Is it time to move beyond sentence classification?

Jeremy Barnes

AIST 2021 - Tbilisi, Georgia

17.12.2021

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

Motivation

Why did I choose this title?

Motivation

- sentiment classification
- topic classification
- language identification
- intent classification (chatbots)

Prevalence in benchmarking models

Prevalence in benchmarking models

- Multitask Benchmarking: GLUE (Wang et al., 2018)

Prevalence in benchmarking models

- Multitask Benchmarking: GLUE (Wang et al., 2018)
- Dynamic Sentiment Analysis Benchmark (Potts et al., 2021)

Prevalence in benchmarking models

- Multitask Benchmarking: GLUE (Wang et al., 2018)
- Dynamic Sentiment Analysis Benchmark (Potts et al., 2021)
- Benchmarking Few-shot performance of Large Language Models (LLMs) (Gao et al., 2021)

Speculative reason for prevalence

Speculative reason for prevalence

- They are easier to annotate

Speculative reason for prevalence

- They are easier to annotate
- Because of this, sentence-level classification datasets are often large - better for deep learning models

Speculative reason for prevalence

- They are easier to annotate
- Because of this, sentence-level classification datasets are often large - better for deep learning models
- Conceptually they allow for simpler train/test procedures

Today's goal

Today's goal

Sentence classification is often not an ideal way to benchmark models.

Problems with sentence-level classification

What values do we care about?

The Values Encoded in Machine Learning Research (Birhane et al., 2021)

What values do we care about?

The Values Encoded in Machine Learning Research (Birhane et al., 2021)

1. performance,
2. generalization,
3. efficiency,
4. researcher understanding,
5. novelty,
6. building on previous work

Accuracy/F1/BLEU

Accuracy/F1/BLEU

These allow for fair comparison of new models across many tasks.

Accuracy/F1/BLEU

These allow for fair comparison of new models across many tasks.

Allows the community to focus on a single number and be happy when the numbers go up.

Accuracy/F1/BLEU

These allow for fair comparison of new models across many tasks.

Allows the community to focus on a single number and be happy when the numbers go up.

"We've achieved superhuman performance on task B!"

Case study: sentiment analysis

- Movie Reviews dataset (Pang et al., 2002)
- Camara Review dataset (Hu and Liu, 2004)
- Subjectivity dataset (Pang and Lee, 2004)
- MPQA Subjectivity dataset (Wiebe et al., 2004)
- Stanford Sentiment Treebank (Socher et al., 2013)

Case study: sentiment analysis

Current SOTA on several of these datasets is incredibly high.

Case study: sentiment analysis

Current SOTA on several of these datasets is incredibly high.

- Stanford Sentiment Treebank binary: 97.5
- Movie Reviews binary: 92.5
- Subjectivity dataset: 95.5

Case study: sentiment analysis

Current SOTA on several of these datasets is incredibly high.

- Stanford Sentiment Treebank binary: 97.5
- Movie Reviews binary: 92.5
- Subjectivity dataset: 95.5

Are the models really that good?

Not really...

Not really...

**Sentiment analysis is not solved!:
Assessing and probing sentiment classification**

Jeremy Barnes, Lilja Øvrelid, Erik Velldal

University of Oslo
{jeremycb,liljao,erikve}@ifi.uio.no

Not really...

Sentiment analysis is not solved!:
Assessing and probing sentiment classification

Jeremy Barnes, Lilja Øvrelid, Erik Velldal
University of Oslo
{jeremycb,liljao,erikve}@ifi.uio.no

We collected a subset of sentences that four models (BOW, BiLSTM, ELMO, BERT) all failed on.

Error types can roughly be divided into the following categories:

Not really...

Sentiment analysis is not solved!:
Assessing and probing sentiment classification

Jeremy Barnes, Lilja Øvrelid, Erik Velldal
University of Oslo
{jeremycb,liljao,erikve}@ifi.uio.no

We collected a subset of sentences that four models (BOW, BiLSTM, ELMO, BERT) all failed on.

Error types can roughly be divided into the following categories:

- annotation related (incorrect label, mixed sentiment)
- data related (non-standard spelling, emoji)
- setup related (negation, modality, amplifiers, polarity shifters, polarity reducers)

Sentiment analysis is not solved!

The sentence-level setup hides the fact that models perform poorly on certain subsets of the data:

Sentiment analysis is not solved!

The sentence-level setup hides the fact that models perform poorly on certain subsets of the data:

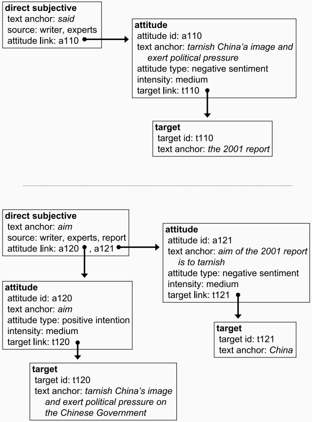
- negation
- modality
- compositional knowledge (amplifiers, reducers)

Simplification

Simplification

MPQA dataset (Wiebe et al., 2005)

Figure 7.3: Private state, attitude, and target frames for sentence 7.18



Simplification

Stanford Sentiment Treebank (Socher et al., 2013)

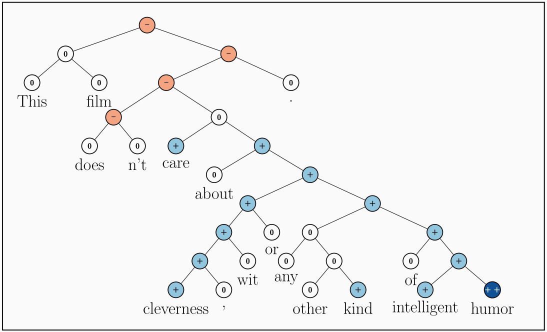


Figure 1: Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (– –, –, 0, +, + +), at every node of a parse tree and capturing the negation and its scope in this sentence.

Simplification

Many other sentiment datasets...

Simplification

They have been converted to sentence classification and further binarized.

An example from language identification

An example from language identification

Error analysis at sentence-level is difficult

Error analysis at sentence-level is difficult

Base model	brilliant	and	moving	performances	by	tom	and	peter	finch
Jain and Wallace (2019)	brilliant	and	moving	performances	by	tom	and	peter	finch
Our adversary	brilliant	and	moving	performances	by	tom	and	peter	finch

Figure 2: Attention maps for an IMDb instance (all predicted as positive with score > 0.998), showing that in practice it is difficult to learn a distant adversary which is consistent on all instances in the training set.

Error analysis at sentence-level is difficult

Base model	brilliant	and	moving	performances	by	tom	and	peter	finch
Jain and Wallace (2019)	brilliant	and	moving	performances	by	tom	and	peter	finch
Our adversary	brilliant	and	moving	performances	by	tom	and	peter	finch

Figure 2: Attention maps for an IMDb instance (all predicted as positive with score > 0.998), showing that in practice it is difficult to learn a distant adversary which is consistent on all instances in the training set.

Although there has been some back and forth on whether this is a useful approach or not

Error analysis at sentence-level is difficult

Base model	brilliant	and	moving	performances	by	tom	and	peter	finch
Jain and Wallace (2019)	brilliant	and	moving	performances	by	tom	and	peter	finch
Our adversary	brilliant	and	moving	performances	by	tom	and	peter	finch

Figure 2: Attention maps for an IMDb instance (all predicted as positive with score > 0.998), showing that in practice it is difficult to learn a distant adversary which is consistent on all instances in the training set.

Although there has been some back and forth on whether this is a useful approach or not

- Attention is not Explanation (Jain and Wallace, 2019)
- Attention is not not Explanation (Wiegreffe and Pinter, 2019)

Sentence-level prediction is not always particularly informative

Sentence-level prediction is not always particularly informative

If we have a model that performs binary sentiment prediction at 97.5 percent accuracy (superhuman level!)...

Sentence-level prediction is not always particularly informative

If we have a model that performs binary sentiment prediction at 97.5 percent accuracy (superhuman level!)...

what would it mean if that model predicts 'positive' for the following sentence?

Sentence-level prediction is not always particularly informative

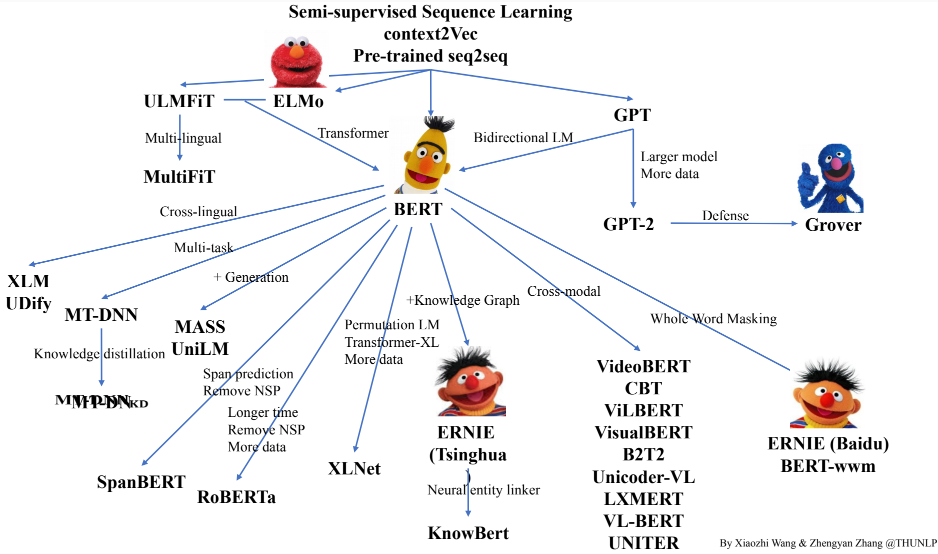
If we have a model that performs binary sentiment prediction at 97.5 percent accuracy (superhuman level!)...

what would it mean if that model predicts 'positive' for the following sentence?

"James went to the store."

Do large language models reduce these problems?

Large language models



Gains in performance

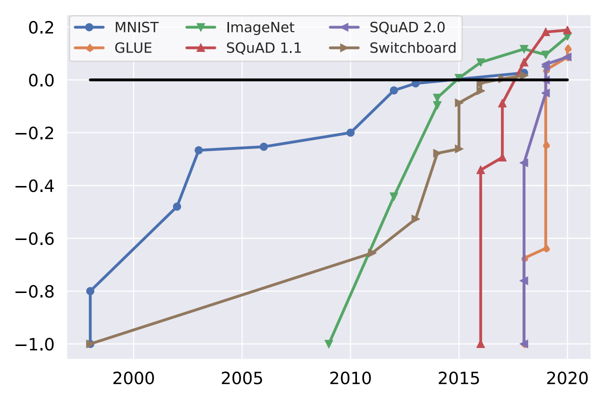


Figure 1: Benchmark saturation over time for popular benchmarks, normalized with initial performance at minus one and human performance at zero.

Kiela et al. (2020) Dynabench

Benchmarking

Sentence classification commonly used in benchmarking large language models.

Benchmarking

Sentence classification commonly used in benchmarking large language models.

Of the tasks used, largest gains usually on sentence-classification tasks.

Better numbers

Gains on SST-2 (binarized sentence classification)

Better numbers

Gains on SST-2 (binarized sentence classification)

- ELMo (from bert paper): 90.4 (Peters et al., 2018)
- byte mLSTM: 91.8 (Radford et al., 2017)
- BERT: 94.9 (Devlin et al., 2019)
- Electra large: 97.1 (Clark et al., 2020)

Few shot learners

Language Models are Few-Shot Learners

Tom B. Brown* **Benjamin Mann*** **Nick Ryder*** **Melanie Subbiah***

Jared Kaplan† **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam** **Girish Sastry**

Amanda Askell **Sandhini Agarwal** **Ariel Herbert-Voss** **Gretchen Krueger** **Tom Henighan**

Rewon Child **Aditya Ramesh** **Daniel M. Ziegler** **Jeffrey Wu** **Clemens Winter**

Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**

Benjamin Chess **Jack Clark** **Christopher Berner**

Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**

OpenAI

(Brown et al., 2020)

Few shot learners

Few shot learners

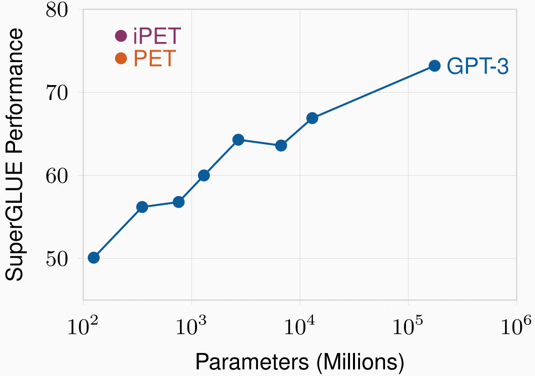


Figure 1: Performance on SuperGLUE with 32 training examples. **ALBERT with PET/iPET outperforms GPT-3 although it is much "greener" in that it has three orders of magnitude fewer parameters.**

(Schick and Schütze, 2021)

Few shot learners

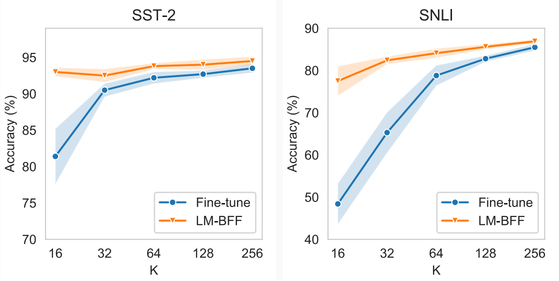


Figure 3: Standard fine-tuning vs our LM-BFF as a function of K (# instances per class). For lower K, our method consistently outperforms standard fine-tuning.

(Gao et al., 2021)

Do these models lead to better generalization?

There is a growing amount of evidence that they have serious limitations.

Do these models lead to better generalization?

There is a growing amount of evidence that they have serious limitations.

Let's take negation as an example.

Do these models lead to better generalization?

There is a growing amount of evidence that they have serious limitations.

Let's take negation as an example.

Large-scale LMs seem to fail completely at handling most negation

Do these models lead to better generalization?

There is a growing amount of evidence that they have serious limitations.

Let's take negation as an example.

Large-scale LMs seem to fail completely at handling most negation

- Ettinger (2020) What BERT Is Not...

Do these models lead to better generalization?

There is a growing amount of evidence that they have serious limitations.

Let's take negation as an example.

Large-scale LMs seem to fail completely at handling most negation

- Ettinger (2020) What BERT Is Not...
- Kassner and Schütze (2020) Negated and Misprimed Probes...

Do these models lead to better generalization?

There is a growing amount of evidence that they have serious limitations.

Let's take negation as an example.

Large-scale LMs seem to fail completely at handling most negation

- Ettinger (2020) What BERT Is Not...
- Kassner and Schütze (2020) Negated and Misprimed Probes...
- Hossain et al. (2020) An Analysis of Natural Language Inference Benchmarks through the Lens of Negation

Do these models lead to better generalization?

There is a growing amount of evidence that they have serious limitations.

Let's take negation as an example.

Large-scale LMs seem to fail completely at handling most negation

- Ettinger (2020) What BERT Is Not...
- Kassner and Schütze (2020) Negated and Misprimed Probes...
- Hossain et al. (2020) An Analysis of Natural Language Inference Benchmarks through the Lens of Negation
- Ribeiro et al. (2020) Beyond Accuracy...

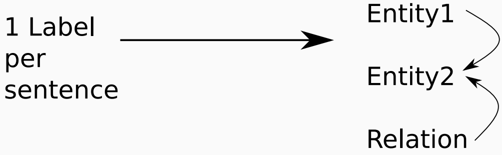
Do these models lead to better generalization?

- Furthermore, papers showing improvements on sentence classification datasets often do not provide any error analysis

- Without these, we cannot know a priori where models still fail

What can we do instead of sentence classification?

Option 1: Evaluation and reformulation of tasks



Structured Sentiment

Given a sentence, find all opinion tuples, where

Structured Sentiment

Given a sentence, find all opinion tuples, where

an opinion tuple consists of 4 elements:

- Holder
- Target
- Expression
- Polarity

Several of these can be implicit.

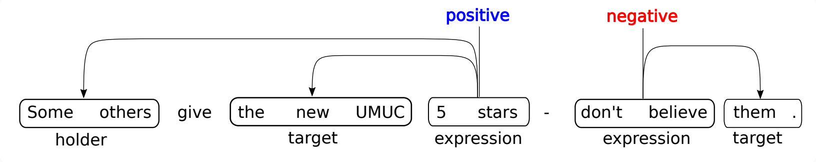
Structured Sentiment

Given a sentence, find all opinion tuples, where

an opinion tuple consists of 4 elements:

- Holder
- Target
- Expression
- Polarity

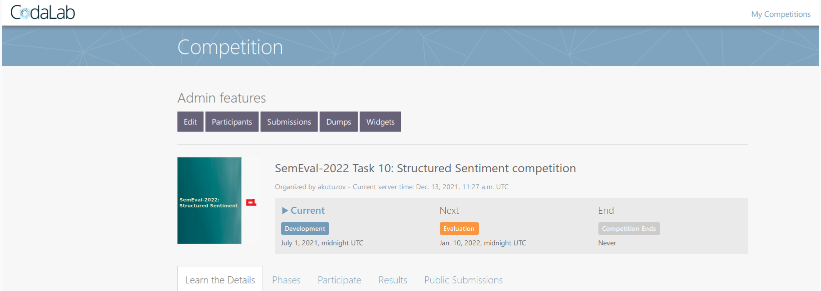
Several of these can be implicit.



Structured Sentiment

Dataset	Languages	# sents.	Ref.
NoReC$_{fine}$	Norwegian	11,437	(Øvrelid et al., 2020)
MultiBooked	Basque, Catalan	~1600	(Barnes et al., 2018)
OpeNER	en, es, it, de, fr, nl	~2500	(Agerri et al., 2013)
MPQA	English	10,048	(Wiebe et al., 2004)
Darmstadt	English	2803	(Toprak et al., 2010)

Structured Sentiment



Advantages and Disadvantages

Advantages:

Advantages and Disadvantages

Advantages:

- more realistic task
- more informative predictions
- easier to perform error analysis
- harder to do well with simple heuristics

Advantages and Disadvantages

Advantages:

- more realistic task
- more informative predictions
- easier to perform error analysis
- harder to do well with simple heuristics

Disadvantages:

Advantages and Disadvantages

Advantages:

- more realistic task
- more informative predictions
- easier to perform error analysis
- harder to do well with simple heuristics

Disadvantages:

- harder to annotate well
- more expensive

Option 2: Creation of challenging datasets

Option 2: Challenging datasets using linguistics!

What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models

Allyson Ettinger

Department of Linguistics University of Chicago
aettinger@uchicago.edu

(Ettinger, 2020)

Option 2: Challenging datasets using linguistics!

Context	$\text{BERT}_{\text{LARGE}}$ predictions
A robin is a ___	*bird, robin, person, hunter, pigeon*
A daisy is a ___	*daisy, rose, flower, berry, tree*
A hammer is a ___	*hammer, tool, weapon, nail, device*
A hammer is an ___	*object, instrument, axe, implement, explosive*
A robin is not a ___	*robin, bird, penguin, man, fly*
A daisy is not a ___	*daisy, rose, flower, lily, cherry*
A hammer is not a ___	*hammer, weapon, tool, gun, rock*
A hammer is not an ___	*object, instrument, axe, animal, artifact*

Table 13: $\text{BERT}_{\text{LARGE}}$ top word predictions for selected NEG-136-SIMP sentences.

Option 2: Challenging datasets using domain knowledge!

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Marco Tulio Ribeiro
Microsoft Research
marcotcr@microsoft.com

Tongshuang Wu
Univ. of Washington
wtshuang@cs.uw.edu

Carlos Guestrin
Univ. of Washington
guestrin@cs.uw.edu

Sameer Singh
Univ. of California, Irvine
sameer@uci.edu

(Ribeiro et al., 2020)

Option 2: Challenging datasets using domain knowledge!

Capability	Min Func Test	INVariance	DIRectional
Vocabulary	Fail. rate=15.0%	16.2%	C 34.6%
NER	0.0%	B 20.8%	N/A
Negation	A 76.4%	N/A	N/A
...			

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT*** Labels: negative, positive, neutral			
Template: I {NEGATION} {POS_VERB} the {THING}.			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
Failure rate = 76.4%			
B Testing **NER** with ***INV*** Same pred. (inv) after removals / additions			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral neg	X
...			
Failure rate = 20.8%			
C Testing **Vocabulary** with ***DIR*** Sentiment monotonic decreasing (↓)			
@AmericanAir service wasn't great. You are lame.	↓	neg neutral	X
@JetBlue why won't YOU help them?! Ugh. I dread you.	↓	neg neutral	X
...			
Failure rate = 34.6%			

Figure 1: CHECKLISTING a commercial sentiment analysis model (G). Tests are structured as a conceptual matrix with capabilities as rows and test types as columns (examples of each type in A, B and C).

Option 2: Challenging datasets using error analysis techniques!

	negation	modals	sarcasm	comparatives	emoji	spelling	...
Reasonable Model	50.0	45.0	63.0	30.0	55.0	14.0	...
Better Model	55.0	48.0	62.0	50.0	58.0	14.0	...
Even Better Model	55.9	46.0	66.2	49.3	69.0	20.4	

(Barnes et al., 2019)

Option 2: Challenging datasets using error analysis techniques!

	negation	modals	sarcasm	comparatives	emoji	spelling	...
Reasonable Model	50.0	45.0	63.0	30.0	55.0	14.0	...
Better Model	55.0	48.0	62.0	50.0	58.0	14.0	...
Even Better Model	55.9	46.0	66.2	49.3	69.0	20.4	

(Barnes et al., 2019)

Option 2: Challenging datasets using adversarial examples!

Dynabench: Rethinking Benchmarking in NLP

Douwe Kiela[†], Max Bartolo[‡], Yixin Nie[⋆], Divyansh Kaushik[§], Atticus Geiger[¶],

Zhengxuan Wu[¶], Bertie Vidgen[‖], Grusha Prasad[⋆⋆], Amanpreet Singh[†], Pratik Ringshia[†],

Zhiyi Ma[†], Tristan Thrush[†], Sebastian Riedel[†‡], Zeerak Waseem[††], Pontus Stenetorp[‡],

Robin Jia[†], Mohit Bansal[⋆], Christopher Potts[¶] and Adina Williams[†]

[†] Facebook AI Research; [‡] UCL; [⋆] UNC Chapel Hill; [§] CMU; [¶] Stanford University
[‖] Alan Turing Institute; [⋆⋆] JHU; [††] Simon Fraser University
dynabench@fb.com

Option 2: Challenging datasets using adversarial examples!

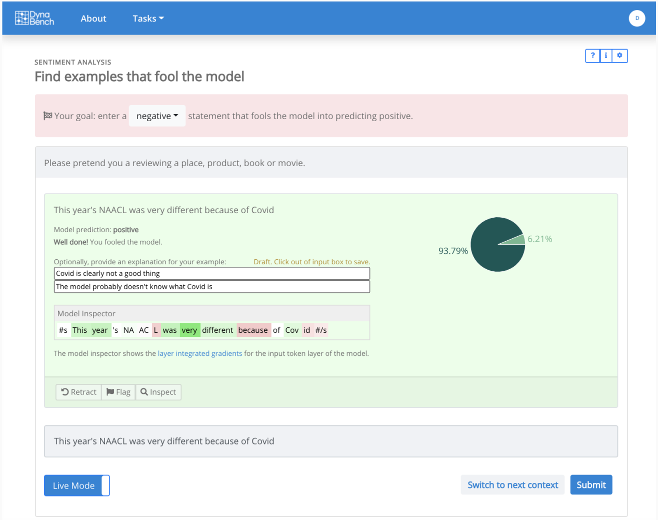


Figure 2: The Dynabench example creation interface for sentiment analysis with illustrative example.

Option 3: Change annotation paradigm

Previous paradigm

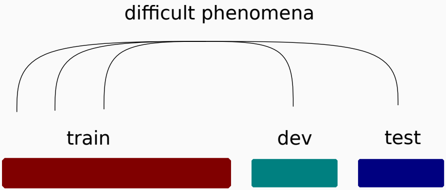
Option 3: Change annotation paradigm

Previous paradigm

New paradigm

Option 3: Change annotation paradigm

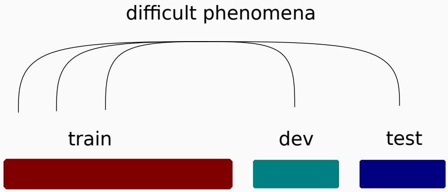
Previous paradigm



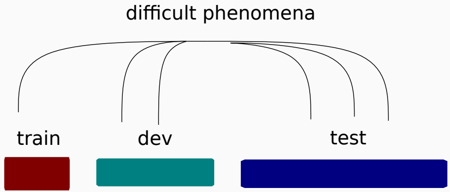
New paradigm

Option 3: Change annotation paradigm

Previous paradigm



New paradigm



Conclusion

Use sentence classification with caution

Performance might not correlate well with downstream performance on other tasks.

Use datasets as originally intended

Avoid simplified versions of data.

if you still really want to do sentence classification...

if you still really want to do sentence classification...

- consider additional kinds of evaluation, i.e.,
  - CheckList,
  - Dynabench,
  - one of the many challenge datasets that have appeared for many tasks

if you still really want to do sentence classification...

- consider additional kinds of evaluation, i.e.,
    - CheckList,
    - Dynabench,
    - one of the many challenge datasets that have appeared for many tasks
- Don't report just performance.
    - With the available data and software, an analysis of model failure and behavior has never been easier.

if you are conducting an annotation project...

if you are conducting an annotation project...

- consider annotating a more complex, realistic version of the task
- try to include other meta-data that will enable testing model behavior further
- concentrate on adversarial curation
- consider concentrating more on creating representative dev/test sets than large training sets

Returning to the original question

Is it time to move beyond sentence classification?

Returning to the original question

Is it time to move beyond sentence classification?

- The final answer to this depends highly on your task...

Returning to the original question

Is it time to move beyond sentence classification?

- The final answer to this depends highly on your task...
- but for many tasks: sentiment analysis, emotion analysis, etc. I would suggest that we move on.

Returning to the original question

Is it time to move beyond sentence classification?

- The final answer to this depends highly on your task...
- but for many tasks: sentiment analysis, emotion analysis, etc. I would suggest that we move on.
- The datasets were often annotated for a different purpose and later simplified for convenience.

Returning to the original question

Is it time to move beyond sentence classification?

- The final answer to this depends highly on your task...
- but for many tasks: sentiment analysis, emotion analysis, etc. I would suggest that we move on.
- The datasets were often annotated for a different purpose and later simplified for convenience.
- We have models and software to perform the full tasks now, no need to simplify.

Returning to the original question

Is it time to move beyond sentence classification?

- The final answer to this depends highly on your task...
- but for many tasks: sentiment analysis, emotion analysis, etc. I would suggest that we move on.
- The datasets were often annotated for a different purpose and later simplified for convenience.
- We have models and software to perform the full tasks now, no need to simplify.

Questions?

Jeremy Barnes

https://jerbarnes.github.io/

jeremy.barnes@ehu.eus

References

Agerri, R., Cuadros, M., Gaines, S., and Rigau, G. (2013). OpeNER: Open polarity enhanced named entity recognition. In *Sociedad Española para el Procesamiento del Lenguaje Natural*, volume 51, pages 215–218.

Barnes, J., Badia, T., and Lambert, P. (2018). MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Barnes, J., Øvrelid, L., and Velldal, E. (2019). Sentiment analysis is not solved! assessing and probing sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 12–23, Florence, Italy. Association for Computational Linguistics.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., and Bao, M. (2021). The values encoded in machine learning research. *arXiv preprint arXiv:*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Gao, T., Fisch, A., and Chen, D. (2021). Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Hossain, M. M., Kovatchev, V., Dutta, P., Kao, T., Wei, E., and Blanco, E. (2020). An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Hu, M. and Liu, B. (2004). Mining opinion features in customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, pages 168–177.

Jain, S. and Wallace, B. C. (2019). Attention is not Explanation. In

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Kassner, N. and Schütze, H. (2020). Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Øvrelid, L., Mæhlum, P., Barnes, J., and Velldal, E. (2020). A fine-grained sentiment dataset for Norwegian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5025–5033, Marseille, France. European Language Resources Association.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In

Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), pages 271–278, Barcelona, Spain.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Potts, C., Wu, Z., Geiger, A., and Kiela, D. (2021). DynaSent: A dynamic benchmark for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language*

Processing (Volume 1: Long Papers), pages 2388–2404, Online. Association for Computational Linguistics.

Radford, A., Jozefowicz, R., and Sutskever, I. (2017). Learning to generate reviews and discovering sentiment.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Schick, T. and Schütze, H. (2021). It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the*

2013 Conference on Empirical Methods in Natural Language Processing, pages 1631–1642, Seattle, Washington, USA.

Toprak, C., Jakob, N., and Gurevych, I. (2010). Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584, Uppsala, Sweden. Association for Computational Linguistics.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. In

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 11–20, Hong Kong, China. Association for Computational Linguistics.