

MASTER

Interpretable Evaluation of Factual Consistency in Abstractive Summarization

Hegger, Daan

Award date:
2023

[Link to publication](#)

Disclaimer

This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain

Interpretable Evaluation of Factual Consistency in Abstractive Summarization

by

Daan Hegger

MSC THESIS

Assessment committee

Member 1 (chair): Assistant Professor, M,
Holenderski
Member 2: Full Professor, M, Pechenizkiy
Member 3: N.A., Al-Serw

Graduation

Program: Computer Science in
Engineering
Capacity group: Interconnected Resource-
aware Intelligent Systems
Supervisor: Assistant Professor, M,
Holenderski
Date of defense: August 14, 2023
Student ID: 1254537
Study load (ECTS): 30
Track: Data Science in Engineering

The research of this thesis has been carried out in collaboration with *Accenture*.
This thesis is public and Open Access.

This thesis has been realized in accordance with the regulations as stated in the TU/e Code of Scientific Conduct.

Disclaimer: The department of Mathematics and Computer Science of the Eindhoven University of Technology accepts no responsibility for the contents of MSc theses or practical training reports.

Interpretable Evaluation of Factual Consistency in Abstractive Summarization

Daan Hegger (1254537)
 Eindhoven University of Technology
 d.hegger@student.tue.nl

Abstract—Despite recent improvements of text summarization models, still many current models create summaries with factual inconsistencies with the source document. Traditional evaluation metrics for text summarization do not account for factual consistency between a summary and its source. Various metrics have been proposed in the last years to detect these inconsistencies automatically. Even though these recent metrics are increasingly correlating with human judgement, most of these proposed methods use deep learning models to determine factual consistency. Here, the model transforms text (the document and the summary) into an embedding, and uses this numeric representation of text to classify whether (a piece of) the summary is consistent with its source document. However, these embeddings make it hard, and often impossible, for humans to interpret the reasoning behind their decision-making process. In this work, we propose a method that uses open information extraction and text comparison methods to evaluate whether facts stated in the summary are consistent with the information in the source document. Even though our metric uses deep learning in the pipeline, to perform semantic role labelling (a form of open information extraction), we perform the final comparison step on a token-level. Since tokens are (parts of) words, readable for humans, using them makes the process more interpretable. Our metric is evaluated on existing manually labelled datasets containing news articles, and summaries generated by summarization models. Our experiments show that using semantic role labelling increases the metric performance over triple extraction. Additionally, we created our own labelled dataset of summaries generated by three popular large language models.

I. INTRODUCTION

Recent advancements in the field of deep learning have led to the development of sophisticated language models, which can assist users in processing and understanding textual data. One specific task that language models have been employed for is text summarization, a process that aims to generate concise and coherent summaries of documents while retaining the most essential information.

Text summarization can be put into two categories: *extractive* and *abstractive*. Extractive summarization involves identifying key segments (such as sentences or paragraphs) directly from the original text and assembling them to form a summary. It doesn't alter the original text, but rather selects the most informative parts. Abstractive summarization aims to express the main ideas in new, condensed terms, much like how a human would summarize a text.

Extractive summaries have the benefit of staying closer to the meaning of the original document, but they often lack coherence, as they are a collection of extracted segments.

Abstractive summaries solve this issue by paraphrasing and merging information from the document.

Despite the potential benefits offered by abstractive summarization, a problem arises in the context of factual consistency. Factual consistency in abstractive summarization means every statement or claim made in the summary should align with the information in the original document and can be verified directly from it, ensuring no distortions or inaccuracies are introduced. When generating abstractive summaries, language models sometimes fail to generate fully consistent text. This may come from various factors, including the model's inherent limitations or biases in the training data. For instance, Cao et al. [1] showed that nearly 30% of the outputs of a (at the time of the study) state-of-the-art neural summarization system suffer from this problem.

The common way abstractive summaries are evaluated, is by evaluating their informativeness using simple textual overlap metrics. Using these metrics, a summary is compared to a human-written reference summary. ROUGE [2] is the mostly used metric to evaluate these models. However, since measuring textual overlap does not consider the semantic meaning of the text, it is possible for these metrics to assign high overlap scores when two pieces of text do not support each other. For example, take the sentences "John is studying" and "John is not studying". When comparing these sentences, the difference is only one word. However, the word "not" negates the second sentence, which makes it contradicting with the first sentence. To solve the evaluation of factual consistency, we either need human evaluation of summaries, which is time-consuming and defeats the purpose of automatic summarization. Or, we need metrics specifically designed to evaluate factual consistency.

In response to the lack of metrics to specifically detect factual inconsistencies, several metrics have been proposed to detect these inconsistencies in generated summaries. These factual consistency metrics perform a comparison between a summary and the source document, and outputs a (predicted) verdict whether the summary contains factual inconsistencies regarding the original document. Even though these recent metrics show better performance in terms of correlation with human factual evaluation, most of them perform their decision-making process by using neural networks or probabilistic methods in major parts of their pipeline.

The use of these neural networks can cause a lack of interpretability: When humans evaluate facts, they consider a fact to be consistent with a document when all its attributes (entities, relations, time/location circumstances, etc.)

are completely verifiable using the document. When using neural networks to make this decision, the text is transformed into an embedding, a vector of floating numbers which the network uses to construct determining features. These features are now hard for humans to understand, and it is also hard to check if the model verifies all attributes of a fact. This can be a problem in domains like journalism or healthcare, where the cost of presenting inaccurate information can be very high.

Another limitation of recent factuality metrics has to do with the rise of large language models (LLMs). Models like OpenAI’s GPT-4 [3] seem to do well on text-based tasks, like summarization. However, current datasets to evaluate factuality metrics do not include summaries generated by these LLMs. This is a problem because (1) we cannot evaluate the performance of factual consistency metrics on summaries generated by these LLMs, and (2) the summaries generated by LLMs show an increased level of abstractiveness compared to the summaries of non large-language summarization models.

Given the described lack of interpretability in recent metrics, and the lack of available data concerning factual consistency in summarization by LLMs, this work aims to address two key research questions:

- How can we automatically provide an interpretable evaluation of the factual consistency of an abstractive summary and its source documents?
- How many factual inconsistencies do the current state-of-the-art large language models produce when summarizing documents?

In this work, we present three primary contributions towards the study of factual consistency metrics:

- 1) An evaluation metric, called **FactMetric**, to detect factual inconsistencies in summaries is proposed and developed. This metric combines information extraction and textual comparison to determine consistency. To support fact verification for abstractive summaries, a graph comparison method is proposed. This method combines all facts of a document into a single graph, and uses path finding to verify whether a fact from the summary can be entailed by the document.
- 2) A labelled dataset of summaries generated by three popular LLMs is created. Here, each sentence in the summary is human labelled, indicating its factual consistency with the source document.
- 3) A comparison of **FactMetric**’s performance against other metrics is presented.

The following parts in this thesis are organized as follows: section II discusses related work. Section III presents the proposed evaluation metric. In section IV, the results regarding the metric and the experiments are presented. In section V, the results are discussed. At last, section VI concludes the work.

II. RELATED WORK

Recently, several approaches have been proposed to assess the factual consistency of abstractive summaries. In this section, we discuss different techniques and metrics used in related work.

We can divide the existing metrics into four categories, based on their verification approach: information-extraction-based, textual entailment based, question answering based, and other metrics.[4]

A. Information-extraction based metrics

In this approach, structured information is extracted from both the document and the summary, which is subsequently used to measure the overlap of information between the summary and the document. The most common representation of structured information from text is through fact triples, in the form of (*subject, relation, object*). Figure 1 presents an example of how the triples and comparison could look.

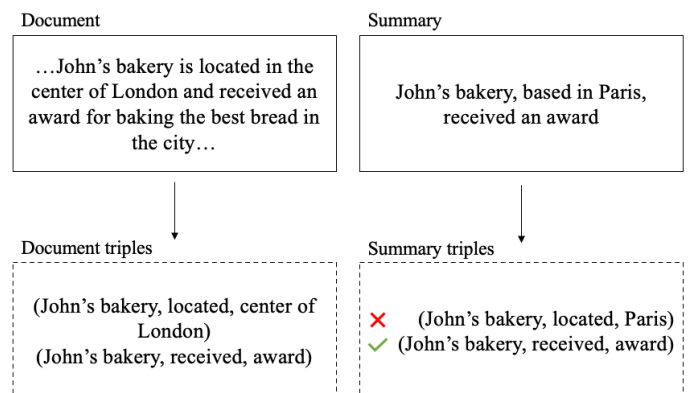


Fig. 1: Example of the triple based factual consistency verification

In this category, we found one metric. Goodrich et al. [5] proposed a metric based on this idea of triple extraction and measuring overlap. To extract the triples, they train their own triple-extraction model based on a transformer model. This model is trained to extract closed-relation triples from text, which means that it is tasked to only extract relations from a known set.

Limitation: While these closed-relations simplify triple comparison, since the possible relations between all the entities are predictable, it limits the types of triples that can be found. Applying this metric to a new setting means re-training this extraction model using custom training data.

FactMetric uses semantic role labelling to extract information from text. The chosen SRL implementation does not operate on a closed-relation set and tries to find all relations in the given text. This makes the metric easier adaptable to other domains.

B. Entailment-based metrics

Textual entailment is a task that involves determining if a piece of text (the hypothesis) logically follows or can be inferred from another piece of text (the premise). Metrics in this category follow from the idea that a summary is factually consistent with its source document if its text can be entailed by the source document.

Falke et al. [6]: Their work proposes to use existing textual entailment prediction models to solve the problem. They found that these out-of-the-box models trained on textual entailment datasets do not yet offer the desired performance for the task, which is their main contribution.

FactCC: Another metric in this category is FactCC, proposed by Kryściński et al. [7]. FactCC is a BERT-based model trained to identify whether sentences are factually consistent with a source document. The model is trained using a weakly-supervised approach. Their training data is collected by first gathering source documents in the same domain as the summarization models that are to be evaluated. Afterward, samples are created by sampling single sentences from these documents. These sentences are passed through a set of textual transformations that lead to changed sentences, with both *consistent* and *inconsistent* labels, depending on the type of transformation(s). They propose two types of transformations. Their semantically invariant transformation is applied by paraphrasing the sentence using backtranslation: translating the sentence to an intermediate language, followed by translating it back to the original one. Sentences transformed by this transformation are given a *consistent* label. For the semantic variant transformations, they propose: Entity and number swapping (replacing entities and number from the sentence with ones sampled from the rest of the document), pronoun swapping, and negating the sentence. A sentence changed by any of these transformations is given an *inconsistent* label. At last, the model is trained by giving it a (transformed) sentence and a document as input, and comparing the predicted (binary) label with the true label. This metric comes with a limitation. Since the training data contains synthetic errors, generated from a fixed set of transformations, the model is only trained to detect these kinds of errors. Furthermore, the training data is not manually verified after creation. This introduces the possibility of faulty transformations. For example, the *semantically invariant* transformation using translation could introduce a factual inconsistency if the translation module makes such a mistake during translation. Also, a *semantic variant* transformation like entity swap could make a factual inconsistent sentence consistent, for example, when a wrong entity is swapped with the right one.

FactGraph: This method by Ribeiro et al. [8] tries to compare the document and the summary by decomposing them first into structured meaning representations (MR). MRs are graphs that describe semantic concepts and their relations in a text. FactGraph encodes these graphs using a graph encoder. It uses the encoded representations of both the summary and the source document, along with the text (in the form of embeddings) of the summary and the source document, to classify whether a summary is consistent with a document. An advantage of FactGraph over FactCC is that FactGraph trained on human-labelled data (compared to the synthetic training data of FactCC). This can make the possible detected errors of FactGraph broader than the possible errors found by FactCC.

Limitations: All methods above share an additional limitation. All their decision-making processes (checking whether a sentence/summary is consistent) is done after transforming text (or a graph) into an embedding. These embeddings are long

vectors with (floating-point) numbers, and the content they represent is not trivially interpretable for humans. This creates a lack of interpretability of the decision-making process as a whole.

While parts of our metric also uses of neural networks, our method performs the final comparison between a (part of) the summary and the document in the form of (structured) text, which is interpretable for humans.

C. Question answering based metrics

This approach is based on the intuition that if questions are asked about a summary, the answers should be the same when using the summary as *available knowledge* and when using the source document as *available knowledge*. When comparing these answers, it is possible to measure the consistency of facts between the document and the summary.

Wang et al. [9] and Durmus et al. [10] respectively propose QAGS and FEQA, two QA-based metrics based on the intuition above. We can break down their method in the following steps:

- 1) A question generation (QG) model generates questions (and the correct answers) about the summary under evaluation.
- 2) A question answering (QA) model answers all questions using the source document as *knowledge* input.
- 3) The answers from the QG and the QA models are compared using textual overlap to compute a factuality score.

Limitations: Since the methods above do not hide their decision-making behind embeddings, but rather base their process on comparing the textual answers of (textual) questions, it has a higher interpretability than the textual entailment models. However, because of the complexity of the (deep neural) QG and QA models, these methods are computationally expensive. Moreover, these metrics can only detect the factual inconsistency if a QG model generated a question about the specific subject.

The neural networks in our metrics are smaller and provide faster inference. Furthermore, our metrics compares all facts from the summary with the document. It is not bound by a QG model to scope the facts under review.

D. Other metrics

Finally, we review two metrics to evaluate summaries in general.

ROUGE: The Recall-Oriented Understudy for Gisting Evaluation[2] (ROUGE) metric is a measure used for evaluating the quality of summaries by comparing them with reference summaries. ROUGE is a set of metrics, encompassing ROUGE-N (which measures the overlap of N-grams between the system and reference summaries) and ROUGE-L (which accounts for longest common subsequences). ROUGE comes in three variations: precision, recall, and F1-score. Take O as the number of overlapping n-grams between system and reference summaries, S as the number of n-grams in the generated summary, and R as the number of n-grams

in the reference summary. Now we can define the ROUGE scores as $ROUGE_{Precision} = \frac{O}{S}$, $ROUGE_{Recall} = \frac{O}{R}$, and $ROUGE_{F1} = \frac{2 * (Precision * Recall)}{Precision + Recall}$.

Even though ROUGE is a popular metric to evaluate generated summaries, it shows a low correlation with human judgement when evaluating factual consistency for abstractive summaries [7, 11]. A possible explanation for the lower correlation can be the fact that when using ROUGE, each N-gram is weighted equally. In reality, all N-grams can represent a different semantic meaning. If this difference is ignored, the metric does not take the meaning of the text into account and cannot determine factual consistency effectively.

BERTScore: BERTScore [12] is another metric used in evaluating the quality of generated text, by leveraging the contextual embeddings from the BERT model. The metric computes the cosine similarity between the BERT contextual embeddings of the generated and reference texts, which captures both the semantic and syntactic similarity. It presents a more nuanced evaluation than metrics based on n-gram overlap alone, as it can understand semantic equivalences even with paraphrasing or word order changes.

BERTScore calculates precision (P), recall (R), and F1-score (F1) based on the contextual embeddings from the BERT model. The scores are computed as follows:

Precision: For each token in the candidate text, find the maximum cosine similarity with any token in the reference text. The precision is the average of these maximum similarities.

Recall: For each token in the reference text, find the maximum cosine similarity with any token in the candidate text. The recall is the average of these maximum similarities.

F1-score: The F1 score is the harmonic mean of precision and recall.

An advantage of using BERTScore over ROUGE is that it can compare texts in a more nuanced way using semantic meaning, compared to only looking at the textual data. However, we do not know what kind of information the BERT model stores in its embeddings. It can be that the embedding misses parts of the text. Because of this, it is not guaranteed to find all factual inconsistencies in all scenarios.

E. Fact extraction

In the context of fact extraction, a fact is a piece of information, usually represented as a relation between entities. For example, “Alice sees Bob on Tuesday”, is a fact where Alice and Bob are two entities, they share a directional relation “see” from Alice to Bob. “on Tuesday” is a time-related attribute of the fact, indicating when the relationship is relevant. Fact extraction is the process of taking unstructured text and converting it into structured representations.

Common fact extraction processes extract facts in the form of triples: a subject, a predicate, and an object. This type of extraction, often referred to as triple extraction, allows for finding all entities in a text that for which there is a relation between them.

Another specific way of extracting facts is using Semantic Role Labelling (SRL). The goal of SRL is to associate words or phrases of a sentence with their respective roles in a given

predicate or event[13]. Examples of semantic roles are *agent*, *patient*, *action*, *location*, *time*, *negation* and so on. Note that since SRL annotates works based on verbs in the sentence. This means multiple structured facts can be extracted from a single sentence, if the sentence contains more than one verb.

F. Graph-based text entailment

Recall the textual entailment problem explained in section II-B. In this section, we will give an overview of graph-based solutions to this problem. These solutions use a graph representation of text to verify whether a hypothesis (H) can be entailed by a text (T). There are two main ways to convert text into a graph-structure: a syntactic and a semantic approach. Syntactic graphs focus on the grammatical structure of sentences. Each node represents a word or a part of speech, while the edges represent the grammatical relationships between them (like subject, object, modifier, etc.). Semantic graphs focus on the meaning of sentences. Nodes represent concepts or entities, while edges represent semantic relationships between them (like is-a, part-of, cause-effect, etc.). Semantic graphs capture the meanings behind the words, their context, and how they interact to convey a complete idea.

Silva et al. [14] propose a composite approach for recognizing text entailment, which analyzes the entailment pair (a text and a hypothesis) to decide whether it must be resolved syntactically or semantically. For solving syntactic entailments, they use a tree edit distance algorithm over a dependency tree representation of T and H. For identifying semantic relationships, they employ an embedding-based navigation algorithm over graph knowledge bases composed of natural language dictionary definitions. By finding paths in these graphs linking T and H, they provide human-readable justifications that show explicitly what is the relationship holding between them.[14]

The advantage of the solution above is the interpretability of the results. Namely, when the method decides a hypothesis can be entailed by a text, a path in the graph is given which indicates the relations leading to the conclusion. Furthermore, this path is translated into a human-readable justification, which makes it accessible for a broader range of users. However, this solution is dependent on an (external) knowledge base. This decision on the knowledge base can affect the quality and the bias of the results of this method.

Levy et al. [15] start their method by observing that facts extracted as open-relation extracted triples are not organized or generalized, which can lead to redundancy or insufficient information when they are queried. To address this issue, they propose the use of entailment graphs. In these graphs, each node represents a proposition and each directed edge signifies an entailment relation, which is a logical connection or implication from one proposition to another. This structure helps to consolidate equivalent propositions and creates a hierarchy from specific to general information. To create this graph, they developed a probabilistic model that can predict entailment. This model is trained on a human made dataset of 30 entailment graphs containing 1.5 million entailment decisions.

The downside of this method is the need to either use or build a training dataset. When using existing sets, it is important to make sure the domain of the used dataset matches the domain of the problem.

III. METHOD

A. Metric overview

Our proposed factual consistency metric, **FactMetric**, works by extracting facts from both the document and the summary using semantic role labelling (SRL). Recall that using triple-extraction, a fact is represented using three attributes: (subject, relation, object). Using SRL, a fact is a more detailed representation of a fact, where attributes are also split into *time*, *location*, and more contextual pieces of information. Section III-B explains how we use SRL to extract facts in our metric. Then, for each fact from the summary (called a claim), the metric uses the list of facts from the document that to verify the claim. For this verification, a verification function is used that verifies each attribute of the claim using token-level equality. We propose two verification functions: one that operates on the list of facts from the document, and one that combines all document-facts into a graph, and tries to verify a claim using path finding in that graph. Section III-C explains the two verification methods in detail. Figure 2 gives a visual overview of the components in **FactMetric**.

Finally, two extensions to the metric are proposed. To handle unresolved pronouns in the summary (e.g., he, she, it), coreference resolution is used as a preprocessing step to replace those pronouns with their named entity. Moreover, since there exist many words with the same meaning, an addition to the verification function is proposed. This addition allows for synonyms’ to be verified in claims and facts. More on coreference resolution and this synonym-aware verification can be found in section III-D and III-E.

B. Extracting facts with Semantic Role Labeling

In our work, we choose to use SRL over triple extraction. Mainly, we observe that SRL provides a more fine-grained representation of facts, owing to its ability to capture and tag more specific roles within a sentence. This can lead to a more accurate representation of the information.

As for the implementation, AllenNLP[16] is used to perform the labelling. AllenNLP is an open-source natural language processing library, built on PyTorch, that provides interfaces for using complex models. The library includes a SRL model based on BERT [17]. The reason for choosing this implementation is because (1) the underlying model provides all the semantic labels we require, and (2) because it provides a high-level interface which improves the ease-of-use. A list of the specific semantic roles that are used in this research can be found in appendix A. As mentioned before, this implementation of SRL does not operate on a closed set of relations, and tries to find all relations between entities in a given text. This contrasts with the triple extraction method used in Goodrich et al. [5], where only a fixed list of relations could be extracted.

Figure 3 shows an example of a sentence annotated by the SRL model of AllenNLP

C. Claim verification

Now, to verify if a sentence from the summary is factually consistent with the document, we need to check if all the claims can be verified by the list of document facts. If the problem of verifying a single claim is solved, the whole sentence is verified as *consistent* if all its extracted claims are *consistent*. In this work, we propose two methods of performing this claim verification.

1) *Single-fact verification*: This method works by finding a document fact $f \in F$ that can back up the claim $c \in C$. The verification function can be described as:

$$verifies(f, c) = \exists_{f \in F} [\forall_{a \in attributes(c)} [f.a = c.a]]$$

Here, F is the list of document facts and C is the list of claims. The function $attributes(c)$ retrieves all attributes of the fact (e.g., *agent*, *action*, *patient*, etc.).

Note that both $f.a$ and $c.a$ contain textual values. To prevent falsely determining the two values are not equal, we perform basic preprocessing steps on both sides. We remove special characters, transform all letters to lower case, and normalize each verb to each stem.

2) *Verification using graph entailment*: Some claims contain information that is spread out over multiple facts in the source document, which poses a challenge to the single-fact verification method. To address this issue, we propose the second method, verification using graph entailment.

This method creates a fact graph from the document fact list, where each node is an entity (an agent or a patient), and each edge represents a verb (the relationship between the agent and the patient). Furthermore, we add all remaining attributes of the fact as additional information to the edge. For example, if the fact contains negation, time or location context, this is encoded in the edge. Essentially, this transforms the document into a graph of interconnected facts, making it easier to trace the relations between various facts in the document.

The claim is then verified by checking if there is a path in the graph going from the claim’s agent to its patient. We can formalize this verification method as follows:

Consider $G(V, E)$ is a fact graph of the documents, where V is the list of entities, and E the list of relations between the entities. Again, we consider $c \in C$ to be a claim from the list of claims C .

We can describe the verification method as the following function:

$$verifies_{graph-based}(G(V, E), c) = \{True, False\}$$

This function will return *True* if all points below hold, and *False* otherwise.

- 1) There exists a path P in $G(V, E)$ from $agent(c)$ to $patient(c)$.

¹<https://demo.allennlp.org/semantic-role-labeling>

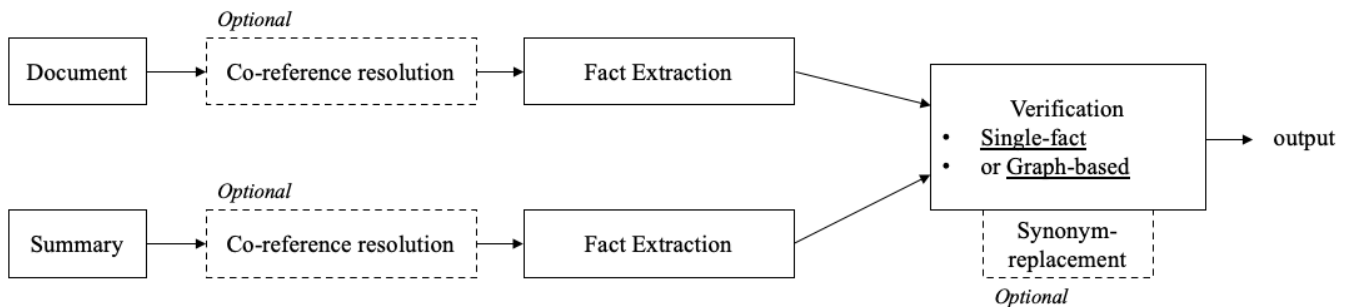


Fig. 2: Overview of the components in **FactMetric**



Fig. 3: SRL annotations by AllenNLP on the sentence “Yesterday, Alice baked a bread in her new bakery”. In this implementation, *ARG0* and *ARG1* are the *agent* and *theme*. *V* is the action, and *ARGM-TMP* and *ARGM-LOC* are the *time* and *location* attributes.¹

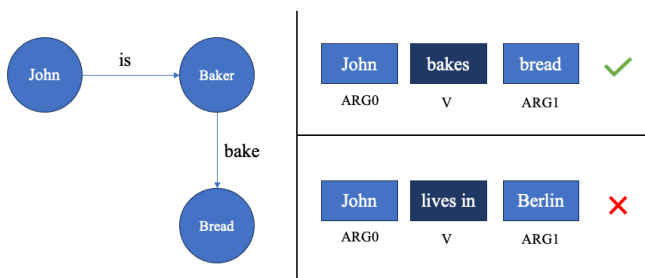


Fig. 4: An example of a fact graph (left), a verified claim (top right), and a claim that cannot be verified (bottom right)

- 2) For all edges $e \in P$, e must either be $relation(c)$, or e must be a form of the verb *to be*

Note that we allow an edge in path P to be a form of the verb *to be*. This decision is made to allow for the verification of claims using indirect or transitive relationships between entities in the document.

Figure 4 shows an example of a fact graph, a verified claim, and a claim that cannot be verified according to our proposed graph-based verification method.

D. Co-reference resolution

When dealing with pronouns in text, it can be hard for fact-based metrics to align entities which are referred by different phrases. In the case of “John is a baker. He is 24 years old”, it might be clear for humans to imply that John is 24 years old. However, when using extracted facts, we are dealing with two different entities: “John” and “He”.

To aid this alignment, we try an optional component to **FactMetric** that alters the text using a co-reference resolution

method, before passing it into the SRL model. Co-reference resolution is the task of determining when two or more expressions in a text (e.g., pronouns, nouns, noun phrases) refer to the same entity. Both the summary and the document are processed by this method (independently) before extracting the facts.

To accomplish co-reference resolution in this research, we leverage the NeuralCoref² system. Developed by Hugging Face, NeuralCoref uses a deep learning approach to the problem. This implementation is chosen because (1) it has an easy programming interface, where the interaction with the underlying neural network is abstracted away, and (2) because its results are comparable to other methods (where the implementation is available).

E. Synonym-aware verification

When language is written, it is possible to write two statements with the same meaning, by using synonyms. Since a simple text equality function cannot see the equal semantic meaning of synonyms, an optional method is proposed to allow different words sharing the same meaning to be counted as equal.

This idea is to expand words into a set of their (mostly used) synonyms. Afterward, instead of comparing only the words directly, check if there is a non-empty intersection in their synonym set. If so, classify the two words as equal. Otherwise, we classify the words as not equal. For example, the words “essential” and “crucial” will be turned into {essential, crucial, fundamental, imperative} and {crucial, bottom-line, essential, imperative}. This synonym expansion step is performed during the verification step, after the facts are extracted.

In our implementation, we retrieve the synonym-set of a word using the WordNet³ database. WordNet is a large lexical database of English words, where nouns, verbs, adjectives, and adverbs are organized into sets of cognitive synonyms. We interface the database using the NLTK⁴ Python library.

F. Evaluation of the metric

To evaluate the performance of a factual consistency metric like FactMetric, a dataset with a set of documents and

²<https://github.com/huggingface/neuralcoref>

³<https://wordnet.princeton.edu/>

⁴<https://www.nltk.org/>

corresponding summaries with ground-truth labels for factual consistency is required. Such a dataset serves as a basis for evaluating the performance of metric.

Each sample in this dataset should contain the following:

Document: This is the original text from which the summary has been made. It serves as the source of truth against which the facts mentioned in the summary will be verified.

Summary: This is a summarized version of the document, which is under review by the metric.

Ground-truth Factual Consistency labels: These labels provide information about whether a given summary (or individual sentences within it) is factually consistent with the document. The labels serve as the benchmark against which the output of FactMetric will be compared.

Creating such a dataset requires a manual effort by domain experts or annotators who can reliably assess the factual consistency of the summaries.

IV. EXPERIMENTAL RESULTS

We evaluate our metric on four human labelled datasets: FRANK, DeFacto, FactCC’s dataset, and our crowd annotated LLMs dataset. This section describes the datasets, implementation details, and empirical results. We also study the performance difference of the metric when including or excluding steps described in the previous section.

A. Evaluation datasets

FRANK[11]: This dataset contains annotations for summaries generated from documents in the CNN/DailyMail[18] and XSum[19] sets. These are two summarization datasets containing English news articles and reference summaries, which are widely used in the field of summarization. Each instance in the FRANK dataset contains an article, model-generated summary, and annotations.

The authors picked random articles from each article set. For each article, a summary is generated by five (CNN/DM) or four (XSum) generated summaries from various summarization models. For the articles coming from the CNN/DailyMail part, a LSTM seq2seq (s2s), Pointer-Generator-Network (pgn), Bottom-Up Summarization (bus), BERT-based abstractive-extractive model (bert_sum), and a jointly pretrained transformer based encoder-decoder model BART (bart), are used to generate the summaries. For the XSum part, a Topic-Aware CNN Model (TConvS2S), a Pointer-Generator Network model (PtGen), a randomly initialized transformer (Trans2S), and a transformer initialized with Bert-Base (BERTS2S), are used to create the summaries. More details about these models are provided in appendix C.

Three crowd-workers annotated each sentence of the summary on a per-sentence level. A sentence can be annotated *consistent* or *inconsistent*. In the case of an inconsistent sentence, the annotator had to choose one out of seven possible error types. This typology of errors consists of a relationship error (RelE), entity error (EntE), circumstance error (CircE), co-reference error (CorefE), discourse link error (LinkE), out-of-article error (OutE) and the grammatical error (GramE). More information on the kinds of errors is written in appendix

B. For our experiment, a sentence is considered factually consistent if the majority of the annotators choose no error (NoE) for the sentence.

For the CNN/DailyMail part, frank contains 175 articles, 875 summaries and 2439 labelled sentences. For the XSum part, there are 170 articles, 683 summaries and 690 labelled sentences.

Statistics on the number of annotated errors on both the CNN/DailyMail and XSum part of FRANK can respectively be found in table I and II

DeFacto [20]: This dataset, like FRANK, consists of source documents, their summaries, and human annotations. In this dataset, 1025 news articles are randomly sampled from the XSum dataset. The generated summaries are outputs from Google’s transformer-based model PEGASUS[21]. The annotations are collected by one annotator per summary. In contrast to FRANK, annotations are made at summary-level. When a summary is annotated as incorrect, the annotator is asked to indicate whether the error is intrinsic (misinterpreting information from the source document) or extrinsic (hallucinate new information or facts not present in the source article)[22]. DeFacto consists of 1075 annotated document/summary pairs, of which 779 are marked as inconsistent.

FactCC evaluation dataset: The model-based metric FactCC by [7] also comes with a manually annotated dataset to evaluate factual consistency. Their published dataset is also used in our work to evaluate the performance of our metric. The set contains 150 summaries generated from the CNN/DailyMail dataset. The summaries come from 9 abstractive summarization models, listed in appendix C. The authors of FactCC annotated the dataset themselves. Contrary to FRANK, they used a binary factuality label for each sentence. In total, it contains 439 annotated sentences, of which 384 are labelled as consistent, and 55 as inconsistent.

Large language models dataset: In addition to the existing factual consistency datasets, a new dataset is created, consisting of summaries generated by three large language models (LLM). This dataset is created (1) to gain insights into the types of errors LLMs make and (2) to see how well factuality metrics generalize to these kinds of summaries. The articles in this set are coming from the CNN/DailyMail dataset. Summaries are made by GPT-4 [3], Falcon⁵, and StableLM⁶. We chose GPT-4 since it is the biggest accessible language models at the time of writing. Falcon and StableLM are chosen because they are both open-source, which makes them available for everyone. Between the two, Falcon is the bigger model, which requires 90GB of GPU RAM for inference. StableLM only requires 16GB, which makes it possible to be used by an average desktop computer. Example summaries and the prompt used to generate them can be found in appendix C.

For the human annotation part, we tried to replicate the annotation process of Pagnoni et al. [11]. We used Amazon Mechanical Turk⁷, a crowd data annotation tool, to let three

⁵<https://falconllm.tii.ae/>

⁶<https://github.com/Stability-AI/StableLM>

⁷<https://www.mturk.com/>

	CircE	CorefE	EntE	GramE	LinkE	OtherE	OutE	RelE	NoE	Total	%-correct
bart	4	5	4	2	5	0	4	5	478	507	94.3
bert_sum	7	16	16	2	4	0	4	5	472	526	89.7
bus	23	24	35	18	23	3	6	17	334	483	69.2
pgn	7	14	15	5	5	0	2	8	404	460	87.8
s2s	33	36	80	24	8	1	15	35	231	463	49.9
Total	74	95	150	51	45	4	31	70	1919	2439	78.7

TABLE I: Number of sentences with errors per model in the CNN/DailyMail part of the FRANK dataset

	CircE	CorefE	EntE	GramE	LinkE	OutE	RelE	NoE	Total	%-correct
BERTS2S	34	3	29	11	3	53	18	20	171	11.7
PtGen	14	6	38	23	6	53	31	5	176	2.8
TConvS2S	27	3	30	5	1	76	21	8	171	4.7
TranS2S	28	3	33	8	6	62	22	10	172	5.8
Total	103	15	130	47	16	244	92	43	690	6.2

TABLE II: Number of sentences with errors per model in the XSum part of the FRANK dataset

annotators label our summaries on a per-sentence level using FRANK’s annotation platform ⁸. In short, the tool works by letting the annotator read a document, and presenting them with various summaries afterward (shown in random order). Per summary, the annotator has to indicate the factual consistency for each sentence in the summary. When the annotator decides the sentence is *inconsistent* with the document, he/she is requested to choose the type of inconsistency.

An overview of the number of sentences with errors per model can be found in table III. Here, summaries by Falcon and GPT-4 show a similar percentage of factual consistency.

Additional to the three crowd-workers, we also annotated the summaries of three models ourselves. To do this, the author of this work annotated 525 (three times 175) summaries himself using the same tool, the crowd workers had to use. The reason we also annotated the data ourselves is because of the low agreement between the crowd workers in our LLM annotation experiment, explained further in section IV-A0d. We choose to annotate the models: GPT-4, StableLM and BART. We chose GPT-4 since it had the highest amount of consistency annotated by the crowd workers. StableLM was chosen because it had the lowest amount of consistency annotated by the crowd workers. At last, we chose to annotate the summaries of BART ourselves because we want to be able to compare the factual consistency of large language models to a non-LLM.

The number of sentences with errors per language model in this *author-annotated* dataset can be found in table IV. We see that, also in our annotations, a high percentage of summaries generated by BART are labelled as consistent. Summaries by GPT-4 show a higher amount of consistency compared to the crowd annotations. At last, StableLM-summaries are labelled less consistent compared to the crowd annotations.

Inter-annotator agreement: Since the FRANK and LLMs datasets are both annotated by more than one annotator (per summary), we have to deal with the possibility of disagreement between annotators on a single sentence. The inter-annotator agreement can be used to quantify the agreement over a dataset. It serves as a measure of consensus among multiple annotators and reflects the consistency of the an-

notation process. A high inter-annotator agreement indicates a low degree of ambiguity in the labels, thereby enhancing the dataset’s reliability for evaluating models. On the other hand, low agreement can highlight potential problems, such as ambiguous instructions or subjective interpretation, which can lead to unreliable results when using the dataset to evaluate a model.

In the case of the FRANK and LLMs dataset, we choose to report three variants of the inter-annotator agreement.

a) Agreement on majority label: First, we compute the agreement on the majority label, A_{label} . This is the percentage of annotations that agree with the majority chosen label. In this computation, the exact chosen error (or no-error) label has to match the majority, to count as an agreement.

Assuming we have n sentences, and $m = 3$ annotators per sentence. We define L_{ij} as the label given to sentence i by annotator j . Furthermore, we define M_i as the majority voted label for sentence i .

Define $f(A_{ij}, M_i)$ to be an indicator function that is 1 when the label from annotator j for sentence i is equal to the majority label for sentence i and 0 otherwise:

$$f(L_{ij}, M_i) = \begin{cases} 1 & \text{if } L_{ij} = M_i \\ 0 & \text{otherwise} \end{cases}$$

Then, the agreement A_{label} can be calculated as the sum of the function $f(A_{ij}, M_i)$ across all sentences and annotators, divided by the product of the number of sentences n and the number of annotators m :

$$A_{label} = \frac{1}{mn} \sum_{i=1}^n \sum_{j=1}^m f(L_{ij}, M_i)$$

b) Agreement considering only error and no-error labels: Second, the agreement on consistency is computed, while transforming all labels to either *no error* (when the label is *NoE*), and *error* (for all other labels). We call this agreement score A_{binary} , and compute is using a slightly changed formula

$$A_{binary} = \frac{1}{mn} \sum_{i=1}^n \sum_{j=1}^m f_{binary}(L_{ij}, M_i)$$

⁸<https://github.com/artidoro/frank-annotation-platform>

	CircE	CorefE	EntE	GramE	LinkE	OutE	PredE	NoE	Total	%-correct
falcon	40	14	20	32	26	103	58	503	796	63.2
gpt4	31	8	24	25	25	82	50	428	673	63.6
stablelm	34	6	15	29	23	208	35	155	505	30.7
Total	105	28	59	86	74	393	143	1086	1974	55.0

TABLE III: Number of sentences with errors per model in the *crowd-annotated* large language models dataset

	CircE	CorefE	EntE	GramE	OutE	RelE	NoE	Total	%-correct
bart	2	0	7	4	1	0	564	578	97.6
gpt4	1	1	2	0	0	0	669	673	99.4
stablelm	6	7	50	20	320	17	85	505	16.8
Total	9	8	59	24	321	17	1318	1756	75.1

TABLE IV: Number of sentences with errors per model in the *author-annotated* dataset

Where we define f_{binary} as:

$$f_{binary}(L_{ij}, M_i) = \begin{cases} 1 & \text{if } b(L_{ij}) = b(M_i) \\ 0 & \text{otherwise} \end{cases}$$

and $b(l)$ as:

$$b(l) = \begin{cases} 1 & \text{if } l = NoE \\ 0 & \text{otherwise} \end{cases}$$

c) *Agreement on type of error, for subset where majority votes any error:* For the last agreement score, we consider only the sentences where the majority labels of a sentence are indicated an error. For this subset, we compute the percentage of annotators that agree on the *type* of error.

This score $A_{label-error}$ is computed the same way as A_{label} , with the difference that only the sentences (indexed by $i \in [0, \dots, n]$) are included for which $M_i \neq NoE$.

d) *Results:* All inter-annotator agreements can be found in table V, and VI. Overall, FRANK has a higher agreement score than the LLMs dataset. We discuss these differences in section V-A.

Given that the agreement on the crowd annotated LLMs dataset is significant lower than the original FRANK annotated dataset, we decided to annotate the LLMs dataset ourselves.

B. Evaluation Method

In this experiment, we regard the evaluation of the factual consistency metric as a binary classification task, where the ground truth is the binary labels assigned to each sentence or summary (based on the granularity of the dataset) by human annotators. These labels signify whether a sentence in the summary is factually consistent with the source document (true) or not (false). The metric we have developed, also generates predictions in a binary format, consistent with this binary ground truth.

For assessing the quality of our binary classification, we utilize four key metrics: Precision, Recall, F1-score, and weighted Accuracy. In this experiment, the class 'inconsistent' is considered the primary class because we want to identify sentences or summaries where factual consistency deviates.

Precision and Recall are measures in classification problems, and they serve distinctive roles. Precision quantifies the proportion of true positive instances out of all instances predicted as positive (in this case, 'inconsistent'). It answers the

question: among all the sentences/summaries that the metric labels as 'inconsistent', how many are truly inconsistent? A high precision indicates that the metric rarely falsely identifies a sentence/summary as inconsistent, providing reliable alerts when inconsistency is predicted.

On the other hand, Recall measures the proportion of true positive instances that are correctly identified as such. In our context, it answers: out of all the sentences/summaries that are truly inconsistent, how many does the metric correctly identify? A high recall suggests the metric is adept at catching inconsistencies present in the summaries.

Taken together, precision and recall provide an informative view of the classifier's performance. A high-precision but low recall indicates that the metric is conservative in its predictions and while it seldom misclassifies, it may miss many true inconsistencies. In contrast, a high recall but low precision suggests the metric is liberal in identifying inconsistencies but may raise many false alarms.

To combine these two metrics into a singular one that balances both, we calculate the F1-score. This harmonic mean of precision and recall presents an overall performance measure when both are considered equally important. The weighed accuracy, the proportion of correct classifications over all instances, corrected by the imbalance of the labels, will also be computed.

C. Metric variations

In this section, we present the different variations of the proposed metric and compare them to two baselines: FactCC and another information-extraction based metric we call *TripleMetric*, evaluated on the discussed datasets.

To evaluate FactCC, we used their implementation on GitHub⁹ and ran their evaluation on our selection of datasets to get the results.

To compare the proposed metric to another information-extraction based metric, we tried to evaluate the metric of Goodrich et al. [5] on the datasets. Since they did not provide an implementation of their metric, we made an approximated replication of their metric using OpenIE triple extraction. In this approach, facts are extracted from sentences as (*subject, relation, object*) triples from both the summary and the document. Then, we label a sentence (of the summary)

⁹<https://github.com/salesforce/factCC>

Model	Agreement of majority label (A_{label})	Agreement on correctness (A_{binary})	Agreement on label if incorrect ($A_{label-error}$)
BERTS2S	67.1%	90.3%	67.3%
PtGen	70.8%	95.5%	72.2%
TConvS2S	74.5%	95.1%	75.2%
TransS2S	69.4%	92.2%	71.1%
bart	89.7%	90.8%	56.3%
bert_sum	89.3%	90.6%	65.7%
bus	81.6%	88.3%	63.8%
pgn	90.1%	91.9%	66.7%
s2s	76.2%	87.8%	66.5%
Average	82.1%	90.7%	68.7%

TABLE V: Inter-annotator agreement for the FRANK dataset

	Model	Agreement of majority label (A_{label})	Agreement on correctness (A_{binary})	Agreement on label if incorrect ($A_{label-error}$)
0	falcon	61.5%	76.9%	41.7%
1	gpt4	61.3%	77.3%	40.8%
2	stablelm	52.4%	77.3%	45.4%
3	Average	59.1%	77.1%	42.8%

TABLE VI: Inter-annotator agreement for the LLMs dataset

as consistent if all its extracted facts can be found in the list of facts from the document.

In the results, we reference to this metric as *TripleMetric*.

For the proposed metric, we present several variations. The basic variant, which we denote as *FactMetric*, consists of performing semantic role labelling, and comparing the facts using *Single-fact verification* as explained in section III-C. The intent here is to provide a base reference of our metric’s performance before applying additional enhancements. To account for the nuances in language use, we add co-reference resolution to *FactMetric*, which is called *FactMetric_{coref}*, in this work. Another modification of our basic metric, *FactMetric_{synonym}*, utilizes synonym augmentation. It expands the scope of the comparison by treating synonymous words as equivalent during the comparison process. As a result, *FactMetric_{coref+synonym}* represents the combination of co-reference resolution and synonym augmentation.

Furthermore, the graph entailment variant of the metric, as described in section III-C2 is also evaluated. In this section, we refer to the basic variant of this metric as *FactMetric_{graph}*. Subsequently, *FactMetric_{graph,coref}*, *FactMetric_{graph,synonym}*, and *FactMetric_{graph,coref+synonym}* are variants of the graph comparison methods where we preprocess text first using co-reference resolution, and use synonym augmentation during the comparison step.

D. Results

The main evaluation results of all variants of the proposed metric, the related metrics FactCC and *TripleMetric* can be found in table VII and VIII.

Our metric’s performance across the FRANK-XSum and DeFacto datasets were particularly high. These findings will be discussed further in section V-B of the discussion. Because of this, we will ignore these results when making further observations in this section. The observations in the other datasets are mentioned below and discussed in section V-C.

Overall, we observe that the base variant, *FactMetric*, yielded better accuracy scores than *TripleMetric* among

all datasets. The recall is (almost) the same in the crowd- and author-annotated dataset, and is lower on the FRANK CNN/DM and FactCC datasets.

Notably, it can be seen that the integration of co-reference resolution into our metric, did not lead to an enhancement in performance as anticipated. In all datasets, using the *single-fact* and *graph* comparison, it decreased the accuracy.

The inclusion of the synonym step did not seem to affect the results. On all datasets, comparing *FactMetric* to *FactMetric_{synonym}* and *FactMetric_{Graph}* to *FactMetric_{Graph,synonym}*, there are no significant differences in results.

Comparing *FactMetric* to *FactMetric_{Graph}*, we see that the *graph* comparison performs worse on the FRANK CNN/DM dataset. On the FactCC dataset, the precision, f1 and accuracy are also worse, while the recall is higher for *FactMetric_{Graph}*. On the crowd- and author-annotated datasets, *FactMetric_{Graph}* scores (almost) equal or lower than *FactMetric*.

Table IX shows the performance of the variants of *FactMetric* on the author-annotated dataset per model. We observe that the precision scores for all variants are low on BART and GPT-4. Between the three models, the metric seems to perform best on StableLM.

V. DISCUSSION

A. Low inter-annotator agreement in LLMs dataset

When looking deeper into the lower inter-annotator agreement in our crowd annotated dataset compared to the FRANK dataset, we see several factors that can explain this difference. Despite sharing most parts of the annotation process such as the source documents, annotation platform, and reward system, there is a difference in summary generation techniques. In our dataset, Large Language Models were used, compared to the non-LLMs used by Pagnoni et al. [11].

After comparing the summaries generated by the models of FRANK with the LLMs (for 20 randomly selected articles, to be found in appendix D) manually, we found that the LLMs generate more abstractive summaries. The LLMs rephrase

Dataset	FRANK								DeFacto				FactCC			
	CNN/DM				XSum				P	R	F1	A	P	R	F1	A
Metric	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1	A
<i>FactCC</i>	0.46	0.52	0.49	0.77	0.95	0.80	0.87	0.77	0.75	0.75	0.76	0.64	0.94	0.90	0.92	0.86
<i>TripleMetric</i>	0.22	1.00	0.35	0.08	0.94	1.00	0.97	0.91	0.74	1.00	0.85	0.63	0.13	1.00	0.22	0.03
<i>FactMetric</i>	0.36	0.78	0.49	0.67	0.95	0.99	0.97	0.92	0.73	0.98	0.84	0.62	0.31	0.83	0.45	0.79
<i>FactMetricCoref</i>	0.22	1.00	0.36	0.08	0.95	1.00	0.97	0.92	0.73	1.00	0.85	0.62	0.12	1.00	0.22	0.03
<i>FactMetricSynonym</i>	0.36	0.77	0.49	0.67	0.95	0.99	0.97	0.92	0.73	0.98	0.84	0.62	0.31	0.83	0.45	0.79
<i>FactMetricCoref+Synonym</i>	0.22	1.00	0.36	0.08	0.95	1.00	0.97	0.92	0.73	0.98	0.85	0.62	0.12	1.00	0.22	0.03
<i>FactMetricGraph</i>	0.29	0.90	0.44	0.50	0.95	1.00	0.97	0.92	0.73	1.00	0.85	0.62	0.22	0.94	0.35	0.64
<i>FactMetricGraph,Coref</i>	0.22	1.00	0.36	0.08	0.95	1.00	0.97	0.92	0.73	1.00	0.85	0.62	0.12	1.00	0.22	0.03
<i>FactMetricGraph,Synonym</i>	0.29	0.90	0.44	0.50	0.95	1.00	0.97	0.92	0.73	1.00	0.85	0.62	0.22	0.94	0.35	0.64
<i>FactMetricGraph,Coref+Synonym</i>	0.22	1.00	0.36	0.08	0.95	1.00	0.97	0.92	0.73	1.00	0.85	0.62	0.12	1.00	0.22	0.03

TABLE VII: Performance of metric variants on datasets FRANK, DeFacto, and FactCC

Dataset	Crowd-annotated (Falcon, GPT-4, StableLM)				Author-annotated (BART, GPT-4, StableLM)			
	P	R	F1	A	P	R	F1	A
Metric	P	R	F1	A	P	R	F1	A
<i>FactCC</i>	0.54	0.51	0.51	0.52	0.86	0.66	0.75	0.69
<i>TripleMetric</i>	0.52	1.00	0.69	0.36	0.24	1.00	0.39	0.09
<i>FactMetric</i>	0.53	0.97	0.68	0.40	0.30	0.99	0.47	0.40
<i>FactMetricCoref</i>	0.52	1.00	0.68	0.36	0.25	1.00	0.40	0.10
<i>FactMetricSynonym</i>	0.53	0.97	0.68	0.40	0.31	0.99	0.47	0.41
<i>FactMetricCoref+Synonym</i>	0.52	1.00	0.68	0.36	0.25	1.00	0.40	0.10
<i>FactMetricGraph</i>	0.52	0.99	0.69	0.37	0.28	0.99	0.43	0.27
<i>FactMetricGraph,Coref</i>	0.52	1.00	0.68	0.36	0.25	1.00	0.40	0.10
<i>FactMetricGraph,Synonym</i>	0.52	0.99	0.69	0.37	0.28	0.99	0.43	0.27
<i>FactMetricGraph,Coref+Synonym</i>	0.52	1.00	0.68	0.36	0.25	1.00	0.40	0.10

TABLE VIII: Performance of metric variants on the crowd-annotated and author-annotated datasets

Model	BART				GPT-4				StableLM			
	P	R	F1	A	P	R	F1	A	P	R	F1	A
Metric	P	R	F1	A	P	R	F1	A	P	R	F1	A
<i>FactMetric</i>	0.03	0.70	0.07	0.69	0.01	1.00	0.01	0.07	0.83	0.99	0.91	0.75
<i>FactMetricCoref</i>	0.02	1.00	0.04	0.00	0.01	1.00	0.01	0.00	0.83	1.00	0.91	0.76
<i>FactMetricSynonym</i>	0.04	0.70	0.07	0.69	0.01	1.00	0.01	0.07	0.83	0.99	0.91	0.75
<i>FactMetricCoref+Synonym</i>	0.02	1.00	0.04	0.00	0.01	1.00	0.01	0.00	0.83	1.00	0.91	0.76
<i>FactMetricGraph</i>	0.03	0.80	0.05	0.44	0.01	1.00	0.01	0.02	0.83	1.00	0.91	0.76
<i>FactMetricGraph,Coref</i>	0.02	1.00	0.04	0.00	0.01	1.00	0.01	0.00	0.83	1.00	0.91	0.76
<i>FactMetricGraph,Synonym</i>	0.03	0.80	0.05	0.44	0.01	1.00	0.01	0.02	0.83	1.00	0.91	0.76
<i>FactMetricGraph,Coref+Synonym</i>	0.02	1.00	0.04	0.00	0.01	1.00	0.01	0.00	0.83	1.00	0.91	0.76

TABLE IX: Performance of FactMetric variants on the author-annotated dataset, grouped by the model that generated the summaries

sentences, combine information, and use more synonyms of words appearing in the source document than the models used in FRANK. This can explain the lower agreement score, since the higher attractiveness can cause confusion about the factual consistency of a sentence. It can be harder for an annotator to determine whether a statement can be exactly verified by a document that uses slightly different language.

Additionally, the annotation process by Pagnoni et al. [11] contains undisclosed variables that could have influenced their higher agreement scores. Such as the amount of bonus rewards or the quantity of continuous evaluation questions an annotator had to get correct to be included in the pool. If these parameters were not entirely matched with ours, it could explain some differences too.

B. High metric performance on the XSum (part of FRANK) and DeFacto dataset

The high performance of all variants of *FactMetric* on the XSum part of FRANK, and DeFacto datasets is worth further investigation. After manually inspecting the generated

summaries of 20 random selected documents (in the two datasets), we see that many are directly copied sentences from the source document. These random selected documents can be found in appendix E.

In the case of these fully extractive summaries, it is trivial for *FactMetric* and *TripleMetric* to correctly classify their factual consistency. Since its sentences are exact copies of sentences in the source document. Because of this, it is hard to draw conclusions on the performance of the metric using this dataset.

C. Performance of FactMetric variants

First, we discuss the higher performance of *FactMetric* compared to *TripleMetric*. The difference between the two metrics is the use of semantic role labelling instead of triple extraction. Since the verification step of both metrics is the same (namely, verifying that each attribute of the fact is equal), we state that the way this SRL model structures sentences into these semantic attributes makes the evaluation of factual consistency easier.

Co-reference resolution: Furthermore, adding co-reference resolution to *FactMetric* decreases its performance on most datasets. One possible explanation is the fact that the co-reference resolution model can make mistakes in resolving references. When a reference is replaced by a wrong entity, the SRL model uses this replaced entity in its extracted structure. This error propagates to the verification method, which can make a mistake because of this. We also need to look at the human annotated labels. In all datasets, these labels are assigned to the unresolved (sentences of) the summaries. After a summary is processed by the co-reference resolution model, we did not re-annotate them. Since the annotations are now the untreated texts, we cannot fully rely on those labels anymore.

Synonym replacement: We expected synonym replacement to increase the performance on abstractive summaries. However, this softer verification step did not increase the performance of *FactMetric*. To find out why this optional step did not increase the performance of our metric, we found that out of all 25,595 unique words in all datasets, 16,400 words had at least one registered synonym in the WordNet dataset. From this, we can state that the WordNet dataset covers the majority of the words in (the domain of) our datasets. Continuing from this, we found that for the FRANK-CNN dataset, out of the 1919 human labelled *consistent* sentences, only 19 sentences are labelled *consistent* by *FactMetric_{synonym}* while *FactMetric* labels them *inconsistent*. For the *author-annotated* dataset, this condition only holds for 2 out of the 1318 sentences. This number tells us that there is a low number of sentences for which synonym replacement made sure the sentence was correctly classified as *consistent*. Given the fact that WordNet covers the majority of the words in our datasets, and the low number of sentences where *FactMetric_{synonym}* correctly predicts *consistent* compared to *FactMetric*, we can state that in these datasets, the summaries do not follow the expected pattern of using synonyms in their sentences.

Graph-based verification: At last, we discuss the lower and equal performance of *FactMetric_{Graph}* compared to *FactMetric*. We implemented this verification function to verify information about a claim that is spread out over multiple sentences in the source document. Our implementation of this graph entailment can explain this lack of improvement over *FactMetric* on the abstractive summaries. Recall that we decided to determine factual consistency of a claim in a graph, by finding a path in the graph where edges are either the verb of the claim, or a form of the verb *to be*. An advantage of this decision is that we only verify claims for which the entity is definitely performing the claimed action. However, as we observed afterward, language can be more abstractive than this. We discuss other methods of graph entailment in section II-F

Per-model performance analysis in the author-annotated dataset: We can find an explanation for the low performance of all variants of *FactMetric* on BART and GPT-4 in the author-annotated dataset in the balance of the classes. Table IV shows only 14 inconsistent sentences in the summaries generated by BART, and only 4 for GPT-4. This very low number of inconsistent sentences makes it hard to say something about the metric’s ability to predict inconsistent sentences.

D. Contradiction in the document

Another limitation of *FactMetric* is determining factual consistent in terms of contradiction. Assume two sentences in the source document contradict each other (e.g., “John is a baker”, and “John is a taxi driver”) and a claim has to be verified concerning this contradiction (e.g., “John is a baker”). Our metric (in all variants) will mark the claim as *consistent*, since there is a fact in the document which verifies it. However, this may not be the ideal outcome in some scenarios.

VI. CONCLUSION

In this work, we proposed a metric where the comparison of facts and claims is interpretable. Within the set of information-extraction based metrics, we showed that semantic role labelling can show improved results over triple extraction. Furthermore, we showed that adding (our chosen implementation of) co-reference resolution did not improve, and sometimes even decreased, the results. Additionally, using WordNet as a dataset for synonym replacement did not improve the results.

Moreover, we studied the factual consistency of summaries generated by large language models. Here, we found that using a crowd platform for this data annotation task can lead to a low agreement between annotators.

To continue on this work, we list the following potential improvements to investigate. At first, there might be potential in choosing a more sophisticated graph entailment method. There could be possibilities to increase the trustworthiness of the labels in crowd annotated datasets, one could look into increasing the number of annotators, or by changing the way an annotator is instructed and evaluated during the annotation process. At last, for some use cases, it might be needed to handle the possibility of contradiction in the source document.

REFERENCES

- [1] Z. Cao, F. Wei, W. Li, and S. Li, “Faithful to the Original: Fact Aware Neural Abstractive Summarization,” 11 2017.
- [2] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” pp. 74–81, 2004. [Online]. Available: <https://aclanthology.org/W04-1013>
- [3] OpenAI, “GPT-4 Technical Report,” 3 2023. [Online]. Available: <https://arxiv.org/abs/2303.08774v3>
- [4] Y. Huang, X. Feng, X. Feng, and B. Qin, “The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey.”
- [5] B. Goodrich, V. Rao, M. Saleh, and P. J. Liu, “Assessing The Factual Accuracy of Generated Text,” 5 2019.
- [6] T. Falke, L. F. R. Ribeiro, P. Ajie Utama, I. Dagan, and I. Gurevych, “Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference,” pp. 2214–2220. [Online]. Available: <https://tudatalib.ulb>
- [7] W. Kryściński, B. McCann, C. Xiong, and R. Socher, “Evaluating the Factual Consistency of Abstractive Text Summarization,” 10 2019.
- [8] L. F. R. Ribeiro, M. Liu, I. Gurevych, M. Dreyer, and M. Bansal, “FactGraph: Evaluating Factuality in

- Summarization with Semantic Graph Representations,” 4 2022.
- [9] A. Wang, K. Cho, and M. Lewis, “Asking and Answering Questions to Evaluate the Factual Consistency of Summaries,” 4 2020. [Online]. Available: <http://arxiv.org/abs/2004.04228>
- [10] E. Durmus, H. He, and M. Diab, “FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 5055–5070, 5 2020. [Online]. Available: <http://arxiv.org/abs/2005.03754><http://dx.doi.org/10.18653/v1/2020.acl-main.454>
- [11] A. Pagnoni, V. Balachandran, and Y. Tsvetkov, “Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics,” *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pp. 4812–4829, 4 2021. [Online]. Available: <https://arxiv.org/abs/2104.13346>
- [12] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” 4 2019.
- [13] D. Gildea and D. Jurafsky, “Automatic Labeling of Semantic Roles.”
- [14] V. S. Silva, A. Freitas, and S. Handschuh, “Exploring knowledge graphs in an interpretable composite approach for text entailment,” in *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 2019.
- [15] O. Levy, I. Dagan, and J. Goldberger, “Focused Entailment Graphs for Open IE Propositions,” pp. 87–97.
- [16] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer, “AllenNLP: A Deep Semantic Natural Language Processing Platform,” pp. 1–6, 6 2019. [Online]. Available: <https://github.com/allenai/allennlp>
- [17] P. Shi, J. Lin, and D. R. Cheriton, “Simple BERT Models for Relation Extraction and Semantic Role Labeling,” 2019.
- [18] R. Nallapati, B. Zhou, C. dos Santos, and B. Xiang, “Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond,” 2016.
- [19] S. Narayan, S. B. Cohen, and M. Lapata, “Don’t Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization,” *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 1797–1807, 8 2018. [Online]. Available: <https://arxiv.org/abs/1808.08745>
- [20] Y. Liu, B. Deb, M. Teruel, A. Halfaker, D. Radev, and A. H. Awadallah, “On Improving Summarization Factual Consistency from Natural Language Feedback,” 12 2022.
- [21] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” 12 2019.
- [22] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On Faithfulness and Factuality in Abstractive Summarization,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 1906–1919.
- [23] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020. [Online]. Available: <https://aclanthology.org/2020.acl-main.703>
- [24] Y. Liu and M. Lapata, “Text Summarization with Pretrained Encoders,” *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 3730–3740, 2019. [Online]. Available: <https://aclanthology.org/D19-1387>
- [25] S. Gehrmann, Y. Deng, and A. M. Rush, “Bottom-Up Abstractive Summarization,” *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 4098–4109, 2018. [Online]. Available: <https://aclanthology.org/D18-1443>
- [26] A. See, P. J. Liu, and C. D. Manning, “Get To The Point: Summarization with Pointer-Generator Networks,” *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, pp. 1073–1083, 2017. [Online]. Available: <https://aclanthology.org/P17-1099>
- [27] A. M. Rush, S. Chopra, and J. Weston, “A Neural Attention Model for Abstractive Sentence Summarization,” *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389, 2015. [Online]. Available: <https://aclanthology.org/D15-1044>
- [28] S. Rothe, S. Narayan, and A. Severyn, “Leveraging pre-trained checkpoints for sequence generation tasks,” *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 264–280, 1 2020. [Online]. Available: https://dx.doi.org/10.1162/tacl_a_00313
- [29] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of the 2019 Conference of the North*, pp. 4171–4186, 2019. [Online]. Available: <https://aclanthology.org/N19-1423>
- [30] A. Vaswani, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, “Attention is All you Need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [31] W. T. Hsu, C. K. Lin, M. Y. Lee, K. Min, J. Tang, and M. Sun, “A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss,” *ACL 2018 - 56th Annual Meeting of the Association*

- for *Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, pp. 132–141, 5 2018. [Online]. Available: <https://arxiv.org/abs/1805.06266v2>
- [32] Y. Jiang and M. Bansal, “Closed-Book Training to Improve Summarization Encoder Memory,” *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 4067–4077, 9 2018. [Online]. Available: <https://arxiv.org/abs/1809.04585v1>
- [33] Y. C. Chen and M. Bansal, “Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting,” *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, pp. 675–686, 2018. [Online]. Available: <https://aclanthology.org/P18-1063>
- [34] W. Kryściński, R. Paulus, C. Xiong, and R. Socher, “Improving Abstraction in Text Summarization,” *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 1808–1817, 2018. [Online]. Available: <https://aclanthology.org/D18-1207>
- [35] W. Li, X. Xiao, Y. Lyu, and Y. Wang, “Improving Neural Abstractive Document Summarization with Structural Regularization,” *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 4078–4087, 2018. [Online]. Available: <https://aclanthology.org/D18-1441>
- [36] R. Pasunuru and M. Bansal, “Multi-reward reinforced summarization with saliency and entailment,” *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 2, pp. 646–653, 4 2018. [Online]. Available: <https://arxiv.org/abs/1804.06451v2>
- [37] H. Guo, R. Pasunuru, and M. Bansal, “Soft layer-specific multi-task summarization with entailment and question generation,” *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, pp. 687–697, 5 2018. [Online]. Available: <https://arxiv.org/abs/1805.11004v1>
- [38] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay, “The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only,” 6 2023. [Online]. Available: <https://arxiv.org/abs/2306.01116v1>
- [39] N. S. Google, “Fast Transformer Decoding: One Write-Head is All You Need,” 11 2019. [Online]. Available: <https://arxiv.org/abs/1911.02150v1>

APPENDIX A SEMANTIC ROLE LABELS

Table X provides a list of attributes and descriptions that the semantic role labelling model can label in its prediction.

Argument	Description
ARG0	Subject/Agent
ARG1	Patient
ARG2	Benefactive/instrument/attribute
ARG3	Ending point
ARGM-TMP	Time
ARGM-LOC	Location
ARGM-PRD	Purpose
ARGM-PRP	Predication
ARGM-MNR	Manner
ARGM-ADV	Adverbial
ARGM-CAU	Cause
ARGM-DIS	Discourse
ARGM-NEG	Negation
ARGM-MOD	Modal

TABLE X: Listing of attributes the semantic role labelling model can output

APPENDIX B FRANK ERROR ANNOTATIONS

Pagnoni et al. [11] define a set of labels to annotate the summary sentences regarding their factual consistency. Table XI provides an overview of all labels and a description. Aside from the errors, the label **NoE** (no error) is used to indicate that a sentence is factually consistent.

APPENDIX C SUMMARIZATION MODEL DETAILS

In this section, we provide brief explanations and sample summaries for the models used in the datasets. For readability, we put all source documents at the end of this section of the appendix.

A. FRANK CNN/DM

The following summaries were generated from document 1.

BART [23]: A denoising autoencoder for pretraining sequence-to-sequence models.

Summary: manchester city players trained at their etihad campus base on wednesday . sergio aguero was back to his menacing best as he took part in shooting practice . yaya toure and samir nasri are reportedly two players city are willing to listen to offers for in the summer .

BERTSum [24]: A BERT-based summarization model using a document-level encoder which can express the semantics of a document and obtain representations for its sentences.

Summary: manchester city were trained on wednesday ahead of their match on sunday at home to west ham united . sergio aguero fine-tuned his finishing skills with some shooting practice . yaya toure and samir nasri are reportedly willing to listen to offers for in the summer . sky sports pundits gary neville and graeme souness were highly critical of toure .

BUS [25]: A Pointer-Generator model with a data-efficient content selector to over-determine phrases in a source document that should be part of the summary.

Summary: manchester city face west ham in the premier league on sunday . manuel pellegrini ’s stars were trained on

	Category	Description	Example
RelE	Relation Error	The predicate in the summary statement is inconsistent with the source article.	The Ebola vaccine was rejected by the FDA in 2019.
EntE	Entity Error	The primary arguments (or their attributes) of the predicate are wrong.	The COVID-19 vaccine was approved by the FDA in 2019.
CircE	Circumstance Error	The additional information (like location or time) specifying the circumstance around a predicate is wrong.	The first vaccine for Ebola was approved by the FDA in 2014 .
CorefE	Coreference Error	A pronoun/reference with wrong or non-existing antecedent.	The first vaccine for Ebola was approved in 2019. They say a vaccine for COVID-19 is unlikely to be ready this year.
LinKE	Discourse Link Error	Error in how multiple statements are linked together in the discourse (for example temporal ordering/causal link). show successful human trials, then sequence the DNA of the virus.	To produce the vaccine, scientists have to show successful human trials, then sequence the DNA of the virus.
OutE	Out of Article Error	The statement contains information not present in the source article.	China has already started clinical trials of the COVID-19 vaccine.
GramE	Grammatical Error	The grammar of the sentence is so wrong that it becomes meaningless.	The Ebola vaccine accepted have already started .

TABLE XI: Overview of error annotations used in FRANK’s annotation process[11]

wednesday ahead of their match . yaya toure and samir nasri are reportedly two players who could leave .

PGN [26]: An LSTM based Pointer-Generator Network with Copy Mechanism.

Summary: manchester city face west ham united at their etihad campus on sunday . manuel pellegrini ’s stars were trained on wednesday ahead of their match . sergio aguero was criticised by sky sports ’ gary neville and graeme souness after city ’s loss at manchester united .

S2S [27]: An LSTM based Sequence-to-Sequence with attention model.

Summary: manchester city host manchester united in the premier league on sunday . sergio aguero and samir nasri could leave the club in the summer . manuel pellegrini ’s side are on sunday at the etihad stadium .

B. FRANK XSum

The following summaries were generated from document 2.

BERTS2S [28]: Transformer encoder-decoder model with parameter sharing, where both encoder and decoder are initialized with the BERT-Base checkpoints [29] and fine-tuned on XSum.

Summary: three schools have been threatened with legal action over an alleged trojan horse affair, the education secretary has said.

PtGen: The same pointer-generator network (PGN) network as above.

Summary: a primary school in birmingham has been rated "inadequate"by a senior ofsted inspector who has been rated "inadequate"by ofsted.

TConvS2S [19]: Topic-Aware Convolution Sequenceto-Sequence model.

Summary: the head teacher of a birmingham school has been placed in special measures after being found guilty of misconduct.

TransS2S [30]: A randomly initialized Transformer encoder-decoder model fine-tuned on the XSum dataset.

Summary: pupils at a school in birmingham have been told they will lose their jobs because of a lack of funding.

C. DeFacto

The following summary is generated from document 3.

PEGASUS [21]: A transformer-based model with an encoder-decoder structure, where it is pre-trained on a task of predicting important sentences (instead of masked words or sentences as in other models), and fine-tuned for the summarization task.

Summary: Kris Boyd is against Scottish clubs playing matches in the United States.

D. FactCC

We could not provide example summaries for each of the models, as the summaries in the dataset did not contain this information.

Hsu et al. [31]: A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

Bottom Up Summarization: described above as *BUS*.

Jiang and Bansal [32]: A sequence-to-sequence model with the aim of improving memorization capabilities of the encoder of a pointer-generator model by adding a 'closed-book' decoder without attention and pointer mechanisms.

Chen and Bansal [33]: A fast summarization model that first selects salient sentences and then rewrites them abstractively (i.e., compresses and paraphrases) to generate a concise overall summary.

Pointer-generator network: described above as *PGN*.

Kryściński et al. [34]: A sequence-to-sequence model where the decoder is decomposed into a contextual network that retrieves relevant parts of the source document. Also, a novelty metric is used that is optimized directly through policy learning to encourage the generation of novel phrases.

Li et al. [35]: A sequence-to-sequence model that leverages the structural information of both documents and multi-sentence summaries to improve the document summarization performance. They do so by importing both structural-compression and structural-coverage regularization into the summarization process to capture the information compression and information coverage properties, which are the two most important structural properties of document summarization.

Pasunuru and Bansal [36]: A reinforcement learning approach with two novel reward functions: ROUGESal and Entail, on top of a coverage-based baseline.

Guo et al. [37]: A summarization model trained using multi-task learning with the auxiliary tasks of question generation and entailment generation, where the former teaches the summarization model how to look for salient questioning-worthy details, and the latter teaches the model how to rewrite a summary which is a directed-logical subset of the input document.

E. Large language models

The following summaries were generated from document 1. All language models below are instructed with the same prompt, described below.

GPT-4 [3]: (Generative Pre-trained Transformer 4) is a multimodal large language model created by OpenAI. This transformer-based model is trained to predict the next token, given a context, and fine-tuned with reinforcement learning.

Summary: Manchester City players trained on Wednesday in preparation for their weekend clash against West Ham United. Striker Sergio Aguero practiced his shooting skills, while Yaya Toure and Samir Nasri participated in stretching exercises. Both Toure and Nasri have been criticized for their recent performances, but they will still be relied on to help the team secure a top four finish this season. Manchester City currently sits fourth in the Premier League with just seven games remaining.

Falcon¹⁰: a large language model trained on the high quality RefinedWeb [38] dataset. Different than other LLMs, multi-query attention [39] is used instead of the normal attention mechanism.

Summary: Manchester City trained on Wednesday ahead of their match on Sunday against West Ham United. Sergio Aguero, who just scored his 99th and 100th Premier League goals, participated in shooting practice. Yaya Toure and Samir Nasri may leave the club in the summer as they are reportedly up for offers. Sky Sports pundits, Gary Neville and Graeme Souness criticized Yaya Toure in the aftermath of Manchester United's 4-2 win against City. Toure, who became a bit of a liability for City in the past two years, is instrumental to their Premier League win last season. However, Manchester City is still looking to cement a top-four finish this season.

StableLM¹¹: a fine-tuned version of Facebook's Llama model¹².

Summary: The soccer news today focuses on the upcoming football matches between Manchester City FC and West Ham United over the next few days, including details about how both teams arrived late in Europe earlier this week following disappointing losses away to Blackburn Rovers and QPR respectively. However, it seems there isn't much else noteworthy or significant happening regarding either team right now aside from these potential clashes.

F. Source documents

Document 1: Manchester City may be still licking their wounds from their derby day mauling at bitter rivals Manchester United on Sunday, but the reigning Premier League champions looked in determined mood as they trained on Wednesday. Preparing for their weekend clash at home to West Ham United, Manuel Pellegrini's stars were all in action at their Etihad Campus base. Fresh from scoring Premier League goals 99 and 100 in his City career, striker Sergio Aguero was back to his menacing best as he took part in shooting practice against goalkeeper Joe Hart. Manchester City players were trained on Wednesday ahead of their match on Sunday at home to West Ham City striker Sergio Aguero fine-tuned his finishing skills with some shooting practice at their Etihad Campus Yaya Toure (left) and Samir Nasri are reportedly two players who could leave the club in the summer. Aguero, who signed a new five-year contract with the club in August, will certainly be one player who will remain at the Etihad next season. The same cannot necessarily be said for Yaya Toure and Samir Nasri - who are reportedly two players City are willing to listen to offers for in the summer. And the midfield duo looked out in the cold at one point during Wednesday's session as they appeared as Nasri underwent some stretching exercises. In the aftermath of Sunday's defeat to United at Old Trafford, Sky Sports pundits Gary Neville and Graeme Souness were highly critical of Toure. Of the 31-year-old, the latter said that he had become 'a bit of a liability' for City in the past two years, despite being instrumental to their Premier League title win last season. Toure was criticised by Sky Sports' Gary Neville and Graeme Souness after City's loss at Manchester United 'When he is playing and he has the ball at his feet, he is a real threat going forward,' the former Liverpool midfielder said. 'When they don't have the ball he doesn't want to take part and come back towards his own goal.' 'He just doesn't take part when City have the ball, they're too easy to play against. And that's a problem they've had for the last couple of seasons. And maybe that's one of the big reasons why Both the Ivory Coast international and Nasri have come under criticism for recent displays on the pitch, but Pellegrini will still be calling on their services as they look to cement a top four finish this season. The 61-year-old's side currently sit fourth in the Premier League with seven games remaining - four points clear of fifth-placed Liverpool and with a far greater goal difference too. City players' heads drop after conceding the fourth goal on Sunday to their cross-city rivals at Old Trafford

Document 2: Birmingham academies Park View and Nansen Primary - rated inadequate by Ofsted - will have their funding agreements terminated. Park View Educational Trust said it may be removed from running the schools. Oldknow Academy and Golden Hillock School, also rated inadequate, were warned they could lose funding unless concerns were addressed. A head teacher at one of the Trojan Horse schools said parents tried to undermine her authority and descended on the school during an Ofsted inspection. The head teacher, who wished to remain anonymous, said she heard parents speaking to an inspector about her. She said she came close to quitting

¹⁰<https://falconnllm.tii.ae/>

¹¹<https://github.com/Stability-AI/StableLM>

¹²<https://github.com/facebookresearch/llama>

her job on a number of occasions but everything "fell into place" when news of the Trojan Horse letter broke."From 2012 I felt alone," she said."I did not know it was happening to other heads. If I had known, I could have rationalised it but at the time I felt embarrassed."I started thinking 'maybe it is me, maybe I'm doing something wrong.'"On Monday, the watchdog published 21 reports into schools allegedly targeted as part of a hardline Muslim takeover.Five of those schools, including Park View and Oldknow, were rated inadequate.'Islamic-themed assemblies'Birmingham City Council said it would work with Saltley School, the only local authority school placed in special measures, and the Department for Education to install a temporary governing board.In a letter to Park View chairman Tahir Alam, education minister Lord Nash said there were "deep concerns" about the way Park View School was run.The letter said: "I have decided under clause 5.6 of the supplementary funding agreement for the academy to give written notice of the Secretary of State's intention to terminate that agreement."Lord Nash criticised the trust for failing to promote "spiritual, moral, social and cultural development of pupils"."I am deeply mindful of the need to eliminate discrimination, advance equality of opportunity and foster good relations," he said.A similar letter about Nansen Primary School, also addressed to Tahir Alam, said it would have its funding arrangement terminated.Both letters invited the trust to respond before 4 July.A spokesman for Park View School said they would challenge the Ofsted report through the appropriate legal channels.They said terminating the funding agreement would mean the trust could be removed from the running of the three schools."What happens to the schools after that would be the decision of the secretary of state, as they are already academy schools under the remit of [the] Department for Education," they said.In a letter to Dr Achmad da Costa, chair of Oldknow Academy Trust, Lord Nash said it was not meeting the requirements set out in its funding agreement."Therefore the Secretary of State is minded to terminate the funding agreement unless all the breaches are addressed promptly, in full, and to his satisfaction," it said.Lord Nash set out a list of actions the trust must take before 4 July.These included ending "less favourable" treatment of non-Muslims, ceasing segregation of boys and girls, and drawing up a plan to ensure a more "broad and balanced" curriculum.Oldknow Academy declined to comment on the letter from Lord Nash.A letter to Tahir Alam regarding Golden Hillock's funding arrangement set out a similar list of demands to be completed within the same timescale.Ofsted chief Sir Michael Wilshaw said "a culture of fear and intimidation has taken grip" in Birmingham schools, while a number of head teachers had reported an "organised campaign to target certain schools in order to impose a narrow, faith-based ideology".Own agendas'The report into Park View School said the academy's work to keep students safe, including from extremism, was inadequate.It said external speakers had not been vetted properly and the school had allowed a guest with known extremist views to speak to students as part of Islamic-themed assemblies.Park View and Oldknow trusts did not comment on the future but one of Saltley School's senior teachers said it hoped to boost its Ofsted rating.Darren Turner was brought in as associate

head teacher on Thursday as part of a partnership scheme with other local schools.He said there had been "upset" among staff and pupils at Saltley's inadequate rating."This is all about moving forward, what happens next," he said."Behind these walls they've got some excellent teachers and they want to get back to doing what they are good at: teaching."When this school gets back to 'good' again we want to go further and become outstanding."Parents at the schools have criticised Ofsted's reports and accused inspectors of bias.Health professional Arshad Malik told the BBC the findings were "alien" to his child's experience at Park View."200eInspectors came with loaded questions," he said."People are trying to use this school to push their own agendas."He said parents would like to invite Education Secretary Michael Gove to the school for an open meeting about Ofsted's findings and to talk about what evidence was found to support them.

Document 3: Celtic and Dundee are exploring the possibility of staging a game in the US, where the Kilmarnock striker has experience with Portland Timbers. "I think there are avenues for the Scottish game to tap into if you look at the American model," said Boyd. "But, for our teams to go there to play, I don't think it's right." A formal proposal is set to be put before the Scottish Professional Football League regarding the Dundee v Celtic match in the next few weeks, with Boston and Philadelphia possible host cities. "It's not for me," Boyd, 32, told BBC Scotland. "With the two of them being Scottish clubs, it's right that they play in Scotland. "It's obviously a big market for Celtic with them being a worldwide club and because Dundee's owners are American, that's why they'll be trying to explore it. "There's a lot of Scottish people living in America. There is interest. "But the majority of Dundee fans will be from Dundee. I don't imagine there'll be a lot in America. With Celtic, there will be. That comes with the size of the club." Boyd spent 2012 in Major League Soccer, scoring seven goals in 22 starts for Portland. "Scottish clubs can go over and see what's going on and try and make things better but bring it back here to try and make the Scottish game better in Scotland," added the former Scotland international. "Football is definitely getting bigger [in the US]. "The majority of clubs are now playing in front of 20,000 every week. Seattle are playing in front of 60,000 - that's as big as Rangers and Celtic! "There's more to take into it - the travelling, the different climates, the time zones. It's a difficult place to go and play football. "And they change the rules every year as well to suit the teams that are struggling because if you win the league then players' wages might go up then the next thing they need is to trade their players to keep themselves under their salary cap. "It would be difficult to manage in and definitely be an owner in but the flipside is that if you're an owner and you're getting 20,000 every week, you know you're still making money."

APPENDIX D

MANUAL EVALUATION ABSTRACTIVENESS LLMs

Document 1

bart: pm announces plans to give millions of workers three days paid leave a year . scheme will only affect firms with 250

or more staff - an estimated 10 million in private sector and five million in public sector . but business leaders criticised the move as ‘ heavy-handed government intervention ’

bert_sum: prime minister says he will amend working times regulations so for three days people can volunteer or serve as a school governor . but business leaders have criticised the move as heavy-handed government intervention . policy will apply to firms with at least 250 staff - an estimated 10million in the private sector and five million in the public sector .

bus: david cameron said he will amend working times regulations so for three days people can volunteer or serve as a school governor . but business leaders criticised the move as heavy-handed government intervention . policy will apply to workers at least 250 staff - an estimated 10 million in the public sector .

pgn: david cameron announced plans to give millions of workers three days paid leave . the prime minister said he will amend the working times regulations so for three days people can volunteer or serve as a school governor . business leaders criticised the move as heavy-handed government intervention ’ , amid doubts about how the policy would be paid for .

s2s: david cameron announced plans to give millions of workers three days paid to be paid . but business leaders criticised move as heavy-handed government intervention . but business leaders criticised move as heavy-handed government intervention .

stablelm: The prime minister of Great Britain is proposing legislation allowing citizens who earn over £1m to grant three days’ fully paid leave throughout April 2021. Business leaders claim such measures may undermine employee safety and free play during annual holidays. Critics question whether tax policies alone justify these generous benefits, pointing out that taxpayers subsidize daycare and child care facilities which cater specifically to low-income families. However, supporters argue any attempt to curtail benefit generosity must respect industry conventions and maintain workplace flexibility while still ensuring

falcon: UK Prime Minister David Cameron has announced plans to include three days of paid leave for volunteering in the Working Time regulations. The policy will apply to firms with at least 250 employees but has attracted criticisms from directors and the Labour party regarding the cost and lack of flexibility. Despite this, the scheme has been praised as a ‘triple win’ for employees, communities, and companies and gained support from Bear Grylls, who noted the benefits of government support for volunteering. Following the announcement, however, Communities Secretary Eric Pickles said companies would not be forced to give staff days-off for volunteering.

gpt4: David Cameron announced plans to give millions of workers three days paid leave a year to volunteer or serve as a school governor. Business leaders criticized the move as heavy-handed government intervention amid doubts about how the policy would be paid for. The policy will apply to workers at firms with at least 250 staff, an estimated 10 million in the private sector and 5 million in the public sector.

Document 2

bart: emma jackson , 28 , moved into a quiet cul-de-sac in hull just over a year ago . but within a week of moving in , her dream home turned into a nightmare . mark ray , 56 , waged a year-long campaign of terror against ms jackson .

bert_sum: emma jackson , 28 , moved into a quiet cul-de-sac in hull just over a year ago . she thought she had found the perfect place to call home . but within a week of moving in , her dream home turned into a nightmare . aggressive neighbour mark ray , 56 , waged a year-long campaign of terror against her .

bus: mark ray , 56 , waged a year-long campaign of terror against ms jackson , 28 , just over a year ago . now ms jackson has been forced to move home because of mr ray but says she lives in terror of her new address . ms jackson was subjected to vile abuse and was stalked by her neighbour .

pgn: emma jackson , 28 , moved into a quiet cul-de-sac in hull just over a year ago . she thought she had found the perfect place to call home . but within a week of moving in , her dream home turned up at all hours of the day . now ms jackson has been forced to move home because of mr ray .

s2s: mark ray , 28 , was met with a barrage of abuse and was stalked by her neighbour . she was met with a barrage of abuse that culminated , a year later . ms jackson ’ s ordeal began less than a week after moving into her ground flat .

stablelm: Emma Jackson, who felt increasingly isolated due to neighbors exhibiting loud noises past midnight, complained to housing authorities regarding their treatment towards her apartment neighbor, Alexander Ray. Despite having received multiple warnings specifying behavior issues including excessive noise levels and harassment, Jackson initially chose to remain indoors rather than confront the issue. However, upon realizing that one other tenant already faced similar issues, Jackson decided to take legal action via the local housing association. Through her complaint process, Jackson details various instances of inappropriate behavior directed toward her such

falcon: Emma Jackson had to leave her dream home a year after moving in because of an aggressive 56-year-old neighbor who watched TV loudly and began to lash out at her. She reported Mark Ray to the local Housing Association and he was ordered to turn down the volume, but after a week, he resumed his previous behavior and subjected Jackson to heinous language and violent abuse. He also threatened her family. Each interaction with Ray made Jackson increasingly paranoid she lost sleep, hair, and suffered stress and depression. She installed CCTV equipment to prove that he had been breaking the terms of the injunction not to approach or speak to her – he was eventually arrested. However, Jackson still lives in fear of Ray discovering her new address.

gpt4: Emma Jackson, a 28-year-old photography student, moved into a quiet cul-de-sac in Hull only to find her dream home turn into a nightmare due to an aggressive neighbor, Mark Ray, who constantly watched TV at high volume. Despite asking him to lower the volume, the situation escalated into verbal abuse, stalking, and violence. Ms. Jackson eventually contacted the local Housing Association and installed a CCTV camera to gather evidence of Mr. Ray breaching an

injunction. In October 2014, a new injunction was issued that would result in Mr. Ray's arrest should he become violent or threatening however, Ms. Jackson chose to move to a new home, fearing him discovering her address. The ordeal lasted over a year and had a severe impact on Ms. Jackson's wellbeing.

Document 3

bart: nbc journalist jeff rossen filmed an airport worker at laguardia airport in new york doing push-ups on the runway after loading a plane . footage shows the employee wearing his high-visibility jacket and gloves while performing the aerobic stunt . the video - filmed april 19 - quickly garnered dozens of ' likes ' on facebook .

bert_sum: nbc journalist jeff rossen filmed an airport worker at laguardia airport in new york doing push-ups on the runway after loading a plane . footage shows the employee wearing his high-visibility jacket and gloves while performing the aerobic stunt . even after his 12th push-up , the fitness-enthusiast shows no sign of slowing down .

bus: nbc journalist jeff rossen filmed an airport worker at laguardia airport in new york doing push-ups on the runway after loading a plane . footage shows the employee wearing his high-visibility jacket and gloves while performing the aerobic stunt . even after his 12th push-up , the fitness-enthusiast shows no sign of slowing down .

pgn: nbc journalist jeff rossen filmed an airport worker at laguardia airport in new york doing push-ups on the runway after loading up a plane . footage shows the employee wearing his high-visibility jacket and gloves while performing the aerobic stunt .

s2s: jeff UNK filmed an airport worker at laguardia airport in new york . footage shows the employee wearing his distinctive jacket and gloves while performing the lightning bolt stunt . the video shows the employee wearing his distinctive jacket and gloves while performing the stunt .

stablelm: An airport worker in Seattle was discovered sleeping on the job when filming took place during a flight despite not showing any signs of slacking off, footage showed him still performing routine physical tasks such as push-ups.

falcon: An NBC journalist filmed an airport worker at LaGuardia Airport in New York doing push-ups on the runway after loading a plane. The worker seems to be fit and continued to exercise even after doing 12 push-ups. The video of the worker was uploaded to social media and received many likes. The incident happened on April 19, and the worker was dubbed as the "hardest working man at LaGuardia."

gpt4: An airport worker at LaGuardia Airport in New York was filmed doing push-ups on the runway after loading a plane. The employee, wearing a high-visibility jacket and gloves, continued his workout despite being observed by NBC journalist Jeff Rossen, who later shared the video on social media. The clip garnered positive attention, with viewers appreciating the worker's dedication to fitness and his job.

Document 4

bart: terry martin , 48 , shot his girlfriend laurice hampton at their home in fort worth , texas , after she asked for half of

the proceeds . but hampton , critically injured from a gunshot wound to her chest , was able to call 911 about 6:30 am saturday to report the shootings before she died . responding officers found the couple inside their master bedroom where martin was pronounced dead from a wound to the head .

bert_sum: terry martin , 48 , shot his girlfriend laurice hampton , 48 . she asked for half of the winnings but hampton was critically injured . police found the couple inside their master bedroom where martin was pronounced dead from a gunshot wound to the head . hampton was taken by ambulance to john peter smith hospital but died a few hours later .

bus: terry martin shot his girlfriend laurice hampton , 48 , after she asked for half of the proceeds . hampton was taken by ambulance to john peter smith hospital but died a few hours later . martin was sentenced in july 2012 to a day in jail and fined \$ 214 for assaulting hampton .

pgn: terry martin , 48 , shot his girlfriend laurice hampton , 48 , after she asked for half of his winnings . but hampton , critically injured from a gunshot wound to her chest , was able to call 911 about 6:30am saturday to report the shootings before she died . hampton was sentenced in july 2012 to a day in jail and fined \$ 214 for assaulting hampton in their home in fort worth , texas .

s2s: terry martin , 48 , shot his girlfriend UNK hampton , 48 , after she asked for half of the money . he was taken by ambulance to john peter smith hospital but died a few hours later .

stablelm: A young woman named Laurice Hampton has been shot multiple times during a fight over money while she demanded what amounted to half of the total sum shown as winning prize in a Texas lottery scratcher game. Police were called shortly after 3 am early Sunday morning when the incident took place at the family's home located near Interstate 25 outside Fortworth, TX around 8 ago. Investigators are currently working on finding out if the wounded victims knew each other well or not yet have any criminal charges filed against them

falcon: Terry Martin shot and killed his girlfriend, Laurice Hampton, after she asked for half of the \$500 winnings from his scratch-off lottery ticket. Hampton called 911 after being shot, but died a few hours later at the hospital. Martin also shot himself and died at the scene. There was a history of violence between the couple, and Martin had previously assaulted Hampton in their home in 2011.

gpt4: Terry Martin, a 48-year-old man, fatally shot his girlfriend Laurice Hampton, also 48, and then himself after an argument over sharing his \$500 scratch-off lottery winnings. Hampton, critically injured, managed to call 911 before passing away. The police found the couple in their Fort Worth, Texas home where Martin was pronounced dead. The couple had known each other since childhood and had a history of violence, with Martin having previously been jailed and fined for assaulting Hampton.

Document 5

bart: blackpool drew 1-1 with reading at bloomfield road on tuesday night . lee clark 's side have been relegated to

league one for the first time in their history . the seasiders ' main reception was pelted with eggs before the game . fans chanted for chairman karl oyston to leave the club .

bert_sum: blackpool drew 1-1 with reading at bloomfield road . lee clark admitted the club have had a ' horrendous season ' to put up with . jamie o'hara scored from the penalty spot to give blackpool a lead . but grant hall scored an own goal just after half-time to leave .

bus: lee clark admitted blackpool face a challenge to win back their stay-away fans . defender grant hall scored an own goal just after half-time to leave lee clark . jamie o'hara scored from the penalty spot to give blackpool a lead .

pgn: blackpool face a challenge to win back their stay-away fans after an evening in which their relegation into sky bet league one was marked by pre-match supporters ' protests against karl oyston . blackpool face a challenge to win back their stay-away fans after rotherham 's win over brighton on easter monday . lee clark admitted blackpool face a challenge to win back lee clark disappointed .

s2s: blackpool drew 1-1 at the bottom of the championship on saturday . lee clark admitted blackpool face a challenge to win back their UNK fans . lee clark admitted blackpool face a challenge to win back their UNK fans .

stablelm: Blackpool FC are facing challenges to restore their reputation among both local and visiting audiences after heavy defeats compared to last year. The team announced they would run league football again while announcing plans to recruit talented youngsters ready to fight past promotion battle this season. However, despite these efforts, many remain unconvinced or frustrated with former manager Lee Clark's management style and want him replaced. Meanwhile, various groups – including blackmarket vultures seeking unclaimed property via land buyouts disgraced

falcon: Blackpool were relegated to Sky Bet League One, which led to pre-match supporters' protests against Karl Oyston. Fans threw eggs at directors' boxes windows and launched a flare, smoke bomb, and firework in front of the main entrance of West Stand, where Oyston and his family usually watch games. The Royals drew 1-1 with Blackpool, who gained their third successive draw, as Jamie O'Hara scored from the penalty spot to give Blackpool a lead during the game. Defender Grant Hall scored an own goal just after half-time to leave Lee Clark disappointed.

gpt4: Blackpool faces a challenge to win back their stay-away fans after being relegated into Sky Bet League One, marked by pre-match supporters' protests against chairman Karl Oyston. Disgruntled fans protested against Oyston, calling for him to leave the club, which became the first of the Football League to be demoted following Rotherham's win over Brighton. Manager Lee Clark acknowledged a "horrendous season" but is hopeful of winning back fans by rebuilding the squad over the summer.

Document 6

bart: three men accused of being agents of the syrian government have been executed in a shocking new video . the barbaric footage , released earlier today , shows the men

kneeling in the middle of a dusty road in deir ez-zor , syria . it comes just one day after the extremist group claim to have shot and beheaded more than 30 ethiopian christians in libya .

bert_sum: warning : graphic content . video shows men kneeling in middle of road in deir ez-zor , syria . comes just one day after extremist group claim to have shot and beheaded more than 30 ethiopian christians in libya . they were blindfolded and transported in a van to the execution point - thought to be deir ez-zo in syria . the eight-minute video shows the men wearing orange guantanamo-style jumpsuits with their hands and feet in chains . after giving information about each of the men in arabic , the men are blindfolded and filmed in the back of a moving van .

bus: three men accused of being agents of syrian government executed in video . barbaric footage shows men kneeling in middle of road in syria . each of the men are forced on to their hands and knees in the street .

pgn: islamic state has claimed . it comes one day after the extremist group claim to have shot and beheaded more than 30 ethiopian christians in libya . each of the men are blindfolded and filmed in the back of a moving van while an isis militant decked in black guards over them .

s2s: video shows men kneeling in the middle of a dusty road in deir UNK , syria . each of the men are blindfolded and filmed in the back of a moving van . the men are blindfolded and transported in a van while they were beheaded .

stablelm: The Syrian opposition has condemned the murder of these three men outside Al Bab in Deira e z oro city last week. However, according to local news sources cited above, a video surfaced yesterday suggesting that some activists within the Muslim Brotherhood affiliated Palestinian Islamist organisation Um al Qassam may have facilitated such killings. This follows recent allegations that members of the organization helped plan and coordinate several attacks inside Egypt's Sinai peninsula aimed at Israeli military personnel during 2013 Ramadan celebrations both senior Egyptian security figures

falcon: ISIS has released a video showing three men accused of being agents of the Syrian government being executed. They were beheaded in a dusty street, with their bodies left on the ground. A sharpened sword was used to behead them and the shocking footage shows it being embossed and sharpened for the executions. ISIS has not confirmed who the men were or when the execution took place. The prisoners were filmed in a dirty underground room before being transported by van to the execution point and were seen barefoot and in shackles. The militant group uses propaganda videos to attract would-be fighters. The battle against ISIS amid a civil war in Syria has resulted in over 200,000 deaths.

gpt4: ISIS has released a video claiming to show the execution of three men accused of being agents of the Syrian government. The footage shows the men kneeling in the middle of a road in Deir ez-Zor, Syria, before being beheaded. This comes just one day after the extremist group claimed to have killed more than 30 Ethiopian Christians in Libya. These videos are often used by ISIS as propaganda to attract would-be jihadi fighters. The battle against the militant group has already claimed the lives of more than 200,000 people.

Document 7

bart: john sims , who had burkitt 's lymphoma , died on saturday , with his new wife lindsey sims by his side . the pair married march 21 at the university of texas m.d. anderson cancer center . after the wedding they set off on a a cross-country road trip last week , hoping to make it to visit john 's friend dave in oregon . but their journey came to an end when john 's condition worsened and they had to go to a hospice in grand junction , colorado .

bert_sum: john sims , who had burkitt 's lymphoma , died on saturday , with his new wife lindsey by his side . the pair married march 21 at the university of texas m.d. anderson cancer center , after john , who was to turn 31 tomorrow , was told his cancer was terminal . after the wedding they set off on a cross-country road trip last week , hoping to make it to visit john 's friend dave in oregon .

bus: john sims , who had burkitt 's lymphoma died on saturday , with his new wife lindsey sims by his side . the pair married march 21 at the university of texas m.d. anderson cancer center , after john , who was to turn 31 tomorrow , was told his cancer was terminal . after the wedding they set off on a honeymoon road trip last week .

pgn: john sims , who had burkitt 's lymphoma , died on saturday . the pair married march 21 at the university of texas m.d . john 's condition worsened and they had to go to a hospice in grand junction , colorado .

s2s: john sims , who had UNK 's lymphoma died on saturday , died on saturday . john sims , who had UNK 's lymphoma died on saturday , was told his cancer was terminal .

stablelm: After working together since becoming engaged over 10 years ago, Ryan Smith later faced devastating news – his close colleague Marissa Carreno passed away unexpectedly earlier this year due to complications caused by diabetes. However, she didn't give up hope so easily despite having struggled with illness for several years, Ryan remains hopeful about what lies ahead. Meanwhile back home in Houston, Lindy Sims suffered another unexpected loss shortly before Christmas when she announced publicly that she is expecting their first child next month. Despite losing

falcon: John Sims passed away from Burkitt's Lymphoma while on his honeymoon with new wife Lindsey. They got married at the University of Texas M.D. Anderson Cancer Center before embarking on a road trip. But John's condition worsened, and they had to go to a hospice in Grand Junction, Colorado, where he eventually died. John had set up a JustGiving page to help with funeral costs and to support his wife and mother. Burkitt's Lymphoma is a rare, aggressive form of non-Hodgkin's lymphoma. Lindsey quit her job as a restaurant manager to care for John. The couple urged people to remember that "there is still love out there," even in difficult circumstances.

gpt4: John Sims, a newlywed who had terminal Burkitt's Lymphoma, passed away during a honeymoon road trip with his wife, Lindsey, by his side. The couple got married on March 21 at the University of Texas M.D. Anderson Cancer Center, after learning of John's terminal condition. They embarked on a cross-country road trip, but had to cut it short

due to John's deteriorating condition, and he was admitted to a hospice in Colorado. Burkitt's Lymphoma is a rare, aggressive form of non-Hodgkin's lymphoma that affects lymphocytes, which are white blood cells found in the blood and lymph glands.

Document 8

bart: nancy perry will no longer teach students at dublin middle school and will retire at the end of the year , city schools superintendent chuck ledbetter announced on tuesday . perry is alleged to have told students that obama is a muslim and that any parent who support him could n't be christian , either . a 12-year-old boy in perry 's class brought the matter to the attention of his father , jimmie scott , who complained to the school . scott said he immediately requested a parent-teacher conference to which perry brought along her husband bill , who sits on the county board of education .

bert_sum: nancy perry will no longer teach students at dublin middle school and will retire at the end of the year . perry is alleged to have told students that obama is a muslim and that any parent who support him could n't be christian , either . a 12-year-old boy in perry 's class brought the matter to the attention of his father , jimmie scott , who complained to the school . ' if your parents voted for obama , they 're evil and i do n't see how your parents could vote for someone that 's muslim , ' he said .

bus: nancy perry will no longer teach students at dublin middle school and will retire at the end of the year . perry is alleged to have told students that obama is a muslim and that any parent who support him could n't be christian . ' if your parents voted for obama , they 're evil and i do n't see how your parents could vote for someone that 's muslim ' he told wmoz .

pgn: nancy perry will no longer teach students at dublin middle school in georgia . she gave students her highly critical opinion of president barack obama . perry is alleged to have told students that obama is a muslim .

s2s: nancy perry , a veteran georgia middle school teacher , gave students her highly critical personal opinion of obama . perry is alleged to have told students that obama is a muslim and that any parent who support him could n't be christian . perry is alleged to have told students that obama is a muslim .

stablelm: Nancy Perry, a Vietnam War veteran teaching English in Georgia recently retired after giving controversial opinions to young adults in her classes regarding their presidents, including telling one child to exclude any religion other than Christianity when questioned. Parents have accused her of spreading "anti-American" sentiments among some attendees during lessons with her. Despite this controversy, Perry plans to continue working once leavingschools due to conflicts within staffing.

falcon: Nancy Perry, a middle school teacher in Georgia, has been removed from the classroom and will retire at the end of the year after reportedly giving her students her highly critical opinion about President Barack Obama, including that he is Muslim and that any parent who supports him couldn't

be Christian. A parent of one of the students brought the matter to the attention of the school after Perry allegedly made the comments in class, leading to her removal. During a subsequent parent-teacher conference, Perry and her husband reportedly showed the parent what he described as anti-Obama internet propaganda. The school district's superintendent issued an apology for Perry's actions and required principals to notify the superintendent if any board member requests to attend a parent-teacher conference.

gpt4: A middle school teacher in Georgia has been removed from the classroom and will retire at the end of the year after giving her highly critical personal opinion of President Barack Obama to her students, which led to some parents complaining. The teacher, Nancy Perry, allegedly told students that Obama is a Muslim, and that any parents who support him couldn't be Christian. The school's superintendent apologized to the student and parents involved, emphasizing that it is not the place of teachers to attempt to persuade students about religious or political beliefs.

Document 9

bart: with budgets for outfits , flowers , transportation , food and other sundries at around \$ 1,000 for the average american high schooler , that means most intend to spend \$ 324 on promposals . that 's according to a survey from visa , which asked 3,000 people aged 18 or older what they expected to shell out for the prom and on what .

bert_sum: promposals are the big trend in high school proms as teens think up more outlandish ways to ask each other to the biggest dance of the year . with budgets for outfits , flowers , transportation , food and other sundries at around \$ 1,000 for the average american high schooler , most intend to spend \$ 324 on the once-simple question : will you go to prom with me ? .

bus: students this year will likely spend of a third of their prom budgets on the elaborate new ritual of promposals . 3,000 people aged 18 or older are expected to shell out for the prom . for a primer on promposals , see previous years ' prom proposals .

pgn: students think up progressively more outlandish ways to ask each other to the biggest dance of the year . with budgets for outfits , flowers , transportation , food and other sundries at around \$ 1,000 for the average american high schooler , that means most intend to spend \$ 324 on the once-simple query .

s2s: more than \$ 1,000 for the average american high UNK is the most popular dance of the year . in 2012 , matthew ho from toronto UNK to his girlfriend , tina , with 1,500 pink helium balloons and a sign that read , ' tina , will you go to prom with me ? '

stablelm: The article discusses various aspects related to formal or informal fundraising activities during high school dances such as prom ceremonies or sweet sixteen parties. Prom season has become increasingly popular among teenagers due to their thought-provoking nature. Various articles discuss different types of proposals including traditional one-night stands or low budget options like homemade ice cream cones or rental planes. Additionally, there are news reports about schools

planning expensive extravagance events called "promposals" featuring large amounts spent on attire alone, decorations, music,

falcon: Promposals, where teenagers come up with creative ways to ask each other to prom, are becoming a big trend in high schools. According to a Visa survey, students will spend around a third of their prom budgets, which are typically around \$1,000, on promposals which can cost over \$300. The survey also shows that there are many creative promposals that take different forms such as using balloons, banners, and pop quizzes. The average Canadian family plans to spend \$508 overall on prom, with \$151 on promposals, while the average US family plans to spend \$919 overall. Dads plan to spend 63% more than mothers (1,160vs.710) and the Northeast region of the US is projected to spend the most overall.

gpt4: Promposals, the trend of elaborate ways to ask someone to attend prom, often make up a third of the total prom budget for high school students in the US. The average American high schooler plans to spend \$324 on the promposal, out of a total budget of around \$1,000, according to a Visa survey. Some students have gone to extreme lengths, spending a lot of money on grand gestures such as renting plane banners, while others have been more creative and low-cost.

Document 10

bart: rand paul , 52 , shared a snapchat video of himself learning how to play poker from the infamous ' king of instagram ' dan blizerian . the pair were playing liar 's poker - a game where players use dollars bills instead of cards and track the serial numbers on the banknotes .

bert_sum: on friday , rand paul shared a 10-second video of himself learning how to play poker from dan blizerian . the video , obtained by daily caller reporter kaitlin collins , is titled ' lessons from dan blotian ' . the pair were playing liar 's poker - a game where players use dollars bills instead of cards and track the serial numbers on the banknotes . the goal of the game , according to investopedia.com , is to bluff the opponents into believing that your bid does not exceed the combined sum of all of the serial number .

bus: senator rand paul shared a snapchat video of himself learning how to play poker from the infamous ' king of instagram ' dan blizerian . the pair were playing liar 's poker - a game where players use dollars bills instead of cards and track the serial numbers on the banknotes . on friday , the libertarian lawmaker is titled ' lessons from dan blizerian .

pgn: senator rand paul shared a snapchat video of himself learning how to play poker from the debauched social media star dan blizerian . the pair were playing liar 's poker - a game where players use dollars bills instead of cards and track the serial numbers on the banknotes . liar 's poker is also the title of a book by bestselling author michael lewis about the dissolute , everything-goes culture of wall street bond salesmen in the 1980s .

s2s: rand paul posted a video of himself using a video of himself learning how to play poker from the infamous ' king of UNK dan UNK ' . paul and UNK were playing oj 's poker - a game where players use money bills instead of cards and track the serial numbers on the notes .

stablelm: In today's news, we're reporting yet again details about a recent scandal involving politicians and celebrities! This week, we caught up with Congressman Ron Paul, a Republican senator and presidential candidate, sharing a fascinating insight about why he believes guns should remain legal despite widespread societal opposition. We also saw some highlights from Janaan Ganiyeh's new movie "The Happiness Club" and got a sneak peak of Supergirl actress Kate McKinnon hosting a Q & A session next Tuesday during lunch

falcon: Senator Rand Paul shared a Snapchat video of himself playing Poker with the infamous Dan Blizerian, a professional poker player and self-described international Playboy. The duo played Liar's Poker, a game that uses cash instead of cards. Blizerian is known for his lavish lifestyle and weapons collection, which he calls his "accessories." Blizerian once hurled an 18-year-old porn star from the roof of his mansion into the pool, leaving her with a broken foot. Blizerian was arrested at Los Angeles International Airport for allegedly trying to manufacture a bomb and was kicked off Snapchat for posting videos of naked women in his bed.

gpt4: Rand Paul, a Republican senator and presidential contender, recently shared a Snapchat video of himself learning how to play poker from the professional poker player, Dan Blizerian. The pair were seen playing Liar's Poker, a game that uses dollar bills instead of cards and involves players tracking the serial numbers on the banknotes.

Document 11

bart: bayern munich beat hoffenheim 2-0 at the allianz arena on saturday . sebastian rode and andreas beck scored for the german champions . bayern are 13 points clear of second-placed vfl wolfsburg . pep guardiola 's side host porto on tuesday in champions league .

bert_sum: bayern munich beat hoffenheim 2-0 in the bundesliga on saturday . sebastian rode scored the opening goal from close range . hoffenheim 's andreas beck added an own goal in stoppage time . borussia moenchengladbach drew 0-0 at eintracht frankfurt on friday .

bus: bayern munich beat hoffenheim 2-0 on saturday in the champions league quarter-final second leg . sebastian rode scored in the minute and hoffenheim 's andreas beck claimed an own goal . borussia moenchengladbach drew 0-0 at eintracht frankfurt on friday .

pgn: bayern munich beat hoffenheim 2-0 on saturday . sebastian rode scored in the 38th minute and hoffenheim 's andreas beck added an own goal in stoppage time . philipp lahm , arjen robben , franck ribery and bastian schweinsteiger scored the opening goal .

s2s: bayern munich beat hoffenheim 2-0 in the bundesliga on sunday . the german champions beat UNK wolfsburg in the champions league . the german champions beat UNK wolfsburg in the champions league . the german champions beat UNK wolfsburg in the champions league .

stablelm: The news article contains details about recent football-related events such as injuries suffered by Bayern Munich club, transfers rumors affecting other clubs like

Hoffeinshöhe (Zurich), competition between different teams during European qualifying rounds, and updates on potential replacements or additions being brought forward due to departures or unexpected issues. Additionally it touches upon general updates regarding sports leagues around Europe, particularly Germany where Bayern have struggled lately despite strong performances under Jupp Heynckes. Overall this report covers

falcon: Bayern Munich, despite a list of absent and injured starters and a recent Champions League game loss to Porto, won 2-0 over Hoffenheim, with Sebastian Rode and Andreas Beck scoring a goal each. With this win, five matches left to play, and a 13-point lead over VfL Wolfsburg, Bayern Munich is closing in on another Bundesliga title.

gpt4: Bayern Munich defeated Hoffenheim 2-0 despite multiple injuries and a midweek Champions League loss. Sebastian Rode scored the first goal, and Hoffenheim's Andreas Beck scored an own goal, resulting in a 13-point lead over second-placed VfL Wolfsburg. Bayern played without several starters, but goalkeeper Manuel Neuer displayed excellent form. Other Bundesliga matches included Bayer Leverkusen's 4-0 victory over Hanover 96, Borussia Dortmund defeating Paderborn 3-0, and Borussia Moenchengladbach's 0-0 draw against Eintracht Frankfurt.

Document 12

bart: wayne rooney suffered a knee injury during manchester united 's defeat by everton at goodison park on sunday . the england international was replaced by robin van persie in the closing stages . louis van gaal confirmed after the game that rooney had sustained the injury but that it was too early to know the severity .

bert_sum: wayne rooney suffered a knee injury during the defeat by everton . the england international was seen receiving treatment on the bench . louis van gaal said it was too early to know the severity of the injury . united suffered back-to-back league defeats for the first time since van gaal took over as everton climbed into the top half .

bus: louis van gaal says wayne rooney suffered a knee injury during the defeat by everton at goodison park on sunday . the england international was replaced by robin van persie in the closing stages of the premier league clash . united suffered back-to-back defeats for the first time since van gaal beat everton .

pgn: wayne rooney suffered a knee injury during the defeat by everton . the england international was replaced by robin van persie in the 88th minute . van gaal said that rooney had sustained the injury but that it was too early to know the severity .

s2s: wayne rooney limped off in the closing stages of the premier league clash . the england international was replaced by robin van persie in the closing stages of the premier league clash . the england international was replaced by robin van persie in the closing stages .

stablelm: The Manchester United manager, Louis van Gaal, announced two new substitutions – one with Robin van Persie replacing Robin van Persie who went down injured

due to illness against Everton, while another substitution saw midfielder Jonathan de Jong step up from midfield to replace Yves Bédelot at centre-half. He then added that there are no known injuries or setbacks currently affecting any player under the age of 40, making this statement both reassuring and optimistic towards their chances going forward. Additionally, former

falcon: Wayne Rooney sustained a knee injury during Manchester United's defeat to Everton and was replaced by Robin van Persie in the closing stages of the Premier League clash. Van Gaal confirmed that Rooney had sustained the injury, but it was too early to know the severity. Manchester United suffered back-to-back league defeats for the first time since Van Gaal took over as Everton climbed into the top half. The defeat leaves United fourth in the race for automatic Champions League qualification, and the Rooney injury is also a concern.

gpt4: Manchester United manager Louis van Gaal confirmed that Wayne Rooney suffered a knee injury during their defeat by Everton at Goodison Park. Rooney was replaced by Robin van Persie during the game and received treatment on the bench. However, the severity of the injury remains unknown.

Document 13

bart: beatrize carrion-moore , 23 , is accused of kicking a deputy in the groin . police were called to boonies bar in west palm beach on friday night . the bar manager told police the woman was ' extremely intoxicated ' she was charged with trespassing after warning , resisting arrest with violence and battery on a law enforcement officer .

bert_sum: police were called to boonies bar in west palm beach on friday night after getting a report about a woman who was offering oral sex in exchange for money . the bar manager also told police the woman , beatrize carrion-moore , was ' extremely intoxicated ' . carrion-moore was previously arrested in february and charged with aggravated battery and resisting an officer without violence .

bus: beatrize carrion-moore is accused of kicking a deputy in the groin and thigh several times before offering him oral sex in exchange for money and refusing to leave the establishment . the bar manager said she had a similar problem with carrion-moore , 23 , the week before . she was charged with aggravated battery and resisting an officer without violence .

pgn: police were called to boonies bar in west palm beach on friday night after getting a report about a woman who was offering oral sex in exchange for money . the bar manager said she had a similar problem with carrion-moore , 23 , the week before , according to wesh . carrion-moore carrion-moore is accused of kicking a deputy in the groin several times before offering him oral sex in exchange for a get out of jail free card .

s2s: UNK UNK is accused of assaulting a deputy in the groin and thigh several times before offering him oral sex . the bar manager said she had a similar problem with UNK , 23 , the week before .

stablelm: The unnamed woman has been reported to police after allegedly offering sexual services in exchange for money during incidents at two bars - Boonies Bar in West Palm Beach and Boozee's Cafe in nearby Broward County. In response, police are investigating allegations of assaulting a deputy, as well as giving physical assistance while trying to remove the alleged perpetrator. Although no arrests have yet been made, there appears to remain much work to be done if these cases turn into criminal proceedings against women attempting to seek

falcon: A woman in Florida was arrested for offering oral sex in exchange for money at a bar. The woman, identified as Beatrize Carrion-Moore, became angry when police were called and refused to leave. During the scuffle with the authorities, she kicked a deputy in the thigh and groin multiple times and continued offering oral sex to the deputy while in the back seat of the patrol car. Carrion-Moore was charged with multiple crimes and has a previous arrest in February for aggravated battery and resisting an officer.

gpt4: A Florida woman, Beatrize Carrion-Moore, was charged with trespassing, resisting arrest with violence, and battery on a law enforcement officer after reportedly kicking a deputy in the groin and thigh multiple times and offering him oral sex in exchange for her release. The incident took place at Boonies Bar in West Palm Beach, where Carrion-Moore was accused of offering oral sex for money and refusing to leave. Carrion-Moore was released from Palm Beach County Jail after posting a \$6,000 bond.

Document 14

bart: marc leishman has withdrawn from this year 's masters chairman billy payne has ruled out starting a women 's masters despite talking about growing the game . jack nicklaus showed he 's still got it at the age of 75 with a hole in one .

bert_sum: marc leishman has withdrawn from this year 's competition to be with his wife . billy payne talked a lot about growing the game but he ruled out any idea of starting a women 's masters . masters chairman billy payne has promised a ' beautification ' project .

bus: marc leishman has withdrawn from this year 's competition to be with his wife . his wife audrey spent part of last week in an induced coma owing to a serious infection . masters chairman billy payne has withdrawn out any women 's masters .

pgn: australian marc leishman has withdrawn from this year 's competition to be with his wife . his wife audrey spent part of last week in an induced coma owing to a serious infection . leishman left augusta taking with him the best wishes of his sport .

s2s: australian marc UNK has withdrawn from this year 's masters . he has withdrawn from this year 's competition to be with his wife . UNK left augusta national in an induced coma owing to a serious infection .

stablelm: Australian athlete Marc Leishman withdrew from the Australian Masters due to health issues and their daughter Audrey was admitted with pneumonia earlier this month

leading to her father leaving the event entirely however, some believe there are plans to transform the course inside and improve upon its natural beauty.

falcon: Marc Leishman withdrew from this year's Masters competition to be with his wife Audrey, who was in the hospital due to serious infection. Billy Payne ruled out starting a women's Masters due to the short member season and the extensive time dedicated to the preparation and conduct of the tournament. Aside from this, Payne promises a beautification project of Augusta National. Bubba Watson's menu for the Champions Dinner on Tuesday was Caesar salad with grilled chicken breast. Arnold Palmer and Niall Horan were seen shaking hands, and a mounted cross-section of the Eisenhower Tree was unveiled. Jack Nicklaus had a hole-in-one at the fourth, but Camilo Villegas was a winner with two aces.

gpt4: Australian golfer Marc Leishman withdrew from the Masters to be with his wife, who was in an induced coma due to a serious infection. Masters Chairman Billy Payne discussed growing the game but ruled out starting a women's Masters, citing the existing time commitment and that Augusta National would likely not host another. Payne also plans to beautify the surroundings outside of Augusta National. Golfer Bubba Watson served a simple meal for the Champions Dinner, and favorite iconic images this week featured Arnold Palmer and Niall Horan together. Finally, the Eisenhower Tree, removed after an ice storm, was commemorated with a mounted cross-section and golfer Jack Nicklaus showed his skills at age 75 during the Par-3 contest.

Document 15

bart: a former heavyweight champ 's gaudy , abandoned mansion . the tallest and fastest " giga-coaster " in the world . a professor of physics asks 100 people to create a composite with facial features they thought were beautiful . a dramatic interview with a famed spiritual leader .

bert_sum: mike tyson lived in a southington , ohio , mansion with an indoor swimming pool . he sold the house in 1999 it 's due to become , of all things , a church . alisyn camerota looked into allegations that bikram choudhury sexually assaulted six former students .

bus: former heavyweight boxer mike tyson sold the house in 1999 . he was part of bikram yoga creator 's religious freedom law . he is a former heavyweight champ .

pgn: the fury 325 can be found at carowinds amusement part in charlotte , north carolina . the home featured an indoor swimming pool , a jacuzzi (with mirrored ceiling , naturally) and an entertainment room large enough for small concerts .

s2s: mike tyson , a former heavyweight boxer , is a former heavyweight boxer . the former heavyweight champ abandoned his home in a UNK , ohio , home . the boxer was a former heavyweight boxer .

stablelm: The text provides news articles related to various topics including roller coasters, famous humans, Spiritual Leaders, and athletes. Articles include personal accounts, interviews, clips from movies, and historical events happening right now. The tone throughout each article is informative

and engaging, making it easier readers have fun reading through them. Furthermore, most articles focus around specific individuals' lives within these fields without overwhelming overkill.

falcon: Some of the best videos of the week include a tour of Mike Tyson's former luxurious mansion in Ohio, the Fury 325 giga-coaster in North Carolina, an interview with former Bikram yoga students accusing creator Bikram Choudhury of sexual assault, a witness account of the Germanwings Flight 9525 crash, a discussion on the controversial Indiana religious freedom law, and a British university physics professor conducting a study on what facial features people find beautiful.

gpt4: The text talks about some of the best videos of the week, featuring a visit to Mike Tyson's abandoned mansion in Ohio, a ride on the world's tallest and fastest "giga-coaster," an interview with Bikram Choudhury and former students accusing him of sexual assault, a witness account of the Germanwings Flight 9525 crash, a discussion on the Indiana religious freedom law with Penn Jillette, and a study on facial features and attractiveness.

Document 16

bart: a new york judge issued an opinion authorizing service of divorce papers on a husband completely via facebook . the idea of service by facebook seems to offend traditional notions of ensuring notification of a defendant of a case against him . online service may be a new frontier , but it 's not unheard of .

bert_sum: danny cevallos : a new york judge authorizing service of divorce papers on a husband via facebook . cevallos : in-person service is not always possible , for obvious reasons . he says it 's defendants – no wait , all of humanity – that ca n't be trusted .

bus: serving people with legal papers is " service of process " in new york . judge : if you are going to sue , you should know about it . he says it 's an opinion authorizing service of divorce papers on a husband .

pgn: a new york judge issued an opinion authorizing service of divorce papers on a husband completely via facebook . serving people with legal papers is an industry and its own body of law premised on one guiding principle .

s2s: a new york judge rules that divorce papers on a husband ca n't be trusted . if you are going to sue someone , you should at least let them know about it . if you 're going to sue someone , you should at least let them know about it .

stablelm: 1. A CNN journalist recently gave testimony during US federal hearings regarding privacy issues surrounding social media platforms such as FB and Twitter leading up to her decision to use technology instead of mailing documents due to cost concerns. 2. Some argue that allowing lawyers through these platforms would lead to faster processing times without any real benefits other than making individuals afraid about their phones potentially ringing when receiving paperwork. 3. Despite popular belief, working oneself into such debt simply to reach outside counsel does come cheaply provided that she

somerset said his dog dragged him back indoors . mr cooper , from UNK , somerset said his dog dragged him back indoors .

stablelm: Terry Cooper lives near a rural community where he recently visited another retiree who encountered one of these beasts during her daily walks around her house block she is now staying away for several weeks because she doesn't want anyone getting too close.

falcon: Terry Cooper, a pensioner from Somerset, was saved by his Jack Russell, Sam, after a badger, the size of a pig with six-inch-long teeth, entered his garden. The badger was accompanied by two cubs. Mr Cooper thinks he could have been attacked if his dog had not been there. The National Farmers' Union advised residents to protect their property from invading badgers. People can use badger-proof fencing, keep lids on their bins, and avoid leaving food or water bowls out in the open.

gpt4: A 79-year-old pensioner, Terry Cooper, was saved by his Jack Russell dog, Sam, from an unusually large badger with six-inch teeth that appeared in his garden alongside two cubs. Fortunately, Cooper was not attacked and didn't require medical attention. Following the encounter, the National Farmers' Union issued advice for residents on protecting their property from badgers, including badger-proof fencing and keeping pet food and water bowls indoors.

Document 19

bart: blackburn were knocked out of the fa cup by liverpool on wednesday . manager gary bowyer believes the club should brace themselves for interest in their players . bowyer has rejected advances for jordan rhodes and rudy gestede in the last two transfer windows . the blackburn boss joked he is considering starting goalkeeper simon eastwood in attack against reading this saturday .

bert_sum: gary bowyer believes the club should brace themselves for interest in their players from top-flight clubs following their run to the fa cup quarter-finals . the championship side were eliminated from the competition on wednesday night by brendan rogers ' liverpool , who booked a semi-final date at wembley at the second time of asking . blackburn strikers jordan rhodes and rudy gestede have both attracted in the last two transfer windows .

bus: sky bet championship side rovers were eliminated from the competition on wednesday night by brendan rogers ' liverpool . blackburn manager gary bowyer believes the club should brace for interest in their players . bowyer is no stranger off interest .

pgn: championship side rovers were eliminated from the competition on wednesday night . brendan rogers ' liverpool booked a semi-final date at wembley at the second time of asking thanks to philippe coutinho 's angled 70th-minute finish . gary bowyer believes the club should brace themselves for interest in their players .

s2s: gary bowyer believes philippe coutinho 's side can win the fa cup . but gary bowyer believes the club should brace themselves for interest . gary bowyer believes the club should brace themselves for interest .

stablelm: The Guardian highlights various aspects related to soccer news such as transfers, performance data, team dynamics - including the ongoing challenge regarding securing sufficient funds due to financial fairplay limitations imposed upon certain teams by national regulators - and how fans continue to demand quality personnel regardless or whether there'll ever materialize any new signings/signing bonuses for these "underdogs" financially struggling teams like Leicester City FC and Burnley.

falcon: Blackburn Rovers manager, Gary Bowyer, predicts that his team's recent FA Cup success could result in many of his players attracting interest from Premier League teams. Bowyer expects other clubs to approach Blackburn to inquire about their players but has stressed the importance of the players maintaining their form. Blackburn's owners have previously refused to let go of their players, but this may change as the club attempts to reduce its debts in order to conform to the Financial Fair Play regulations.

gpt4: Blackburn Rovers' manager, Gary Bowyer, has said that the club should prepare for interest in their players from top-flight clubs after their performance in the FA Cup quarter-finals. Although eliminated from the competition by Liverpool, Rovers had previously beaten Premier League sides Swansea and Stoke. Bowyer believes that the players' ability to perform at a high level has likely caught the attention of bigger clubs planning their summer signings. In the past, he has rejected advances for players such as Jordan Rhodes and Rudy Gestede but expects even more interest in his team in the coming months.

Document 20

bart: karim benzema has been ruled out of real madrid 's clash with malaga . the frenchman limped out of training on friday with a knee problem . carlo ancelpotti says benzema will be fit for the return game on wednesday . james rodriguez and isco are likely to replace the frenchman .

bert_sum: carlo ancelpotti has revealed karim benzema will miss real madrid 's clash with malaga on saturday . the frenchman picked up a knee injury in tuesday 's 0-0 champions league quarter-final first leg draw at atletico madrid . real are two points behind la liga leaders barcelona , who host fourth-placed valencia earlier on saturday .

bus: carlo ancelpotti has revealed karim benzema will miss real madrid 's la liga game against malaga on saturday . the frenchman limped out of training with a knee problem . real madrid striker benzema has scored 15 goals during the la liga campaign .

pgn: carlo ancelpotti has revealed karim benzema will miss real madrid 's la liga game at home to malaga on saturday . the frenchman limped out of training with a knee problem . real madrid striker karim benzema has been ruled out of his side 's la liga match against malaga .

s2s: carlo ancelpotti has revealed karim benzema will miss real madrid 's game at home to malaga on saturday . the frenchman has been ruled out of the club 's la liga match against atletico madrid on saturday . carlo ancelpotti 's side face malaga on saturday .

stablelm: Karim Benzema's decision regarding Carlos Tevez's future came under scrutiny following reports indicating he might take an indefinite leave from playing due to surgery on his left Achilles tendon. It was suggested as early as yesterday afternoon fans would continue hoping Diego Lopez suffered no setbacks so they can see Benoit Assou-Ekotto taking over sooner rather than later. However, it seems unlikely Luis Enrique or the coaching staff will let go despite assurances made last week stating "he

falcon: Karim Benzema will not be playing for Real Madrid in their La Liga match against Malaga as he picked up an injury during a Champions League match against Atletico Madrid. The striker is expected to recover in time for the return game. Isco is likely to take his place and James Rodriguez will operate as a central attacking midfielder. Real Madrid is currently two points behind La Liga leaders Barcelona and Isco is expected to help fill Benzema's void in the team.

gpt4: Carlo Ancelotti has confirmed that Karim Benzema will not play in Real Madrid's La Liga game against Malaga on Saturday due to a knee injury sustained in the Champions League quarter-final against Atletico Madrid. However, Benzema is expected to be fit for the return game on Wednesday. In his absence, Isco is likely to start alongside James Rodriguez in Saturday's match. Real Madrid is currently two points behind La Liga leaders Barcelona.

APPENDIX E

MANUAL EVALUATION EXTRACTIVENESS XSUM

Article 1

Defence attaché Capt Oleg Kornienko presented Ushakov medals to the men of the Arctic Convoys. Able seaman Charles Lewis, 91, from Calmore, Hampshire said he was "honoured and delighted". The Foreign Office initially did not allow Russia to honour the veterans as it broke the rules on foreign medals. The Russian embassy said the convoys allowed Russian soldiers to defeat the Germans on the Eastern Front. Mr Lewis said: "I certainly don't see myself as a hero - far from it - and I think all my friends and colleagues, many of whom sadly are no longer with us, would regard us as simply doing our bit for the war effort." Although I have to say that Churchill's famous quote about the Arctic convoys' missions through the North Sea in winter as being 'the worst journey on earth' did strike a chord - then and now." They consisted of vessels which travelled from Britain to northern Russia to resupply the Soviet Union during World War II. Under constant threat of attack by German U-boats and aircraft, the craft also had to deal with severe cold, storms, and ice floes. By May 1945, the Arctic route had claimed 104 merchant and 16 military vessels. Thousands of Allied seamen lost their lives. Source: National Maritime Museum. Minister-Councillor Alexander Kramarenko from the Russian Embassy said: "It [the ceremony] is done for them because they have been expecting for a very long time to be recognized." It is done for us as it was important for us." The Arctic convoy was a material expression of our wartime alliance and they provided weapons and ammunition that armed us to defeat the Germans on the Eastern front. "We and Britain were in it together." More than

3,000 men died during the maritime campaign that Winston Churchill was said to have called the "worst journey in the world". The Russian Embassy wrote to survivors in May 2012 that it intended to award them with the Medal of Ushakov as a symbol of the country's gratitude. But the move was blocked by the Foreign Office as its rules stated British soldiers could not receive a foreign medal if the act happened more than five years ago. The British Government created the Arctic Star in 2012 following a long campaign for the convoy veterans to be recognised. But campaigners believed veterans should still be able to receive the Russian accolade. The Foreign Office allowed an "exception to the rules" last year and President Vladimir Putin presented the first medals during his visit to London on 16 June 2013. The Ushakov medal is awarded to sailors who have displayed courage in the course of defending Russia or its interests.

BERTS2S: a world war two veteran has been presented with medals awarded by Russia for their role in the Arctic convoys.

TConvS2S: the Russian ambassador to the Russian ambassador has been awarded the freedom of the country.

PtGen: the Arctic embassy in the Arctic have been given the chance to honour Russian soldiers for the first time since the Arctic convoys missions.

TransS2S: a world war two veteran has been awarded medals for Russian military veterans after a world war two veteran's war two aircraft.

Article 2

The police allege 28-year-old Nazrul, also called Nazu, was the man who raped the 74-year-old nun. The man, allegedly a Bangladeshi citizen, was arrested at a railway station in Kolkata (Calcutta). The incident caused outrage in India and led to street protests in many cities across the country. This marks the sixth arrest in the crime. Police have identified eight suspects in the attack, of which two still remain untraced." All the five accused arrested earlier have named Nazrul as the person who raped the nun," a police officer told the BBC. Dilip Kumar Adak, deputy inspector general of the state's Criminal Investigation Department, told the AFP news agency that police acted on a tip off that Nazu, who had fled to Bangladesh after being identified from CCTV footage of the attack, was due to return to Kolkata by train. In May, the police said they had arrested Milon Sarkar and described him as the leader of the gang which attacked the convent. They said it was not clear what role the man had played in the case. During the attack in Ranaghat town on 14 March, money was stolen from the convent school and the building ransacked, before the nun was raped in the convent itself.

BERTS2S: Indian police have arrested a man in connection with the rape of a nun in the eastern city of Kolkata (Calcutta) in March.

TConvS2S: a man has been arrested in connection with the rape and murder of a man in the northern Indian state of Uttar Pradesh.

PtGen: a man has been arrested in connection with a gang rape in a convent in Kolkata (Calcutta), police have said.

Trans2S: police in the indian city of kolkata (calcutta) have charged a man with raping a nun in their home town of kolkata (calcutta).

Article 3

Her Royal Highness took the baton at the Hamburg Symphony Orchestra's new home, the Elbphilharmonie concert hall. It was part of an event where 250 schoolchildren were introduced to music in front of the Duke and Duchess. The Royal party are due back in the UK on Friday evening, after a five-day tour that began in Poland on Monday. On the last day of the tour, the Duchess, who comes from a musical family and played the flute at school, took the musicians through the first notes of Beethoven's fifth symphony. The royal couple then listened to a performance of a symphony especially adapted for young people, and then went on stage to learn more about the orchestra. Earlier, Catherine and William walked around the stage accompanying six children trying out different instruments, including a violin, flute, timpani [kettle drums] and trombone. Later the Duke of Cambridge gave Prince George and Princess Charlotte a guided tour of a helicopter at the Airbus factory in Hamburg. Prince George tried on a pilot's helmet while Princess Charlotte played with buttons in the cockpit. The Duke flew helicopters when he was an air ambulance pilot. Antoine van Gent, head of flight testing at Airbus, said: "The Duke was very relaxed showing his children the professional aircraft he uses." George was excited, with the first helicopter he wanted to sit in the cockpit and then he wanted to sit in the next one, he already knew there was a difference between them." Charlotte was less impressed after taking a tumble on the concrete after leaving the aircraft but Mr van Gent said she enjoyed playing in the pilot's seat.

BERTS2S: the duke and duchess of cambridge have taken part in a concert in the german city of hamburg.

TConv2S: the duke and duchess of cambridge have joined the duke and duchess of cambridge for the first time.

PtGen: the duchess of cambridge has paid tribute to the duke and duchess of cambridge, george charlotte, who has died at the age of 88.

Trans2S: the duke and duchess of cambridge have taken part in a final tour of poland for their children's service.

Article 4

Wenger, whose club contract runs to the summer, has been linked with the job. Southgate succeeded Sam Allardyce in September and oversaw England's 2-0 win against Malta as well as the disappointing 0-0 draw with Slovenia on Tuesday. Wenger said Southgate "has accumulated experiences and is a good observer of what's going on in the national team." Of his own links to the role, the 66-year-old Frenchman added: "I focus on my job, that's what it is about." Former England international Southgate, 46, has a World Cup qualifier against Scotland at Wembley on 11 November and a friendly against Spain four days later to persuade the Football Association to give him the job on a permanent basis. Wenger continued: "We're in a job where you get questioned, especially if you have less experience, and the only answer he can give is by

showing the decisions he makes are right." "There's always a demand for big names but the most important thing is the person's competency. He has a good opportunity and I'm happy he has a chance to show he has the quality." Speaking before his side's game against Swansea on Saturday, Wenger added that 21-year-old defender Hector Bellerin would be offered a new contract by the club. The Arsenal right-back has been linked with a move, with both Barcelona and Manchester City said to be interested, but Wenger said he was "surprised" by the speculation surrounding the Spain international, who progressed through Barcelona's La Masia academy before joining Arsenal in 2011. "He has nearly three years to go and we always try get the players to the level where we think they deserve to be. So we try to extend, certainly, his contract," said Wenger. "For me, his long-term future is here." Wenger added that midfielder Aaron Ramsey, 25, will miss "one or two" more games, having injured a hamstring during Arsenal's opening-day Premier League defeat by Liverpool on 14 August. Subscribe to the BBC Sport newsletter to get our pick of news, features and video sent to your inbox.

BERTS2S: england manager arsene wenger has backed garth southgate to be the next england manager.

TConv2S: arsenal manager arsene wenger says he is "very happy" about his future as manager of the club.

PtGen: arsenal manager arsene wenger has confirmed he is happy with the role of the club's role in the premier league.

Trans2S: arsenal manager arsene wenger says his future at the club is "not good enough", despite speculation linking him with a permanent job.

Article 5

We are aware that this interactive video may not work on some older browsers. You can find a non-interactive version of the content below. In this interactive video, Robert Opie - the founder of the Museum of Brands, Packaging and Advertising in London - outlines some of the main reasons why we change what we buy. The arrival of the domestic freezer, of washing powder, and of cheaper imports all affected the sort of items on sale and popular among UK consumers. The interactive elements to this video also explore how ownership and ranking of some of the biggest brands in the UK is changing. The video works best in full screen. Click on the links below to read the extra information used in this video in a non-interactive format: What in the world does China own? Top tips on freezing food, BBC Good Food Annual survey ranking the top 20 consumer brands in the UK. Museum of Brands, Packaging and Advertising. Let us know what you think of this new format. Required field. If you are happy to be contacted by a BBC journalist please leave a telephone number that we can contact you on. In some cases a selection of your comments will be published, displaying your name as you provide it and location, unless you state otherwise. Your contact details will never be published. When sending us pictures, video or eyewitness accounts at no time should you endanger yourself or others, take any unnecessary risks or infringe any laws. Please ensure you have read the terms and conditions. Terms and conditions

BERTS2S: the bbc has released a series of information about the uk's food market and how they have changed since the start of the decade.

TConvS2S: we've been using your questions to tell us what you want to know about the west midlands.

PtGen: the world's biggest food video may not be able to buy a domestic food freezer, according to a video released by the bbc.

TransS2S: we all know what we know about the world of food and technology.

Article 6

Stuart Campbell was arrested in the west of England on Friday following a complaint from a woman in south London. She had made allegations of harassment taking place over a two-year period. Mr Campbell, who was released on bail, said it concerned some tweets and insisted they were not threatening. He accused the media of "innuendo" designed to encourage "speculations". The blogger, a former computer games reviewer who was born in Stirling but lives in Bath, has been a vocal campaigner for Scottish independence and launched the Wings Over Scotland blog in 2011. On Friday he tweeted that he would be posting less frequently than usual because of "reasons totally outwith my control (don't ask)". End of Twitter post by @WingsScotlandA spokesman for the Metropolitan Police said: "Police are investigating an allegation of online harassment." The allegation was made after a woman, aged in her 30s, attended a south London police station. The harassment is said to have taken place over the past two years. "Mr Campbell has been bailed, pending further inquiries, to a date in mid-September. In a statement on the Wings Over Scotland website, Mr Campbell responded to a report of his arrest which appeared in The Herald newspaper. He said that piece "has been written for maximum innuendo to allow the wildest speculations on social media - which are of course duly taking place - but the alleged events relate entirely to some tweets from our Twitter account, none of which have been deleted and all of which are still publicly visible." "Nothing more sinister or serious than some tweets has occurred or been alleged to have occurred. None of the tweets involved are in ANY way threatening, not even in a joking sense. That's all we'll be saying on the subject at this time."

BERTS2S: police have arrested a blogger on suspicion of online harassment over an allegation of online harassment.

TConvS2S: a west london politician has been arrested on suspicion of stalking and harassment.

PtGen: a police blogger has been arrested in connection with online harassment, the metropolitan police has told the bbc.

TransS2S: a pro-independence blogger has been arrested in connection with an alleged tweet posted on twitter.

Article 7

Media playback is not supported on this device World champion Hamilton started from the back, suffered a broken front wing on lap one and did five pit stops. The Briton faces the

prospect of trying to overhaul a 36-point deficit in the remaining 18 races of the season. Ferrari's Sebastian Vettel came second despite colliding with team-mate Kimi Raikkonen on the first lap. The incident damaged both cars, with Raikkonen fighting back to fifth behind the Red Bulls of Daniil Kvyat and Daniel Ricciardo. Raikkonen's rise through the field included passing Hamilton in the closing stages, Ferrari's decision to use soft tyres for the final stint as opposed to the mediums chosen by Mercedes giving the Finn an advantage as the Briton fought in vain to try to pass the Williams of Felipe Massa. Hamilton will express his confidence at being able to erode Rosberg's points lead and will expect his awful luck of the first three races to turnaround - but the fact remains that no driver who has won the first three races of a season has failed to go on to claim the World Championship. The race was shaped by an incident-packed opening, with a number of collisions on the first lap leading to the introduction of the safety car on lap four. By that stage, Hamilton and Raikkonen had already pitted once to replace front wings damaged in first-corner collisions. Raikkonen moved right into Vettel's path as the German moved left to ensure he was not tagged by an aggressive dive down the inside by Kvyat. Hamilton was then turned into by Sauber's Felipe Nasr as he avoided Raikkonen's damaged car swerving back on to the circuit. Vettel confronted Kvyat about the incident in the pre-podium room after the race, saying: "If I don't go left, you crash into us and we all three go out - you are like a torpedo." Kvyat replied: "We didn't crash." Vettel responded: "You didn't. You were lucky this time." Kvyat shrugged: "I'm on the podium you're on the podium." In the post-race interviews, Kvyat added: "You see the gap, you go for it on the inside. You see one car it is hard to see both. It was a risky move, I agree with Seb but you have to take risks and I am on the podium." Listen to Vettel's post-race rant At the front, Ricciardo had made a stunning start to take the lead from pole-sitter Rosberg at the first corner. In the process his left-rear wheel had been tagged by Rosberg's front wing. Whether it was that or the debris on the track, Ricciardo suffered a left-rear puncture on lap three, allowing the Mercedes to pass. Ricciardo would almost certainly not have been able to hold off Rosberg in any case, but his removal from the lead left the German serene out front. Mercedes decided not to stop him under the safety car and when the race re-started on lap nine, Rosberg simply drove off into the distance. Listen to the first-lap madness Hamilton, meanwhile, was in the wars. Mercedes stopped him twice under the safety car, doing only one lap on the super-soft tyres in an attempt to ensure he could use only the softs for the rest of the race. So Hamilton had done three stops within the first six laps and he was lying third from last at the restart. He began his fightback from there, with the Ferraris doing the same - Vettel starting from 11th and Raikkonen not far behind Hamilton, who said his progress had been affected by damage to the car in the incident with Nasr. "The car was pretty damaged," said the 31-year-old afterwards. "I'm sure it was some aero components but I think it was the suspension as well because it was flexing all over the place." Mercedes F1 boss Toto Wolff said the front wing had damaged the leading edge of the floor and the barge boards when it came off and

that the car had lost a "lot of aerodynamic performance". After a further switch to the medium compound tyres to last the race, Hamilton was not able to recover as much ground as the Ferraris and faces an uphill battle if he is to stop Rosberg winning his first world title this year. Chinese Grand Prix race results Chinese Grand Prix coverage details

BERTS2S: lewis hamilton's title hopes suffered a major blow as he crashed out of the hungarian grand prix.

TConvS2S: lewis hamilton set the pace in final practice at the season-ending abu dhabi grand prix.

PtGen: nico rosberg beat mercedes team-mate lewis hamilton in the first practice session for the first formula 1 race of the season.

TransS2S: kimi raikkonen headed a ferrari one-two in final practice at the german grand prix.

Article 8

The venture's backers include Tesla Motors and SpaceX CEO Elon Musk, Paypal co-founder Peter Thiel, Indian tech giant Infosys and Amazon Web Services. Open AI says it expects its research - free from financial obligations - to focus on a "positive human impact". Scientists have warned that advances in AI could ultimately threaten humanity. Mr Musk recently told students at the Massachusetts Institute of Technology (MIT) that AI was humanity's "biggest existential threat". Last year, British theoretical physicist Stephen Hawking told the BBC AI could potentially "re-design itself at an ever increasing rate", superseding humans by outpacing biological evolution. However, other experts have argued that the risk of AI posing any threat to humans remains remote. A statement on OpenAI's website said the venture aims "to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return". "It's hard to fathom how much human-level AI could benefit society, and it's equally hard to imagine how much it could damage society if built or used incorrectly." The statement said AI "should be an extension of individual human wills and, in the spirit of liberty, as broadly and evenly distributed as is possible safely". It said only a tiny fraction of the \$1bn pledged would be spent in the next few years.

BERTS2S: google has pledged \$ 1bn (£800m) to research and research into artificial intelligence (ai).

TConvS2S: ai weiwei, the founder of ai giant ai weiwei, has launched a \$ 1m (£ venture) research project.

PtGen: open ai weiwei has unveiled a \$ [UNK] (£ [UNK]) venture prize for the spirit of liberty.

TransS2S: one of the world's biggest physicists has warned that artificial intelligence could be used in humans.

Article 9

It works by looking for a combination of "markers" in the blood which are different in healthy people and those with the disease. Delegates at the Alzheimer's Research UK Conference heard that the University of Nottingham is now developing a quick and easy test to do in clinics. It could mean much earlier diagnosis and better treatments, they said. The

test uses some proteins that have been strongly linked with Alzheimer's disease, such as amyloid and APOE. But through careful analysis of blood from people with the disease, as well as those with early-stage memory problems, the researchers detected some other markers that were suggestive of the disease. Most notably, some proteins related to inflammation seem to have been added to increase the power of the test. Prof Kevin Morgan from the University of Nottingham said they still had to validate the test and it could be a decade before it was used in patients. But he added that the combination of markers they had found was looking very promising. "Our findings are exciting because they show that it is technically possible to distinguish between healthy people and those with Alzheimer's using a blood test." As blood tests are a fast and easy way of aiding diagnosis, we are really encouraged by these findings and the potential they hold for the future." He said there were several ways the test could benefit patients, including giving people a definitive diagnosis, which was not always possible at the moment. It could also direct future therapies to make sure patients were getting the most appropriate treatment, he explained. Potentially, it could be a "cheap and easy pre-screen" test which enabled Alzheimer's to be picked up before symptoms appeared, he said. "The way we see it working is you can test people and it will tell them if they have the all-clear, or if they are medium- or high-risk." If they are medium-risk, they can be monitored closely and high-risk patients can be referred to a specialist for more in-depth testing. "Dr Eric Karran, director of Research at Alzheimer's Research UK, said: "Giving people with dementia an accurate diagnosis is not always easy, and so building up our armoury of diagnostic techniques is vital." While there is still some way to go before a test like this could become available, the results are promising. "When used alongside other diagnostic techniques, a blood test like this could be a real help."

BERTS2S: a blood test that can diagnose alzheimer's and alzheimer's disease has been developed by scientists.

TConvS2S: scientists have developed a blood test that can detect the lives of patients with alzheimer's disease.

PtGen: blood tests have been made in clinics to clinics with alzheimer's disease, a study suggests in the journal nature of research suggests.

TransS2S: an international test for alzheimer's disease has been developed by researchers in nottingham.

Article 10

However, there were people in the TV industry who thought the cast would struggle to find work once their time on the show was up. Some of the actors were warned that they wouldn't be able to shake off their sulky, angry, teen characters. But when you look at the careers of Dev Patel and Jack O'Connell for example, they couldn't have been more wrong. Kaya Scodelario was one of the longest-serving members of the cast, appearing in series one to four as Effy Stonem. She went on to appear in sci-fi thriller Maze Runner and is now in Australia filming the next instalment of Pirates of The Caribbean with Johnny Depp. Nicholas Hoult was the only member of the main cast who had serious

acting experience. He'd starred in *About a Boy* with Hugh Grant in 2002. He has since gone on to play Beast in the *X-Men* prequels and will soon be seen starring opposite Kristen Stewart in a film called *Equals*. Dev Patel's first job after playing Anwar Kharral in series one and two, took him from London to India and to the Oscars with *Slumdog Millionaire*. He told *Newsbeat*, he never would have got that role without *Skins*. "My career started off because Danny Boyle's daughter was a fan of the show, that is how I got an audition for *Slumdog*." He has two films out in cinemas at the moment, *The Second Best Exotic Marigold Hotel* with Dame Judi Dench and *Chappie*, starring alongside Sigourney Weaver and Hugh Jackman. Jack O'Connell, who played Cook in series three and four, was hand-picked by Angelina Jolie to star in her most recent movie *Unbroken*. The 24-year-old also picked up the Rising Star Award at this year's Baftas. Fans of *Game of Thrones* will be familiar with Joe Dempsie and Hannah Murray who play Gendry and Gilly respectively. It is, of course, not unusual for an actor from a British television show to find success in Hollywood films. What is unusual is the amount of stars that were discovered from one single teen drama on a digital channel. So how were the makers of *Skins* able to spot so many talented young actors? Co-creator Bryan Elsley told *Newsbeat* they made a conscious decision to hold open auditions. "There are really good kids working from agencies and drama schools but just on that one occasion we decided to go a different way," he explained. "We felt if we had young people playing characters the same age as themselves, and those actors weren't too trained, it would have a more natural feel." It sounds quite risky, but Bryan said the show was such a small project to begin with, they were able to take that approach. Jack O'Connell credits the whole team for the actors' future success. "I think they knew if they were going to identify that genre, that era, that generation, they needed people on the ground level," he said. "A writing team that understood, they got that right first and then I guess they chose to cast exciting people." "I feel very fortunate to be involved in the show but I don't think it was a coincidence that we all came through there." Piers Wenger is the head of drama at Channel 4 and he thinks it was the opportunities the cast had to stretch themselves that helped nurture their talent. "They had the opportunity to do properly complex [roles], rather than just two dimensional ones," he said. "It is extraordinary, in the past soap operas had been the sort of places where young people had been able to do their first job." "That would have been your only break, but time's are changing." "I think thanks to shows like *Skins*, unconventional approaches to casting are becoming more and more common." Follow @BBCNewsbeat on Twitter, BBCNewsbeat on Instagram and Radio1Newsbeat on YouTube

BERTS2S: the stars of tv's game of thrones have been surprised by the news that the stars of the latest series, game of thrones, will

TConvS2S: it's been a difficult year for hollywood fans.

PtGen: the story of game of game of thrones has gone on display at the top of the uk box office.

TranS2S: as the cast of the x factor draws to a close, there's a look at some of the stars who have decided to play in the next

Article 11

Share this with [Email](#) [Facebook](#) [Messenger](#) [Twitter](#) [Pinterest](#) [WhatsApp](#) [LinkedIn](#) Copy this link John Johnston, 27, admitted taking the vehicle without consent and driving while over the legal alcohol limit. He was spotted by a CCTV operator committing the offence and stopped by police in Buccleuch Street. At Dumfries Sheriff Court he was ordered to carry out 100 hours of unpaid work and given the driving ban. Johnston, who had moved from Belfast to Dumfries, committed the offence in September this year. He had been staying in homeless accommodation in the town and took the roller from Brooms Road. He then drove it across town in a bid to get home from a night out drinking with a friend. "I was just trying to find my way home," he told the court at an earlier appearance.

BERTS2S: a homeless man has been banned from driving for three years after he drove away from a pub in dumfries.

TConvS2S: a man has been given a 12-month suspended jail sentence after he was caught speeding in the street in dumfries and galloway.

PtGen: a man who admitted causing the death of a homeless man by dangerous driving has been given a 12-month suspended prison sentence.

TranS2S: a man has been banned from driving for five years after he drove through a dumfries town centre car on the outskirts of belfast.

Article 12

Gareth Colfer-Williams, 25, died last week at his home in Swansea, the city at the centre of an epidemic of the disease which has reached 942 cases. But the examination was unable to establish whether measles was the main cause of his death. An inquest will be opened and adjourned on Tuesday to allow further tests. Since the outbreak began in November, 83 people have needed hospital treatment for the illness. Mr Colfer-Williams' mother, Angela Colfer said her son had recently been treated at Swansea's Morriston Hospital for asthma. A few days after he was discharged he became unwell and developed a rash. Mrs Colfer said her son went to see an out-of-hours GP on 17 April with a rash "from head to foot" but not on his arms. The following morning, Mr Colfer-Williams was found dead at his flat. Public Health Wales said on Friday that laboratory tests confirmed a diagnosis of measles but further tests were needed to determine the cause of death. The family said they had no idea how Mr Colfer-Williams got measles as no close family member has it, including his three-year-old daughter who has been vaccinated.

BERTS2S: a post-mortem examination has concluded that a man died from measles and measles.

TConvS2S: the family of a five-year-old boy who died after suffering serious head injuries have said they are "devastated" by the death of his son.

PtGen: the family of a woman who died after a measles outbreak have said she is no close to his death.

TranS2S: an inquest into a swansea man's death has opened.

Article 13

Share this with [Email](#) [Facebook](#) [Messenger](#) [Messenger](#) [Twitter](#) [Pinterest](#) [WhatsApp](#) [LinkedIn](#) Copy this link Avon and Somerset Police closed the Lower Bristol Road in Bath at Rackfield Place outside the First depot after the crash at 03:00 GMT. The A36 road closure, which was lifted at 14:00 GMT, had caused disruption and delays to commuters and bus services during rush hour. Investigations are under way into how the collision happened. Local Live has all the latest travel news First said that while parts of its services had resumed a normal schedule, it would take a few hours for all services to return to normal. Managing director James Freeman said: "We are hoping we will have a full service running in time for the evening commute." We would like to thank the public for their support and understanding and staff for going beyond their call of duty to help manage the situation and recovery."

BERTS2S: bus firm first first has reopened after a crash which left one of its vehicles with serious injuries.

TConvS2S: a man has been taken to hospital after a crash involving a car and a lorry in bristol.

PtGen: a man has been killed in a crash on the m4 in berkshire, south wales police has said.

TransS2S: a major road has reopened in bristol following a crash involving two commuters.

Article 14

The stone got past the elephant's fence and a ditch separating the animal and visitors, the zoo said in a statement. The girl was taken to hospital and died within a few hours, the zoo added. The zoo statement said the enclosure met international standards and said "this kind of accident is rare, unpredictable and unusual". Africa Live: More on this and other stories The statement went on (in French) to point out two other recent incidents in the US: Phyllis Lee, Scientific Director of the Amboseli Trust for Elephants, says that targeted throwing of stones and branches by elephants is very unusual. "It can happen when elephants are frustrated or bored. In my opinion, it's unlikely the elephant was directly targeting the girl - but exhibiting frustration. You can't predict what animals in captivity will do." The moments after the girl was struck at Rabat Zoo on Tuesday were filmed by a bystander and uploaded onto YouTube. The video shows the elephant waving its trunk behind a fence and swerves round to show a stone on the ground. Metres away people are gathered around the girl, holding her head and stroking her leg.

BERTS2S: an elephant has been hit by a stone at a zoo in western france after it was hit by a tree.

TConvS2S: an elephant has died after being hit by a elephant at a zoo in the us state of idaho.

PtGen: a 12-year-old girl has died after being hit by a elephant at a zoo in rabat, rabat, texas.

TransS2S: an elephant has been killed by an elephant at a zoo in paris.

Article 15

Share this with [Email](#) [Facebook](#) [Messenger](#) [Messenger](#) [Twitter](#) [Pinterest](#) [WhatsApp](#) [LinkedIn](#) Copy this link John Clancy,

Labour councillor for Quinton, beat closest rival Penny Holbrook in Monday night's ballot. Ahead of the vote, he said providing new homes and improving existing social housing should be the authority's "top priority". He will officially replace Sir Albert Bore on 1 December. Mr Clancy, who has bid to become leader of the city council several times, was chosen in the ballot by the authority's 78 Labour councillors. Updates on this story and others on Birmingham and Black Country Sir Albert announced he would stand down as head of the self-styled "largest local authority in Europe". A close call for a man waiting in the wings for more than a decade. John Clancy has become leader of the Labour group after five attempts - and by just one vote. He will officially become leader of Birmingham City Council on 1st December when it's ratified at full council. An English teacher for many years, he's no stranger to working in the public sector but he also had a spell as a solicitor, as well as a business lecturer. His big message at a hustings last week was the council under his leadership can provide smarter financial decisions, political openness and free school meals for all primary age children. Some critics from within the party are already saying they're worried he won't have the clout or the vision to make the huge and rapid improvement needed to keep government commissioners at bay. The former Labour group leader's departure followed two other senior party figures, who resigned after criticising the way city was being run. Ian Ward, the deputy leader of the local authority and councillor Barry Henley also stood for the position. Mr Clancy, a former teacher and lecturer, has also pledged to widen access to free school meals for primary school children and expand regeneration outside the city centre. Birmingham City Council has found itself under scrutiny from a government-appointed independent panel. The panel was appointed at Sir Bob Kerslake's recommendation to oversee changes at the council. Sir Bob's study was commissioned after a series of reports focussed on the council during an inquiry into the so-called Trojan Horse letter.

BERTS2S: the leader of birmingham city council has been elected as the city's new leader.

TConvS2S: the former leader of birmingham city council has been elected as the new leader of birmingham city council.

PtGen: the new leader of the labour party has been elected as the new leader of birmingham city council, replacing sir albert bore.

TransS2S: a birmingham councillor has been appointed as the leader of the city council.

Article 16

The law also permits 12-year-olds to be contracted to work for others. But they need parental authorisation. Vice-President Alvaro Garcia Linera said the new legislation reflects the needs of Bolivia, one of the poorest countries in South America. It also sets harsher punishment for violence against children. The law sets a sentence of 30 years in jail for child homicide. The measure, approved by Congress earlier this month, was signed into law by Mr Garcia in the absence of President Evo Morales, who was travelling to Brazil. "President Evo [Morales] intervened to make sure we found a balance between the reality and the law, between rights and international

treaties," added Mr Garcia. More than 500,000 children already work to supplement the family income in Bolivia according to the United Nations Children's Fund (Unicef). Many work cleaning shoes and selling food in stalls in La Paz and other Bolivian cities. But others face extreme conditions in mines and in agriculture fields. Mr Morales's socialist government hopes the law will help eradicate extreme poverty in Bolivia. The International Labour Organisation says children under the age of 15 should not be allowed to work. But it allows a minimum working age of 14 for developing countries. "It would have been easier to pass a law in line with international conventions, but it would not be enforced because Bolivia's reality has other needs and characteristics," said Mr Garcia. The ILO says it is studying the legislation to decide whether it breaches international regulations on child labour.

BERTS2S: bolivia's parliament has approved a law banning child labour in the country's capital, la paz.

TConvS2S: bolivian president evo morales has signed a law banning children's rights in latin america.

PtGen: bolivia's president evo morales has signed a new law banning children from children under the age of 15.

TranS2S: bolivia's president evo morales has signed a law requiring children to be banned from work in the country.

Article 17

Share this with [Email](#) [Facebook](#) [Messenger](#) [Messenger](#) [Twitter](#) [Pinterest](#) [WhatsApp](#) [LinkedIn](#) [Copy this link](#) The latest threats bring the number of incidents to 69 in 27 states in the past month, according to the JCC Association of North America. No bombs were found at any of the centres targeted with telephone calls. Last week, 27 Jewish community centres in 17 US states reported receiving hoax bomb threats. All of the threats made on Monday, along with previous incidents in January, turned out to be false and Jewish centres have since reopened with normal services resumed. Responding to the latest incidents, White House press secretary Sean Spicer said that US President Donald Trump had made it "abundantly clear" that such actions were "unacceptable." "Hatred and hate-motivated violence of any kind have no place in a country founded on the promise of individual freedom," the statement, which was shared on Twitter by NBC News correspondent Peter Alexander, said. Mr Trump's daughter Ivanka - who has converted to Judaism and whose husband is Jewish - also condemned the threats. Former Democratic presidential candidate Hillary Clinton said that action must be taken to prevent further incidents. In a post on Twitter, Mrs Clinton wrote: "JCC threats, cemetery desecration and online attacks are so troubling and they need to be stopped. Everyone must speak out, starting w/@POTUS." The threats were made to the Jewish community centres through calls that were both pre-recorded and live, with suspects using voice-disguising technology, according to Israeli newspaper Haaretz. Since the beginning of the year, there have been reports of threats to centres in New York, New Jersey, Ohio, Florida, Massachusetts, Maryland, Michigan, Minnesota, Delaware, Connecticut, Alabama, California, Maine, Tennessee, South Carolina, Missouri, Wisconsin, Texas and Kansas. The JCC Association of North America, a network

of health and education centres, has since been reviewing its security plans. In an earlier statement the FBI said that along with the US justice department it was investigating possible civil rights violations in connection with the threats. The Anti-Defamation League, a Jewish anti-bigotry organisation, said in a statement that it was "deeply disturbed" by the latest threats to the Jewish community. Meanwhile, more than 100 headstones have been damaged at a Jewish cemetery in St Louis, Missouri, local media report. In the Canadian city of Toronto, Mayor John Tory has condemned anti-Semitic hate notes left outside the homes of Jewish residents.

BERTS2S: a number of jewish community centres in the us have been threatened by hoax bomb threats, the white house says.

TConvS2S: us president donald trump has condemned threats against jewish community centres in response to a spate of attacks in the us.

PtGen: the number of jewish community centres in the us and us has fallen to its lowest level in recent weeks, the white house has said.

TranS2S: jewish and jewish centres across the us have been targeted by jewish threats, the white house has said.

Article 18

Share this with [Email](#) [Facebook](#) [Messenger](#) [Messenger](#) [Twitter](#) [Pinterest](#) [WhatsApp](#) [LinkedIn](#) [Copy this link](#) The firm said it had signed a Letter of Intent to enter into exclusive negotiations with Liberty House Group. More than 1,700 people are employed in the division, which has factories in Rotherham and Stocksbridge. Steel union Community said it welcomed news of negotiations following "months of unnecessary stress and concern". More on this and other South Yorkshire stories The union's general secretary Roy Rickhuss said: "This is a positive step for the UK steel industry however there remain huge challenges which government must address." The union said it would be seeking urgent talks with Liberty House Group and would be asking what their plans were for investment, protecting jobs and providing decent pensions for members in retirement. Tata Steel's UK boss Bimlendra Jha said the announcement was "an important step forward". "We now look forward to working with Liberty on the due diligence and other work streams so that the sale can be successfully concluded," he said. The Speciality Steels unit makes high-end components for the automotive, aerospace and oil industries. In April, Tata sold its long-products division, based in Scunthorpe, to Greynull Capital, a UK-based investment firm.

BERTS2S: tata steel has confirmed it is in talks with the company about selling its long products division.

TConvS2S: tata steel has announced it is to close its operations in the uk.

PtGen: tata steel has said it is planning talks with liberty house group in the wake of the sale of tata steel.

TranS2S: tata steel has said it is in talks with the uk's largest steel group to sell its operations to the uk.

Article 19

18 March 2014 Last updated at 09:22 GMT Richard Blake-Smith and his partner Anna beat Vytautas Kirkliuskas of

Lithuania, carrying his wife, Neringa Kirliauskiene. Rich said: "We put in a whole week's work, running around the athletics track near where we live - and it's paid off!" "We were pushed hard, especially by the Lithuanians there." Competitors must tackle a 380-metre course, hurdling over haystacks, scrabbling up a steep slope, and dodging water pistols. You don't need to be married to take part. The race saw all combinations of competitors - men carrying women, men carrying men and one woman carrying a woman. Rich and Anna will now compete at the World Championships in Finland in July.

BERTS2S: a british man and his wife have won one of the world's most successful men's athletics championships.

TConvS2S: the wife of one of the world's most famous competitors athletes has said she is "shocked" by the lack of success in the sport.

PtGen: the world's elephants have been celebrating the world's elephants race, which has been at the centre of the world's olympic games.

TransS2S: a pair of paralympian pistols has set up a marathon on track in a bid to break the world championships in lithuania.

Article 20

In an interview with newspaper Die Zeit, Hitzlsperger said he was coming out to advance discussion of a topic that was "simply ignored" in German football. He is the most prominent footballer to publicly reveal his homosexuality and his former team-mates were quick to praise his announcement. Arsenal and Germany forward Lukas Podolski wrote on Twitter: "Respect to Thomas Hitzlsperger! That was a courageous and correct decision and an important step." And retired defender Arne Friedrich, who played with Hitzlsperger at Euro 2008, wrote: "Proud of you Thomas. A good decision and in my opinion the right time." There was also support for Hitzlsperger from leading figures in German football administration. "Finally, a footballer has the courage to make his sexuality public," said Theo Zwanziger, the former president of the German Football Association. Bundesliga president Reinhard Rauball said in a statement: "The decision of Thomas Hitzlsperger to be the first prominent footballer to make public his homosexuality is - even after his active career - a huge and courageous step, and surely leads the way in the fight against homophobia." Gary Lineker @GaryLineker Congratulations to Thomas Hitzlsperger on bravely being the first player to have played in the PL to 'come out' Joseph Barton @Joey7Barton 15m Thomas Hitzlsperger has shown a lot of courage today. Sad times when people have to wait till they retire from their chosen profession.....before the feel other people will judge them solely on who the human being is. Shame on all of us as a society. John Amaechi OBE @JohnAmaechi Congratulations to Thomas Hitzlsperger - welcome to the club! Bayern Munich, the club where Hitzlsperger began his career as a junior, also tweeted their support for the midfielder's "brave step". Hitzlsperger's announcement is seen as a watershed moment in Germany, which like England has struggled with the problem of homophobia in football. Zwanziger admitted that football was "now as ever a hard business, where unfortunately

an open discourse about homosexuality is still not taken for granted". Current Germany captain Philipp Lahm created controversy in 2011 when he wrote in his autobiography that he would encourage gay footballers to keep quiet about their sexuality. "I would not encourage any gay professional footballer to come out," he wrote at the time. "I would fear that he would end up like Justin Fashanu [the English footballer who killed himself after revealing that he was gay]." The last German footballer to come out was second tier player Marcus Urban, who told Die Welt in 2007 that he knew of at least three gay footballers in the Bundesliga who were too scared to reveal their sexuality. German Chancellor Angela Merkel's spokesman Steffen Seibert praised Hitzlsperger for going public. "It's good that he's spoken about something which is important to him," Seibert told a media briefing. "We live in a country where nobody should be worried about making his sexuality known or concerned about intolerance." We judge footballers on whether they have behaved well and worthily on and off the pitch and I believe that both are the case with Mr Hitzlsperger. "Openly Gay German politician Volker Beck said he hoped Hitzlsperger's announcement would encourage others to do the same. "I hope that Hitzlsperger's courageous step contributes to finally breaking down this taboo," he said. "It's now the job of the German FA to embolden active players [to do the same]." And Zwanziger said he was optimistic that Hitzlsperger's announcement would lead to greater openness about sexuality in German football. He said: "I'm confident that sexual orientations will soon not be an issue in football at all."

BERTS2S: german footballer thomas hitzlsperger has come out as gay.

TConvS2S: former german footballer lukas hitzlsperger has announced his retirement from international football.

PtGen: thomas hitzlsperger has become the first german footballer to have his sexuality played in the german premier league, the football association has said.

TransS2S: the football association says it is "deeply disappointed" and "saddened" that footballer thomas thomas has quit football.