Language & AI

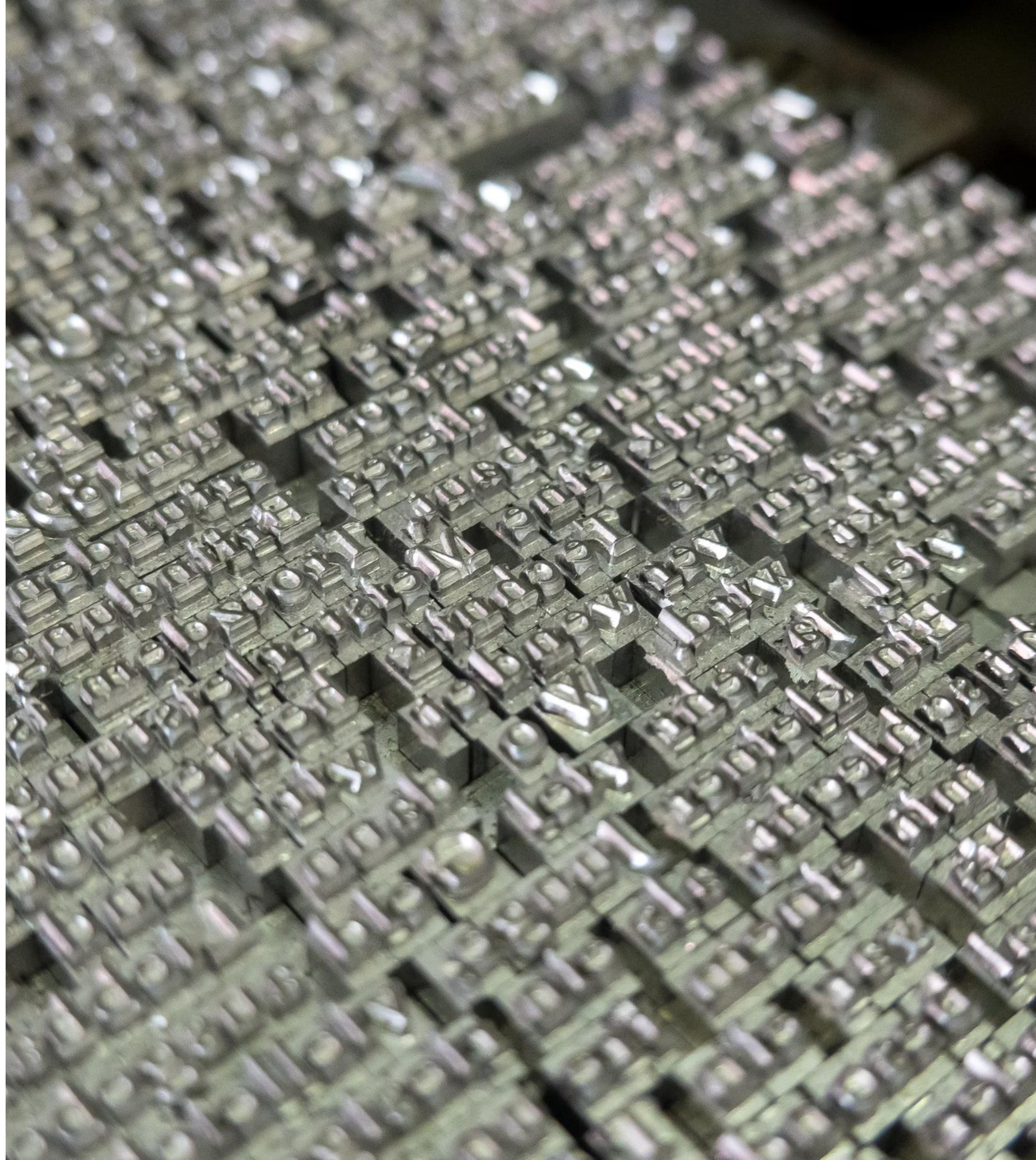
Lecture 3

Dr.ir. Marijn ten Thij, Nov 27th 2025

Video Lecture

Classification & Language Modeling

- What are Language Models?
- How do we evaluate them?
- Toolbox of (simple) classifiers.
- Evaluation:
  - Generalization.
  - Optimization.
  - Stratification.



Predictors

Classifiers vs. Language Models

- Classifiers use labels to predict the data, LMs use the data to predict the data (self-supervised).
- Predictors all require representation methods to convert text into a computation space. LMs learning close to Naive Bayes.
- Evaluation with different metrics; LMs use (average) perplexity.

Tokenization

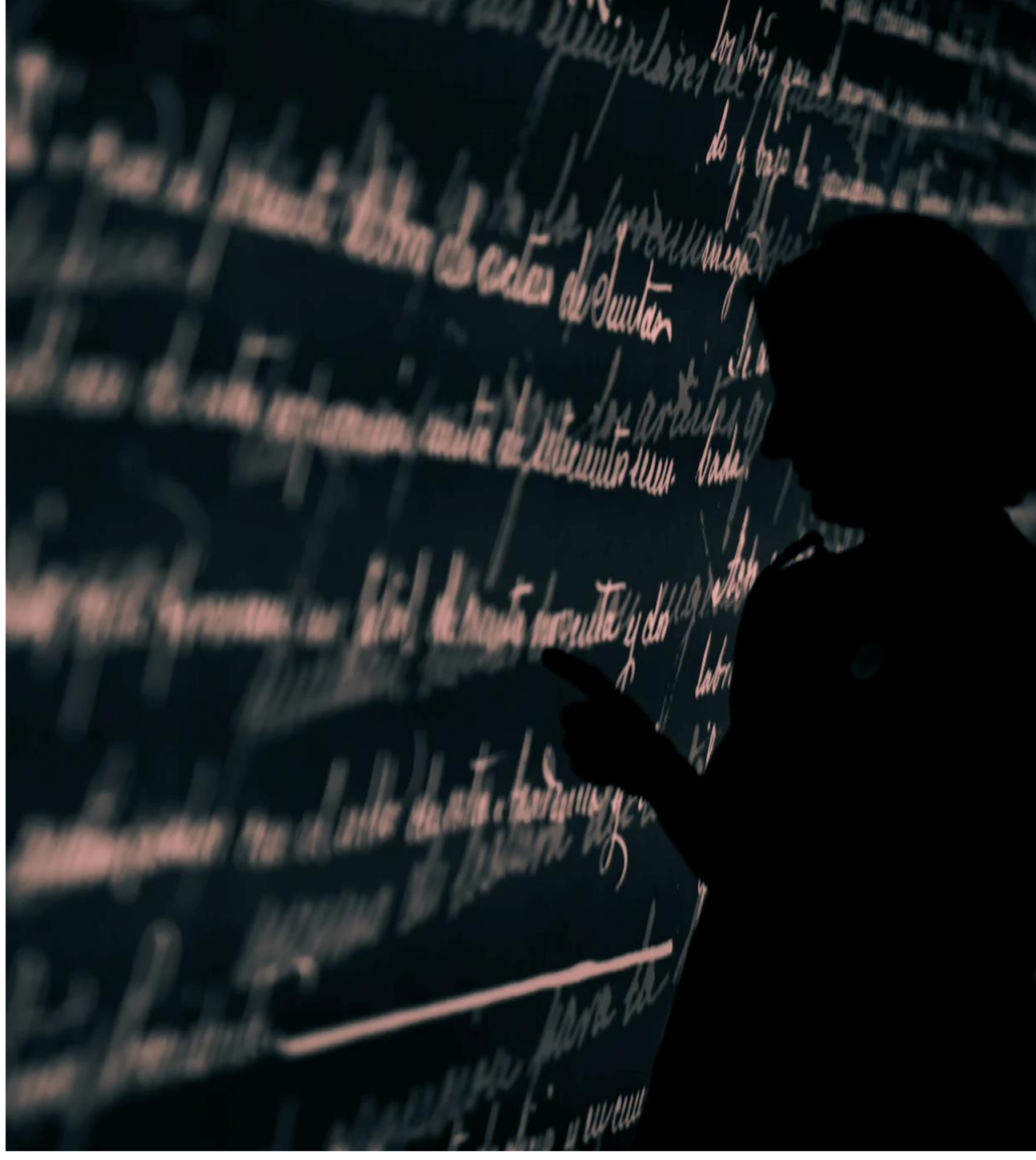
Important For All Things NLP

- Why?
  - Common denominator NLP pipelines.
  - Better representations = less noise = better models.
  - Data preparation is where most errors introduce themselves.

LM Application

More Than Probabilities

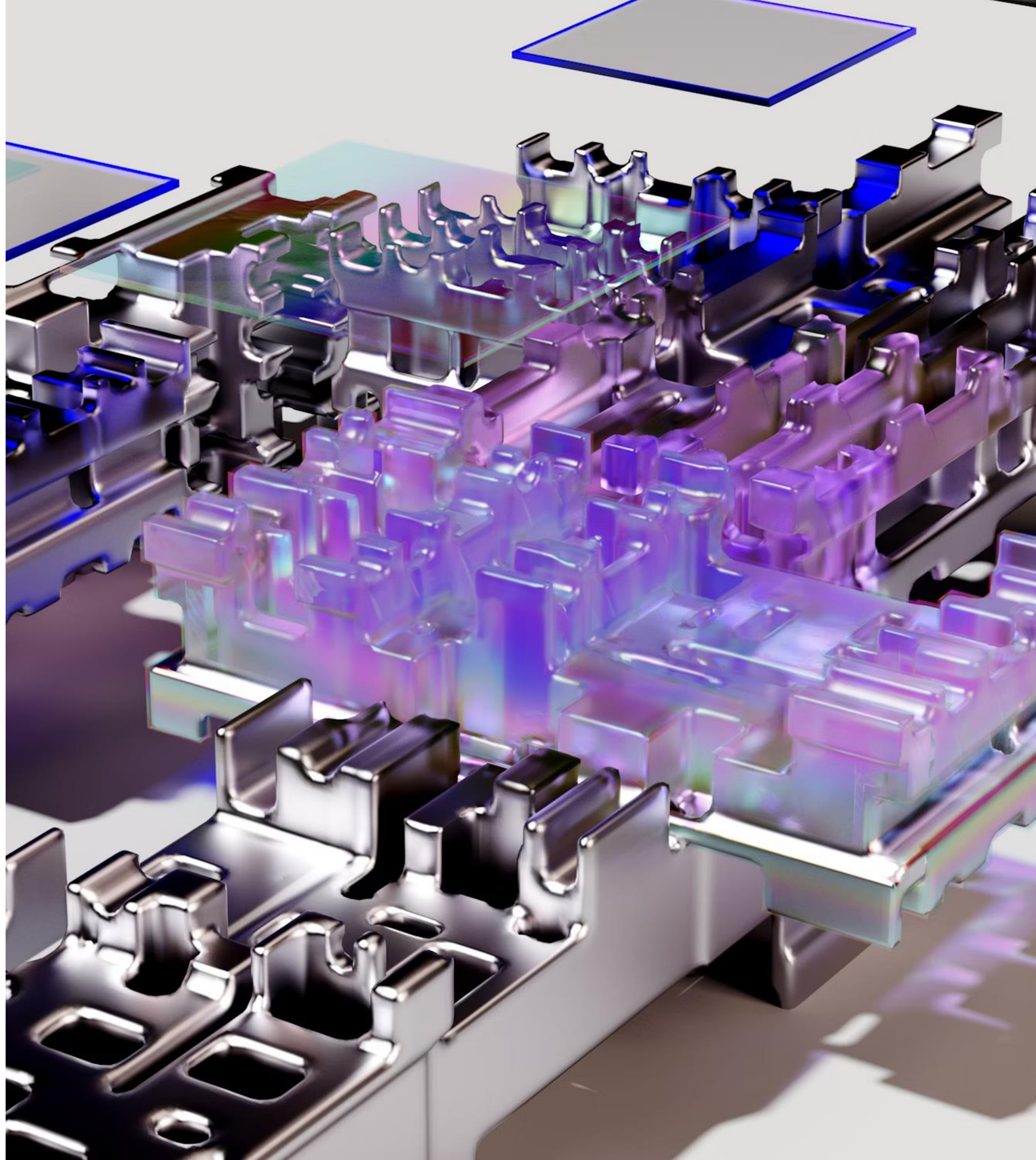
- Can be used to extract features / represent words (covered later).
- Might drive decisions (spell checking informed by perplexity).
- May be used for error analysis, post-hoc correction, etc. (especially in generation).



This Lecture

Thinking About Models

- Pipeline overview.
- Generalization importance and notable components.
- Qualitative evaluation.
- Use case.



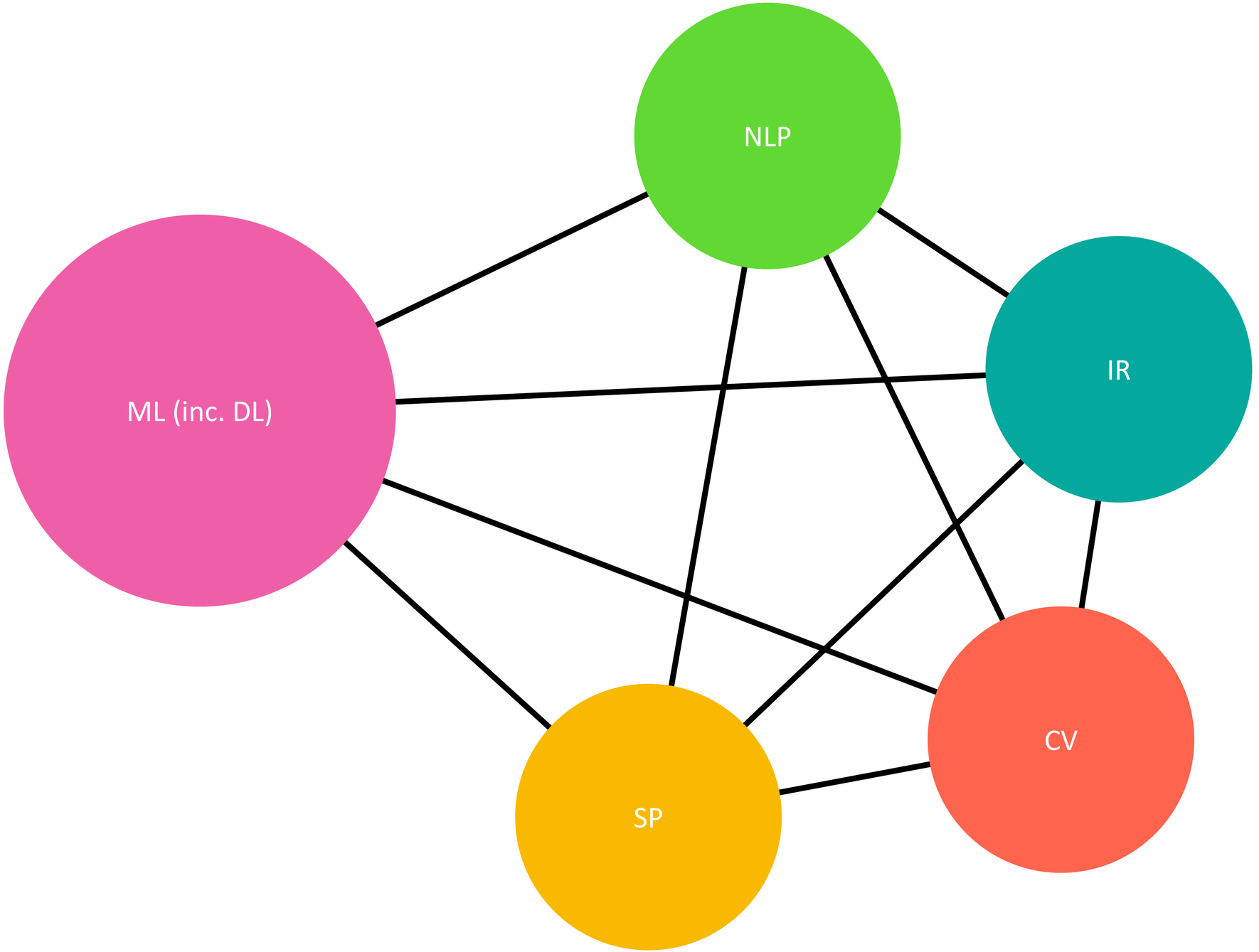
Experimental Context

Assembling the Pipeline

NLP and ML

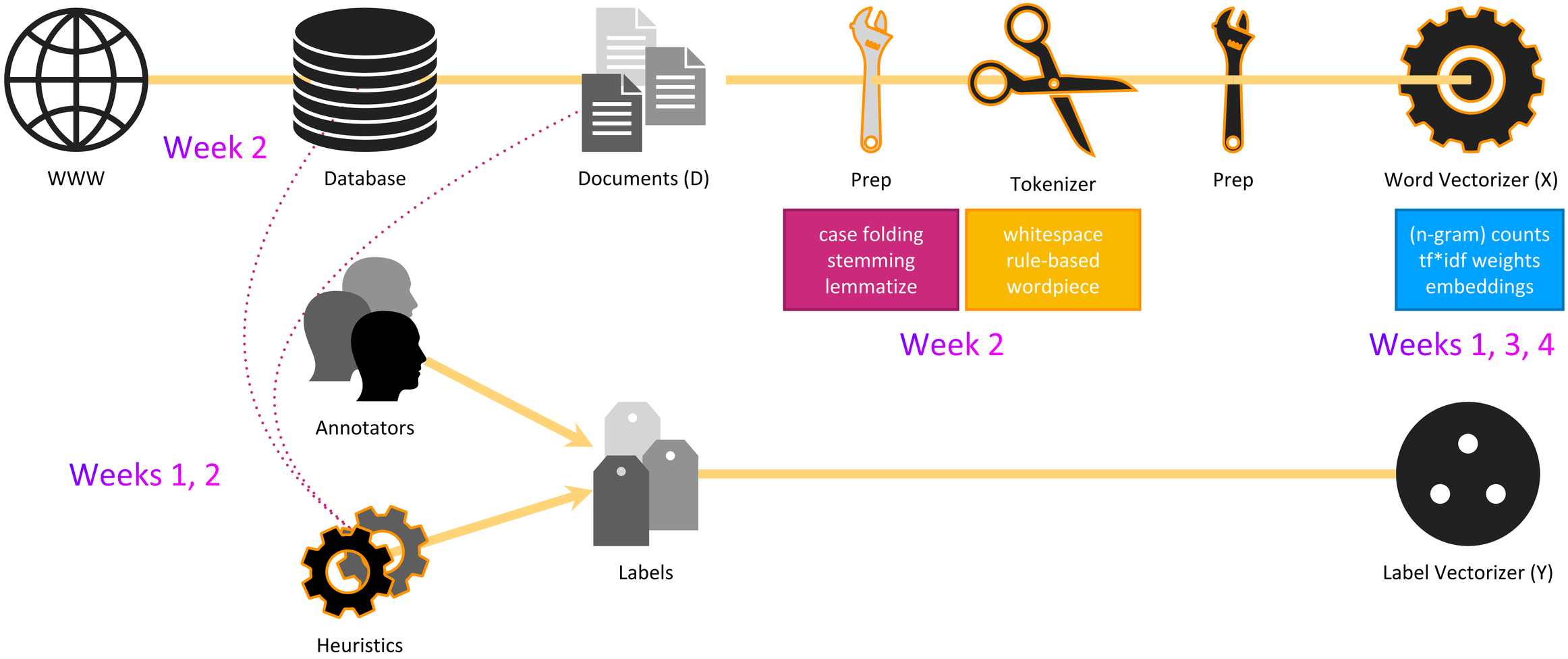
🐓 and 🥚

- “NLP is used to convert language to matrices.”
- “NLP offers a variety of methods to produce features for ML algorithms.”
- “NLP is its own set of algorithms to either classify or represent language data.”
- ?


NLP
ML (inc. DL)
IR
SP
CV

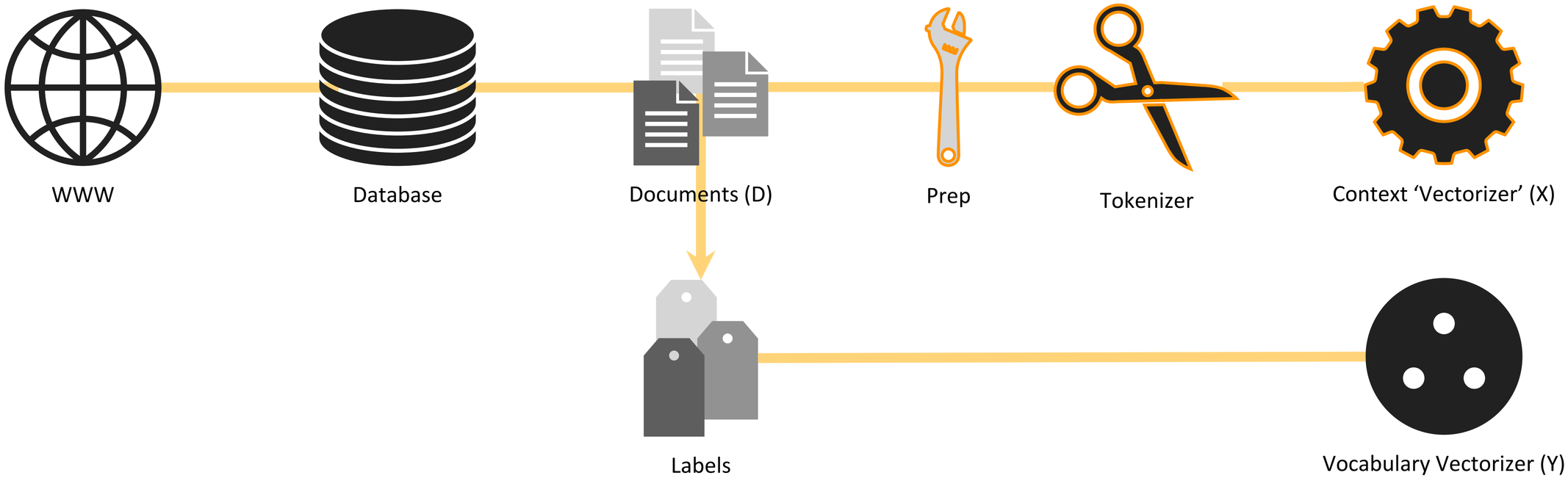
Pipeline Thus Far

And What Are Experimental Parameters



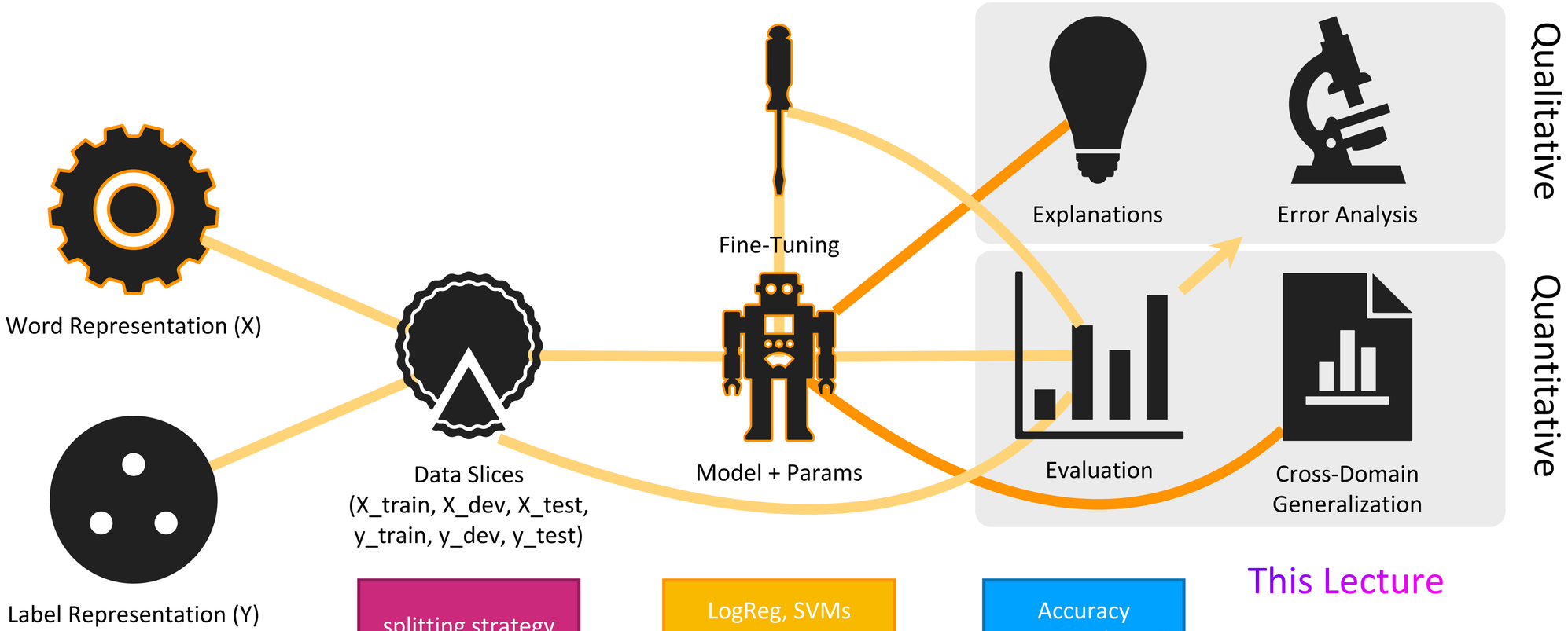
Language Models

Slightly Different Pipeline



Pipeline Thus Far Cont.

And What Are Experimental Parameters



The (Assignment) Task

Decides

- Ask yourself:
  - Do I think my preprocessing steps will add anything or remove information?
  - Consider using both types of representations!
- All considerations: assess **quantitatively**, and diligently!

Any Questions?

(So far)

Quantitative Analyses

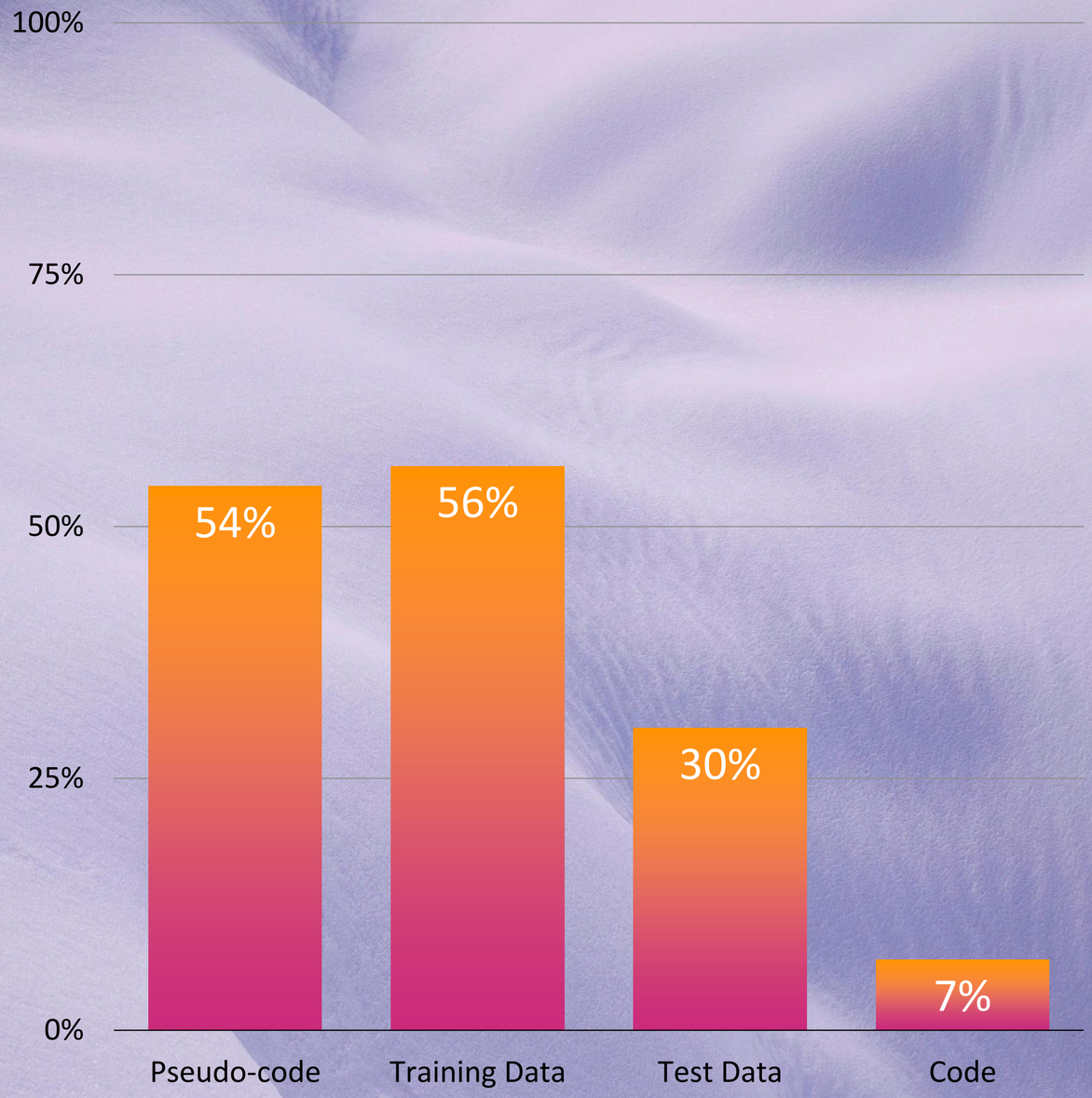
Why One Ought to Be Stringent


Reproducibility
Why?

Reproducibility
In Machine Learning

- Over 90% of papers don’t share their code.
- Most work with open-source libraries and open access data.
- The entire contribution of one’s work is the model, under the assumption that it can be reproduced.

Why do we write papers?



Gundersen, O.E., & Kjensmo, S. (2018). State of the art: reproducibility in artificial intelligence. In *Thirty-second AAAI conference on artificial intelligence*.

Human Error

Publish or Perish

- Always keep in mind any bug in the pipeline might compromise your results.
- For NLP, mostly data reader side.
- Don’t let this deter you. Errors are common. Admitting them is rare.

Generalization

Convincing Yourself Your Models Work

- k-fold Cross Validation gives you an indication of sample sensitivity.
- A baseline (e.g. majority baseline / no-skill baseline / ZeroR) of gain (heuristic?).
- Your test set should be unseen data.
  - Often research simply samples a test set without much thought.
  - If you re-use data, the test set might have been compromised.
- Typically stops there. However:
  - How do we know your model works ‘in the wild’?
  - Cross-domain generalization: domain can be very broad.

Choosing the Right Metric

What Fits and is Fair?

- Accuracy only works if your problem is balanced.
- Precision and Recall might be tricky to interpret:
  - High Precision and Low Recall: all ‘positive’ instances correct, but few predicted in total (many positives labeled negative).
  - Low Precision and High Recall: many ‘positive’ instances predicted in total, but few correctly (many negatives labeled positive).
- F_1 score also not the holy grail: do we look at micro (mean over classes), macro (unweighted mean per class), positive class (what is important)?

Splitting and Sampling Anomalies

What Fits and is Fair?

- What do the instances in my data represent?
- Does shuffling the data matter?
- Does the distribution of my data matter?
- Could I mess something up by not accounting for the data?

- The goal is to make prediction tasks strict, and test sets difficult, not to achieve high evaluation scores.

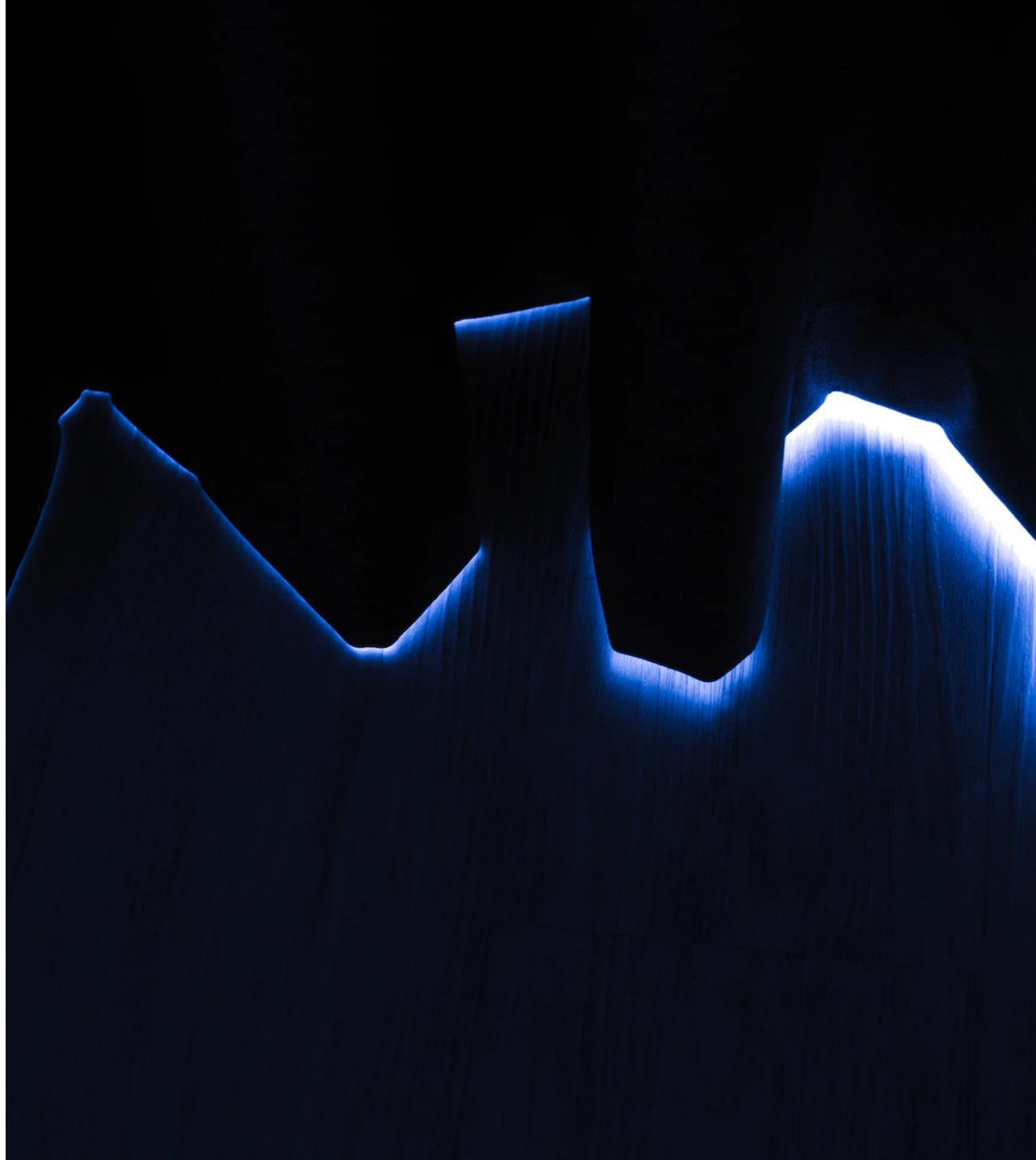
Qualitative Analyses

Looking Beyond Results

Check Your Models

Some Have Batteries Included

- Simple ways to ‘explain’ models when they are not ‘black boxes’.
  - Naive Bayes encodes probabilities / associations.
  - Decision Trees (simplified) provide a chain of rules.
  - k-NN provides information about neighbors.



Check Your Results!

You Are the Expert

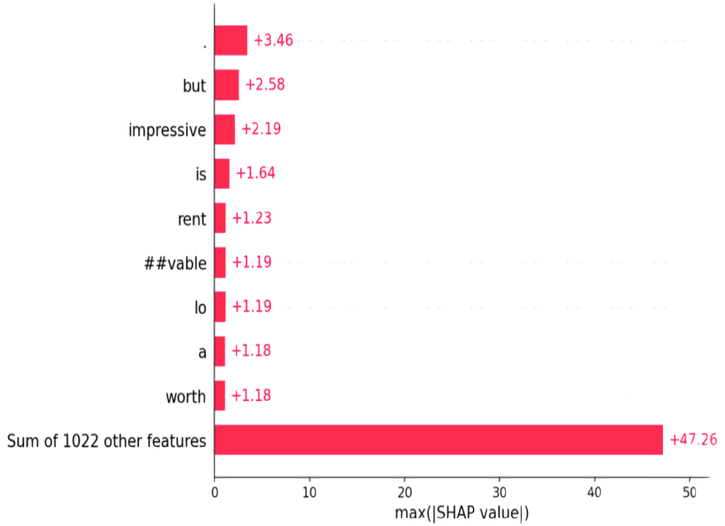
- All you need:
  - The text.
  - The predicted label.
  - The true label.
  - Maybe confidence scores.
  - A keen eye.

Use Explanation Tools

Often Slow, and Results May Vary

- We covered LIME.
- Omission scores: delete a word, measure impact.
- SHAP (SHapely Additive exPlanations) also relevant.
- Interesting to try multiple: these are generally model-dependent and interact with noise the same way.

This is easily the most underrated film inn the Brooks cannon. Sure, its flawed. It does not give a realistic view of homelessness (unlike, say, how Citizen Kane gave a realistic view of lounge singers, or Titanic gave a realistic view of Italians YOU IDIOTS). Many of the jokes fall flat. But still, this film is very lovable in a way many comedies are not, and to pull that off in a story about some of the most traditionally reviled members of society is truly impressive. Its not The Fisher King, but its not crap, either. My only complaint is that Brooks should have cast someone else in the lead (I love Mel as a Director and Writer, not so much as a lead).



Contrast Sets
(De)bug Your Models

- Test sets might be biased.
- Your model might learn simple rules having nothing to d ow with the task.
- Change your test set minimally to meaningfully change predictions.
- Fix model accordingly, or?

Evaluating Models' Local Decision Boundaries via Contrast Sets

Matt Gardner[★◇] Yoav Artzi[Γ] Victoria Basmova[◇♣] Jonathan Berant[◇♠]
Ben Bogin[♠] Sihao Chen[♡] Pradeep Dasigi[◇] Dheeru Dua[□] Yanai Elazar[◇♣]
Ananth Gottumukkala[□] Nitish Gupta[♡] Hanna Hajishirzi[◇△] Gabriel Ilharco[△]
Daniel Khashabi[◇] Kevin Lin[+] Jiangming Liu[◇†] Nelson F. Liu[¶]
Phoebe Mulcaire[△] Qiang Ning[◇] Sameer Singh[□] Noah A. Smith[◇△]
Sanjay Subramanian[◇] Reut Tsarfaty[◇♣] Eric Wallace[+] Ally Zhang[Γ] Ben Zhou[♡]
[◇]Allen Institute for AI [Γ]Cornell University [♣]Bar-Ilan University
[♠]Tel-Aviv University [♡]University of Pennsylvania [△]University of Washington
[□]UC Irvine [+]UC Berkeley [†]University of Edinburgh [¶]Stanford University
mattg@allenai.org

Abstract

Standard test sets for supervised learning evaluate in-distribution generalization. Unfortunately, when a dataset has systematic gaps (e.g., annotation artifacts), these evaluations are misleading: a model can learn simple decision rules that perform well on the test set but do not capture the abilities a dataset is intended to test. We propose a more rigorous annotation paradigm for NLP that helps to close systematic gaps in the test data. In particular, after a dataset is constructed, we recommend that the dataset authors manually perturb the test instances in small but meaningful ways that (typically) change the gold label, creating *contrast sets*. Contrast sets provide a local view of a model's decision boundary, which can be used to more accurately evaluate a model's true linguistic capabilities. We demonstrate the efficacy of contrast sets by creating them for 10 diverse NLP datasets (e.g., DROP reading comprehension, UD parsing, and IMDb sentiment analysis). Although our contrast sets are not explicitly adversarial, model performance is

Original Example:

Two similarly-colored and similarly-posed chow dogs are face to face in one image.

Example Textual Perturbations:

Two similarly-colored and similarly-posed **cats** are face to face in one image.

Three similarly-colored and similarly-posed chow dogs are face to face in one image.

Two **differently-colored but** similarly-posed chow dogs are face to face in one image.

Example Image Perturbation:

Two similarly-colored and similarly-posed chow dogs are face to face in one image.

Figure 1: An example contrast set for NLVR2 (Suhr and Artzi, 2019). The label for the original example is TRUE and the label for all of the perturbed examples is FALSE. The contrast set allows probing of a

Any Questions?

(So far)

Content Warning: next part deals with toxic content. Slides contain (partly censored) profanity, talk about online harassment, and related unpleasant topics.


Story Time

Discovering a Flawed Task

… And Finding Evidence

- Started in 2014 on the AMiCA Project
- Content moderation: one of the tasks focused on cyberbullying detection.
- In-house data was being annotated.
- Do some prior work exploration and testing the waters.

Lang Resources & Evaluation (2021) 55:597–633
https://doi.org/10.1007/s10579-020-09509-1

Check for updates

ORIGINAL PAPER

Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity

Chris Emmery[1,2] · Ben Verhoeven[2] · Guy De Pauw[2] · Gilles Jacobs[3] · Cynthia Van Hee[3] · Els Lefever[3] · Bart Desmet[3] · Véronique Hoste[3] · Walter Daelemans[2]

Accepted: 24 September 2020 / Published online: 16 November 2020

Abstract The detection of online cyberbullying has seen an increase in societal importance, popularity in research, and available open data. Nevertheless, while computational power and affordability of resources continue to increase, the access restrictions on high-quality data limit the applicability of state-of-the-art techniques. Consequently, much of the recent research uses small, heterogeneous datasets, without a thorough evaluation of applicability. In this paper, we further illustrate these issues, as we (i) evaluate many publicly available resources for this task and

The work presented in this article was carried out in the framework of the AMiCA (IWT SBO-project 120007) project, funded by the government agency for Innovation by Science and Technology (IWT).

✉ Chris Emmery
cmry@pm.me

Lots of Reading
Lots of Frowning

- Tendency seemed to be working on single messages annotated in a variety of ways.
- Standard features: lexicons, n-grams, topics, classifiers typically SVMs.
- Metrics seemed all over the place: accuracy, micro F1-score, generally all were quite high.
- Chris got skeptical.

Platform	Pos	Neg	Max	F_1
Kongregate	42	4802	[31]	.920
Slashdot	60	4303	[31]	.920
Myspace	65	1946	[31]	.920
Formspring	369	3915	[42]	.779
YouTube	2277	4500	[14]	–
Myspace	415	1647	[54]	.776
Twitter	684	1762	[55]	.780
Myspace	311	8938	[12]	.350
YouTube	449	4177	[12]	.640
Twitter	220	5162	[7]	.726
Twitter	194	2599	[7]	.719
Ask.fm	3787	86419	[17]	.465
Instagram	567	1387	[18]	.750
Twitter	2102	5219	[54]	.719

Know the Literature

Not Only NLP / ML

- A power imbalance between bully and victim.
- The harm is intentional.
- It is repeated over time.
- Complex acts: *flaming*, *outing*, *harassment* (repeated), *exclusion*, *denigration* (gossip), and *impersonation*.

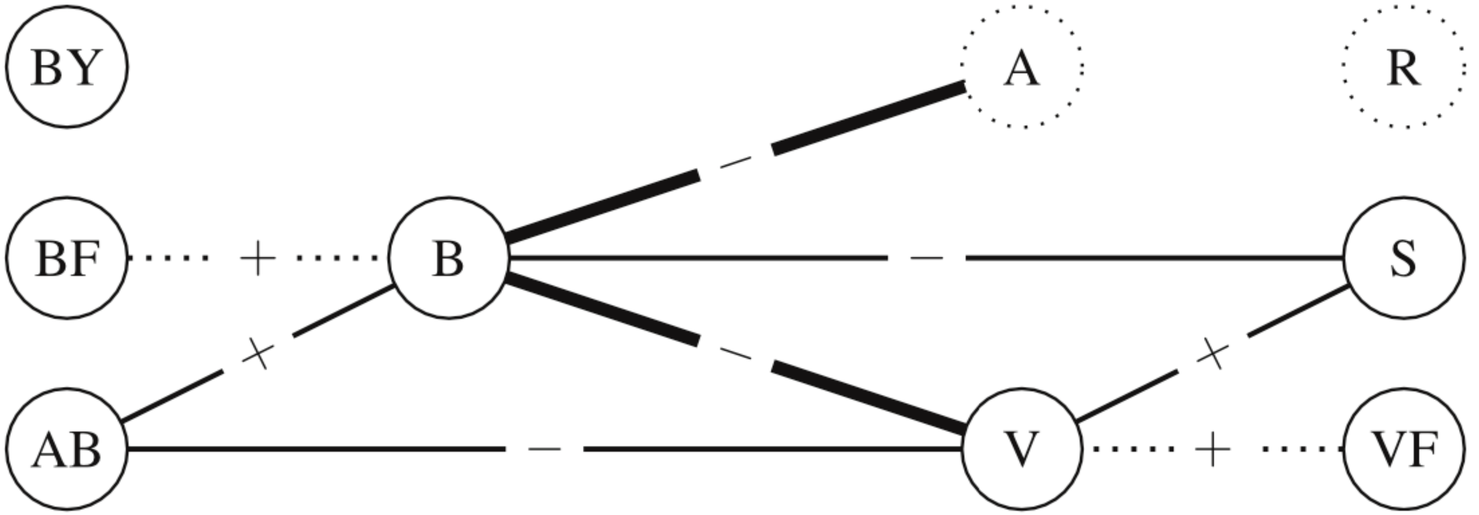


Fig. 1 Role graph of a bullying event. Each vertex represents an actor, labeled by their role in the event: bully (*B*), victim (*V*), bystander (*BY*), reinforcer (*AB*), assistant (*BF*), defender (*S*), reporter (*S*), accuser (*A*), and friend (*VF*). Each edge indicates a stream of communication, labeled by whether this is positive (+) or negative (−) in nature, and its strength indicating the frequency of interaction. Dotted edges indicate nonparticipation in the event, and vertices those added by Xu et al. (2012) to account for social-media-specific roles

Table 1 Fictional example of a cyberbullying conversation. Lines represent sequential turns

Line	Role	Message	Bully	Type
1	V	me and my friends hanging out tonight! :)		Neutral
2	B	@V lol b*tch, you dont have any friends.. ur fake as sh*t	✔	Curse, insult
3	AB	@B haha word, shes so sad	✔	Encouragement
4	VF	@V you know it girl		
5	S	@V dont listen to @B, were gonna have fun for sure!		Defense
6	V	@B shut up @B!! nobody asked your opinion!!!!		Defense
7	A	@B you are a f*cking bully, go outside or smt		Insult
8	B	@V @S haha you all so dumb, just kill yourself already!	✔	Insult, curse
9	A, R	@B shut up or ill report you		
10	B	@A u gonna cry? go ahead, see what happens tomorrow!	✔	Threat

Roles are noted as described on Page 4 (under the eponymous paragraph), if the message can be considered bullying by ✔, and types according to Van Hee et al. (2015)

Reality

--- *Bretschneider et al.*
1, @doiseydoisey1 No because you know I'm right haha Barry really s*cks
1, "@heyitssmecarol l*ser ;D
1, 2 times for you p*ssies cause you know who #1
--- Kontostathis et al.
1, "'Make the lie simple, keep repeating it, and eventually the will believe it.' - Hitler. well said Iran is not a friend to the US. I suggest you so call American's who believe that they are take a vacation there. Then the truth will be revealed to you. Iran is a nice place for a Muslim terrorist
--- *Xu et al.*
1, Damm I think I'm slick bully this dude always offering sh*t to avoid me for roasting his *ss LoL
1, @chelseabieber95 duuuuude. cyber bully much? I just read some of the tweets. how did it even start
1, @katie_ogden @boybandslut @StephWall27 @BAMitsSHAZ @xstephhh_ I was not bullying her! She's trying to steal Jean Martyn from meee! :'(xx
--- *Dadvar et al.*
1, "215 hurry the f*ck up we gota go! nova lol 157th yea boy 101 it was only 7 roes troooooool someone should get rid of these f*ckers if they hate america why do they still live in america
--- Van Hee et al.
1, "I'm not r*t*rded okay, I just get stressed very easily especially when my hair doesn't work;) x"
1, Your a f*cking sl*t look at your Halloween costume what is wrong with you your 14 not 16
1, Ahaha d*ck head I'm 13 not 14 so shut the f*ck up you anonymous c*nts so f*cking come of anon and ill f*cking knock you out you d*ckhead and my Halloween costume is amazing
1, Ahahahahhaha f*ck off .... Are you a d*ckhead

Practical Limitations

A Priori Skepticism

- Datasets capture little to no context and have no temporal information.
- Communication is of unknown direction and weight.
- Role in interaction and relations towards victim are unknown.

So… this shouldn’t work, right?

Hypotheses
/ Research Questions

- There's too little data, too diverse, therefore **(i)** the samples are underpowered in terms of accurately representing the strong language variety amongst platforms.
- BoW works fairly well, but there's a noticeable ceiling. Easy linguistic cues probably only cover curse words etc., so **(ii)** positive instances might therefore be biased — only reflecting a limited bullying dimension.
- Current access is obviously limited, but (iii) collecting synthetic data might actually prove useful.

Table 3 Corpus statistics for English and Dutch cyberbullying datasets, list number of positive (Pos, bullying) and negative (Neg, other) instances, Types (unique words), Tokens (total words), average number of tokens per message (Avg Tok/Msg), number of emojis and emoticons (Emote), and swear word occurrence per neutral (SweaN), and positive (SweaP) instance

	Pos	Neg	Types	Tokens	Avg Tok/Msg		Emote	SweaN	SweaP
D_{twB}	237	5258	12K	78K	14	($\sigma = 8$)	961	277	867
D_{frm}	1025	11,742	21K	348K	27	($\sigma = 29$)	3322	1228	2871
D_{msp}	426	1627	13K	803K	391	($\sigma = 285$)	931	1447	3730
D_{ytb}	417	3045	52K	827K	239	($\sigma = 252$)	3662	2606	8705
D_{ask}	5001	89,404	63K	1,017K	12	($\sigma = 23$)	17,362	4839	12,191
D_{twX}	281	4654	19K	86K	18	($\sigma = 8$)	1344	74	502
D_{tox}	15,279	144,226	220K	12,924K	81	($\sigma = 121$)	11,876	13,732	22,404
D_{ask_nl}	8675	70,557	58K	776K	10	($\sigma = 15$)	16,905	2025	2299
D_{sim_nl}	2330	2681	7K	55K	11	($\sigma = 16$)	434	682	194
D_{don_nl}	152	211	2K	7K	20	($\sigma = 24$)	33	47	19

Emojis were detected with https://github.com/NeelShah18/emot. Swears were detected with reference lists: for English these were taken from https://www.noswearing.com and the Dutch were manually composed.

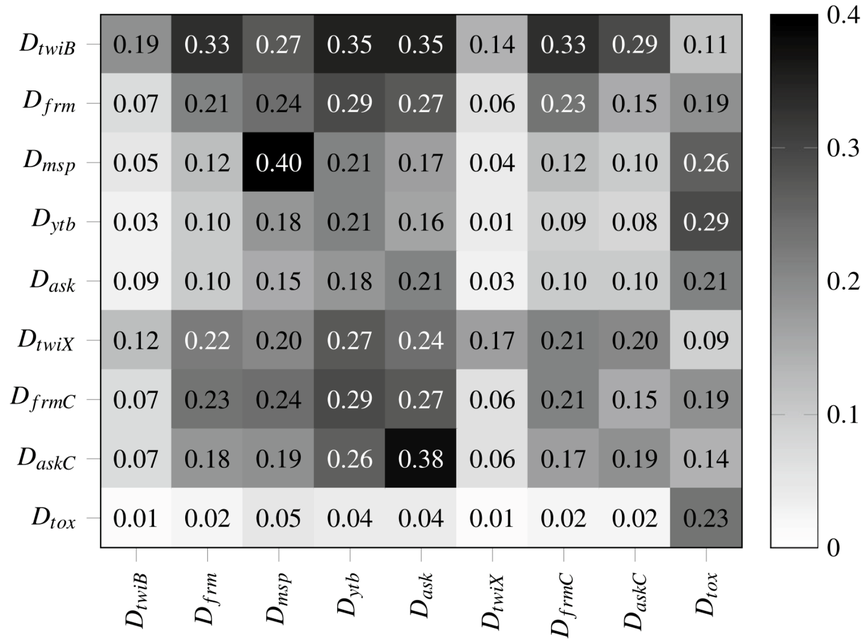


Fig. 2 Jaccard similarity between training sets (y-axis) and test sets (x-axis)

Table 5 Cross-corpora positive class F_1 scores for Experiment I (T1), II (T2), and III (T3)

Train	T1						Avg	T2		T3
	D_{twB}	D_{frm}	D_{msp}	D_{ytb}	D_{ask}	D_{twX}		C_{frm}	C_{ask}	D_{tox}
D_{twB}	.417	.308	.000	.122	.298	.051	.153	.131	.158	.349
D_{frm}	.423	.454	.042	.379	.418	.041	.321	.682	.259	.465
D_{msp}	.120	.176	**.941**	.324	.168	.043	.197	.364	.185	.185
D_{ytb}	.074	.160	.375	.365	.138	.000	.183	.338	.197	.140
D_{ask}	.493	.444	.211	**.421**	**.561**	.139	.351	.389	.357	.584
D_{twX}	.049	.131	.184	.175	.077	**.508**	.205	.496	.325	.082
D_{all}	**.524**	**.473**	**.941**	.397	.553	.194	**.557**	**.780**	.570	.587
C_{frm}	.152	.253	.143	.286	.136	.126	.214	.758	.400	.372
C_{ask}	.286	.237	.359	.244	.356	.107	.310	.582	**.579**	.280
D_{tox}	.343	.373	.449	.335	.443	.149	.389	.628	.539	**.806**

Models are fitted on the training proportion of the corpora row-wise, and tested column-wise. The out-of-domain average (Avg) excludes test performance of the parent training corpus. The best overall test score is noted in bold, the best out-of-domain performance in gray

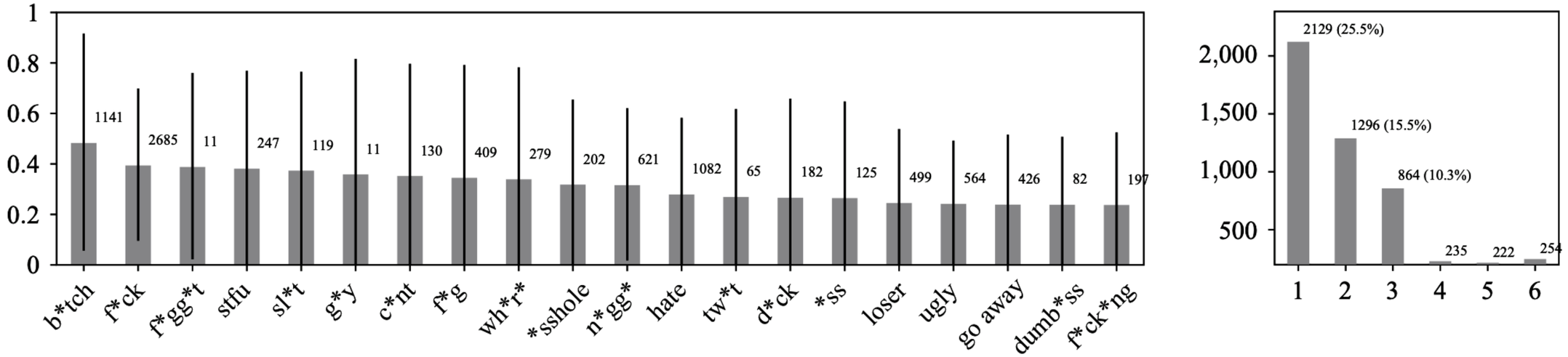


Fig. 3 Left: Top 20 *test set* words with the highest average coefficient values across all classifiers (minus the model trained on D_{tox}). Error bars represent standard deviation. Each coefficient value is only counted once per test set. The frequency of the words is listed in the annotation. Right: Test set occurrence frequencies (and percentages) of the top 5000 highest absolute feature coefficient values

Table 6 Examples of uni-gram weights according to the `baseline` SVM trained D_{all}, tested on D_{twB} and D_{ask}

y	$\hat{y}$	D_{twB}	D_{ask}
👍	👎	about to leave this school library	bigerrr ? how much ? its gon na
		and take my *ss homeeee	touch the sky ? a wonder d*ck ?
👎	👎	you p*ss me off so much .	r u a r*t*rd liam mate f*ck off
👎	👍	@username i will skull drag you across campus .	h* of me xoxoxoxoxoxoox

Words in red are associated with bullying, words in green with neutral content. The color intensity is derived from the strength of the SVM coefficients per feature (most are near zero). Black boxes indicate OOV words. Labels are divided between the gold standard (y) and predicted ($\hat{y}$) labels, 👎 for bullying content, 👍 for neutral

Table 8 Overview of different architectures (Arch) their in-domain positive class F_1 scores for Experiment I (T1) and II (T2), the out-of-domain average for D_{all} (*all*), and D_{tox} (*tox*)

Arch	T1						Avg		T2		T3
	D_{twB}	D_{frm}	D_{msp}	D_{ytb}	D_{ask}	D_{twX}	*all*	*tox*	C_{frm}	C_{ask}	D_{tox}
`baseline`	.417	.454	.941	.365	**.561**	**.508**	**.557**	**.389**	**.758**	.579	**.806**
NBSVM	.383	**.486**	.925	**.387**	.476	.396	.551	.385	.703	.604	.797
BiLSTM*	.171	.363	.938	.152	.504	.400	.440	.349	.609	.507	.762
BiLSTM+	.188	.396	**.951**	.160	.438	.341	.417	.337	.541	.505	.737
BiLSTM	.182	.341	.905	.148	.463	.246	.479	.356	.608	.522	.774
CNN*	**.500**	.276	.790	.133	.462	.438	.364	.350	.000	.306	.753
CNN	.444	.416	.816	.000	.498	.438	.464	.342	.000	.610	.754
CNN★	.444	.419	.816	.000	.499	.375	.460	.362	.000	**.647**	.774
C-LSTM*	.000	.421	.875	.095	.000	.000	.449	.329	.094	.425	.757
C-LSTM	.000	.019	.829	.000	.066	.000	.463	.355	.095	.518	.761
C-LSTM★	.000	.057	.853	.075	.008	.000	.278	.358	.296	.506	.756

Baseline model (and scores) is that of Table 4. Reproduction results of Agrawal and Awekar (2018) are denoted by *, their oversampling method by +. Our tuned model versions have no annotation, character level models are denoted by ★

Any Questions?

(So far)

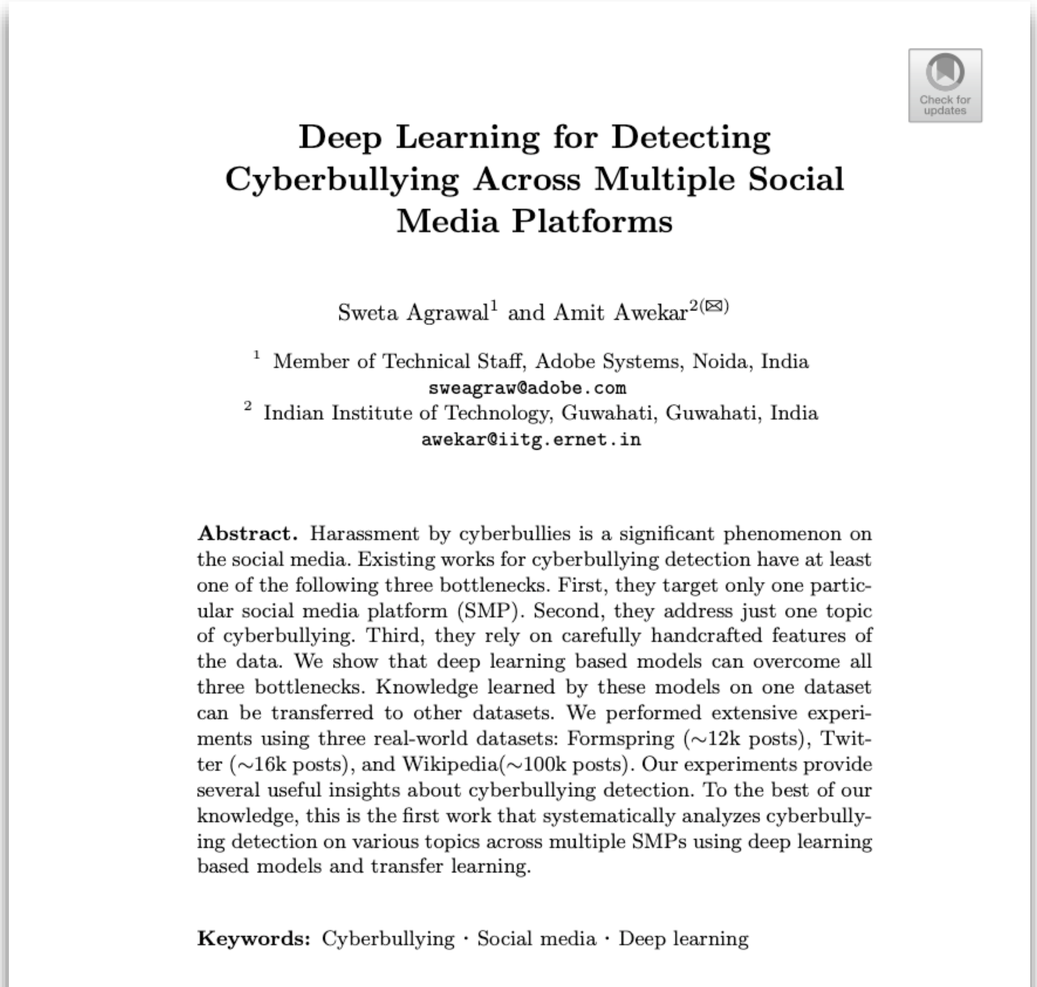
It Gets Better

Or Worse, Actually

Table 4. Effect of oversampling bullying posts using BLSTM with attention

Dataset	Label	P			R			F1		
		Random	Glove	SSWE	Random	Glove	SSWE	Random	Glove	SSWE
F	Bully	0.52	0.56	0.63	0.40	0.49	0.38	0.44	0.51	0.47
F+	Bully	0.84	0.85	0.90	0.98	0.97	0.91	0.90	0.90	0.91
T	Racism	0.67	0.74	0.76	0.73	0.76	0.77	0.70	0.75	0.76
T+	Racism	0.94	0.90	0.90	0.98	0.95	0.96	0.96	0.93	0.93
T	Sexism	0.65	0.86	0.83	0.64	0.52	0.47	0.65	0.65	0.59
T+	Sexism	0.88	0.95	0.88	0.97	0.91	0.92	0.93	0.91	0.90
W	Attack	0.77	0.81	0.82	0.74	0.67	0.68	0.76	0.74	0.74
W+	Attack	0.81	0.86	0.87	0.91	0.89	0.86	0.88	0.88	0.87

400+ citations 🚀



Check for updates

Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms

Sweta Agrawal[1] and Amit Awekar[2(✉)]

[1] Member of Technical Staff, Adobe Systems, Noida, India
sweagraw@adobe.com
[2] Indian Institute of Technology, Guwahati, Guwahati, India
awekar@iitg.ernet.in

Abstract. Harassment by cyberbullies is a significant phenomenon on the social media. Existing works for cyberbullying detection have at least one of the following three bottlenecks. First, they target only one particular social media platform (SMP). Second, they address just one topic of cyberbullying. Third, they rely on carefully handcrafted features of the data. We show that deep learning based models can overcome all three bottlenecks. Knowledge learned by these models on one dataset can be transferred to other datasets. We performed extensive experiments using three real-world datasets: Formspring (∼12k posts), Twitter (∼16k posts), and Wikipedia(∼100k posts). Our experiments provide several useful insights about cyberbullying detection. To the best of our knowledge, this is the first work that systematically analyzes cyberbullying detection on various topics across multiple SMPs using deep learning based models and transfer learning.

Keywords: Cyberbullying · Social media · Deep learning

“In [2] and [3], the authors discuss the limitations of our oversampling method (Section 4.2) in that, the way oversampling is currently handled may lead to overfitting. We found their claims/criticisms valid and important but we haven't conducted any new experiments to explicitely[sic] test and improve upon the limitations. Based on these studies, we no longer claim that our models provide state of the art results on the dataset untill[sic] further experiments are carried to study the effect of oversampling on the representations learned.”

Replication of Agrawal et al.

Below are my observations and code*. The analyses were conducted on commit `ed823d0`.

> *These are taken from my e-mail correspondence on March 7th and 14th 2019, without context nor respons

Observation I

What follows is my analysis of the assumed error: This notebook was my main initial focus, which I ran locally. the `get_data` function; specifically the `oversampling_rate` part, followed by the `get_train_test` function splitting part. As (to my understanding) shuffling and splitting of the data is not controlled for (the `train_tes` `scikit-learn` shuffles by default), chances are oversampled instances bleed into the test set. This led me t of the positive (oversampled) instances in the train set are also in the test set (and are thus seen during trainin

[See b]elow [for] the code I added to the bottom of the notebook to confirm this:

```
import numpy as np

x_text, labels = get_data(data, oversampling_rate=3)
data_dict = get_train_test(data,  x_text, labels)

train_set = set([str(list(x)) for x in data_dict['trainX']])
test_set = set([str(list(x)) for x in data_dict['testX']])

print("overlapping instances train/test:", len(train_set & test_set))
print("nr. instances test set:", len(test_set))
print("nr. instances train set:", len(train_set))

train_pos = [x for x, y in zip(data_dict['trainX'], data_dict['trainY']) if np.argmax(y) == 
test_pos = [x for x, y in zip(data_dict['testX'], data_dict['testY']) if np.argmax(y) == 1]
pos_train = set([str(list(x)) for x in train_pos])
pos_test = set([str(list(x)) for x in test_pos])

print("unique test instances:", len(pos_test - pos_train))
```

> The notebook can be found in our repository under `DNNs_repl.ipynb` (older version at the time, current ve Observation II).

Which gives me a unique test instances number (which are not in the train set) of 1. This subsequently leads m test instances are actually seen during training, and the model does not need to learn any transferable feature can also be inferred in the paper at the transfer learning experiment, where significant improvement is achieve methods that ignore the trained model, solely using the embeddings / weights, and thusly allowing the new m instances during training.

Observation II

> Removed some additional reply context.

If we forego the featurizer part of the code, i.e. commenting out the below part from `get_train_test`, it does still give the same overlap (this time with raw documents in `trainX` and `testX`):

```
def get_train_test(data, x_text, labels):

    ...

    # vocab_processor = learn.preprocessing.Voc
    # vocab_processor = vocab_processor.fit(x_t

    # trainX = np.array(list(vocab_processor.tr
    # testX = np.array(list(vocab_processor.tra

    trainY = np.asarray(Y_train)
    testY = np.asarray(Y_test)

    # trainX = pad_sequences(trainX, maxlen=max
    # testX = pad_sequences(testX, maxlen=max_

    trainY = to_categorical(trainY, nb_classes=
    testY = to_categorical(testY, nb_classes=NU

    data_dict = {
        "data": data,
        "trainX" : X_train,
        "trainY" : trainY,
        "testX" : X_test,
        "testY" : testY,
        # "vocab_processor" : vocab_processor
    }

    return data_dict
```

(I slightly altered my added bit at the bottom):

```
pos_train = set(train_pos)
pos_test = set(test_pos)
len(pos_train & pos_test)
```

... yielding 202 items, and:

```
sample_positive_instance = list(pos_train & pos_test)[0]
```

... taking the 0th common item, and confirming:

```
sample_positive_instance in pos_train
True

... and:

```python
sample_positive_instance in pos_test
True
```

Therefore, I do still think my raised issue persists. I have attached my version of the DNN notebook with a few extra checks if you want to take a look at it.

> Can be found in `DNNs_repl.ipynb` (these were some changes).

As [] oversampl[ing is applied] by copying instances from the entire dataset---before splitting---those few positive instances that end up in test were also oversampled (and end up in [both train and test]). This is not fixed by disabling shuffle on the sklearn data split (I confirmed), but by only oversampling on `X_train`. I have also attached a notebook where I've adapted the code to only oversample on train, which decreases the test performance to a positive F1 score of 0.33 (granted, for the single run / set of parameters I ran).

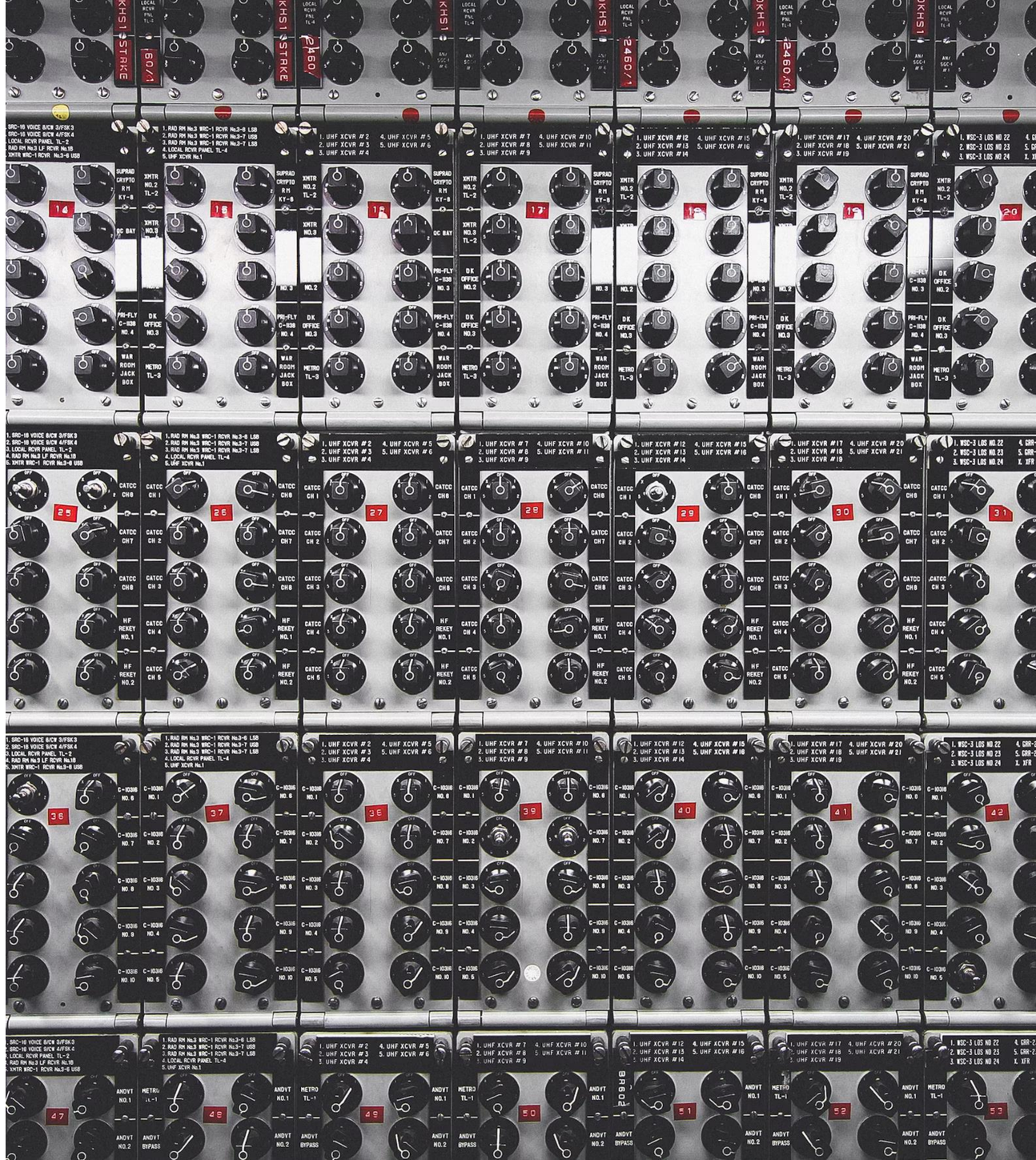
> The full replication correction can be found in our repository under `DNNs_oversample_train.ipynb`.

> [square brackets] are minor edits of the original e-mail.
```

Main Take-Away

(Your) Errors Are Everywhere

- Data
- Code
- Pipeline
- Evaluation



Any Questions?