Assignment 3: Quantivative and Qualitative evaluation of Generative Models for summary generation

Matteo Manias 1822363

Abstract

Generation of summaries using the LLM model **MINERVA-350M** fine-tuned on the **CC-NEWS** dataset is evaluated in this report with standard metrics like ROUGE as a quantitative metric useful to compare the performance of different fine-tuning steps and the quality of the generated summaries. Further qualitative analysis is provided to overcome the inherent limitations of ROUGE, and quantitative evaluation metrics

1 CC-news dataset

The dataset is used for fine-tuning the Model on text summarization on news articles, the data used for fine-tuning in the CC-NEWS dataset is as follows:

- **text**: the original text to be summarized
- **gold summary** : the human generated summary of the text
- **Label**: the title of the article

2 MINERVA-350M model

The model used for fine-tuning on summary generation is the **MINERVA-350M** which is a small but fast and efficient model suitable for fine-tuning on small datasets.
The model is pretrained on italian and english text.

3 Fine-tuning on the CC-NEWS dataset

Fine-tuning refers to the process of supervised training of a pre-trained model, generally in an unsupervised fashion for LLMs, on a new dataset. This application of transfer learning allows the model to retain the general representations learned during pre-training and use them efficiently on the new dataset, which may not be large or general enough to train a model from scratch.
The model is fine-tuned for 1000 additional steps on the CC-NEWS dataset and checkpoints at 500 steps and 1000 steps are considered for evaluation.

4 Evaluation Metrics

Given the intrinsic nature of natural language, the evaluation of LLM generated content proves a challenging task. Considered and implemente metrics are base on correspondence between the generated summary and the human generated summary at a word level and sequence of word level (ROUGE), or similarity computed at a sentence level (BERTScore). Additional metrics are implemented (bleu, meteor) based on more sophisticated algorithms that take into account more complex relationships between words and sentences.

- **ROUGE** : Recall-Oriented Understudy for Gisting Evaluation,
-
  - **ROUGE-1** : measures the overlap of unigrams.
  - **ROUGE-2** : measures the overlap of bigrams.
  - **ROUGE-L** : measures the longest common subsequence between.
- **BLEU**: Bilingual Evaluation Understudy, it is a weighted geometric mean of all modified n-gram precisions, multiplied a the brevity penalty
- **METEOR**: Metric for Evaluation of Translation with Explicit ORdering,The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. It also has several features that are not found in other metrics, such as stemming and synonymy matching, along with the standard exact word matching.
- **BERTScore**: BERTScore is a metric that computes the similarity between two sen-

tences based on the cosine similarity computed on the BERT embeddings of the sentences.

4.1 Limitations of standard metrics

The standard metrics like ROUGE, BLEU, METEOR are based on the correspondence between the generated summary and the human generated summary Theese approaches are hindered by the following limitations:

- **Lack of semantic understanding**: Most metrics primarily focus on surface-level matching without understanding the broader content and meaning of the summary.
- **Lack of informativeness** : Metrics struggle to evaluate whether key information from the original text has been included or omitted
- **Emphasis on lexical matching**:Metrics like BLEU and ROUGE reward exact word matches, which makes them ill-suited for abstractive summarization, where paraphrasing and semantic equivalence are more important.
- **Lack of coherence and fluency**: None of the metrics sufficiently assess the coherence, readability, or overall fluency of the summary

4.2 Usefulness of quantitative Metrics

Nevertheless, these metrics are still useful for comparing the performance of different fine-tuning steps and the quality of the generated summaries, as they are easy to compute and can enstablish a preliminary baseline evaluation of the model.

5 Qualitative Evaluation

To overcome the limitations of the standard metrics, a qualitative evaluation often proves useful. The qualitative evaluation is based on human judgement and can provide insights specifically into the factual consisency, fluency, coherence and relevance.
In the specific evaluated examples, the models seem to output as summary a truncated version of the text.
This poor performance can be attributed to the small size of the dataset or the the model's inability to capture the context of the text as it is a relatively small model.
Further train time and a larger model could improve the performance of the model.
Nevertheless that the model produces fluent and readable summaries, but the summaries are often incomplete or lack key information by comparison to the golden summary.

- **Example A**: The model succesfully generates a summary that is identical to the golden summary @500 and @1000 finetuning steps.
  The output thus suffices the qualitative criterion of factual consistency, fluency, coherence and relevance, However the golden summary corresponds to the first sentence of the text, so this might be an easier example to summarize.
- **Example B**: The model struggles to generate a coherent sentence @500 steps and misses the point of the article as no metion of the controvertial nature of the structure is made nor it is specified that the structure is a wind turbine.
  The model improves significantly @1000 steps at generating a complete and coherent sentence, but still fails to include the information about the location of the new structure and the arguments of the local residents and provides wrong information about the height of the wind turbine.
- **Example C**: The model extract the first sentence of the text @500 steps which broadly corresponds to the golden summary and is a valid and factually correct summary.
  The model struggles significantly @1000 steps, the generated is very similar to the sentences present in the text, so it is fluent and readable but it provides wrong information about locations.

Metric	500 steps	1000 steps
ROUGE-1	34.5951	36.20
ROUGE-2	22.8455	24.54
ROUGE-L	32.0077	34.03
ROUGELSUM	32.258	34.31
BLEU	30.68	30.26
METEOR	32.28	35.96
BERTScore	0.880	0.880

Table 1: Evaluation metrics for the fine-tuning steps at 500 and 1000 steps over a subset of 100 samples, rouge and meteori metrics improve significantly on the 1000 steps model.

Cherry-picked example A	
text	All babies in Scotland due from tomorrow (August 15) will be gifted a Box full of essential items aimed at tackling inequality and promoting health. The boxes are a strong signal of the Scottish Government's determination that every child, regardless of their circumstances should get the best start in life. Each Baby Box contains a large number of items which are not only practical but designed to help tackle inequality and improve health. The Box itself also doubles up as a safe sleep space, awarded British Safety standard accreditation as a crib for domestic use. Mark McDonald Minister for Childcare and Early Years said: "We are committed to doing everything we can to give every baby born in Scotland the best possible start in life and the Baby Box is just one of the range of measures we are using to help babies and parents thrive in the crucial early months. "The Box includes a large number of items which are not only practical but designed to help tackle inequality and improve health. It can also be used as a safe sleep space and has been awarded British Safety standard accreditation as a crib for use at home. "We will continue to listen to feedback as the Baby Box reaches more families and work with parents and healthcare professionals to keep the contents under review. "The national roll-out is the result of months of hard work and engagement with healthcare professionals, stakeholders and parents and I would like to thank everyone involved in helping us reach this momentous occasion. "The Baby Box has certainly captured the public's imagination and we are extremely proud to introduce it to Scotland." Chief Medical Officer Dr Catherine Calderwood said: "All the evidence shows that the early years are crucial for children's development. What happens then can be linked to outcomes much later in life. So we know that measures undertaken in the 0-3 years age range have the opportunity to make the biggest impact. "That is why we have been working hard to enhance the existing infrastructure available to support families in these crucial early years from before birth all the way up to school age and beyond. "Over and above the practical benefits the items within the Baby Box provide, the box itself also offers healthcare professionals a unique opportunity to introduce expectant parents to a wide range of health promotion information." All babies due on or after 15 August will be eligible for a Baby Box.
gold summary	All babies in Scotland due from tomorrow (August 15) will be gifted a box full of essential items aimed at tackling inequality and promoting health.
generated summary @500 steps	All babies in Scotland due from tomorrow (August 15) will be gifted a Box full of essential items aimed at tackling inequality and promoting health..
generate summary @1000 steps	All babies in Scotland due from tomorrow (August 15) will be gifted a Box full of essential items aimed at tackling inequality and promoting health.

Cherry-picked example B	
text	JOHNSON and JOHNSON has unveiled controversial plans for a second wind turbine, towering up to 100 metres high over Castletroy. The company, which employs hundreds of people through its Vistakon subsidiary in the National Technology Park, has confirmed it has submitted a planning application to Limerick City and County Council for the structure. It's anticipated it will be at a height of 99 metres high, with a rotor radius of 51.5m wide. In a statement, the firm has insisted it is the "second and final" wind turbine it has planned for its facility here – and pointed out that it will provide "significant environmental benefits". But the proposals are already set to attract objections from a community left reeling from the erection of a similar structure back in 2015. Fine Gael councillor Michael Sheahan said he is "deeply concerned" about the proposals. "We already have one here. We now might have a second one on the way. The visuals of this will be very off-putting for people. In the country, you're quite a bit away from these wind farms. But in this particular area, the existing turbine overlooks Mulcair Drive and the Mountshannon Road," he said. The location for the new structure is to the north-west of the existing turbine, at a site known as "Castletroy and Rivers" in Plassey, the planning notice states. Cllr Joe Pond, Fianna Fail, said he would have liked it if the company had engaged with local residents ahead of seeking planning permission. "I'd say local residents will be dead set against it. To them, the current turbine is an eyesore in the community. The noise alone, I'm being told by residents in the Mulcair Drive is not nice," he said. "I'm all for a factory being self-sufficient in its energy ratings, but surely to goodness, something can be done to reduce the noise." An Taisce made a submission on the first turbine, outlining concerns over its "visual impact and intrusion on the landscape". A spokesperson for the Limerick branch of the heritage body this week said An Taisce would look at the impact of the turbine on the local environment, as well as the benefits it would bring in terms of helping Ireland meet its commitment to cut greenhouse gas emissions. Johnson and Johnson says the turbine will generate up to three megawatts of electricity.
gold summary	JOHNSON and JOHNSON has unveiled controversial plans for a second wind turbine, towering up to 100 metres high over Castletroy. The company, which employs hun...
generated summary @500 steps	The company has submitted a planning application to Limerick City and County Council for the structure.
generate summary @1000 steps	Wind turbine plans for a second wind turbine, towering up to 100 metres high over Castletroy

Cherry-picked example C	
text	THE man who shot innocent Limerick man Shane Geoghegan nine years ago is seeking to have his conviction overturned by the European Court of Human Rights (ECHR), it has been confirmed. Barry Doyle, aged 32, who has addresses in Ballinacurra Weston and Portland Row in Dublin was sentenced to life imprisonment more than five years ago after he was convicted, following a retrial, of murdering the 28-year-old in Dooradoyle on November 9, 2008. During the trial, a jury at the Central Criminal Court heard that Doyle admitted during garda interviews that he shot Mr Geoghegan in a what was a case of mistaken identity. Doyle, who had been ordered to shoot another man by criminal figure John Dundon, was arrested on February 24, 2009 and taken to Bruff garda station. He was questioned more than a dozen times over three days and the admissions, the jury heard, were made during his 15th interview. In previous appeals before the Court of Appeal and the Supreme Court, lawyers representing the hitman argued the admissions should not have been admitted as evidence during the original trial because they were obtained in a "non-voluntary" manner as a result of threats and inducements and psychological oppression by gardai. This, they claimed, was because Doyle was promised by gardai that his girlfriend would be released if he admitted his role in the killing. The Court of Appeal rejected the appeal in 2015 while the Supreme also rejected the appeal in a 6-1 majority decision which was handed down in January. According to the Irish Sun, lawyers representing Doyle have now initiated proceedings at the ECHR. In submissions, they state the confessions were made as a "result of a process of inducement, persisting misinformation and erosion of his (Doyle's) right to silence". It's not yet known when the latest appeal will be heard."
gold summary	THE man who shot innocent Limerick man Shane Geoghegan nine years ago is seeking to have his conviction overturned by the European Court of Human...
generated summary @500 steps	The man who shot Shane Geoghegan nine years ago is seeking to have his conviction overturned by the ECHR, it has been confirmed.
generate summary @1000 steps	The man who shot Shane Geoghegan nine years ago is seeking to have his conviction overturned by the ECHR, it has been confirmed.Barry Doyle, aged 32, who has addresses in Ballinacurra Weston and Portland Row in Dublin was sentenced to life imprisonment more than five years ago after he was convicted, following a retrial, of murdering the 28-year-old in Ballinacurra Weston and Portland Row in Dublin.