Data Balance Analysis for Lithuanian and English NLP

VYTAUTAS MAGNUS UNIVERSITY
Faculty of Informatics

AUTHORS: Deividas Valiuška · Justina Mandravickaitė · Milita Songailaitė · Tomas Krilavičius



Is your model cheating?

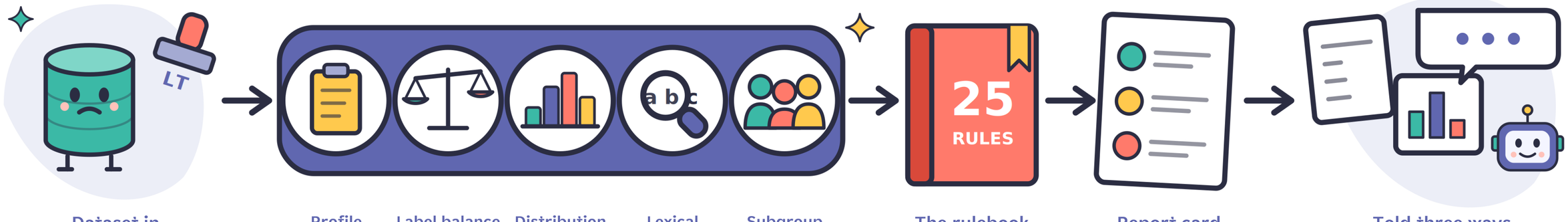
It aced the test. But did it learn the task, or the quirks of its dataset?

Datasets hide shortcuts: test items copied from training, words that give the label away, suspiciously long texts. A model learns them instead of the task, and its score looks better than the model is.

We built a check-up for datasets that runs before training and reports every defect with its evidence and a fix.

HOW THE CHECK-UP WORKS

Cheap, interpretable and done before any training: no model needed to find the defects



Dataset in
training · validation · test. Annotators can stamp on language or time

Profile
sizes, lengths, copies

Label balance
even classes?

Distribution balance
splits alike?

Lexical leakage
giveaway words

Subgroup balance
fair to groups?

The rulebook
25 thresholds, each with its source

Report card
problem · evidence · fix: ok · warning · critical

Told three ways
a summary, charts, and a plain-language LLM explanation

MEET THE USUAL SUSPECTS

Caught by the check-up in a Lithuanian hate-speech corpus (27,357 comments) and a Lithuanian fake-news corpus

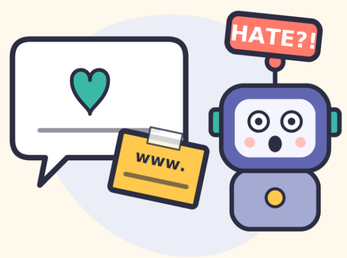
SUSPECT #1
The Copycat



46
validation and test comments that also sit in training

Crime: the model has already seen the exam. Every copied test item makes the score look better than it is.

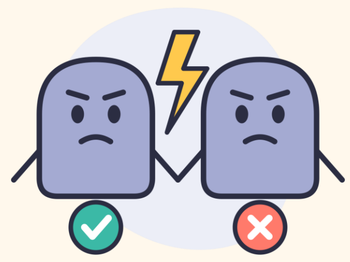
SUSPECT #2
The Giveaway



99.2%
of harmless comments called 'hate' once site text is added (1.2% without)

Crime: scraped website text sits in 9.4% of 'hate' comments, and the classifier learned it. A neutral sentence does nothing (0.7%).

SUSPECT #3
The Quarrelling Twins



4
groups of near-identical comments with opposite labels

Crime: the same comment, posted more than once, labelled differently. Whatever the model answers, it is marked wrong.

SUSPECT #4
The Photocopier



43%
of the 'balanced' version are exact copies

Crime: the classes were balanced by copying rare comments. More rows, no new information, and a 3.6× imbalance underneath.

SUSPECT #5
The Master of Disguise



2
giveaway words visible only when all their forms are counted together

Crime: Lithuanian words change their endings. At 20+ forms, word counts miss a giveaway word every time. Counting lemmas always catches it.

SUSPECT #6
The Chatterbox



449 vs 91
median words: texts labelled true vs the rest

Crime: in the fake-news corpus, text length alone predicts the label (AUC 0.82). No reading needed.

DOES THE DETECTIVE WORK?



100%
planted shortcuts found

3 / 320
false alarms on clean data

ρ = 1.00
bigger defect, stronger verdict

6 kinds of defects planted in 80 clean synthetic datasets. The usual word test (Monroe log-odds) missed every shortcut in the larger class. Ours missed none.

AFTER THE CLEAN-UP



5 → 0
site words in top 'hate' words

46 → 0
leaked test comments

4 → 0
conflicting-label groups

Each fix came straight from the report: strip the site text, drop the leaked comments, settle the conflicting labels.

NEXT ON THE CASE



Ask the experts. Experts judge whether each finding is real and useful.



Crash-test the model. Plant even stronger defects, see when the score starts to lie, and set the warning levels there.



New case: names. The same name in training and test, or 'a place' in one text and 'a person' in the next.

Check your data before you trust your score.

Co-funded by the European Union
These results are part of the project "LLM model training and management platform – LLM4u", project No. 02-020-K-0031.
KRILAS
Lingo data