Machine Learning - Exercise 3: NLP - Next Word Prediction

Group 26 - TU Wien

Bianca Apostolescu, Lu Chen, Matthias Glinzner

1 Datasets

1.1 Dataset 1

The Fake News Dataset is a collection of news articles that were labeled as either ‘Real’ or ‘Fake’. The dataset contains 6335 rows/observations along with 4 columns/features:

1. Title - The title of the news article.
2. Text - The full text of the news article.
3. Label - Classification of the news article, either ‘REAL’ or ‘FAKE’.
4. Subject - The subject of the news article.

We conduct our next word prediction (NWP) task based on the column ’Text’.

1.2 Dataset 2

The dataset that was chosen contains extracts from the book ”The Adventures of Sherlock Holmes”, by Sir Arthur Conan Doyle. It has multiple stories from the aforementioned book and a total of 12175 rows. The next word prediction task is based on these extracts.

2 Implemented Methods

2.1 DistilGPT - 2

DistilGPT-2, an evolution of the renowned GPT-2 language model, represents a significant stride in the realm of transformer-based natural language processing.

The decision to employ DistilGPT-2 for the next word prediction task is rooted in its balanced combination of efficiency, contextual understanding, and versatility. Its reduced size ensures practical deployment, while its pre-trained knowledge empowers it to generate contextually appropriate and semantically meaningful predictions.

2.2 LSTM

The Long Short-Term Memory (LSTM) is a recurrent neural network (RNN) architecture extensively employed in Deep Learning that addresses the vanishing gradient problem, a challenge in training traditional RNNs. It stands out in its ability to capture long-term dependencies, rendering it well-suited for tasks involving sequence prediction.

In contrast to conventional neural networks, LSTM integrates feedback connections, enabling it to handle complete sequences of data rather than individual data points. This characteristic enhances its efficacy in comprehending and forecasting patterns within sequential data, such as time series, text, and speech.

The main reason for employing LSTM for the next word prediction task is their efficacy in capturing dependencies within sequences spanning considerable distances, a pivotal aspect for comprehending contextual nuances in NLP applications. Therefore, considering that the NWP task necessitates the analysis of word sequences and that LSTMs uphold a memory cell that preserves contextual information across intervals, this feature enhancing the network's capacity to comprehend and harness contextual intricacies, the LSTM model was chosen.

3 Evaluation

3.1 Performance Metrics

3.1.1 Perplexity

The first performance metric that was chosen is perplexity. It is a quantitative measure whose role is to assess the predictive performance of a language model. Mathematically, perplexity is calculated as the exponentiation of the entropy, which measures the average surprise or uncertainty associated with predicting the next word in a sequence. Equation 1 displays the formula for perplexity, where $H(w)$ is the entropy of the system. Following this, the formula of entropy is presented in Equation 2.

Regarding the evaluation of this metric, lower perplexity values indicate that the model is more certain and less surprised by the observed sequences, reflecting better predictive performance.

$$Perplexity(w) = 2^{H(w)} \tag{1}$$

$$H(w) = \sum_{i=1}^{N} p_i * \log_2(p_i) \tag{2}$$

3.1.2 BLEU-N Score

The second chosen evaluation metric is the Bilingual Evaluation Understudy (BLEU-N Score). It is a metric used for evaluating the quality of machine-generated text, especially in the context of NLP. The BLEU Score computes a precision-based score that measures the similarity between the machine-generated output and a set of reference texts. In Sequence-to-Sequence tasks, a single candidate may have multiple correct or reference outputs. Therefore, the selection of references is a crucial aspect, and it is essential to include all potential references. Considering its characteristics, the BLEU Score works in the following way: it evaluates the precision of n-grams (contiguous sequences of n items, usually words) in the generated text compared to the reference texts. It considers precision for multiple n-gram lengths, typically up to 4. Also, to address the issue of generating overly short outputs to maximize n-gram matches, BLEU includes a brevity penalty. This reduces the score if the generated output is significantly shorter than the reference texts. Depending on the number of analyzed n-grams, the BLEU Score formula can differ. However, the generalized formula is shown by Equation 4, where BP is the result given by the brevity penalty and p_n is the geometric average of the precision scores for the considered n-grams. The formula for the BP is presented in Equation 3, where c is the word candidate and r is the word reference.

The BLEU score, being a precision-based metric, varies between 0 and 1. A higher value closer to 1 indicates a more accurate prediction. While achieving a perfect score of 1 is challenging, a score exceeding 0.3 is generally considered commendable.

$$BP = \begin{cases} 1 & \text{if } c > r, \\ e^{1-r/c} & \text{if } c \leq r. \end{cases} \tag{3}$$

$$BLEU(N) = BP * exp(\sum_{n=1}^{N} 1/N * \log(p_n)) \quad (4)$$

3.2 Real and Fake News Dataset

3.2.1 Results

Table 1: Performance metrics for The Adventures of Sherlock Holmes Dataset

Train-Test Split [%]	Perplexity-Testset	Perplexity-Validset
70-30-10	41.11	37.22

3.2.2 Manual analysis

For the manual analysis, five prompts were used as input:

1. "But what is the point"
2. "Politicians have to"
3. "Never have I ever"
4. "Help! My dog ate"
5. "Alien abduction - my yard"

They were chosen to incrementally increase the difficulty for the prediction to find suitable continuations. Keep in mind though, that this evaluation of quality is entirely subjective: we analysed the predictions in terms of (felt) plausibility, grammatical correctness as well as how natural they were to read.

The following sentences resulted from the first prompt:

1. But what is the point in the fight over what to do? We must build on that momentum," Cruz said, using the hashtag #RealClearPolitics, saying it was totally a good idea to send Trump a message.
2. But what is the point of a system where you're going to be in a race, where you're going to go up there with people like me who don't like me," says Sanders.

3. But what is the point? It would require the kind of clarity that Trump, and most of the rest of his advisers have been arguing about, from the beginning, that America needs an affirmative action strategy that helps combat radical Islam.
4. But what is the point in this election cycle? Yes, the American people, including so many other Americans, know that the political revolution will be not over. It is going to be a very deep and important lesson of this election.
5. But what is the point of it? The problem is, the US will be in a position, in spite of the world's great leadership leadership. I've said that before and in all my own interviews since I went to Washington.
6. But what is the point of being a senator? Let us know your question! And be sure to let us know your favorite Senators in the comments section below! (CNN) There are two critical issues in the Republican Party:
7. But what is the point of taking the money out of the middle class? This is the real question. So the idea is that we should start with middle-class families and put a dollar in the middle (or at least the middle)
8. But what is the point of the political experience to tell us about the nature of our campaigns?" There are other issues beyond such as political experience that have deep roots in our history:
9. But what is the point of a campaign where you're making the biggest tax cut in your lifetime? If I were, I think I would make a big money — I think it would be big money, certainly
10. But what is the point of creating something new at this stage?" A CNN poll published last week found that nearly half of Americans support a more democratic election of "the lesser of two evils"

The sentence with the earliest occurrence of a clearly wrong structure is number 4; the repetition of "America(s)" happens at the 13th added word. Apart from that, sense only starts to crumble beyond the 20th word, which despite the fact that the model was trained on mostly political news, is quite impressive. Some sentences even manage to look real until termination.

The second prompt is thematically similar to the first:

1. Politicians have to be prepared.” He says, “We must build a strong economy.” In fact, he adds, he notes that Trump is no stranger to criticizing the economy.
2. Politicians have to be cautious too lest we find ourselves in a bubble like the one most in the world. If Sanders’ support for Wall Street deregulation has come from some Wall Street liberal base or a conservative, then it should come from
3. Politicians have to work with the voters, too. “Democrats need to find a way to convince voters they can win,” Clinton said. “And what they must tell voters is that they have an opportunity so easily to
4. Politicians have to figure that out,” Romney said a week ago. “I know this is not going to work, but the truth is that we are, at worst, a far more dysfunctional government than we have been in a very
5. Politicians have to understand why we’re all so sick of these political elites — especially in the Democratic Party. As the Democratic Party has tried to make it easier for us, the party is also going to need to take something new and revolutionary
6. Politicians have to be careful and careful how they look for their money and what they have to pay to make a difference. They have to be careful to get a good deal of their money into the pockets of wealthy Americans. They have to play a
7. Politicians have to think about how much to do with people’s political and economic well-being. What the public actually wants, is a system, not a politics designed to work. They don’t know how long that wait will
8. Politicians have to learn to embrace what they see in their elected officials on policy issues, their political careers, their family and all of their political experience. “We need to embrace the vision to succeed in a country where we care about
9. Politicians have to be careful not to call them ”spokesmen.” They should have used their names on the campaign trail. That’s the problem, she argues. ”It undermines the effectiveness of Trump’s vision of the United States as
10. Politicians have to make the case that the political left is a disaster, and that the parties have made clear that a major party would be a

disaster, and they have spent a lot of money. So they are not going to give a big bang

It performs slightly worse, though; the first occurrence of an error happening as early as the fourth word (Ex. 6). Another error at word six can be observed in sentence 7, as well as two errors beyond the 10th word in sentences 2 and 8.

The third prompt starts to move away from politics:

1. Never have I ever felt better?" "My daughter is still working on it. So much is going on that I just want to tell the whole truth in a safe moment," said Kiesler. "She's worked as an intern and
2. Never have I ever had a hand in America's elections — one I can afford to have for the next six months?" Rubio told conservative radio host Joe Scarborough on Wednesday, "I don't need to write this to tell the people I can't.
3. Never have I ever seen him in a white dress? His hair does have a bright shiny red-coloured light that glows in his eyes after the jump as if he was a real person. One of them was a woman wearing his
4. Never have I ever said a favor or a negative word?" Trump has repeatedly asserted that he's opposed the Iraq War, even if the United States invaded Iraq on the spot. "No, I'm not opposed to the Iraq War
5. Never have I ever experienced? I don't have to believe that I've been there. I have seen the same thing happening to my daughters at the age of seven and have seen in other kids. The same thing happens
6. Never have I ever seen them," Trump said. "I think most of the time. I don't really know if they're good." At the outset, Cruz did not offer any more details
7. Never have I ever heard of a politician who could stop this corruption." Trump, then a Republican from Florida who is the latest to announce his candidacy, said he would consider another run. "I am not saying I am going to
8. Never have I ever been accused of being a bad guy by some folks, if anyone's ever gotten past the standard standard and it was clear to me that being a good guy for the sake of the country has the potential to improve

9. Never have I ever heard of someone with a higher sense of urgency about this thing," Trump told CNN. Trump, as he was about to address another major rally in Cleveland — New York, on Feb. 16, 2016 — went

10. Never have I ever been in the White House?" Obama tweeted on Monday. "When they get their chance, they will let the establishment get a little too involved, as is the case." The president has yet to endorse Clinton, suggesting that

With a few exceptions we notice that the model wasn't trained for prompts like this one. The longest a faulty sentence goes on is to the tenth word (Ex. 8). An interesting thing happens with the tenth sentence: the otherwise wrong-seeming question mark is given a correct context as the sentence goes on. And it is also worth mentioning that two of the sequences (6 and 9) are completely coherent.

The last two prompts are there to provide fringe cases. Number four produced:

1. Help! My dog ate a lot of food, as I can no longer stand it," she said. She later returned to the home, but didn't want to get into her daycare. But she still

2. Help! My dog ate right away! She was a fun little girl – this was the first dog I've had to do this job. You don't know what he eats, but I went to the doctor on a very good day and took one

3. Help! My dog ate all three in the night, then when my mom got off with some popcorn I knew everything was alright. I went out for a walk and ate all three. I knew nothing about our dog, but he had this crazy new

4. Help! My dog ate too much, I took some vitamins and added a lot of dairy," said Brian Pardig, owner of the dog's owner's home in St. Elizabeth.

5. Help! My dog ate my sandwich at the end of that lunch and he still is!" "It's a very big fight!" the mom told the Daily News. "He always says he ate something good here in the middle of the night

6. Help! My dog ate her treats at the mall but my cat ate at a store and they got sick of it! The mall food is so amazing! The only questions I have were: Would you give a kid food for a baby?

7. Help! My dog ate, we all ate!" – Michelle B. Weiss, D-Madison, Wisconsin, USA. "What is your dog's diet? Is his diet a diet you eat when he's there, when
8. Help! My dog ate all the calories in our house – all in food! I've used to spend a few minutes cooking for 2 months because that is when I thought it was best and I've been vegan for the past few
9. Help! My dog ate some of the veggies, he also ate the protein and protein it contained. He's going to have to use our food to live to his full potential. They also need to cook rice that is made with soybeans
10. Help! My dog ate a bag of bacon. I was like, "Wait!" to him; I couldn't get along, but the food was nice. The owner, who did my best to not let the dog go, told his

And finally number five:

1. Alien abduction - my yard signs, the street signs on the block, all the trash from the recycling bin of mine and so many tons of junk. My yard signs. People are talking about dumpsters, garbage bags, etc.
2. Alien abduction - my yard was destroyed. It was one of those that made me angry, like the girl I had had seen hanging from a bridge, from that bridge to the other side of the road!" A white-eyed man
3. Alien abduction - my yard! The last time I remember seeing the man in the car, I saw that they just left me in the garage, and that I don't know who caused that to happen," she said.
4. Alien abduction - my yard," she told the Guardian. "It's very hard to trust what she says. She said things that haven't been proven through forensic analysis or forensic DNA." But the FBI declined
5. Alien abduction - my yard, on the hill, across the street, on a bridge to my house." Two years or so ago, I had some trouble walking in the neighborhood and not getting to work, but it seemed to work
6. Alien abduction - my yard," she said at a news conference in Manhattan on Thursday. The suspects – not me – were found, according to an NYPD document. "My mother and I were sitting in their car," the police report

7. Alien abduction - my yard. The car had just hit me and my feet and that's when I turned up and stopped for help from the FBI, which is very frustrating and a shame as to why they found me," she said.
8. Alien abduction - my yard" and her wife are all in a dark apartment. "I'm a woman of love and love and I want to make sure there's nothing I'm gonna lose," she said. The victim said
9. Alien abduction - my yardage - and the other thing I had thought I could do for my child. My husband and I were having problems eating our foods while they were on holiday. I tried to leave the apartment complex and we had problems sleeping but
10. Alien abduction - my yard is a beautiful home But it just isn't a home. "It's just a beautiful, beautiful, beautiful home" writes Dr. Michael A. Gettner, an

In both cases, the semblance of correctness breaks down very quickly. Still, when viewed in the context of, e.g., an autocomplete function, where only the next word has to be predicted even the last two prompts lead to useable results.

3.3 The Adventures of Sherlock Holmes Dataset

3.3.1 Results

Considering the complexity of the task, the dataset was evaluated using an LSTM model. The original text was split into words which were then split into a train and a test set - 80% of the data for training and 20% for testing the performance of the model.

Regarding the evaluation, both manual and automated methods were used. The subjective assessments of the predictions are discussed in the subsequent Section (3.3.2). Considering the automated methods, three evaluation metrics were used to assess the performance of the model: perplexity, the BLEU Score, and the ROUGE Score. The final results are displayed in Table 2.

Table 2: Performance metrics for The Adventures of Sherlock Holmes Dataset

Train-Test Split [%]	Perplexity	BLEU Score	ROUGE-1 Score	ROUGE-2 Score	ROUGE-L Score	Training Time [s]
70-30	1.21	0	0.0795	0.0001	0.0795	605.47
80-20	1.21	0	0.0861	0.0002	0.0861	1234.06
90-10	1.19	0	0.0848	0	0.0848	572.92

3.3.2 Analysis

The manual analysis for this dataset was done using the same criteria as before. However, due to the limitations of the approach and training set, fewer samples were examined. Further, only the next three words after a prompt were predicted (shown in bold face).

1. I will leave if they **Lestrade! bearing. kind.**
2. The culprit **remarked us. gave**
3. Watson **shown. Mrs. wall.**
4. No one but I **should wish to**
5. You lost **street, houses." questions**
6. By eleven o'clock **I heard the**
7. a middle-sized man **grass. Of in**
8. Mr. Holmes **gloomily. upon his**
9. Mr. Duncan **veil, and I**

Clearly, only two instances make sense as a whole (4, 6) whereas none of the remaining examples have a meaningful second word except for 8, if viewed in the context of a continuing sentence and ignoring the lower case at the start of the next sentence. If reduced to the first predicted word, sentence number 2 also makes sense.

Overall, this model doesn't perform as well as the first.

4 Conclusion

When looking at perplexity values, the second method, performed on the Sherlock Holmes dataset clearly outperforms the first. Nevertheless, in terms of manual evaluation, the first method gives pretty good results. Unlike the second, where even single-word prediction isn't always that accurate. This is also reflected in the second method's BLEU-score of 0 and the relatively poor ROUGE-scores.

Overall, the combination of method and data set for the first implementation proved more robust.