HSLU Lucerne University of Applied Sciences and Arts

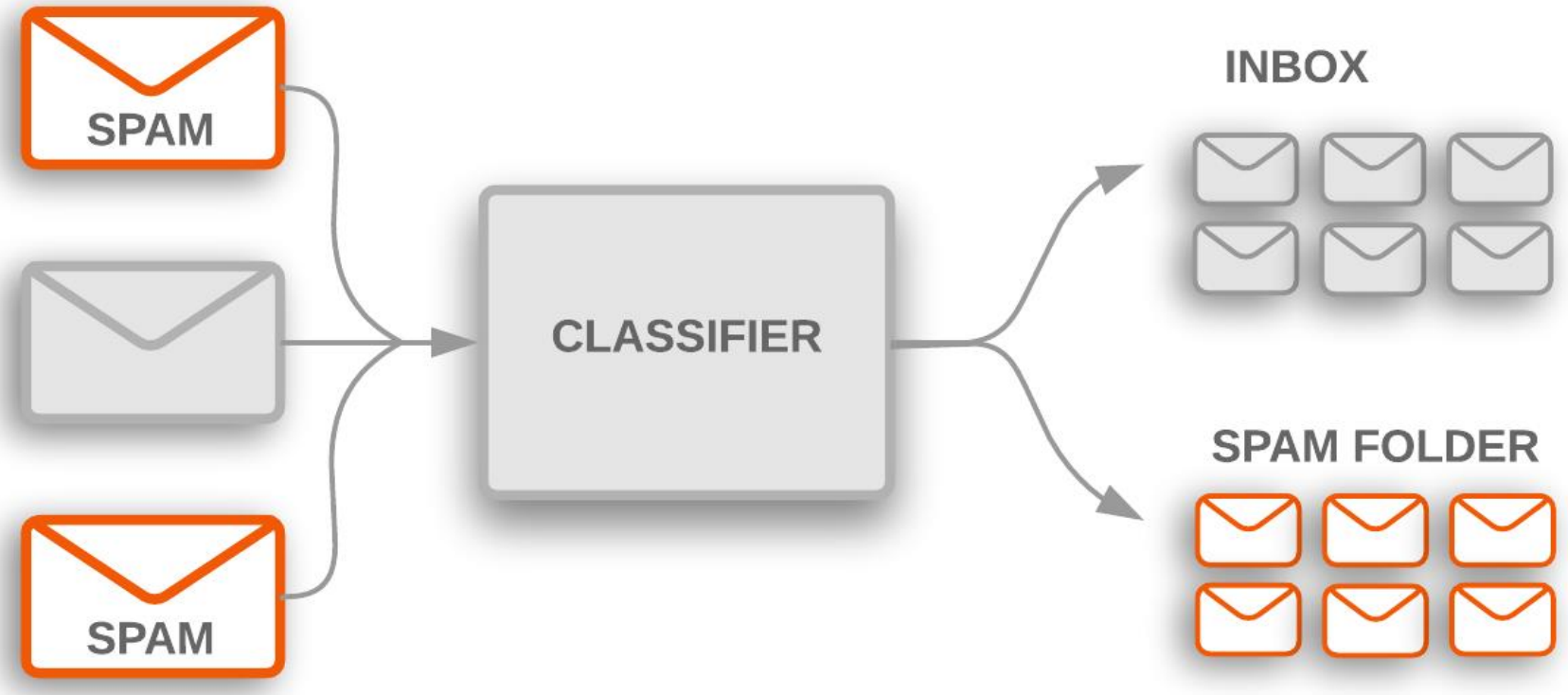
Text Classification

NLP
Andreas Marfurt

Information Technology
08.05.2025

FH Zentralschweiz

Motivation



https://developers.google.com/machine-learning/guides/text-classification

Motivation

Top reviews from the United States

Toni K. Hensley

Great light!

Reviewed in the United States us on November 8, 2022

Color: Pink | **Verified Purchase**

Very warm and gives out great light. My husband and I walk very early (4am) and this really ramps up our safety as senior citizens.

Helpful | Report abuse

Customer reviews

4.5 out of 5

9,107 global ratings

5 star	73%
4 star	16%
3 star	6%
2 star	2%
1 star	3%

How customer reviews and ratings work

By feature

Brightness	4.6
Comfort	4.5
Value for money	4.5

See more

Motivation

```
>>> from transformers import pipeline
>>> classifier = pipeline(“sentiment-analysis”)
>>> classifier("Very warm and gives out great light. My husband and I walk very early (4am) and this really ramps up our safety as senior citizens.")
[{'label': 'POSITIVE', 'score': 0.9997850060462952}]
```

Top reviews from the United States

Toni K. Hensley

Great light!

Reviewed in the United States us on November 8, 2022

Color: Pink | **Verified Purchase**

Very warm and gives out great light. My husband and I walk very early (4am) and this really ramps up our safety as senior citizens.

Helpful | Report abuse

Customer reviews

4.5 out of 5

9,107 global ratings

5 star	73%
4 star	16%
3 star	6%
2 star	2%
1 star	3%

How customer reviews and ratings work

By feature

Brightness	4.6
Comfort	4.5
Value for money	4.5

See more

Overview

- General Text Classification
- Sentiment Analysis
- Natural Language Inference

Material taken and adapted from *Speech and Language Processing* by Dan Jurafsky and James H. Martin (link)

General Text Classification

Text Classification

- Input
  - Document d
  - Set of classes $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$
- Output
  - Predicted class c

Applications

- Spam Detection
- Topic Identification
- Language Identification
- Malicious Input Detection
- Grammatical Correctness

Classify 1 sentence

- Duplicate Detection
- Paraphrase Detection

Compare 2 sentences

Hand-coded Rules

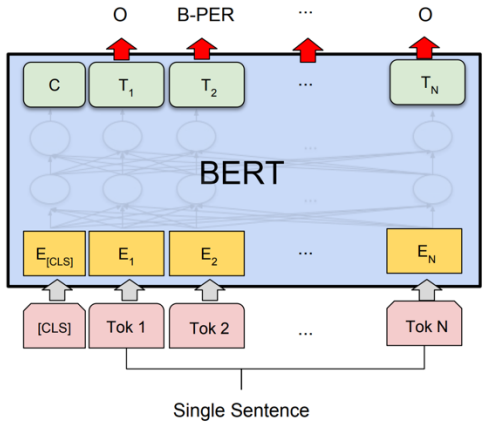
- Manually defined based on features (words and other)
  - is_spam = black-list-address OR (“dollars” AND “you have been selected”)
- Can have high accuracy
  - Carefully defined rules by expert
- Building and *maintaining* is expensive

Supervised Learning

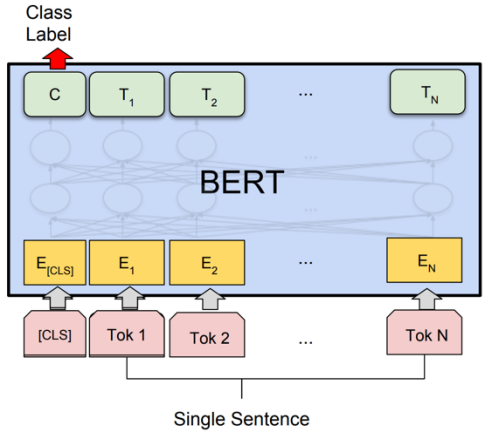
- Training data: document-class pairs (d_i, c_i)
- Can use any classifier
  - Naïve Bayes
  - Logistic regression
  - SVM
  - k-nearest neighbors
  - Neural networks

BERT for Text Classification

- Classifying each token



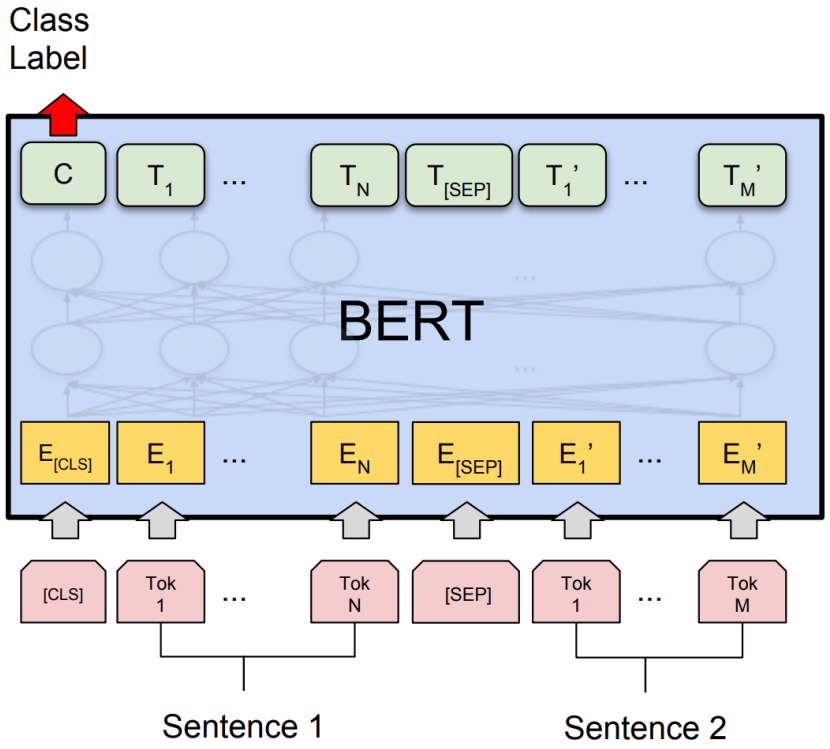
- Classifying the entire sentence



Devlin et al., 2019

BERT for Text Classification

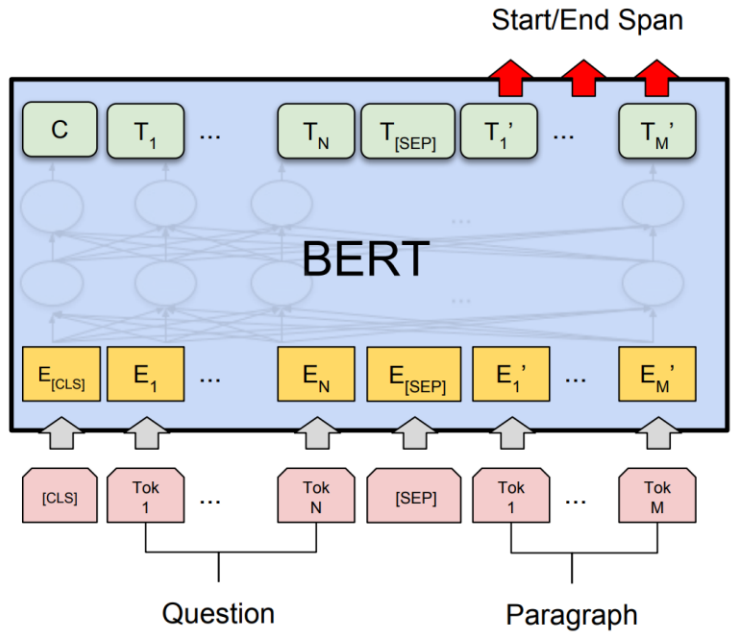
- Classifying one sentence with respect to another



Devlin et al., 2019

BERT for Text Classification

- Extractive question answering was also just a classification task:
  - Classification of best start/end position



Devlin et al., 2019

Sentiment Analysis

Sentiment Analysis

Also known as:

- Opinion mining
- Opinion extraction
- Sentiment mining
- Subjectivity analysis

Resources:

- Sentiment analysis and text classification slides from Speech and Language Processing by Dan Jurafsky and James H. Martin (2021)
- Sentiment analysis tutorial by Bing Liu (2011)

Customer Reviews

- Movies

  + Hilarious comedy. I laughed so much.

- Restaurants

  - Terrible service and the food was overpriced.

Sentiment Analysis

- Movie reviews
- Product reviews
- Elections
- Public sentiment towards a product/policy
- Used for predictions about product success/stock market development/election results

Typology of Affective States

- **Emotion**: evaluation of a major event (short-term)
  - angry, sad, joyful, fearful, ashamed, proud, elated
- **Mood**: change in subjective feeling (diffuse, non-caused, low-intensity, long-duration)
  - cheerful, gloomy, irritable, listless, depressed, buoyant
- **Interpersonal stances**: affective stance toward another person in a specific interaction
  - friendly, flirtatious, distant, cold, warm, supportive, contemptuous
- **Attitudes**: beliefs/dispositions towards objects/persons (long-term)
  - liking, loving, hating, valuing, desiring
- **Personality traits**: stable personality dispositions and typical behavior tendencies
  - nervous, anxious, reckless, morose, hostile, jealous

Typology of Affective States

- **Emotion**: evaluation of a major event (short-term)
  - angry, sad, joyful, fearful, ashamed, proud, elated
- **Mood**: change in subjective feeling (diffuse, non-caused, low-intensity, long-duration)
  - cheerful, gloomy, irritable, listless, depressed, buoyant
- **Interpersonal stances**: affective stance toward another person in a specific interaction
  - friendly, flirtatious, distant, cold, warm, supportive, contemptuous
- **Attitudes**: beliefs/dispositions towards objects/persons (long-term)
  - liking, loving, hating, valuing, desiring

Sentiment analysis

- **Personality traits**: stable personality dispositions and typical behavior tendencies
  - nervous, anxious, reckless, morose, hostile, jealous

Sentiment Analysis

… is the detection of attitudes
(long-term beliefs/dispositions towards objects/persons)

Definition of attitudes by Jurafsky and Martin (2021):

- Holder (source) of attitude
- Target of attitude
- Type of attitude (like, love, hate vs. ++/+/o/-/--)
- Text containing the attitude

Sentiment Analysis

Definition for *opinions* by Liu (2011) adds aspect a and time t

An opinion is a tuple (e, a, s, p, t)

- e: entity about which the opinion is expressed
- a: aspect of the entity targeted by the opinion (optional)
- s: sentiment expressed (positive or negative)
- p: person expressing the opinion (most often: the author)
- t: moment when the opinion is stated (optional)

Example Movie Review

Fantastic Beasts: The Secrets of Dumbledore (2022)

7/10 Stars

NOTHING FELT EARNED.

13 April 2022

Though the style part was enjoyable on it's own; the cinematography, special effects, costumes, score, set design and much of the acting was brilliant. It felt great to lose myself into the feel of the wizarding world again. But the plot and the writing felt unanchored in anything real or relatable. Nothing felt earned. Events just happened without much anticipation or appreciation of their significance, and I didn't feel invested much in any of the plot lines or characters. I'm not expert enough to pinpoint exactly what went wrong, but I know something did.

https://www.imdb.com/review/rw8056446/?ref_=tt_urv

Negation

- Can change positive to negative sentiment:
  I like this movie.
  I don’t like this movie.
- Can also change negative into positive sentiment:
  Don’t dismiss this film.
  You won’t get bored.
- Early idea: add “NOT_” to every word after a negation:
  I don’t like this movie. → I don’t NOT_like NOT_this NOT_movie .
  - These become separate words
  - Must add a negated version for every word in the vocabulary
  - Doesn't work with double negation

Tokenization for Sentiment Analysis

- Preserve capitalization: emphasis/screaming
- Preserve numbers: dates, scores
- Don’t split punctuation/special characters to preserve emoticons:

  :-) =O (ﾉ°□°)ﾉ ⌒ ┻━┻

- Emojis: even more special characters 😃
- Preserve usernames on Twitter/Reddit (@user, u/user)

Features

- Emoticons and Emojis are very strong features!
- Some words are very strong features as well

  → Build a lexicon of positive/negative words (together with how strongly positive/negative they are)

\+ admirable, beautiful, confident, dazzling, ecstatic, favor, glee, great, :-), :D

– awful, bad, bias, catastrophe, cheat, deny, envious, foul, harsh, hate, :(, >:-(

Sentiment Lexicons

- Multiple sentiment lexicons are available:
  - The General Inquirer (Stone et al., 1966, homepage, free for research use)
    - 1915 positive and 2291 negative words
    - Grouped according to several categories: strong vs. weak, active vs. passive, pleasure vs. pain, virtue vs. vice, …
  - MPQA Subjectivity Lexicon (Wilson et al., 2005, homepage)
    - 2718 positive and 4912 negative words
    - Annotated with intensity (strong/weak)
  - SentiWordNet (Baccianella et al., 2010, homepage)
    - Groups WordNet synsets (concepts) into positive/negative/objective
  - VADER (Hutto and Gilbert, 2014, homepage)
    - Especially for social media (includes abbreviations and slang)
    - 7500 words, score from -4 to +4
  - AFINN (Nielsen, 2011, homepage)
    - 2477 words on phrases, for tweets

with Python package

Sentiment Lexicons

- Select the right one for your domain
- Useful when little/no training data
- How to use:
  - Count the positive/negative words in your text
  - Classify as positive if:
    - More positive than negative words:

$$\frac{\text{count}(\text{pos})}{\text{count}(\text{neg})} > 1$$

    - Positive words must outweigh negative words (choose a factor larger than 1)

Sentiment Analysis with BERT

- If you have enough data available for your language & domain:
  - Use a pretrained BERT model for single sequence classification
  - Finetune it on your data
  - This should outperform sentiment lexicons
- If you don’t have enough data, compare with the transfer performance of a model trained on different (but similar) data
  - Should be the same task
  - The closer the data to your domain, the better

Sentiment Analysis Datasets

- IMDB movie reviews (Maas et al., 2011, homepage)
  - 25k train and 25k test reviews with positive/negative sentiment
- Sentiment polarity/scale and subjectivity datasets, also from IMDB movie reviews (Pang and Lee, 2002-2005, homepage)
  - Subjectivity: 5000 subjective vs. 5000 objective sentences
- Amazon product reviews (Ni, 2018, homepage)
  - 233.1 million reviews
- SST: Stanford Sentiment Treebank (Socher et al., 2013, homepage)
  - Movie reviews from RottenTomatoes
  - 11855 sentences with positive/negative labels
  - 215,154 phrases labeled with polarity (--, -, 0, +, ++)
  - Part of the GLUE benchmark

Evaluation

- Accuracy

- If you care more about one class than the other (e.g. detecting negative sentiment reviews on your product), then use precision/recall/F1 score

In-class exercise: Sentiment Analysis

Natural Language Inference

Natural Language Inference (NLI)

- Also called (Recognizing) Textual Entailment (RTE)
- Sentence 1: Premise
- Sentence 2: Hypothesis
- Task: Given 2 sentences, determine if the first sentence
  - entails (we can conclude H from P)
  - contradicts (H contradicts P)
  - is neutral towards (P and H can both be true)

  the second sentence.

Natural Language Inference

- Sometimes viewed as a fundamental task of natural language understanding
  - Included in many NLU benchmarks (GLUE, SuperGLUE, DecaNLP)
  - Requires logical and commonsense reasoning
  - Many tasks can be formulated as an NLI task *(we have seen this before…)*

Tasks formulated as NLI:

- Paraphrase: s1 entails s2 AND s2 entails s1
- Summarization: source document entails summary
- Information retrieval: retrieved documents entail query
- Question answering: answer entails question

Natural Language Inference

- Assumption: premise and hypothesis talk about the same event

  Honda's sales fell by 5 percent contradicts *Honda's sales rose by 5 percent*

These could both be true, but not at the same time, for the same event.

- Assumption: they describe events/facts about our world (and not some hypothetical other world)

  Turtle contradicts *linguist*

In our world, turtles can't be linguists.

Strict Logical Reasoning vs. Commonsense Reasoning

- *Twitter reported that its CEO resigned* entails *Twitter's CEO resigned*
  - The company could be lying (can't use strict logical inference)
  - Question to ask yourself when deciding
    - "Does the premise justify an inference to the hypothesis?"
    - "Would a human reading the premise infer that the hypothesis is most likely true?" – Ido Dagan (2005)
  - In this case, we would say that yes, the company reporting it is sufficient information to assume that the CEO resigned
  - For more: discussion of the definition by Manning (2006)

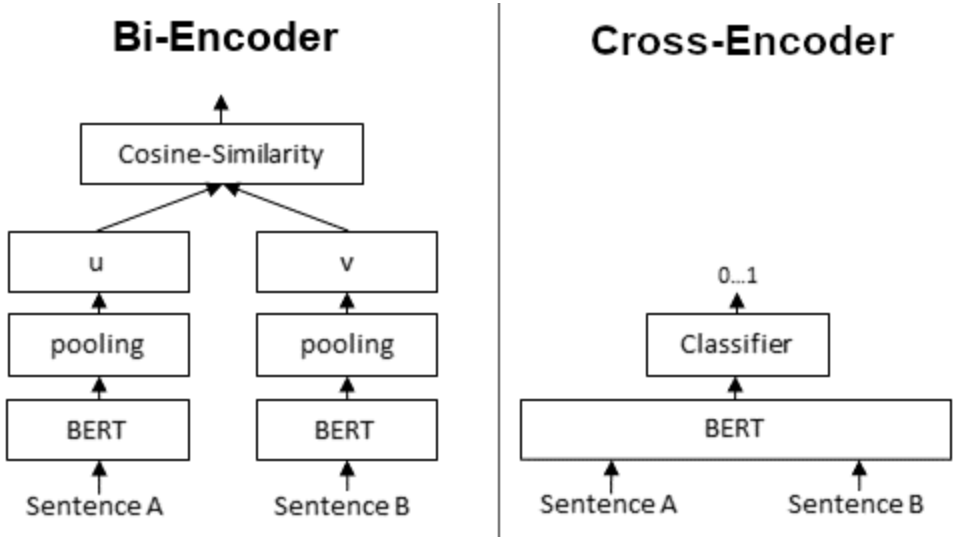
Required Capabilities

- Lexical entailment (cat vs. animal/dog)
  - cat entails animal, but contradicts dog
- Quantification (all, most, some, fewer than eight)
- Lexical/scope ambiguity (polysemy: bank)
- Modality (might, should)
- Common sense

Other sentence classification tasks are simpler: sentiment analysis, sentence similarity

NLI Models

- Bi-encoders: Process premise and hypothesis separately → sentence vectors → compare
- Cross-encoders: Process premise and hypothesis jointly
  - Can use cross-sentence information, e.g. from self-attention



https://www.sbert.net/examples/applications/cross-encoder/README.html

Bi-encoders

- More general goal: create useful sentence representations
  - Should contain all relevant information for downstream task
- Older methods use an RNN (Chen et al., 2017) or CNN encoder (Yoon et al., 2018)
  - Sentence representations for premise (p) and hypothesis (h)
  - Compare the two representations: [p, h, p ⊙ h, p – h]
  - Feed to classifier
- Now: Finetune pretrained encoder-only models
  - Cosine similarity between [CLS] output representations in BERT/RoBERTa

Cross-encoders

- ESIM: Enhanced Sequential Inference Model (Chen et al., 2017)
  - Compute word representations for each sentence separately (they used BiLSTMs in 2017)
  - For each word in s_1: compute attention weights to s_2, then create an attention-weighted vector of s_2's word representations
  - Concatenate with original word vector
    - Sometimes: Also compute difference/element-wise multiplication between the two vectors and concatenate them all
  - Run another BiLSTM on the concatenated vectors to get the final outputs
- Finetuning BERT: feed premise and hypothesis in the format [CLS] p [SEP] h [SEP] → use [CLS] token to classify

Interaction

NLI Datasets

- SICK: Sentences involving compositional knowledge (Marelli et al., 2014)
  - Restricted task: No named entities, no idioms
  - Sentences created by rules from image/video captions
  - 10,000 examples labeled for entailment and semantic similarity
  - Example:
    P: The brown horse is near a red barrel at the rodeo
    H: The brown horse is far from a red barrel at the rodeo
    Label: contradiction

NLI Datasets

- SNLI: Stanford NLI Corpus (Bowman et al., 2015)
  - Premises from image captions (Flickr 30k), hypotheses created by crowd workers
  - 570k examples
  - Neural networks started performing well on NLI with this amount of data
  - Example:
    P: A black race car starts up in front of a crowd of people.
    H: A man is driving down a lonely road.
    Label: contradiction

NLI Datasets

- MNLI: Multi-Genre NLI Corpus (Williams et al., 2018)
  - Extend SNLI to 10 genres
    - Face-to-face, Telephone, 9/11, Travel, Letters, Oxford University Press, Slate, Verbatim, Government and Fiction
  - Premises from OANC, hypotheses from crowd workers
  - 433k examples
  - Example (telephone genre):
    P: yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual
    H: August is a black out month for vacations in the company.
    Label: contradiction

SNLI/MNLI Annotation Artifacts

- Easy solution to create a contradiction: Add a negation (“not”)
  - Premise: ? *(can be ignored)*
    Hypothesis: Someone is not crossing the road.
    Label: contradiction
- Easy solution to create an entailment: Add a generalization (“is outside”)
  - Premise: ?
    Hypothesis: Someone is outside.
    Label: entailment
- Why are they created?
  - Crowd workers get paid by the number of examples that they can do per time → Quick solution earns the most
  - From the crowd workers’ point of view: if they spend a lot of time per example, their hourly salary on e.g. Amazon Mechanical Turk will be *extremely* low

SNLI/MNLI Annotation Artifacts

- Models learn to use the shortcut and neglect the actual task
  - Newer research shows this is not actually the case
- Can train a hypothesis-only baseline model to see how well it performs without looking at the premise
  - SICK
    - Majority class: 56.76%
    - Hypothesis-only model: 56.76%
  - SNLI
    - Majority class: 33.82%
    - Hypothesis-only model: 69.17%
  - MNLI
    - Majority class: 35.45%
    - Hypothesis-only model: 55.52%

Didn't find a shortcut

Big improvement over majority class (but full models get 80-90% accuracy)

NLI Challenge Sets

"Breaking NLI systems with sentences that require simple lexical inferences" (Glockner et al., 2018)

- Replace just a few words in the premise according to a predefined rule → Rule determines the label

Premise/Hypothesis	Label
The man is holding a saxophone The man is holding an electric guitar	contradiction[1]
A little girl is very sad. A little girl is very unhappy.	entailment
A couple drinking wine A couple drinking champagne	neutral

NLI Challenge Sets

"Breaking NLI systems with sentences that require simple lexical inferences" (Glockner et al., 2018)

- Replace just a few words in the premise according to a predefined rule → Rule determines the label
- Previously strong models perform very bad on this new dataset

Model	Train set	SNLI test set	New test set	Δ
Decomposable Attention (Parikh et al., 2016)	SNLI	84.7%	51.9%	-32.8
	MultiNLI + SNLI	84.9%	65.8%	-19.1
	SciTail + SNLI	85.0%	49.0%	-36.0
ESIM (Chen et al., 2017)	SNLI	87.9%	65.6%	-22.3
	MultiNLI + SNLI	86.3%	74.9%	-11.4
	SciTail + SNLI	88.3%	67.7%	-20.6

Adding Explanations

e-SNLI: Adds explanation to SNLI dataset (Camburu et al., 2018)

- Explanations from crowd workers
  - Highlight the relevant parts in the premise and hypothesis
  - Give a natural language explanation for the label
- Example:

Premise: An adult dressed in black holds a stick.
Hypothesis: An adult is walking away, empty-handed.
Label: contradiction
Explanation: Holds a stick implies using hands so it is not empty-handed.

- Task: Model should generate explanation as well

Applications for NLI Models

- Fact extraction and verification
  - Can factoid claims be supported/refuted using evidence from Wikipedia
  - Dataset: FEVER (Thorne et al., 2018)
- Hallucination/factuality detection in summarization
  - Does the news article entail the summary?
- Multi-hop reasoning (logical reasoning over multiple steps)
  - NLI models were integrated as part of the final model
  - Datasets: MultiRC, OpenBookQA