Evaluating NLP Models

Mayank Singh

ACM-IKDD Summer School on Data Science, IIT Gandhinagar, 8th July 2022

A Brief Recap

- **NLP Tasks**
- **The Traditional NLP pipeline**
- **The Modern NLP pipeline**
- **Cleaning and Processing Text**
- **A walkthrough of Modern NLP pipeline (lab session)**

Let’s look at some popular evaluation metrics

Text Classification Tasks

S.No.	Text	Actual Class
1	Very user friendly...long lasting battery....very clear display	1
2	Worse graphics, won't keep a wireless connection,overall not satisfied	0
3	We have had no issues with this tablet. Love it!TY	1
4	so far. A few not so great things, modt of them managed to resolve them though.	0
5	I brought this tablet during the Black Friday sale fast forward to Christmas when my 6 year opens it he was happy I set it up for him two days later it crashed never came back on. Blank screen. He was disappointed and so was I .	0
6	Great tablet/e-reader. You can do a lot on this bad boy!!	1
7	This is an older tablet. I probobly expected too much. If I'm going to deal with lagginess I prefer my ipad2.	0
8	Really expected a chord with it for the price we paid. Very dissapointed.	0
9	I bought 3, better than we ever dreamed for money spent	1
10	We bought this for my mom. She is very satisfied with this product.	1

Ten reviews from Amazon with sentiment class labels

1: Good Product/Service
0: Bad Product/Service

Text Classification Tasks

S.No.	Text	Actual Class	Predicted Class
1	Very user friendly...long lasting battery....very clear display	1	1
2	Worse graphics, won't keep a wireless connection,overall not satisfied	0	1
3	We have had no issues with this tablet. Love it!TY	1	0
4	so far. A few not so great things, modt of them managed to resolve them though.	0	0
5	I brought this tablet during the Black Friday sale fast forward to Christmas when my 6 year opens it he was happy I set it up for him two days later it crashed never came back on. Blank screen. He was disappointed and so was I .	0	1
6	Great tablet/e-reader. You can do a lot on this bad boy!!	1	0
7	This is an older tablet. I probobly expected too much. If I'm going to deal with lagginess I prefer my ipad2.	0	0
8	Really expected a chord with it for the price we paid. Very dissapointed.	0	1
9	I bought 3, better than we ever dreamed for money spent	1	0
10	We bought this for my mom. She is very satisfied with this product.	1	1

We passed these reviews through a trained sentiment classifier and got the predictions

Text Classification Tasks

S.No.	Text	Actual Class	Predicted Class	Category
1	Very user friendly...long lasting battery....very clear display	1	1	TP
2	Worse graphics, won't keep a wireless connection,overall not satisfied	0	1	FP
3	We have had no issues with this tablet. Love it!TY	1	0	FN
4	so far. A few not so great things, modt of them managed to resolve them though.	0	0	TN
5	I brought this tablet during the Black Friday sale fast forward to Christmas when my 6 year opens it he was happy I set it up for him two days later it crashed never came back on. Blank screen. He was disappointed and so was I .	0	1	FP
6	Great tablet/e-reader. You can do a lot on this bad boy!!	1	0	FN
7	This is an older tablet. I probobly expected too much. If I'm going to deal with lagginess I prefer my ipad2.	0	0	TN
8	Really expected a chord with it for the price we paid. Very dissapointed.	0	1	FP
9	I bought 3, better than we ever dreamed for money spent	1	0	FN
10	We bought this for my mom. She is very satisfied with this product.	1	1	TP

True Positive: Correct prediction for class 1

True Negative: Correct prediction for class 0

False Positive: Incorrect prediction for class 0

False Negative: Incorrect prediction for class 1

Text Classification Tasks

S.No.	Text	Actual Class	Predicted Class	Category	Metrics
1	Very user friendly...long lasting battery....very clear display	1	1	TP	
2	Worse graphics, won't keep a wireless connection,overall not satisfied	0	1	FP	
3	We have had no issues with this tablet. Love it!TY	1	0	FN	**Precision** = TP/(TP+FP) = 2/(2+3) = 2/5
4	so far. A few not so great things, modt of them managed to resolve them though.	0	0	TN	**Recall** = TP/(TP+FN) = 2/(2+3) = 2/5
5	I brought this tablet during the Black Friday sale fast forward to Christmas when my 6 year opens it he was happy I set it up for him two days later it crashed never came back on. Blank screen. He was disappointed and so was I .	0	1	FP	**F-Score** = 2*P*R/(P+R) = 2/5
6	Great tablet/e-reader. You can do a lot on this bad boy!!	1	0	FN	**Accuracy** = (TP + TN)/ (no. of examples) = (2+2)/10 = 2/5
7	This is an older tablet. I probobly expected too much. If I'm going to deal with lagginess I prefer my ipad2.	0	0	TN	
8	Really expected a chord with it for the price we paid. Very dissapointed.	0	1	FP	
9	I bought 3, better than we ever dreamed for money spent	1	0	FN	
10	We bought this for my mom. She is very satisfied with this product.	1	1	TP	

True Positive: Correct prediction for class 1

True Negative: Correct prediction for class 0

False Positive: Incorrect prediction for class 0

False Negative: Incorrect prediction for class 1

Spreadsheet is available at: https://docs.google.com/spreadsheets/d/1MlRx8GA0HCrYRZWQhfj0PFMk7K3vxEH4lVk2NAkcr74/edit#gid=0

Summarization Tasks

- Extrinsic: the summary quality is judged on the basis of how helpful summaries are for a given task.

- Intrinsic: based on analysis of the summary itself
  - It involves a comparison with the source document
  - How many main ideas of the source document are covered by the summary in comparison with an abstract written by a human.

Summarization Tasks

ROUGE (Recall Oriented Understudy for Gisting Evaluation)

Given a document D, and an automatic summary X:

- **Have N humans produce a set of reference summaries of D**
- **What percentage of the bigrams from the reference summaries appear in X?**

$$ROUGE-2 = \frac{\sum_{S \in \{RefSummaries\}} \sum_{bigrams \in S} count_{match}(bigrams)}{\sum_{S \in \{RefSummaries\}} \sum_{bigrams \in S} count(bigrams)}$$

Summarization Tasks

ROUGE (Recall Oriented Understudy for Gisting Evaluation)

Automatic Summary: the cat was found under the bed
Reference Summary: the cat was under the bed

Summarization Tasks

ROUGE (Recall Oriented Understudy for Gisting Evaluation)

Automatic Summary: the cat was found under the bed
Reference Summary: the cat was under the bed

Generated Summary Bigrams: the cat, cat was, was found, found under, under the, the bed
Reference Summary Bigrams: the cat, cat was, was under, under the, the bed

$$\text{ROUGE-2} = \frac{4}{5}$$

Summarization Tasks

ROUGE (Recall Oriented Understudy for Gisting Evaluation)

Automatic Summary: the cat was found under the bed
Reference Summary: the cat was under the bed

Generated Summary Bigrams: the cat, cat was, was found, found under, under the, the bed
Reference Summary Bigrams: the cat, cat was, was under, under the, the bed

$$\text{ROUGE-2} = \frac{4}{5}$$

ROUGE-1, ROUGE-L, ROUGE-S, ROUGE-SU??

Machine Translation Tasks

BLEU (Recall Oriented Understudy for Gisting Evaluation)

Given a machine generated text D (candidates), and an reference translations X:

- **Have N humans produce a set of reference translations X**
- **What percentage of the n-grams in D are present in reference translations X?**

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

Machine Translation Tasks

BLEU: **Precision based metric**

ROUGE: **Recall-based metric**

Text Generation

Perplexity Metric

Perplexity is the inverse probability of the test data, normalized by the number of words

$$PP(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

Text Generation

Perplexity Metric

Perplexity is the inverse probability of the test data, normalized by the number of words

$$PP(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

How do you compute the probability of a sentence?

The Probability Of A Sentence

- A sentence is a sequence of tokens
  $$W = w_1 w_2 w_3 w_4 w_5 w_6 w_7 \ldots w_n$$
- $P(W) = P(w_1 w_2 w_3 w_4 w_5 w_6 w_7 \ldots w_n)$
- $P(w_1 w_2 w_3 w_4 w_5 w_6 w_7 \ldots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2 w_1)\ldots P(w_n|w_{n-1} \ldots w_1)$

The Probability Of A Sentence

- A sentence is a sequence of tokens
  $W = w_1 w_2 w_3 w_4 w_5 w_6 w_7 \ldots w_n$
- $P(W) = P(w_1 w_2 w_3 w_4 w_5 w_6 w_7 \ldots w_n)$
- $P(w_1 w_2 w_3 w_4 w_5 w_6 w_7 \ldots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2 w_1)\ldots P(w_n|w_{n-1} \ldots w_1)$
- **A bigram assumption:** A word only depends on its previous word
- $P(w_1 w_2 w_3 w_4 w_5 w_6 w_7 \ldots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2)\ldots P(w_n|w_{n-1})$

The Probability Of A Sentence

- A sentence is a sequence of tokens
  $$W = w_1 w_2 w_3 w_4 w_5 w_6 w_7 \dots w_n$$
- $P(W) = P(w_1 w_2 w_3 w_4 w_5 w_6 w_7 \dots w_n)$
- $P(w_1 w_2 w_3 w_4 w_5 w_6 w_7 \dots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2 w_1)\dots P(w_n|w_{n-1} \dots w_1)$
- A bigram assumption: A word only depends on its previous word
- $P(w_1 w_2 w_3 w_4 w_5 w_6 w_7 \dots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2)\dots P(w_n|w_{n-1})$

How do you compute the probability of a token?

The Maximum Likelihood Estimate

Bigram Probability

$$P(w_i|w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

Revisiting Perplexity

Perplexity Metric

Perplexity is the inverse probability of the test data, normalized by the number of words

$$PP(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

Perplexity Metric with Bigram Assumption

Perplexity is the inverse probability of the test data, normalized by the number of words

$$PP(W) = \left(\prod \frac{1}{P(w_i|w_{i-1})}\right)^{\frac{1}{N}}$$

Behavioral Testing of NLP Models

Behavioral Testing of NLP Models

Sentiment Analysis

	Test *TYPE* and Description	**Failure Rate (%)**					**Example test cases & expected behavior**
						RoB	
Vocab.+POS	***MFT:*** Short sentences with neutral adjectives and nouns	0.0	7.6	4.8	94.6	81.8	The company is Australian. neutral That is a private aircraft. neutral
	MFT: Short sentences with sentiment-laden adjectives	4.0	15.0	2.8	0.0	0.2	That cabin crew is extraordinary. pos I despised that aircraft. neg
	INV: Replace neutral words with other neutral words	9.4	16.2	12.4	10.2	10.2	@Virgin should I be concerned that → when I'm about to fly ... INV @united the → our nightmare continues... INV
	DIR: Add positive phrases, fails if sent. goes down by > 0.1	12.6	12.4	1.4	0.2	10.2	@SouthwestAir Great trip on 2672 yesterday... You are extraordinary. ↑ @AmericanAir AA45 ... JFK to LAS. You are brilliant. ↑
	DIR: Add negative phrases, fails if sent. goes up by > 0.1	0.8	34.6	5.0	0.0	13.2	@USAirways your service sucks. You are lame. ↓ @JetBlue all day. I abhor you. ↓
Robust.	***INV:*** Add randomly generated URLs and handles to tweets	9.6	13.4	24.8	11.4	7.4	@JetBlue that selfie was extreme. @pi9QDK INV @united stuck because staff took a break? Not happy 1K.... https://t.co/PWK1jb INV
	INV: Swap one character with its neighbor (typo)	5.6	10.2	10.4	5.2	3.8	@JetBlue → @JeBtlue I cri INV @SouthwestAir no thanks → thakns INV
NER	***INV:*** Switching locations should not change predictions	7.0	20.8	14.8	7.6	6.4	@JetBlue I want you guys to be the first to fly to # Cuba → Canada... INV @VirginAmerica I miss the #nerdbird in San Jose → Denver INV
	INV: Switching person names should not change predictions	2.4	15.1	9.1	6.6	2.4	...Airport agents were horrendous. Sharon → Erin was your saviour INV @united 8602947, Jon → Sean at http://t.co/58tuTgli0D, thanks. INV

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList, ACL 2020

Behavioral Testing of NLP Models

Sentiment Analysis

Test *TYPE* and Description		Failure Rate (%)					Example test cases & expected behavior
			G	a		RoB	
Temporal	***MFT:*** Sentiment change over time, present should prevail	41.0	36.6	42.2	18.8	11.0	I used to hate this airline, although now I like it. pos In the past I thought this airline was perfect, now I think it is creepy. neg
Negation	***MFT:*** Negated negative should be positive or neutral	18.8	54.2	29.4	13.2	2.6	The food is not poor. pos or neutral It isn't a lousy customer service. pos or neutral
	MFT: Negated neutral should still be neutral	40.4	39.6	74.2	98.4	95.4	This aircraft is not private. neutral This is not an international flight. neutral
	MFT: Negation of negative at the end, should be pos. or neut.	100.0	90.4	100.0	84.8	7.2	I thought the plane would be awful, but it wasn't. pos or neutral I thought I would dislike that plane, but I didn't. pos or neutral
	MFT: Negated positive with neutral content in the middle	98.4	100.0	100.0	74.0	30.2	I wouldn't say, given it's a Tuesday, that this pilot was great. neg I don't think, given my history with airplanes, that this is an amazing staff. neg
SRL	***MFT:*** Author sentiment is more important than of others	45.4	62.4	68.0	38.8	30.0	Some people think you are excellent, but I think you are nasty. neg Some people hate you, but I think you are exceptional. pos
	MFT: Parsing sentiment in (question, "yes") form	9.0	57.6	20.8	3.6	3.0	Do I think that airline was exceptional? Yes. neg Do I think that is an awkward customer service? Yes. neg
	MFT: Parsing sentiment in (question, "no") form	96.8	90.8	81.6	55.4	54.8	Do I think the pilot was fantastic? No. neg Do I think this company is bad? No. pos or neutral

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList, ACL 2020

Behavioral Testing of NLP Models

Quora Question Pair

	Test *TYPE* and Description	Failure Rate		Example Test cases & expected behavior	
			RoB		
Vocab.	***MFT:*** Modifiers changes question intent	78.4	78.0	{ Is Mark Wright a photographer? | Is Mark Wright an accredited photographer? }	≠
Taxonomy	***MFT:*** Synonyms in simple templates	22.8	39.2	{ How can I become more vocal? | How can I become more outspoken? }	=
	INV: Replace words with synonyms in real pairs	13.1	12.7	Is it necessary to follow a religion? Is it necessary to follow an organized → organised religion?	INV
	MFT: More X = Less antonym(X)	69.4	100.0	{ How can I become more optimistic? | How can I become less pessimistic? }	=
Robust.	***INV:*** Swap one character with its neighbor (typo)	18.2	12.0	{ Why am I getting → gettnig lazy? | Why are we so lazy? }	INV
	DIR: Paraphrase of question should be duplicate	69.0	25.0	Can I gain weight from not eating enough? Can I → Do you think I can gain weight from not eating enough?	=
NER	***INV:*** Change the same name in both questions	11.8	9.4	Why isn't Hillary Clinton → Nicole Perez in jail? Is Hillary Clinton → Nicole Perez going to go to jail?	INV
	DIR: Change names in one question, expect ≠	35.1	30.1	What does India think of Donald Trump? What India thinks about Donald Trump → John Green?	≠
	DIR: Keep first word and entities of a question, fill in the gaps with RoBERTa; expect ≠	30.0	32.8	Will it be difficult to get a US Visa if Donald Trump gets elected? Will the US accept Donald Trump?	≠

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList, ACL 2020

Behavioral Testing of NLP Models

Quora Question Pair

	Test *TYPE* and Description	Failure Rate		Example Test cases & expected behavior
			RoB	
Temporal	***MFT:*** Is ≠ used to be, non-duplicate	61.8	96.8	{ Is Jordan Perry an advisor? | Did Jordan Perry use to be an advisor? } ≠
	MFT: before ≠ after, non-duplicate	98.0	34.4	{ Is it unhealthy to eat after 10pm? | Is it unhealthy to eat before 10pm? } ≠
	MFT: before becoming ≠ after becoming	100.0	0.0	What was Danielle Bennett's life before becoming an agent? What was Danielle Bennett's life after becoming an agent? } ≠
Negation	***MFT:*** simple negation, non-duplicate	18.6	0.0	{ How can I become a person who is not biased? | How can I become a biased person? } ≠
	MFT: negation of antonym, should be duplicate	81.6	88.6	{ How can I become a positive person? | How can I become a person who is not negative } ≠
Coref	***MFT:*** Simple coreference: he ≠ she	79.0	96.6	If Joshua and Chloe were alone, do you think he would reject her? If Joshua and Chloe were alone, do you think she would reject him? } ≠
	MFT: Simple resolved coreference, his and her	99.6	100.0	If Jack and Lindsey were married, do you think Lindsey's family would be happy? If Jack and Lindsey were married, do you think his family would be happy? } ≠
SRL	***MFT:*** Order is irrelevant for comparisons	99.6	100.0	{ Are tigers heavier than insects? | What is heavier, insects or tigers? } =
	MFT: Orders is irrelevant in symmetric relations	81.8	100.0	{ Is Nicole related to Heather? | Is Heather related to Nicole? } =
	MFT: Order is relevant for asymmetric relations	71.4	100.0	{ Is Sean hurting Ethan? | Is Ethan hurting Sean? } ≠
	MFT: Active / passive swap, same semantics	65.8	98.6	{ Does Anna love Benjamin? | Is Benjamin loved by Anna? } =
	MFT: Active / passive swap, different semantics	97.4	100.0	{ Does Danielle support Alyssa? | Is Danielle supported by Alyssa? } ≠

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList, ACL 2020

Behavioral Testing of NLP Models

https://github.com/marcotcr/checklist

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList, ACL 2020

Email: singh.mayank@iitgn.ac.in
Webpage: https://mayank4490.github.io/