Sibylvariant Transformations for Robust Text Classification

Fabrice Harel-Canada
Muhammad Ali Gulzar
Nanyun Peng
Miryung Kim

https://github.com/UCLA-SEAL/Sibyl
Invariant (INV) Transformations

Nearly all transformations are constrained to **preserve** the source label.

**EDA Synonyms**

- *i cannot believe how good this movie is* → *i cannot fathom how pleasant this movie is* → *positive*

**Checklist Contractions**

- *i cannot believe how good this movie is* → *i can’t believe how good this movie is* → *positive*

**Key Point**

Limits the amount of change you can inject, reducing input space coverage + diversity.

What if we could knowably change the label and inject more diversity?
Sibylvariant (SIB) Transformations

Jointly transform the input and label

- **Sibylvariant Transformations**
  - **Sibyl Concept2Sentence**
    - extract relevant concepts
      - stupid
      - worse
    - antonymize(concepts) + invert label
      - intelligent
      - better
    - generate new data from concepts

**Key Point**

SIB injects more diversity into your dataset!

- **negative**
  - no comment - stupid movie, acting average or worse... screenplay - no sense at all... SKIP IT!

- **positive**
  - I am better than the rest of the world because I am more intelligent.

- **positive**
  - no comment - stupid movie, acting average or worse... screenplay - no sense at all... SKIP IT!

- **positive**
  - I am better than the rest of the world because I am more intelligent.
Unified Framework for Data Transforms

Transforms

**Invariants**
Label preserving

**Sibylvariants**
Label changing

**Transmutations**
100%A → 100%B
Transform one class into another while retaining elements of the original.

**Mixture Mutations**
P%A + (1-P)%B → 100%AB
Mixing 2 or more classes together to create a new input with a distributional soft label (e.g. [0.5, 0.5]).
Sibyl Tool

41 INV + SIB transforms (17 new)

- Concept2Sentence
- ConceptMix
- TextMix
- SentMix
- WordMix
- Demojify
- Emojify
- AddEmoji
- ChangeAntonym
- AddNegation

Task determines type!

ex. ChangeAntonym

“I love pizza” → “I hate pizza”

SIB for sentiment analysis
INV for grammaticality

Sibyl transforms are configured as either INV or SIB for 5 different tasks: sentiment analysis, topic classification, grammaticality, similarity, entailment
The characters are unlikeable and the script is awful. It's a waste of the talents of Deneuve and Auteuil.

**negative**  
[1, 0]

I think it's one of the greatest movies which are ever made, and I've seen many... The book is better, but it's still a very good movie!

**positive**  
[0, 1]

The book is better, but it's still a very good movie! It's a waste of the talents of Deneuve and Auteuil. I think it's one of the greatest movies which are ever made, and I've seen many... The characters are unlikeable and the script is awful.

[0.438, 0.562]

**Key Point**

Creating mixtures helps the model differentiate with greater nuance.
Adaptive SIB Training

- Periodically assess model performance by class
- Generate more examples by targeting commonly confused classes
  - ex. mix “sports” topics with “politics” more often

Key Point
SIB enables a new kind of training that leads to improved performance
Evaluating Effectiveness of SIB vs. INV

Generalization
Does training on SIB-augmented data improve model accuracy?

Defect Detection
How effective are SIB-transformed tests at inducing misclassifications?

Robustness
Does training on SIB data make models more robust to attack?

Systematic Evaluation
- 6 datasets (3 sentiment, 3 topic)
- 11 transformation pipelines (2 randomly sampled INV / SIB transforms)
- 3 levels of resource availability (10,200,2500)
- 216 models
- 30m training inputs
- 480k tests
- 3.3k adversaries
Results: SIB vs. INV

Generalization

89% of the time
Model Accuracy
SIB > INV

Defect Detection

83% of the time
# of Misclassifications
SIB > INV

Robustness

11x more often
Robustness
SIB > INV
How does SIB help?

- SIB diversifies datasets more than INV to improve input space coverage
- SIB data may support margin maximizing decision surfaces
Conclusion: SIB complements INV

Transmutations
Mixture Mutations
Adaptive SIB Training

Sibyl Tool
Taxonomized 41 transforms (17 new) + packaged in tool

Evaluation
SIB transforms outperform INV ones
89% more accuracy
83% more defects
11x more robust

> pip install sibyl-tool

https://github.com/UCLA-SEAL/Sibyl