CrisisFACTs Track

GROUP 24

Reddy Rishitha Reddy - 12041220
Kammela Aditya Vardhan - 12040740
G.S.V Raghavendra - 12040540
Velluru Shyam - 12041700

The **CrisisFacts Summarisation** task automatically summarises key facts and details related to a crisis from various sources like **twitter, reddit, news** etc. It uses natural language processing techniques to identify important information, such as crisis type, location, casualties, and actions taken. The summarised information is presented in an easy-to-understand format, aiding decision-making and **emergency response efforts.**

ADVANTAGES

- **Quick and Efficient :** By distilling large amounts of information into brief, concise statements, responders can easily prioritize and respond to critical situations.
- **Improved Decision-making:** Having clear and concise summaries of the most important facts helps responders quickly understand the key issues, and identify priorities for action.
- **Improved Communication:** Summarising crisis facts into concrete points allows for better communication between emergency responders, government agencies, and the public.

DATASET DESCRIPTION

- By using the Crisis Facts dataset and the folder structure of crisisfacts/<eventNo>/<day>, data is extracted for specific events and days.
- By using ir_datsets library we get the dataset-queries for particular event-day pair .
- The PyTerrier library is used to rank the texts extracted for a particular query, and to identify the most relevant information for a given query.

BASELINE APPROACH

- It uses the PyTerrier library and concatenate the top-scoring texts into a single string which is a simple and easy-to-implement method for accessing information from the Crisis Facts dataset.
- Provides a broad overview of the information available in the dataset.
- Useful for gaining a general understanding of a crisis situation and identifying key areas where more detailed analysis may be necessary.

NOVELTY

Summarisation using

- **T5**
- **TextRank**
- **GPT 2**
- **BART**
- **TIMESTAMP**

T5 - Text-to-Text Transfer Transformer

- T5 is a state-of-the-art language model developed by **Google** that has been trained on a massive amount of text data, making it highly effective at natural language processing tasks such as text summarisation.
- T5 uses a text-to-text approach, where it is trained to map an input text to an output text. This approach allows T5 to perform a variety of tasks using the same architecture and training procedure, including text summarisation.
- T5 has achieved impressive results on text summarisation tasks, outperforming previous state-of-the-art models on several benchmark datasets

LIMITATIONS

- Limited interpretability
- High computational requirements
- Dependency on training data quality

GloVe using TextRank

- GloVe model is a popular unsupervised learning algorithm used for word embedding, which can be used to represent words as vectors in a high-dimensional space.
- TextRank is a graph-based algorithm that ranks the importance of sentences in a text based on their similarity and centrality to other sentences.
- Combining GloVe model with TextRank algorithm can improve the quality of text summarization by capturing both the semantic meaning and the structure of the input text.

LIMITATIONS

- Limited contextual information
- Bias towards high-frequency words
- Faster and cheaper. Good for real-time, low-resource summarisation

GPT-2

- **GPT-2** is a state-of-the-art language model that can generate high-quality human-like text summarization.
- GPT-2 can be fine-tuned for different summary lengths, depending on the specific requirements of the application.
- GPT-2 can generate summaries that are coherent and well-structured, providing a comprehensive overview of the input text.

LIMITATIONS

- Limited tokens.
- May miss context for large texts.
- Quality of summaries heavily dependent on quality of input text.

BART MODEL

- BART is a pre-trained **sequence-to-sequence** model that can be fine-tuned on a variety of downstream tasks, including text summarisation.
- BART uses a combination of masked language modelling and denoising autoencoding during pre-training, **which enables it to generate more coherent and fluent summaries compared to T5 and GPT-2**. This results in summaries that are more readable and easier to understand for human readers.
- This flexibility allows BART to perform well on a wide range of summarisation tasks, making it a more versatile solution compared to TextRank.

RESULTS

```
summary("001","injuries")
```

```
['2017-12-07', '2017-12-08', '2017-12-09', '2017-12-10', '2017-12-11', '2017-12-12', '2017-12-13', '2017-12-14', '2017-12-15']
2017-12-07

crisisfacts/001/2017-12-07 documents: 7288/? [00:02<00:00, 2437.15it/s]

16:02:07.407 [main] WARN org.terrier.querying.ApplyTermPipeline - The index has no termpipelines configuration, and no control configuration is found. Defaulting to global termpipeline
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `attention_mask` to obtain reliable results.
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
**********************************************************************************************************************


Summary:
'A San Diego County Sheriff's Department deputy suffered minor injuries while directing traffic. Two civilians are being treated for burn injuries suffered in #LilacFire and are being
taken to a hospital, https://t.co/Q2EOJFPVvI Headon collision between Sheriff's vehicle and private vehicle in thick smoke #LilacFire  Moderate damage, no injuries reported UPDATE: #L
ilacFire now 2000 acres, 0% containment.    Two civilians are being treated for burn injuries suffered in https://t.co/Dp4y94slym Strong winds and low humidity fueling the Lilac Fire
burning in Bonsall. To add insult to injury the Red Flag Warni https://t.co/FlpsJPuTSV  CAL FIRE/SAN DIEGO COUNTY FIRE (@CALFIRESANDIEGO) December 8, 2017  Four residents sustained bu
rn injuries as they were evacuating their home, Cal Fire said. Besides the civilian injuries, one firefighter suffered smoke inhalation. It was not clear if it was a factor in the inj
ury. The cause of the fire is under investigation. Firefighters are working to contain the blaze, which is burning at a rate of 1,000 acres per hour.\n\nThe fire has been burning for
more than a week now, and is expected to continue for at least another week, according to the California Department of Forestry and Fire Protection (CDFW). The fire started on Saturda
y, Dec. 7, at the home of a man who had been living with his wife and two young children in a home on the south side of San Luis Obispo. According to local media reports, the man, who
has not been<'
```

THANK YOU