Sentiment Classification of Machine Summarized Text

Roy Elkabetz [*]
Rafael Ltd, IL-31021 Haifa, Israel
elkabetzroy@gmail.com

Abstract

The amount of sentiment-contained text in the shape of comments, reviews, and blog posts available on the web is rapidly growing every year. the ability to synthesize sections of large volumes of texts into a concise and informative summarization format while preserving the objective and subjective information, including sentiment, in the text will enable the review of a large amount of information in a reasonable time. Here I examine the idea of classifying machine summarized text with the objective of exploring the capabilities of state-of-the-art machine summarization and its biases towards objective information in short summarizations.

1 Introduction

In recent years, there has been an increasing interest in open-ended language generation thanks to the rise of large transformer-based language models trained on enormous corpora, such as OpenAI's famous GPT2 and GPT3 models, Google's Xlnet and BERT models, and many more. These state-of-the-art NLP models are being used in a wide variety of tasks, naming a few, sentence completion, sentiment analysis, summarization, question answering and language translation. In some cases, these models are used as an "of the shelf" tool, wherein other times they are being tweaked and fine-tuned with transfer learning in order to fit a specific task. In 2020 Google released its Text-to-Text-Transfer-Transformer (T5) NLP model with the goal of creating a single model that could be used in any NLP tasks after fine-tune it with transfer learning methods. This model is a Transformer based model (same as all the ones mentioned above) and it was trained on the Colossal Clean Crawled Corpus which is Google's "cleaner" version of the Common Crawl corpus. In the paper Raffel et al. [2019] the authors reviled the state-of-the-art performance of the T5 model on benchmark datasets for many NLP tasks including summarization and sentiment classification.

The main motivation for this work was to exploit the exceptional performance of the T5 model to try to answer the next question;

"Does a state-of-the-art machine summarizer has any bias towards objective vs subjective information in a given text?

In order to answer this question, I used transfer learning to fine-tune the T5 model into a state-of-the-art summarizer. Then, I used the tuned model in a summarization pipeline to summarize the IMDB sentiment classification dataset with varying summary lengths, ending up with more than 10 datasets in different text lengths. Last, I trained a simple text classifier with the IMDB (complete text) reviews and compared the performance of that model over the different summarized IMDB datasets, a flowchart of the project steps is illustrated in Fig. 1.

[*] project github repository

Preprint. Under review.

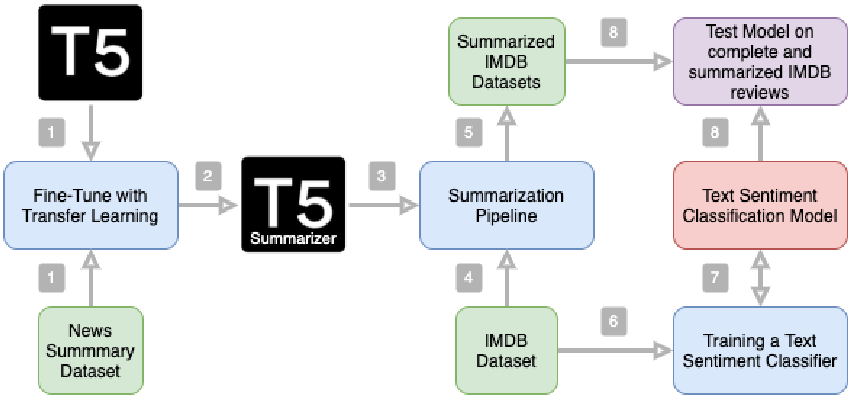
T5
1
Fine-Tune with Transfer Learning
1
News Summmary Dataset
2
T5 Summarizer
3
Summarized IMDB Datasets
5
Summarization Pipeline
4
IMDB Dataset
8
Test Model on complete and summarized IMDB reviews
8
Text Sentiment Classification Model
7
6
Training a Text Sentiment Classifier

Figure 1: The project flowchart: (1) Fine-tune Google's T5 text model using Transfer Learning. (2) Having a state-of-the-art coherent T5-summarizer. (3) Constricting a Summarization Pipeline using the T5-summarizer. (4) Transferring the IMDB dataset to the summarization pipeline for. (5) Constructing and processing the summarized IMDB reviews into a batch of different datasets. (6, 7) Training a Text Classifier on the complete IMDB reviews. (8) Comparing performance of trained Classifier on complete and summarized IMDB reviews.

2 Related Work

The vast majority of work done in text summarization consists of capturing the main information in a piece of text can be divided into two main methods, Extractive summarization, and Abstractive summarization. Extractive Summarization consists of summaries made of the most important words or sentences in a piece of text, which can be done through multiple approaches: reinforcement learning (2), pretrained Bidirectional Encoder Representations from Transformers (BERT) Devlin et al. [2018] and recurrent neural networks that use multiple different encoders and decoders Cheng and Lapata [2016], Isonuma et al. [2017] are some of them.

Abstractive summarization consists of reproducing important material in a new way after interpretation and examination of the text and generating a new shorter text that conveys the most critical information from the original one, such as in Raffel et al. [2019].

From these two approaches, it seems that the more effective one in our case would probably be the abstractive summarization approach. This technique is not bounded to lengths and templates in the original text and potentially can span the relevant information in a less constrained way. Thus, would be more convenient for summaries with variable lengths.

In the domain of text classification, a simple and efficient baseline for sentence classification is to use a bag of words (BoW) to represent the sentences and train a linear classifier, e.g., logistic regression or an SVM Joachims [1998]. However, linear classifiers are limited in their generalization in the context of large output space where some classes have very few examples. A more advanced approach is to use "Deep Learning" in the form of sequential networks such as LSTM, non-sequential deep CNN, or even the combination of both as C-LSTM Zhou et al. [2015].

3 Datasets

I used two kinds of datasets in this work. For the T5 transfer-learning task I used the *News Summary* dataset from *Kaggle*. The *News Summary* dataset is comprised from six fields in total, where I was using only the "ctext" and "text" fields which stands for the complete text extracted from the news website and its human summarized version respectively. Here is an example of a single data sample from the this dataset:

```
Text: administration was forced to retreat within 24 hours of issuing the circular that made
it compulsory for its staff to celebrate Rakshabandhan at workplace.?It has been decided
to celebrate the festival of Rakshabandhan on August 7.  In this connection, all offices/
departments shall remain open and celebrate the festival collectively at a suitable time
wherein all the lady staff shall tie rakhis to their colleagues,?  the order, issued on
August 1 by Gurpreet Singh, deputy secretary (personnel), had said.To ensure that no one
```

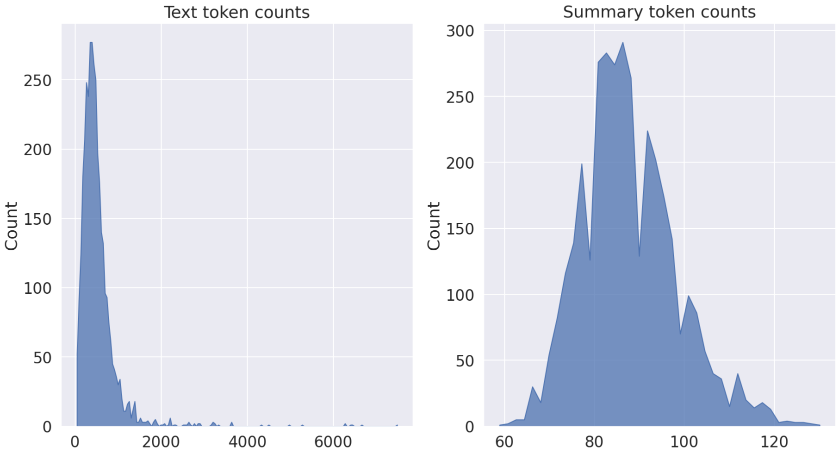

Figure 2: Token count histograms of the News Summary train dataset. The T5 model can take inputs of up to 512 tokens, so any sequence with more token was truncated in the training process.

skipped office, an attendance report was to be sent to the government the next evening.The two notifications ? one mandating the celebration of Rakshabandhan (left) and the other withdrawing the mandate (right) ? were issued by the Daman and Diu administration a day apart. The circular was withdrawn through a one-line order issued late in the evening by the UT?s department of personnel and administrative reforms.?The circular is ridiculous. There are sensitivities involved. How can the government dictate who I should tie rakhi to? We should maintain the professionalism of a workplace? an official told Hindustan Times earlier in the day. She refused to be identified.The notice was issued on Daman and Diu administrator and former Gujarat home minister Praful Kodabhai Patel?s direction, sources said.Rakshabandhan, a celebration of the bond between brothers and sisters, is one of several Hindu festivities and rituals that are no longer confined of private, family affairs but have become tools to push politic al ideologies.In 2014, the year BJP stormed to power at the Centre, Rashtriya Swayamsevak Sangh (RSS) chief Mohan Bhagwat said the festival had ?national significance? and should be celebrated widely ?to protect Hindu culture and live by the values enshrined in it?. The RSS is the ideological parent of the ruling BJP.Last year, women ministers in the Modi government went to the border areas to celebrate the festival with soldiers. A year before, all cabinet ministers were asked to go to their constituencies for the festival.

Summary: The Administration of Union Territory Daman and Diu has revoked its order that made it compulsory for women to tie rakhis to their male colleagues on the occasion of Rakshabandhan on August 7. The administration was forced to withdraw the decision within 24 hours of issuing the circular after it received flak from employees and was slammed on social media.

From The token count statistics of the text and summary data (see Fig. 2) we see that although there are some very long text samples, 64% of the text samples are less than 512 tokens long, while 99% of the summary samples are less than 128 tokens long. This is important because the T5 tokenizer can take as an input sequences of no more than 512 tokens (longer sequences are truncated) and the T5 model can generate sequences no longer than 128 tokens. Also notice that all text summaries are longer than 60 tokens which could affected our results. I will discuss it in more depth in the results section. The dataset has an overall size of 4396 samples splitted between

- Train: 3560
- Validation: 396
- Test: 440

For the text classification task I used the *IMDB* reviews dataset. This dataset is comprised of 50000 IMDB text reviews and their positive/negative labels splitted evenly between train and test. Here is an example of a single data sample from the IMDB dataset:

Label: neg

Review: I rented I AM CURIOUS-YELLOW from my video store because of all the controversy that surrounded it when it was first released in 1967. I also heard that at first it was seized by U.S. customs if it ever tried to enter this country, therefore being a fan of films considered "controversial" I really had to see this for myself.<br /><br />The plot is centered around a young Swedish drama student named Lena who wants to learn everything

Table 1: Summary generator parameters

parameter	with Sampling	without Sampling
length penalty	0.5	1.0
repetition penalty	2.5	2.5
do sample	True	False
top p	0.2	1.0
top k	0	50 (default)
early stopping	False	True
num beams	1	2

```
she can about life.  In particular she wants to focus her attentions to making some sort
of documentary on what the average Swede thought about certain political issues such as
the Vietnam War and race issues in the United States.  In between asking politicians and
ordinary denizens of Stockholm about their opinions on politics, she has sex with her drama
teacher, classmates, and married men.<br /><br />What kills me about I AM CURIOUS-YELLOW
is that 40 years ago, this was considered pornographic.  Really, the sex and nudity scenes
are few and far between, even then it's not shot like some cheaply made porno.  While my
countrymen mind find it shocking, in reality sex and nudity are a major staple in Swedish
cinema.  Even Ingmar Bergman, arguably their answer to good old boy John Ford, had sex
scenes in his films.<br /><br />I do commend the filmmakers for the fact that any sex shown
in the film is shown for artistic purposes rather than just to shock people and make money
to be shown in pornographic theaters in America.  I AM CURIOUS-YELLOW is a good film for
anyone wanting to study the meat and potatoes (no pun intended) of Swedish cinema.  But
really, this film doesn't have much of a plot.
```

In the process of preparing the data I used my T5 summarizer to summarize few of the IMDB reviews with different summarization lengths for the classification testing process. Because the summarization process is time consuming (few seconds per sample) my summarized-IMDB dataset is comprised of the original (label, text) data of 1000 samples (500 positive and 500 negative) and their summaries with 9 different maximal summarization lengths $\{100, 90, 80, 70, 60, 50, 40, 30, 20\}$.

I repeated this summary generation process twice in two different methods: with and without sampling. The sampling parameter corresponds to the *do sample* value in Table .1, and is responsible for random sampling of words in the process of sequence generation. In short, using sampling means the generator randomly chooses the next word w_t according to its conditional probability $w_t \sim P(w|w_{t-1})$, rather than just selecting the one with the highest probability. The other parameters that have been used in each method are listed in Table 1. I will not get into more details of each parameter objective here. For a great blog post explaining the sampling method along with all other Hugging-Face's "generate" class parameters, check out von Platen [2020]. Finally, I ended up with 18 different summarized-IMDB datasets, 9 with sampling, and 9 without. Examples of samples from the two datasets can be found in Appendix A. Token count histograms of all the summarized IMDB datasets are illustrated in Fig. 3. We notice that each "maximal length" value indeed defines a limit on the number of generated tokens per sequence, with a mean token count of a few dozens of tokens shorter per value, i.e. for maximal length of 100 tokens, the mean token count is 70 in both methods.

4 Methods

In this section, I will describe the model architectures and the training processes that have been used through that work, starting with the T5 model and continuing to the text classifier.

4.1 T5 model

4.1.1 Architecture

The T5 model is a sequence-to-sequence NLP model which has an encoder-decoder architecture (see Fig. 4) with both parts consists of multiple Transformer layers. The encoder has a BERT-like architecture with a stack of bidirectional self-attention layers (see Devlin et al. [2018]), while The decoder part has a language-modeling architecture consists of a stack of causally masked self-attention layers. A comprehensive explanation of the self-attention mechanism can be found in Vaswani et al. [2017]. To actually generate words, the output of the decoder is sent to a final linear layer, which is followed by a Softmax function.

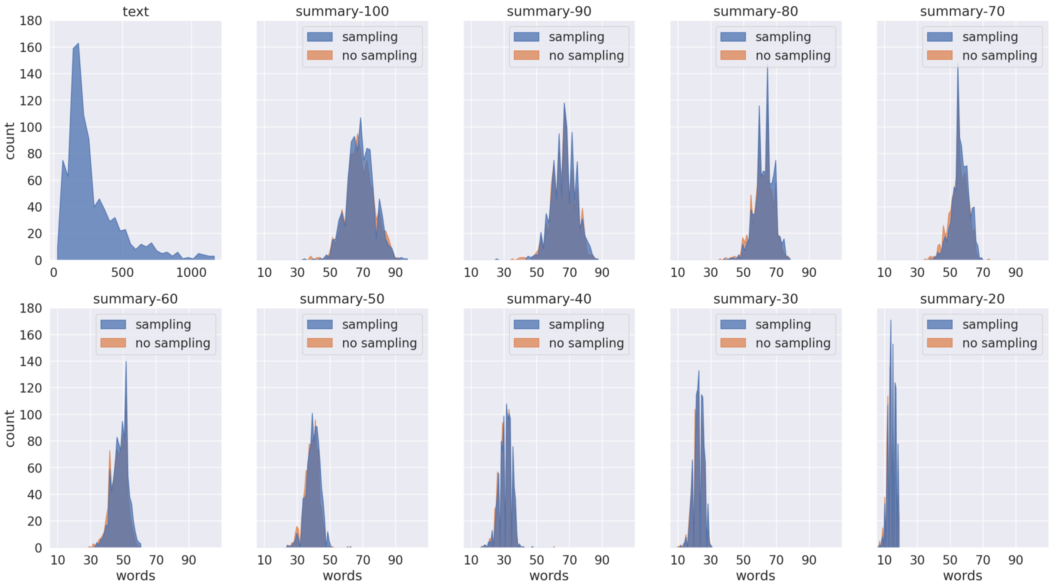

Figure 3: Word count histograms of the two IMDB datasets. The title of each plot describes the *maximal length* parameter value set to the T5 summarizer. We notice that the maximal length parameter of 100 causes a generated average word count of 70 words per summary. In general, we notice the mean word count for each dataset is typically a few dozens of words less than the maximal length value that was set.

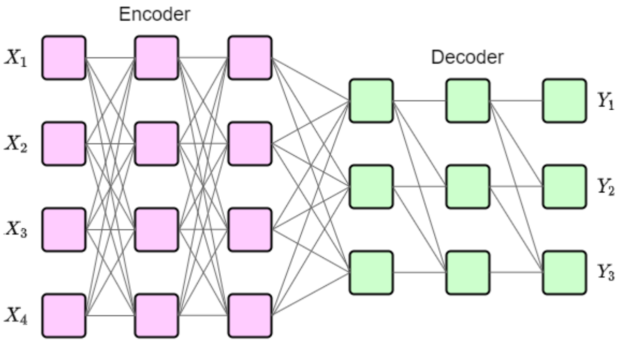

Figure 4: The T5 Encoder-Decoder architecture.

4.1.2 Transfer Learning

I used transfer learning to fine-tune the T5 model on all of its layers, training the model with (text, summary) pairs from the *News Summary* dataset. Each training iteration consists of tokenization and numericalization of the input text (in batches) using the pre-trained T5 tokenizer. Then, the numeric text were fed into the encoder followed by a probabilistic teacher-forcing for text generation in the decoder. In teacher-forcing style, the target sequence is appended with the "End Of Sentence" (EOS) token such that the trained model would know when to halt text generation. The model uses a cross-entropy loss function computed between the conditional probability of the next word generated to the "true" word in the target sentence.

4.1.3 Summary Generator

The way T5 actually works is that after each token is produced, that token is added to the sequence of inputs. That new sequence becomes the input to the model in its next step. This is an idea called “auto-regression” and it is based on the assumption that the probability distribution of a word sequence can be decomposed into the product of conditional next word distributions as follows

Table 2: Text classification hyperparameters list

Epochs	35
Batch size	32
Learning rate	5.0
Scheduler rate (epochs)	1
Scheduler γ	0.5
Embedding dimension	64

$$P\left(w_1, w_2, \ldots, w_T | w_0\right) = \prod_{t=1}^{T} P\left(w_t | w_0, w_1, \ldots, w_{t-1}\right) \tag{1}$$

where w_0 being the first word usually taken as the "Start Of Sentence" (SOS) token. The length T of the word sequence is usually determined on-the-fly and corresponds to the time step $t = T$ the EOS token is generated from $P\left(w_t | w_0, w_1, \ldots, w_{t-1}\right)$. At each time step, the model outputs a conditional probability vector for all the words in its vocabulary that used to generate the following word in the sequence. There is a wide range of methods for selecting the next word from this vector of probabilities, starting with a greedy-search which selects the word of highest conditional probability, followed by the beam-search method which keeps track of the k most prominent sequences probability-wise, and so on. There are few more methods which I will not get into. For a more comprehensive explanation about all the text generation methods I refer the interested reader to the great blog post in von Platen [2020].

4.2 Text Classifier

4.2.1 Architecture

For the text classification task, I used a shallow neural network with an embedding-bag layer followed by a single fully-connected layer with no activation, and finally a binary classification layer with a single unit. The model's architecture is illustrated in Fig. 5. Each sequence $[w_0, w_1, \ldots, w_N]$ of words is tokenized and converted to an array of indices $[i_0, i_1, \ldots, i_N]$ derived from the locations of the words in the tokenizer vocabulary (same as for the T5 model). The array is then sent into an embedding-bag layer where each of its values are being transformed into a vector $\mathbf{e}_k$ in a low dimensional vector space - the embedding space. The collection of embedded vectors correspond to a single sequence are than averaged out

$$[i_0, i_1, \ldots, i_N] \rightarrow [\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_N] \rightarrow \frac{1}{n} \sum_{j=0}^{N} \mathbf{e}_j = \tilde{\mathbf{e}} \tag{2}$$

and finally $\tilde{\mathbf{e}}$ is sent to the next layer in the network.

4.2.2 Training Process

The model was trained on non-summarized IMDB reviews, taken from the IMDB-train dataset. The validation and test sets were constructed by splitting the IMDB-test dataset into $(24,000,\ 1000)$ samples respectively. The 1000 samples for the test was then duplicated and summarized using the T5 summarizer, ending up with a total of 18 test sets of different summarization parameters, see Sec. 3 for the full description. The loss criterion used was the binary cross-entropy loss function

$$Loss\left(\hat{y}, y\right) = -y \log\left(\hat{y}\right) - \left(1 - y\right) \log\left(1 - \hat{y}\right) \tag{3}$$

where $y, \hat{y}$ are the true and predicted class labels respectively. The training loop was carried with an SGD optimizer combined with a discrete learning-rate scheduler. There was no need for a formal hyperparameters optimization, where "good enough" classification accuracy (in the 90% range) was reached with few iterations of manually tweaking the training parameters. The hyperparameters that were chosen are listed in Table 2.

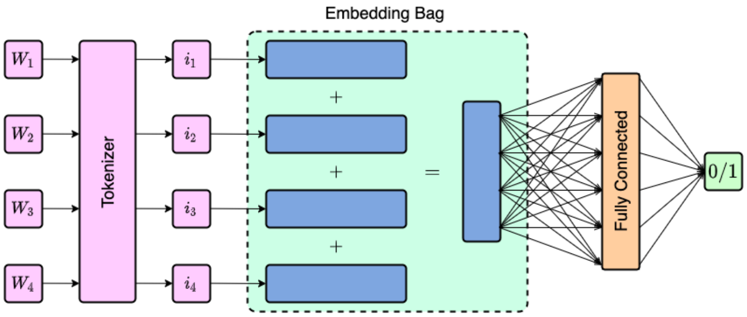

Figure 5: The text classifier architecture was used for sentiment classification. The model is comprised of a single embedding-bag layer followed by a linear fully-connected layer followed by a fully-connected layer with a single unit (binary layer). The W_k variables stand for single words / tokens where i_k are their index representation in the model's vocabulary.

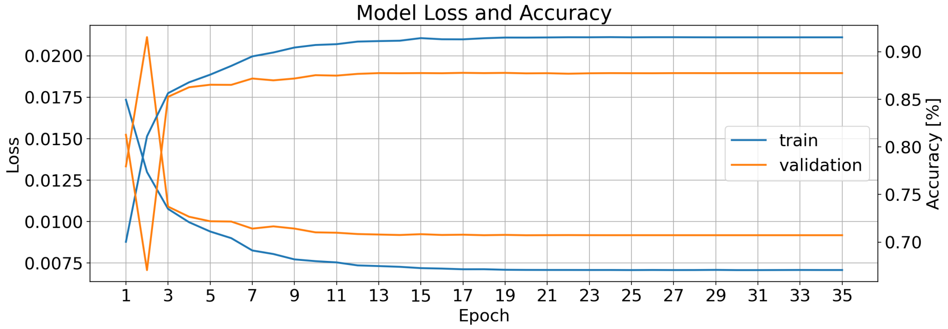

Figure 6: Loss and Accuracy measured through the training process of the text classifier model.

5 Results

The text classifier loss and accuracy measured through the training process are illustrated in Fig. 6. Notice that around 20 epochs the loss stops decreasing and there is no more improvement in train and validation accuracy in following iterations. Therefore, we can say with confidence that the model is converged to its final state given the chosen hyperparameters.

Next, I checked the classifier performance on the summarized test sets, the results are illustrated in Fig. 7. Although there is a large difference between the with/without sampling generation methods (see Appendix A) for an example), we see that the overall performance of the model has no significant difference. Furthermore, the classifier performance is dropping around summaries with a maximal length of 60 words for both sampling methods. In Fig. 2 we have seen that all the News Summary summaries used for the T5 training are longer than 60 words. This, could potentially, inject a bias in the T5 summarizer training process towards generating summaries that are longer than 60 words, which could result in a performance deterioration for shorter summaries generations. Besides that, it is clear that the classification performance is dropping with shorter summaries. The high 89.1% classification accuracy on the non-summarized test set (all the test sets are generated from the same 1000 samples) indicates that the performance deterioration is not a matter of classifier quality rather than the poor quality of text summarization generated by the T5 model.

6 Discussion and Future Work

In this project, I attempted to find the relation between the length of a summarized text piece generated with a state-of-the-art machine summarizer to the amount of subjective information preserved in the machine summarization process. To quantify the amount of subjective information in a text piece I used a sentiment text classifier, and through its performance in classifying correctly the given text, I infer indirectly the quality of the summarization process. From the results I got, it seems that there is

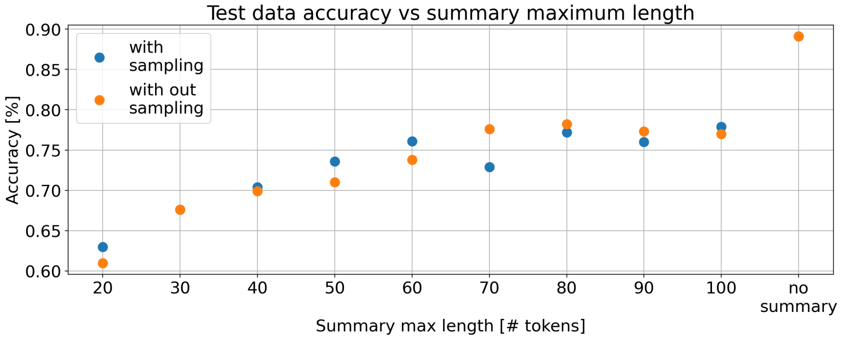

Figure 7: The classifier accuracy over the test sets. Each point in the graph describes the classifier's accuracy of predicting the correct classes for the same 1000 test samples sharing the same maximal summary length. The orange data belong to the test sets generated without "sampling" and the blue data belong the ones generated with sampling (see Sec. 3 for description).

a tight connection between the length of the summary to the amount of subjective information kept in the summarized text through the summarization process. We see that shorter summary pieces contain less subjective information which makes them harder to classify in a positive/negative framework. The results are not unambiguous, that is because of a bias towards summaries of lengths longer than 60 words which have been used in the transfer learning process of the T5 fine-tune. To verify if this bias is the main cause of the results we got, we need to use a dataset of shorter summaries in the transfer learning process of the T5 model. The project github repository.

References

J. Cheng and M. Lapata. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*, 2016.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

M. Isonuma, T. Fujino, J. Mori, Y. Matsuo, and I. Sakata. Extractive summarization using multi-task learning with document classification. In *Proceedings of the 2017 Conference on empirical methods in natural language processing*, pages 2101–2110, 2017.

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

P. von Platen. *How to generate text: using different decoding methods for language generation with Transformers*, 2020. `https://huggingface.co/blog/how-to-generate` [Accessed: Whenever].

C. Zhou, C. Sun, Z. Liu, and F. Lau. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*, 2015.

A Appendix

The main difference between the results of the two with/without sampling methods is that for a non-sampling summary generation the generated summary is identical up to a maximal length truncation. For the sampling method, this is not the case, where each text piece generated is different from the other. An example from the "with-sampling" dataset:

Dataset type: with sampling

Label: neg

Text: If you are in search of a masochistic thrill, rent this movie, and show it to a group of your friends sober. This movie is just plane lame, but there not completely without value. The brief tits are nice, and there is one victim’s death that is funny as hell. Other than that, this is straight garbage. But it is still better than "Grim" or "Spookies"

summary 100: A movie that is just plane lame, but there’s no value. The brief tits are nice and the one victim’ death in this movie is funny as hell. The film has been rated PG-13 (Madi) and PG-13 (Spookie). It also has a comedy scene wherein the man die-off with a few seconds of time to kill himself.

summary 90: Rent this movie, and show it to a group of your friends sober. The brief tits are nice, the brief spit is nice, but there’re one victim’d death that is funny as hell. It is also better than "Grim" or Spookier films. The film has a short flitting scene with a bit more comedy than "Grim".

summary 80: A movie that is rated ’Grim" or "Spookies", has been rated ’Grim’ and the short tits are nice. The brief-time adolescent death of one victim, who was killed by a man in his life is funny as hell. This film also features a video wherein the characters are seen

summary 70: A movie that is rated ’Grim" or "Spookies", which are better than the short tits and titties, there was one victim’d death. The film has been rated ’Grimm’ (Grim) for being sober), while showing it to a group of

summary 60: A film about a man’s death that is funny as hell, rent this movie and show it to thugs sober. The brief ties are nice, the brief fits were nice, but there was one victim’d death which is humorously hilarious...it is

summary 50: A movie that is rated ’Grim" or "Spookies", which are better than the short tits and titties, but there is one victim’? death that is funny as hell. The film

summary 40: A movie that is rated ’Grim" or "Spookies", which are also better than the short tits and titties, this is just plane la

summary 30: Rent this movie, and show it to a group of your friends sober. The brief tits are nice, the brief ties

summary 20: A movie about a man’s death that is funny as hell, rent this movie and

Next is an example from the "without-sampling" dataset:

Dataset type: without sampling

Label: neg

Text: If you are in search of a masochistic thrill, rent this movie, and show it to a group of your friends sober. This movie is just plane lame, but there not completely without value. The brief tits are nice, and there is one victim’s death that is funny as hell. Other than that, this is straight garbage. But it is still better than "Grim" or "Spookies"

summary 100: Rent this movie, and show it to a group of your friends sober. The brief tits are nice, and there is one victim’s death that is funny as hell. The short tits are nice, and there is one victim’s death that is funny as hell. This movie is just plane lame, but there is not complete without value.

summary 90: Rent this movie, and show it to a group of your friends sober. The brief tits are nice, and there is one victim’s death that is funny as hell. The short tits are nice, and there is one victim’s death that is funny as hell. This movie is just plane lame, but there is not complete without value.

summary 80: Rent this movie, and show it to a group of your friends sober. The brief tits are nice, and there is one victim’s death that is funny as hell. The short tits are nice, and there is one victim’s death that is funny as hell. This movie is just plane lame, but there is not complete without value.

summary 70: Rent this movie, and show it to a group of your friends sober. The brief tits are nice, and there is one victim’s death that is funny as hell. The short tits are nice, and there is one victim’s death that is funny as hell. This movie is just plane lame,

summary 60: Rent this movie, and show it to a group of your friends sober. The brief tits are nice, and there is one victim’s death that is funny as hell. The short tits are nice, and there is one victim’s death that is funny as

summary 50: Rent this movie, and show it to a group of your friends sober. The brief tits are nice, and there is one victim's death that is funny as hell. The short tits are nice, and the

summary 40: Rent this movie, and show it to a group of your friends sober. The brief tits are nice, and there is one victim's death that is funny as hell.

summary 30: Rent this movie, and show it to a group of your friends sober. The brief tits are nice, and there is one

summary 20: Rent this movie, and show it to a group of your friends sober. The brief