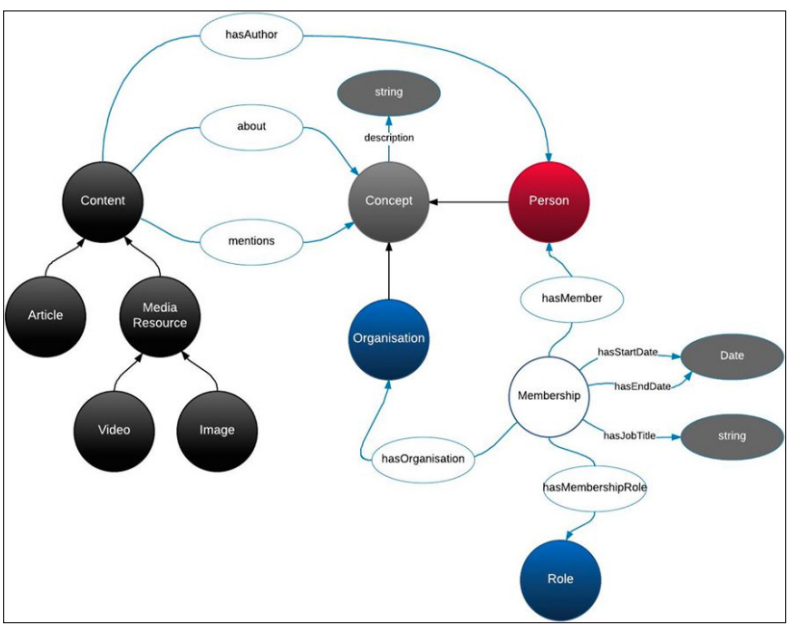

Figure 2. Simplified version of part of FT's ontology

The diagram shows:

– *an article is a type of content (as is a video and an image);*
– *an article has an author, who is a person;*
– *a person is in membership with an organisation and consequently has a job title, start/end date and a job role;*
– *an article can be about or can mention any concept, including a person or an organisation.*

concept duplication is being dealt with by concordance, which maps IDs to each other. Concordance is not just a means to deal with an internal data problem but is also an opportunity to concord with different external data sets. Created a robust solution to use open IDs across all of our datasets. There are loads of opportunities by doing this, for instance:

– we connect our locations to Geoname locations, enabling automatic mapping possibilities using Geonames mapping data;
– we connect to political databases, which allows us to show how MPs (Members of Parliament in the UK) voted and what constituencies they are in, which is particularly useful when covering elections.

Thus, the schemeless graph database is leveraged to a knowledge graph (Kejriwal et al. 2021.) holding a rich domain knowledge base (fig. 3).

Enriching content with comprehensive metadata is significant, given that machines (software applications), rather than humans, increasingly consume media. Attaching metatags at its creation ensures that the content is interpreted in the context set by its creators and guarantees the integrity of the extracted facts. The enriched content enables effectively applied analytics for various purposes – from content recommendation to events extraction.

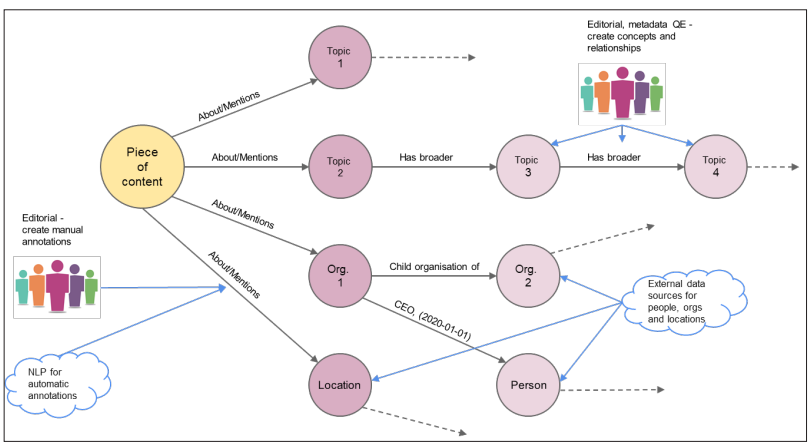

Figure 3. Ritch metadata organized in a Knowledge Graph (KG)

By writing the current paper, the described knowledge base contained about 30,414,000 concepts, representations of 1,172,000 content items and 134,267,000 relationships between these data points. Daily the knowledge graph receives around 35,000 automatic data updates from various sources. The vast data volume and the automated nature of data consumption require extensive data quality monitoring.

3. Knowledge base quality management

The overall data quality in the referred knowledge base depends on maintaining the quality of the automatically ingested data as well as having a mechanism to monitor that quality. The quality of the concept data highly depends on its sources – the data provider is responsible for completeness and accuracy. Such data, for example, is all the organisations in the knowledge base and all the relationships between them – the organisations necessary for the media and how they relate to each other. There are open sources of knowledge like Wikidata (reff. to https://www.wikidata.org/wiki/Wikidata:Main_Page) with extensive data about different domains, Geonames (reff. to https://www.geonames.org/) with location data, etc. A proprietary knowledge base can take full advantage of those open data sources with quality monitoring mechanisms. Also, multiple paid data sources in different domains can be incorporated into a knowledge base.

As opposed to the quality of the concept data, which is mainly the responsibility of the data provider, there are domain-specific data points in the knowledge graph. Their quality should be the organisation's responsibility for operating on and creating the data. Such domain-specific data is the metadata used to classify media content.

4. Intelligent metadata via Natural Language Processing (NLP)

Handling large document stores relies on the help of intelligent metadata. Intelligent metadata, we call our approach to innovative information management that unifies information across different sources based on context, not on the system or database in which the data is stored.

The process of relating media content to real-world concepts is enriching the content with metadata or annotating the content. Examples of annotating an article are finding the topics of the article or extracting all the people, organisations, and locations mentioned in it.

Natural Language Processing (NLP) techniques are heavily utilised to automate the processes for extracting metadata from human-written text such as media content. Two approaches from the NLP field – a Named Entity Extraction (NER) and topics classification - are commonly adopted for annotating content with metadata (Lane et al. 2019).

NER finds entities mentioned in a human-written text. Then those entities are classified into different types. Look at the following sentence – "Remember the

phrase, “It’s the economy, stupid?” It was coined by James Carville, strategist of US President Bill Clinton’s successful 1992 campaign against George H W Bush.” NER algorithm should recognise “James Carville”, “Bill Clinton”, and “George H W Bush” as entities of type Person.

As an enhancement to the classical NER, there is an additional technique known as entity linking, entity disambiguation. After the named entities are found in a text, they are linked to unique concepts from a knowledge base to be assigned a unique identity. The knowledge base referred to in this paper is the knowledge graph. We should stress the importance of the entity disambiguation. The knowledge graph and its ontology represent the semantics of the specific domain. So, linking the articles with the knowledge graph concepts puts them in the context for further data analysis.

If we go back to our previous example of “Clinton”. The knowledge graph has many concepts with that label – from President Clinton, the politician Hillary Clinton to the Iowa City of Clinton. If the word “Clinton” is mentioned in a media article, it should be recognised as the correct concept and linked to it. The actual value from NER is unlocked by linking the content with the right concept from the world in the required context.

Topic classification is a different task in the NLP world. It solves the problem of assigning a text to one category or a list of categories. Those categories could be the editorially curated news topics in the media domain. As opposed to the problem definition of NER, in topic classification, the extracted topic doesn’t have to be mentioned in the text at all. The semantics of the article, though, should refer to the topic. The topics’ taxonomies are tight in the specific context in which the media operates. It represents the view of journalists on the world news agenda. Data categorisation based on such domain-specific criteria again allows linking the data with the needed context of the media business. Categorising news by topics which are not relevant to the news agenda diminishes the value of the classification.

An NLP software system extracts any metadata – named entities and topics.

5. Metrics for NLP system performance

Using metrics different from error percentage is prevalent to measure the performance of NER or text classification algorithms (Shmueli et al. 2020). A very well-adopted pair of metrics are precision and recall.

To calculate any algorithm performance metric, including precision and recall, a version of the problem’s output, considered the ground truth, should be in place for comparison with the work produced by the evaluated algorithm. The problem’s output referred to in this paper is the content annotations. There are strict definitions of precision and recall, but we will put those in the context of annotating media content.

For a media article, there is a set of ground truth annotations, A_{true}, and a set of annotations extracted by the NLP system we are evaluating, A_{NLP}. To calculate precision and recall for an article is helpful to calculate the following values:

– true positives are the annotations which are both in the set of A_{true} and A_{NLP}, or these are the correctly found annotations by the NLP system;
– false positives are the annotations which are not in A_{true} but are in A_{NLP}, or these are the incorrectly extracted annotations by the NLP system;
– false negatives are the annotations which are in A_{true} but not in A_{NLP}, or these are the annotations not found by the learning by the NLP system;
– TP_A – the number of true positives annotations in article *A*;
– FP_A – the number of false positives annotations in article *A*;
– FN_A – the number of false negative annotations in article *A*;
– P_A – the precision of the annotations in article *A*;
– R_A – the recall of the annotations in article *A*.

Then the precision and the recall for article *A* are calculated using the equations:

$$Precision_A = \frac{TP_A}{(TP_A + FP_A)}, \; Recall_A = \frac{TP_A}{(TP_A + FN_A)}.$$

The precision of the learning algorithms will be the fraction of true positive annotations of all annotations predicted from the learning algorithm. Recall of the learning algorithm will be the fraction of true positives annotations of all annotations that are actually positive. For most learning algorithms, there is a trade-off between precision and recall. Techniques which improve precision can reduce recall and vice versa.

One challenge with precision and recall is that you are evaluating the performance of a learning algorithm or NLP system using two different metrics. Since there is usually a trade-off between precision and recall, it is hard to systematically monitor and compare the system's performance using two metrics. One commonly used way to combine precision tooling and practices allowing evaluations on any media content at any time, which is part of the automatic annotation process. The on-demand evaluation could have different output and recall is to use a metric called F1-score. F1-score is a way to connect precision and recall by emphasising whether these values are lower. In that way, very low precision or recall, both cases we would like to avoid, will be penalised in the final metric.

$$F_{1_A} = \frac{1}{\frac{1}{2}(\frac{1}{Precision_A} + \frac{1}{Recall_A})} = 2\frac{Precision_A Recall_A}{Precision_A + Recall_A}$$

The equation above represents the harmonic mean of precision and recall.

These equations are used to evaluate the performance of an NLP system on a single article. But usually, the performance is evaluated towards a set of media

articles. For averaging across a corpus of content pieces, we can take different approaches:

– Micro averaging treats the entire corpus as a big document to calculate precision, recall and F1;

– Macro averaging takes the average over each precision, recall and F1.

Let's consider a corpus of documents D, where $|D|$ is the number of documents in the corpus.

$$Precision^{Macro} = \frac{\sum_{A \in D} Precision_A}{|D|}, Recall^{Macro} = \frac{\sum_{A \in D} Recall_A}{|D|}$$

$$F_1^{Macro} = \frac{2 * Precision^{Macro} * Recall^{Macro}}{Precision^{Macro} + Recall^{Macro}}, Precision^{Micro} = \frac{\sum_{A \in D} TP_A}{\sum_{A \in D}(TP_A + FP_A)}$$

$$Recall^{Micro} = \frac{\sum_{A \in D} TP_A}{\sum_{A \in D}(TP_A + FN_A)}, \quad F_1^{Micro} = \frac{2 * Precision^{Micro} * Recall^{Micro}}{Precision^{Micro} + Recall^{Micro}}$$

All these metrics and their averaging approaches could also be applied per annotation type. For instance, you can measure the performance on the annotations of type Person only

6. Strategies for monitoring NLP system performance

Equipped with the metrics defined above, we utilise them to perform data quality monitoring and support data quality enhancements. The dynamics of media data require monitoring the performance of the NLP algorithms on several levels (fig. 4).

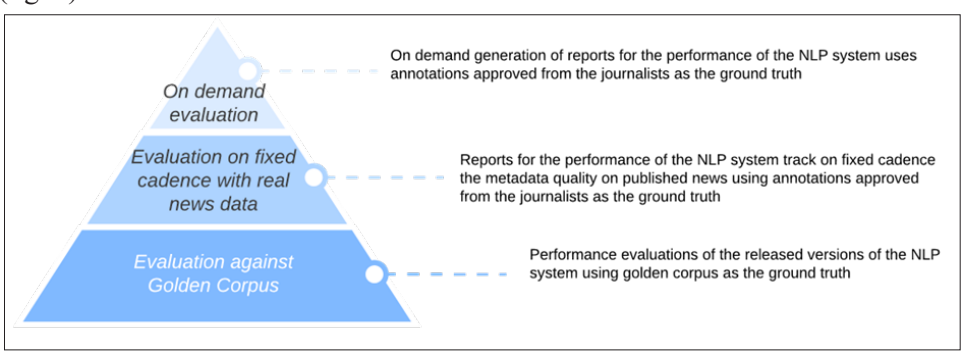

Figure 4. Multilevel performance monitoring

The ability to evaluate how well the NLP system performs at any given time is required when working with a dynamic and vast knowledge base (Huyen 2022). The on-demand evaluation refers to formats, and quality reports are being used by the system referred to in this paper. The quality reports could run on specific media

content or from particular periods. The NLP system produces the annotations evaluated in the reports. The ground truth is the annotations are reviewed by the journalists and accepted as correct.

The quality report should have overall statistics for the performance of the NLP system on the chosen set of documents. The preferred approach is having statistics for overall performance on all annotation types and statics per concept type. You may notice from the summary report in this paper (fig. 5) that different concept types perform differently, which is important to consider. The topic annotations are extracted from topics classification algorithms which operate differently from the NER algorithms. Though the NLP system is regarded as one system extracting all kinds of metadata, separation per annotation types or learning algorithm types is extremely useful for identifying where a potential problem may be hidden. From experience, topic classification is much more challenging and requires more analysis and attention than extracting entities like people, organisations, and locations. The changes in the news agenda require rapid changes in emerging topics which should be closely monitored, and the algorithms for topics extraction retrained appropriately to reflect such changes.

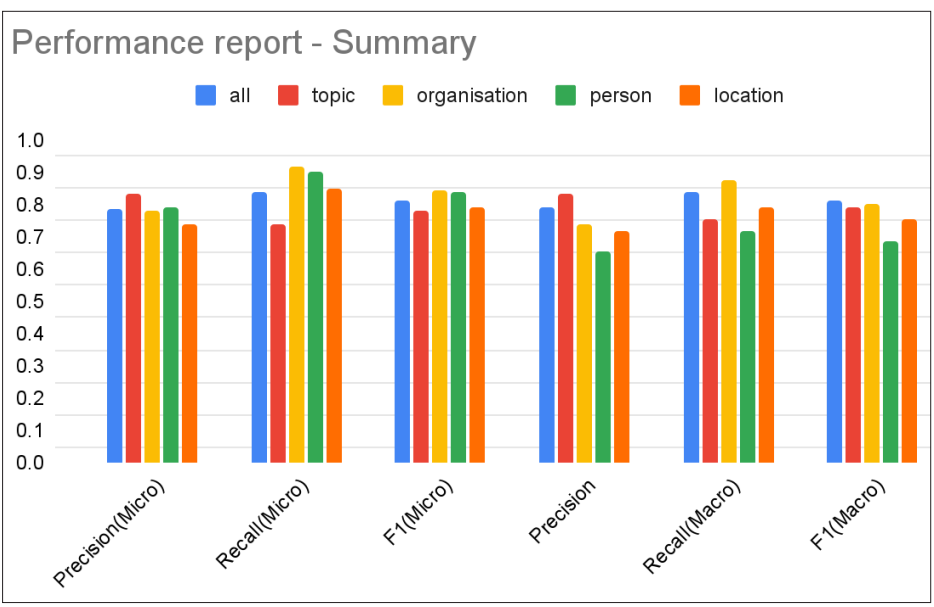

Figure 5. Performance report visualisation – summary view

Several aspects of the media business directly benefit from flexible on-demand metadata quality reporting mechanisms. For media organisations, content for important world topics should have an excellent quality of its metadata to be easily

discoverable. Tracking the metadata quality on all the news about the "Coronavirus pandemic" was a priority task for the business just a year ago. The flexibility of on-demand quality reporting easily satisfies this need. Another example is tracking the metadata quality on scoop content - another high-priority task for the media organisation, which is usually needed immediately.

Furthermore, the metadata quality reports could integrate a summary of metrics and lists of false positives and negatives, standard errors, suggestions for improvement and enrichment of the data. The quality reporting system referred to in this paper suggests concepts that could be added to the knowledge base because they were found in the news but did not exist in the knowledge base.

There also should be a systematic approach for monitoring the quality of the annotations of each published article. The reporting system used for on-demand monitoring could also be utilised for this task by creating reports regularly, including all the media content being published. Again, the ground truth annotations are the annotations which were reviewed by the journalists and accepted as correct. The cadence in which the regular reporting is tracked is biweekly (fig. 6).

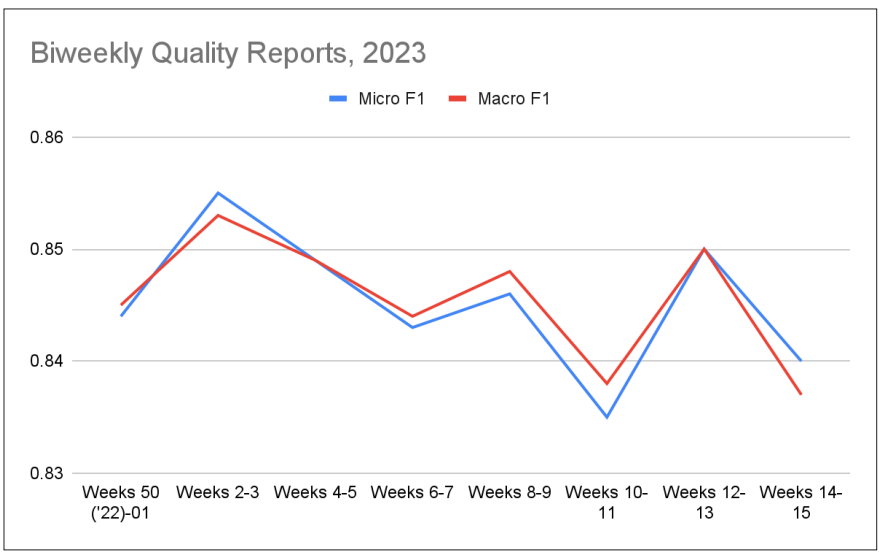

Figure 6. Performance report visualisation – overtime view

In such regular reports, the desired trend will be slightly fluctuating micro and macro F1-scores. The system's overall performance should not be susceptible to changes in the news or changes in the data of the knowledge base. As you can observe, for the first weeks of 2023, the evaluated system kept its F1 score between 0.83 and 0.86.

Measuring the automatic annotations against the editorial is crucial. That is how we can ensure that the evaluation encapsulates the view of the creators of the content. Therefore, the metrics are used in the correct context to truly represent the performance of the NLP algorithms for solving concrete media organisation business needs.

Measuring against human output on production data is one of many ways to address the problem of evaluating NLP system performance. A golden corpus of annotations could be created and supported. It consists of articles annotated by professional human annotators following annotation guidelines. A media organisation should have defined guidelines for annotating its content. Such policies encapsulate what is essential from a metadata perspective for the business.

The data in a Golden corpus is considered to have higher quality as it is created by professional annotators and according to the given annotation guidelines. Because of its higher quality, the Golden corpus is used for retraining the NLP machine learning models and overall evaluation of the NLP system.

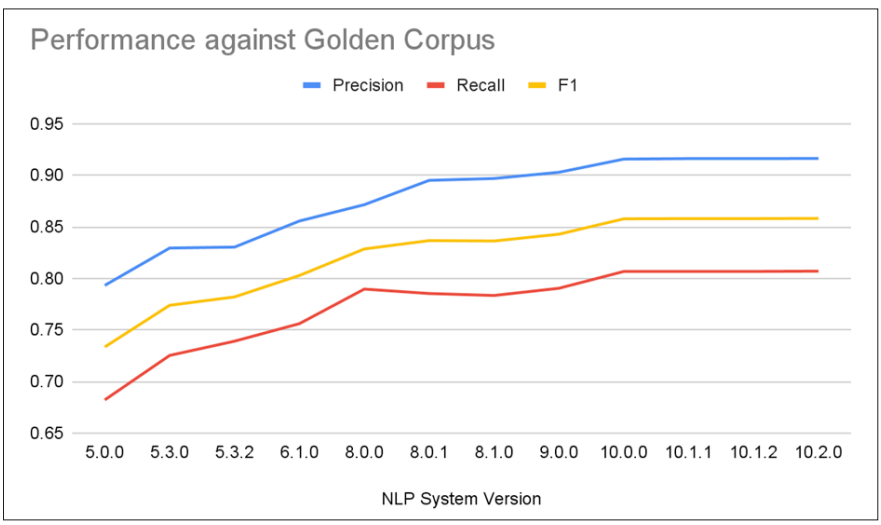

Figure 7. Performance report visualisation – model versions view

The desired trend when evaluating incremental versions of an NLP system is increasing precision, recall and F1-score for each following version (fig. 7). Any algorithm changes between the versions should not affect the overall trend.

7. Conclusions

This paper presents a framework for structuring large document stores with intelligent metadata, focusing on a proprietary knowledge graph that ingests concepts from external data providers and internal class taxonomies. The proposed

solutions have been successfully implemented and tested at The Financial Times (FT), a world-leading media company, where media content serves as a core data asset.

Enriching content with comprehensive metadata is vital for efficient information retrieval and knowledge discovery. The knowledge graph, a graph database following a proprietary ontology, transforms the large document store into a knowledge base, allowing explicit and implicit connections between content and other company data, such as user data. This knowledge base enables the application of advanced analytical models, content recommendations, and the implementation of data-driven virtual assistants.

The paper emphasizes the significance of maintaining a high-quality internal knowledge base as a prerequisite for applied analytics in the era of AI services. Incorporating generally available pre-trained large language models into a domain-specific context requires the availability and discoverability of domain-specific data. Metadata quality is essential for delivering significant business value through applied analytics.

The intelligent metadata is generated through Natural Language Processing (NLP) techniques, including Named Entity Extraction (NER) and topic classification. Monitoring the performance of the NLP system is crucial, and we introduced metrics like precision, recall, and F1-score to evaluate the system's accuracy in extracting metadata from media content. On-demand quality reports and regular reporting mechanisms are employed to track the metadata quality and identify areas for improvement.

Overall, the framework presented in this paper demonstrates effective knowledge management in a complex business environment, showcasing the benefits of enriching content with intelligent metadata and the importance of maintaining a high-quality knowledge base. The methodologies and practices discussed here apply not only to the media industry but to various other domains with vast collections of textual data. By leveraging intelligent metadata and AI technologies, businesses can enhance their analytics capabilities, make better data-driven decisions, and deliver more relevant and personalized customer experiences.

REFERENCES

ALEXANDROPOULOS, P., 2020. *Semantic Modelling for Data*. Sebastopol, CA: O'Reilly, ISBN 978-1-492-05427-6, pp. 14 – 32.

GOSNELL, D., BROECHELER, M., 2020. *The Practitioner's Guide to Graph Data*. Sebastopol, CA: O'Reilly, ISBN 978-1-492-04407-9, pp. 2-9.

HUYEN, C., 2022. *Designing Machine Learning Systems*. Sebastopol, CA: O'Reilly, ISBN 978-1-098-10796-3, pp. 150 – 188.

KEJRIWAL, M., KNOBLOCK, C., SZEKELY, P., 2021. *Knowledge Graphs*. Cambridge, MA: The MIT Press, ISBN 978-0-262-04509-4, pp. 21 – 44.

LANE, H., HOWARD, C., MAX HAPKE, H., 2019. *Natural Language Processing in Action*. Shelter Island, NY: Manning, ISBN 978-1-617-29463-1, pp. 339 – 361.

SHMUELI, G., BRUCE, P., GEDECK, P., PATEL, N., 2020. *Data Mining for Business Analytics*. Hoboken, NJ: Wiley, ISBN 978-1-119-54984-0, pp. 126 – 155.

STRENGHOLT, P., 2020. *Data Management at Scale. Sebastopol*, CA: O'Reilly, ISBN 978-1-492-05478-8, pp. 265 – 285.

✉ **Penko Ivanov**
ORCID iD: 0009-0003-7953-109X
Principal Business Analyst
The Financial Times
9, Moskovska St.
1000 Sofia, Bulgaria
E-mail: penko.ivanov@gmail.com

✉ **Elitsa Pavlova**
ORCID iD: 0009-0003-4743-6880
Senior Software Engineer and Tech Lead
The Financial Times
9, Moskovska St.
1000 Sofia, Bulgaria
E-mail: elitsa.i.pavlova@gmail.com