Oxford Internet Institute, University of Oxford

Assignment Cover Sheet

Candidate Number	1047904
Assignment	Introduction to NLP
Term	HT
Title/Question	Same, same but different? An exploratory investigation of topical differences between ML- and AI-related research abstracts on arXiv
Word Count	4993(excl. tables & captions)

By placing a tick in this box ☑ I hereby certify as follows:

(a) This thesis or coursework is entirely my own work, except where acknowledgments of other sources are given. I also confirm that this coursework has not been submitted, wholly or substantially, to another examination at this or any other University or educational institution;

(b) I have read and understood the Education Committee's information and guidance on academic good practice and plagiarism at `https://www.ox.ac.uk/students/academic/guidance/skills?wssl=1`.

(c) I agree that my work may be checked for plagiarism using Turnitin software and have read the Notice to Candidates which can be seen at: `http://www.admin.ox.ac.uk/proctors/turnitin2w.shtml`, and that I agree to my work being screened and used as explained in that Notice;

(d) I have clearly indicated (with appropriate references) the presence of all material I have paraphrased, quoted or used from other sources, including any diagrams, charts, tables or graphs.

(e) I have acknowledged appropriately any assistance I have received in addition to that provided by my [tutor/supervisor/adviser].

(f) I have not sought assistance from a professional agency;

(g) I understand that any false claims for this work will be reported to the Proctors and may be penalized in accordance with the University regulations.

Please remember:

- To attach a second relevant cover sheet if you have a disability such as dyslexia or dyspraxia. These are available from the Higher Degrees Office, but the Disability Advisory Service will be able to guide you.

Same, same but different? An exploratory investigation of topical differences between ML- and AI-related research abstracts on arXiv

1047904

1 Introduction

The question of "What is AI?" has not lost in prominence ever since the research discipline's inception in the 1950s (Martınez-Plumed et al., 2018). Investigating two decades worth of publications from prominent Artificial Intelligence Conferences (*AAAI* and *IJCAI*) and the *AAAI's AI topics* database spanning 50 years of domain knowledge, Martínez-Plumed et al.(2018) undertake a serious attempt to decipher the nature of Artificial Intelligence (AI). A supplementary finding that the researchers merely acknowledge in passing inspires this research: Among the 12 AI topics identified, Machine Learning (ML) emerges as the only subject witnessing a quasi-exponential growth in document mentions since the new millennium. A short-lived hype or a transition to independence? Publication statistics on arXiv, an open-access archive comprising more than 1.8 million publications across STEM fields (arXiv, 2021a) suggest an interesting evolution.

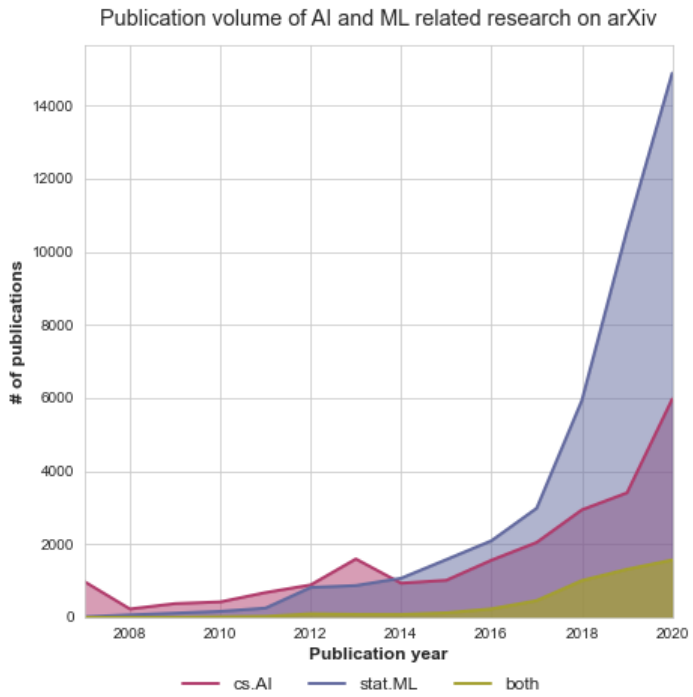

Figure 1: Publication volume of AI- and ML-related research on arXiv (2007-2020)

In recent years, publications affiliated with either ML ("stat.ML") or AI ("cs.AI") are on course for growth whereas submissions overtly positioned at the intersection ("both") remain comparatively scarce (Figure 1). Rather than asking the long-standing question about what AI is for yet another time, we are inclined to explore whether and to what extent AI and ML differ from one another. The seemingly in-

dependent evolution that research subsumed under either notion undergo on *arXiv* gives rise to this question.

1.1 Objective

Using a dataset of 16,000 *arXiv* paper abstracts from 2017 to 2020 as input, this research aims to disambiguate the subjects of AI and ML. Successfully predicting subject and year from abstract texts would provide incriminating evidence that AI and ML thematically differ from each other but also continuously evolve within their respective subject boundaries. A possible ineptitude to distinguish both subjects yet furnishes an equally valuable realisation in that researchers currently affiliating themselves with either discipline would be encouraged to overcome typological hurdles in favour of reuniting two research streams that are more intertwined than current publication dynamics on *arXiv* suggest at first glance.

1.2 Related Work

Acknowledging the fluid definitional boundaries of AI, Martínez-Plumed et al.(2018) circumscribe the discipline by 9 facets delineated by "edges". By allocating ML within the *techniques edge* under the framework's *functionality facet*, the researchers stipulate a hierarchical relationship between the two subjects. AI emerges as an abstract umbrella notion housing numerous, easier-to-articulate, sub-domains such as ML. The subject of ML is perceived to maintain tight-knit topical relations to statistics, optimization and probability theory. Accordingly, neighboring reviews portray ML as an expedient verbalized through technical jargon and succinct model terminology. In this spirit, Boutaba et al.(2018) extract a total 70 modelling techniques to map out the topical boundaries of ML.

Whereas the former two studies perform end-to-end reviews of existing publications, the scholarly perception that *"(. . .) an abstract summarises the essential contents of a particular knowledge record and is a true surrogate of the document"*(Cleveland & Cleveland, 1983, p. 104) has motivated research to ground inquiries into the nature of research disciplines on the exploration of structural and lexical abstract properties. Literature employs the notions of "moves" and "steps" when exploring the anatomy of abstracts. According to Ruiying and Allison (2003, p. 370), *"[t]he concept of Move captures the function and purpose of a segment of text at a more general level, while Step spells out more specifically the rhetorical means of realizing the function of Move"*.

Qualitatively analysing the structural features of a small body of abstracts per-

taining to protozoology research, Cross and Oppenheim (2006) uncover a five-move pattern according to which abstracts make reference to existing research, purpose, pursued methodology, results and discussion. The researchers portray abstracts as condensed document representations granting swift yet coherent access to the main points distributed in the parent text. Melander et al. (2011) uncover that rhetoric and linguistic properties of abstracts are not only a product of culture but also vary by discipline. Analyzing abstracts pertaining to research in biology, medicine and linguistics across 3 languages, within- and between-discipline differences are recorded across subjects.

Departing from in-depth, small-scale abstract appraisals, another stream of research resorts to quantitative methods to mine patterns from larger corpora. With a primary interest for AI-related publications, Uasan (2001) assesses lexical properties based on word and n-gram frequencies while inquiring into syntactical features using Tapanainen and Järvinen's (1997) *non-projective dependency tagger* which, beyond identifying parts-of-speech, uncovers partial dependency word-relations such as noun-verb pairs. Unfortunately, the researcher does not action his proclaimed idea to contrast the former against abstracts pertaining to five other disciplines included in the sample. Proposing a model for automated semantic metadata annotation, Paraschiv et al.(2015) move beyond descriptive appraisals by subjecting abstracts in Educational Research to various NLP techniques including *LSA, LDA* and ontology distance measurements using *WordNet*.

While the aforementioned studies convey the impression that research abstracts do not only bundle subject-specific cues but also consist of distinct, sequentially arranged building blocks, research by Pang and Lee (2004) motivates the thought that such cues might concentrate in certain abstract regions. Predicting polarity for movie reviews, the researchers conclude that classification accuracy considerably varies across sections composing a longer text.

In sum, study motivations outlined within the introduction and findings surfaced in related studies prompt 3 research questions:

RQ_1: Do AI- and ML-related research paper abstracts differ?

RQ_2: Have AI- and ML-related research changed over the recent past?

RQ_3: Are certain abstract portions more indicative of either discipline than others?

2 Methodology

2.1 Data Source and Sampling

arXiv research paper metadata is selected for analysis. The full dataset granting access to publications' abstract and supplementary fields including authors, title and category tags is retrieved from *Kaggle.com* and subsequently filtered to the temporal and topical domains of interest.

Abstracts tagged as "stat.ML" (statistics - Machine Learning) and "cs.AI" (computer science - Artificial Intelligence) are retained for analysis. Since publications on *arXiv* are affiliated with a primary category and an unconstrained number of secondary categories (arXiv, 2021b), only abstracts whose category tag array contains either one of the tags but not both are subset to assemble a contrastive classification baseline. The original idea to constrain analyses only to those abstracts carrying either tag as their primary category was abandoned for data sparsity reasons but most importantly to anticipate a classifier which learns to distinguish between the parent domains of "computer science" and "statistics" as opposed to disambiguate between the fields of AI and ML. Sampling for tag appearances irrespective of positional hierarchy is hence intended to capture the interdisciplinary character of both subjects in interaction with other *STEM* domains.

While publications on either subject date back to 2007, the sample is deliberately constrained to the years 2017 to 2020 where publication volumes consistently surpassed 2000 articles per year and subject. This is to inform a contemporary analysis grounded on a sufficiently large sample. To resolve class imbalances, "stat.ML" publications were down-sampled. The final dataset comprises 16,000 observations (i.e. 2,000 abstracts per subject and year).

2.2 Dependent and Independent variable selection

Abstract subject ("cs.AI" and "stat.ML") as well as abstract subject faceted by publication year (2017-2020) represent the dependent variables. Topical differences between subjects and years are initially investigated by applying a mixture of exploratory analyses to unigrams, bigrams and trigrams.

One-word and two-word sequences are subsequently used for abstract vector construal in preparation of 3 classification experiments. The training set comprises an average of 14,018 unigrams and 155,555 bigrams per class. With standard deviations of 600 uni- and 4025 bigrams respectively, feature diversity distributes fairly equally across class labels (Figure 2).

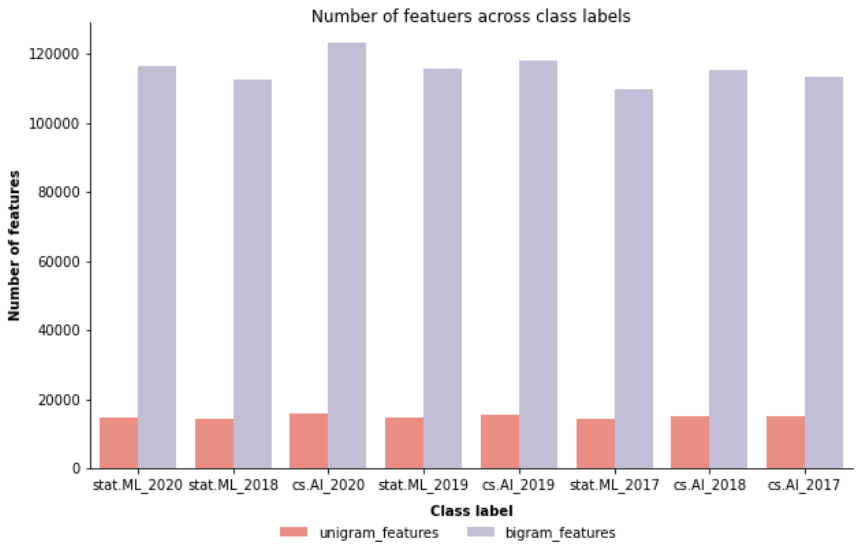

Figure 2: Unique unigram and bigram feature counts across classes

Appreciating input features in relation to the 2 prospective gating experiments rendering only certain text portions available for classification, Figure 3 depicts uni- and bigram feature set size as a function of gates applied. Lexical diversity grows in quasi-linear fashion as larger abstract portions are considered (Figure 3, Subplot 1) whereby different, yet same-sized abstract regions appear equivalently diverse (Figure 3, Subplot 2).

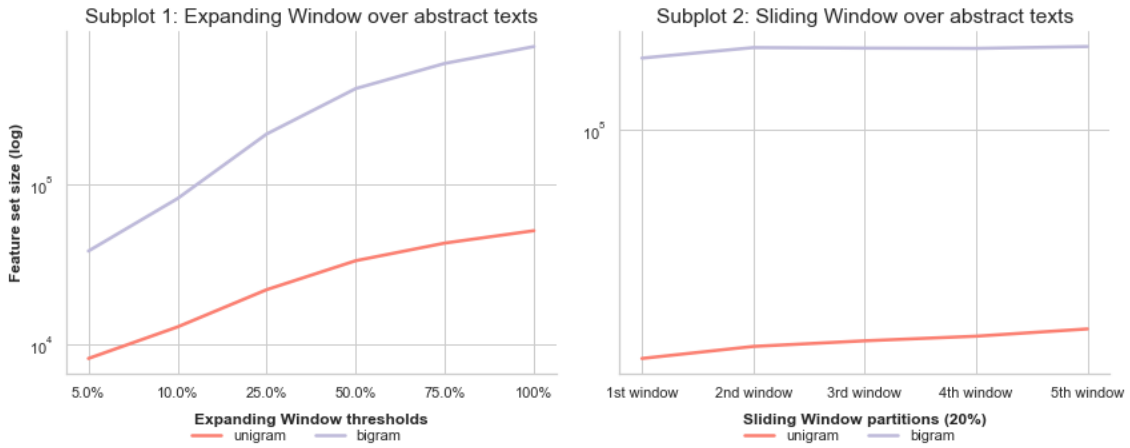

Figure 3: Unique unigram and bigram feature counts contingent on gates

ML- and AI-related abstracts appear near-normally distributed with the mid-50% of abstracts ranging between 122 and 193 words (Table 1) across both subject categories (Figure4 , Subplot 1). Faceting subjects by publication year reconfirms this

narrative (Figure4, Subplot 2).

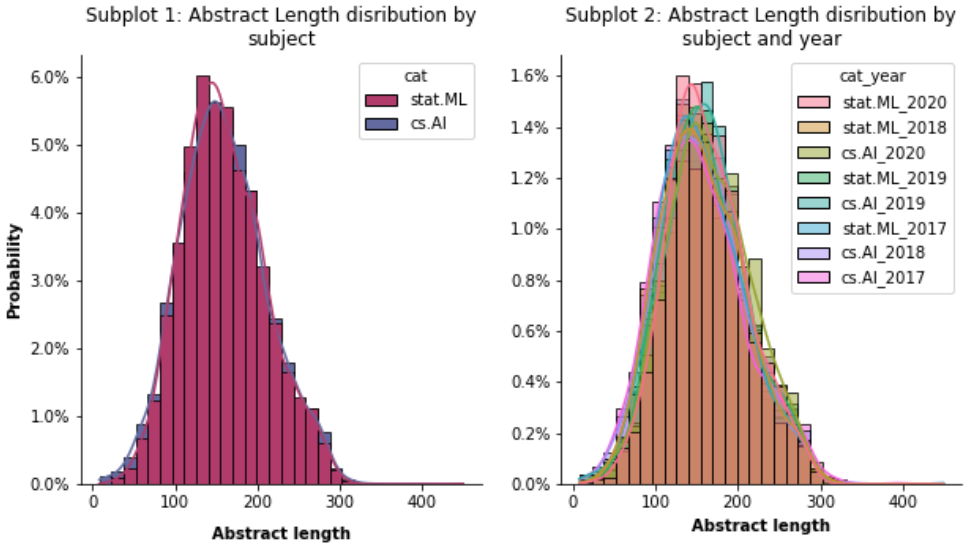

Figure 4: Distributions of abstract length by subject (l.) and subject by year (r.)

Subject	count	mean	std	min	25%	50%	75%	max
stat.ML	6400.0	159.2	50.2	21.0	123.0	155.0	193.0	449.0
cs.AI	6400.0	158.7	52.6	8.0	122.0	155.0	193.0	394.0

Table 1: Summary statistics on abstract length per subject

While the brevity of abstracts might bound our ability to appreciate the identities of ML and AI in their entirety, abstracts serve as promising proxy considering their role as time-saving devices to elucidate audiences about key take-aways (Martın, 2003).

2.3 Data Analysis

Research abstracts are subjected to a series of exploratory analyses and classification experiments. Figure 5 provides an overview of analytical methods applied.

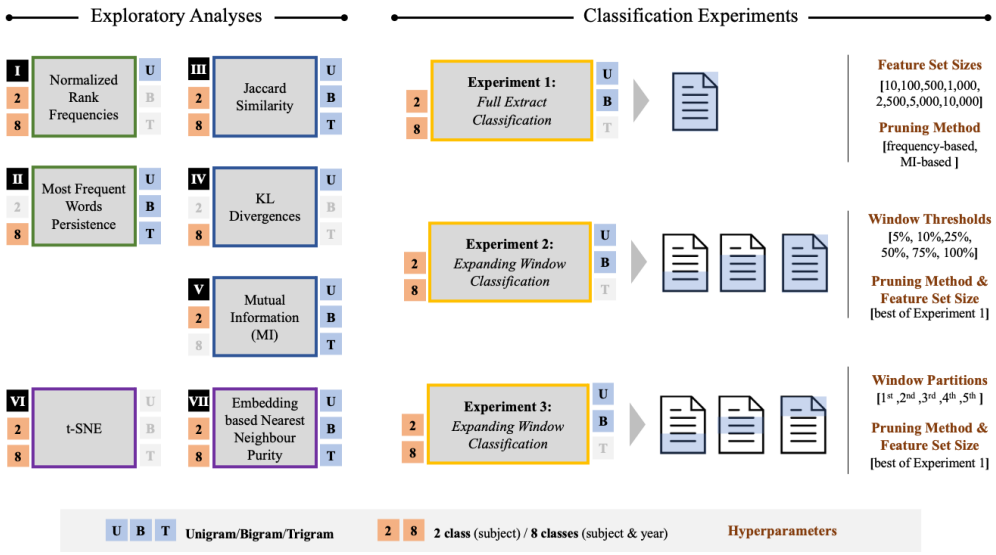

Figure 5: Overview of research methodology

In advance, abstract texts are tokenized and lower-cased. In addition, blank spaces, punctuation and stop words are removed. The pre-processed corpus is split into training, validation test set following a 80% (12,800 observations) - 10% (1,600 observations) - 10% (1,600 observations). Sampling is stratified by label and year to preserve temporal dynamics within the resulting splits. Random seeds are set to enable replicability.

2.3.1 Exploratory Analyses

Normalized word rank-frequencies (I) are computed to contrast word token distributions underlying AI- and ML-related abstracts by years. To inquire into the persistence of frequently occurring words within subjects, the % temporal congruencies of the top 20 unigrams, bigrams and trigrams are measured year-on-year as well as between 2017 and each subsequent year (II). Thereafter, several information theoretical measures are meant to quantify lexical similarities between and within the two subjects. First, we calculate pairwise *Jaccard* similarities for the unigram vocabulary of all 8 classes (III). Second, *Kullback-Leibler* divergences are calculated in consideration of word ranks and identities to assess distributional similarities for unearthed intersections (IV). Third, *Mutual Information* is measured in order to

surface the uni-, bi- and trigrams most indicative of ML or AI (V). Next, we average 300-dimensional pre-trained *fasttext* word embeddings into abstract vector representations. Abstract vectors are dimensionality reduced using *t-SNE* to explore spatial dynamics in the 2D space (VI). Lastly, nearest neighbour purity is assessed based on pairwise cosine similarities (VII). The just-mentioned analyses are applied to the training set only.

2.3.2 Classification Experiments

3 classification experiments are conceived. Each experiment aims to predict subject (2-class classification) as well as subject and year (8-class classification) based on abstract text unigrams or bigrams but differs in regards of hyper-parameters of interest. All things equal, training performant models is a means to an end in this study. The primary objective is to disambiguate the fields of AI and ML by investigating resulting model coefficients as well as to draw conclusions from recorded hyper-parameter dynamics. Figure 5 outlines the cornerstones of each experiment.

The following 3 sections briefly profile each experiment with respect to the hyper-parameters involved:

Experiment 1 - Full Abstract Classification

In response to RQ_1 and RQ_2, experiment 1 inquires into lexical differences between AI and ML and within each subject over time.

Feature pruning based on most frequent words is contrasted with *Mutual Information* (MI) based pruning. With n representing the total set of available unigrams or bigrams, the former selects $k < n$ features based on word frequencies in the training corpus, the latter retrieves $k < n$ features deemed maximally informative about output classes. Drawing on Manning et al.(2008), *MI*-based feature selection is implemented in accordance with the following mathematical equation:

$$I(A;S) = \sum_{e_t \in (0,1)} \sum_{e_s \in (0,1)} P(A = e_t, S = e_s) log_2 \frac{P(A = e_t, S = e_s)}{P(A = e_t) * P(S = e_s)} \quad (1)$$

Random variable A indicates the presence or absence of term t in a given abstract whereas random variable S indicates whether the abstract in question pertains to subject class s. If a term t appears exclusively in abstracts pertaining to subject s, $I(A;S)$ will be maximal.

Having ranked all features, $k/2$ features most informative for either ML or AI are selected for classification. To evaluate joint effects of feature selection method and feature set size on classification performance, both approaches are evaluated over a discrete array of 6 thresholds k ranging from 10 to 10,000 features.

Experiment 2 - Expanding Window Classification

In response to RQ_3, the impact of progressive text availability on classification performance is investigated. The classification exercise is repeated for a discrete array of 5 expanding window thresholds ranging from 10% to 100% of each abstract's text made available. Feature set size and pruning method are determined based on experiment 1 outcomes.

Experiment 3 - Sliding Window Classification

In response to RQ_3, we investigate whether subject matter cues clutter in specific sections of paper abstracts. Abstracts are partitioned into 5 segments each comprising roughly 20% of text (divisional remainders are concatenated to 5^{th} segment). The classifier's capacity to disambiguate between classes is evaluated for each partition respectively. Again, feature set size and pruning method draw on experiment 1 results.

2.3.3 Classifier Choice

Multinomial linear regression (MLR) is chosen to operationalize all experimental setups. The target variable ranges over 2 (subject) or 8 classes (subject and year). Given a sparse abstract vector $a^{(i)}$ with features $[x_1, x_2, ..., x_k]$ where x_k counts the number of occurences of the k^{th} feature in $a^{(i)}$, subject s of all available classes S is assigned for which $p(y = s|a)$ is maximal. We use *softmax* for output normalization purposes and *cross-entropy loss* as loss function.

The decision to opt for *MLR* grounds on several motives.

Model interpretability. The ability to retrieve the most predictive words for each class caters to our research goal to explore subject matter boundaries between AI and ML.

Word embedding scarcity. Pre-trained *fasttext* word embeddings are found to undercover the training vocabulary by 41%. The most frequent out of vocabulary words comprise abbreviated technical jargon such as "dnns" (271 occurrences), "lstm" (325 occurrences) or "mnist" (254 occurrences). While missingness only corresponds to

4% of total word frequencies, undercovered words are highly topical. Discarding such instances would artificially deflate lexical diversity and hence jeopardize endeavours to characterize and differentiate between subjects. At the same time, investigating, *fasttext*'s capacity to approximate unknown words using character n-gram vector representations ((Bojanowski, Grave, Joulin, & Mikolov, 2017) yielded doubtful results. In accord with recent scholarly voices calling its reconstruction abilities into question (Sasaki, Suzuki, & Inui, 2019), cosine similarity spot checks between embeddings of seen words and sub-word-based embedding reconstructions pertaining to unseen, yet semantically related words fall into the single to lower double-digit percentage range (Table 2).

existing embedding	reconstructed embedding	cosine similarity
model	model's	0.0984
activation	relu	0.0508
sigmoid	relu	0.0904
explanation	explainability	0.1921

Table 2: Cosine similarity comparisons between reconstructed embeddings and existing embeddings belonging to semantically related words

In absence of necessary computational resources to re-train such embeddings, sparse vector constructions in the context of *MLR* offer an avenue to include the training vocabulary in its entirety.

Experiment execution within computational constraints. Conceived experiments call for substantial hyper-parameter tuning. Yet, computational requirements and longer run-times of more involved model architectures such as *LSTM*s or *BERT* do not reconcile well with available time and resource constraints. In result, the light-weight architecture of *MLR* emerges as actionable compromise to address the aforementioned tensions between analytical demands and computational limitations.

3 Results

Analysis results are subsequently presented. Interpretations of the same are selectively provided but remain largely reserved for the discussions chapter concluding this report.

3.1 Exploratory Analysis Outcomes

Distributional differences for high-frequency words between ML and AI emerge from normalizes rank frequency plots which alludes to subject-specific word usage (Figure 6).

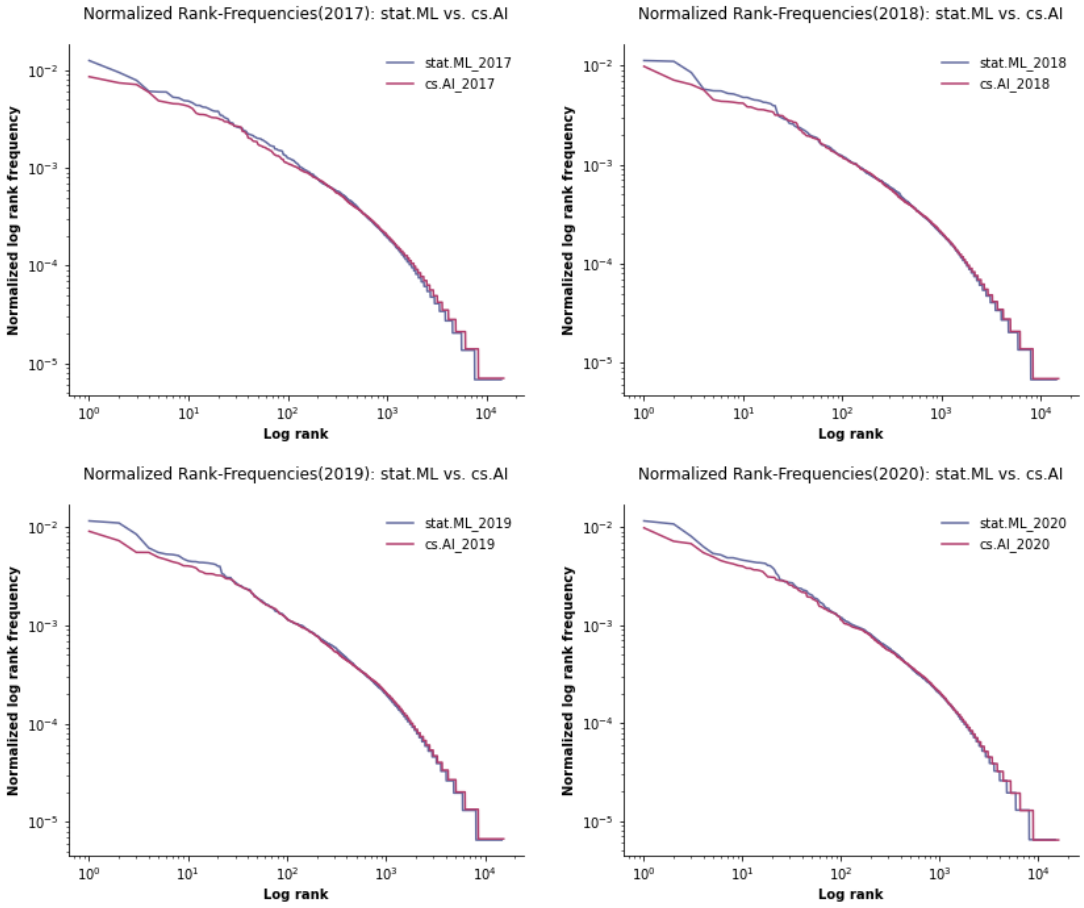

Figure 6: Distributions of abstract lengths by subject (l.) and subject by year (r.)

Specifically, high-frequency words show higher counts for ML-related abstracts between 2017 and 2020. Yet, no salient within-subject changes surface over time.

Zooming into the persistence of most frequent words per subject, temporal volatilities become apparent for both, ML and AI-related abstracts. Whereas more than

three-quarters of the 30 most frequent bigrams in 2017 remain on the leader board in 2020, notable decay emerges for trigrams. Persistence rates of 50% and 57% for ML and AI respectively hint at readjustments of topical focus within both subject matters (Table 3). Comparatively lower year-on-year changes for bi-and trigrams indicate that suspected changes may be gradual in nature (Table 4).

	Persistence from 2017 baseline		
Top 30 n-grams per subject	2017-2018	2017-2019	2017-2020
cs.AI_unigrams	83.0%	83.0%	80.0%
stat.ML_unigrams	87.0%	87.0%	87.0%
cs.AI_bigrams	83.0%	80.0%	77.0%
stat.ML_bigrams	70.0%	70.0%	77.0%
cs.AI_trigrams	63.0%	60.0%	50.0%
stat.ML_trigrams	73.0%	63.0%	57.0%

Table 3: Change in set of most frequent n-grams for AI and ML ensuing from 2017

	Year-on-Year Persistence		
Top 30 n-grams per subject	2017-2018	2018-2019	2019-2020
cs.AI unigrams	83.0%	97.0%	97.0%
stat.ML unigrams	87.0%	100.0%	97.0%
cs.AI bigrams	83.0%	77.0%	87.0%
stat.ML bigrams	70.0%	90.0%	70.0%
cs.AI trigrams	63.0%	67.0%	63.0%
stat.ML trigrams	73.0%	80.0%	70.0%

Table 4: Year-on-Year change in set of most frequent n-grams for AI and ML

Applying *MI*, n-grams most indicative of either AI or ML surfaces lexical differences. Unigrams informative about ML pertain to technical terminology on data structures, distributions (e.g. *matrix, gaussian, distribution*) and optimization (e.g. *gradient, convergence*) whereas unigrams characteristic of AI are more abstract (e.g. *reasoning, logic, language*) and relational (*human, agent*) in nature. Bigrams and trigrams indicative of ML allude to analytical techniques (e.g. generative adversarial networks, principal component analysis) and procedural concepts (e.g. *numerical experiment, gaussian process, empirical risk minimization*) portraying ML as a means to an end. On the other hand, AI-centric word sequences appear more purposive,

hinting at use cases and application scenarios (e.g. *knowledge representation, neural machine translation*). Appendix 1 provides an overview of the top 10 most indicative n-grams for both subjects.

While *Jaccard* similarities of 27.9%, 8.8% and 1.2% between AI and ML for unigram, bigrams and trigrams respectively hint at subject-bound lexica in the aggregate, examining pairwise *Jaccard* similarities on unigram vocabulary in due regard of publication year reveals that within-subject boundaries blur when faceted by time. As AI-related abstracts published in 2020 ("cs.AI_2020") appear to be almost as similar to 2019 publications (35.5%) as they are to those abstracts dating back to 2018 (35.2%) and 2017 (35%), temporal proximity manifests itself at most in the decimal place of the similarity metric (Figure 7). The same pattern emerges for abstracts pertaining to ML.

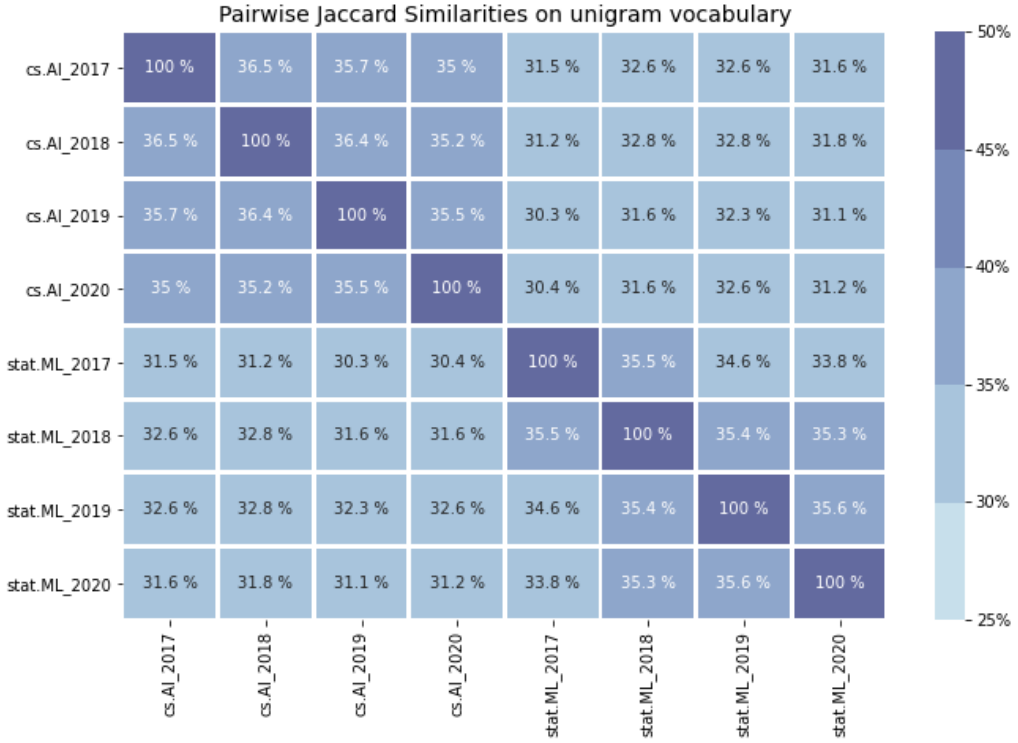

Figure 7: Pairwise *Jaccard* similarities for unigram vocabulary

Quantifying pairwise differences between unigram probability distributions on shared vocabulary, one-to-many *KL-divergence* plots reconfirm previous findings (Figure 8). While divergences from 0.07 (e.g. "stat.ML 2018" to "stat.ML 2019") up to 0.34 (e.g. "cs.AI_2020 to stat.ML_2017") prove generally low considering that *KL-divergence* values range from 0 to $+\infty$, between-subject divergences (i.e. "cs.AI" vs. "stat.ML") are much more pronounced relative to within-subject divergences

(e.g. "cs.AI_2017" to "cs.AI" for all other years) for any given combination. At the same time, pairwise within-subject distances are largely invariant. Subject-based averages computed for each subplot reinforce this narrative. Solely calculated on shared vocabulary, recorded (dis-)similarities furnish a lower bound.

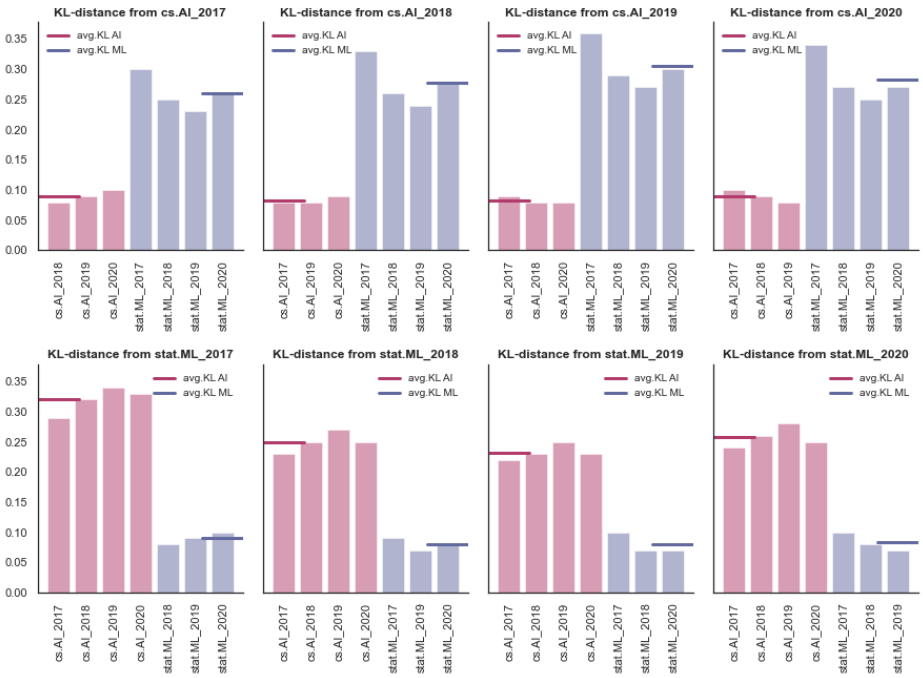

Figure 8: One-to-all *KL-divergences* based on unigram vocabulary

Spatial allocations of dimensionality-reduced abstract-level embeddings across two *t-SNE* plots reconfirm previous findings.

Whereas both subjects occupy distinct although not mutually exclusive areas (Figure 9), faceting subjects by year reveals substantial overlay between classes pertaining to the same subject (Figure 10).

Lastly, a random sample of 160 observations stratified by subject and year (i.e. 20 observations per class) is taken to measure average top 5 nearest neighbour homogeneity per class in an attempt to de-clutter dense visual representations with sober arithmetic (Figure 11). We exhaustively score each sample observation against the sample remainder using pairwise *cosine similarity*, retain the 5 most similar candidates and average results by class label in hindsight.

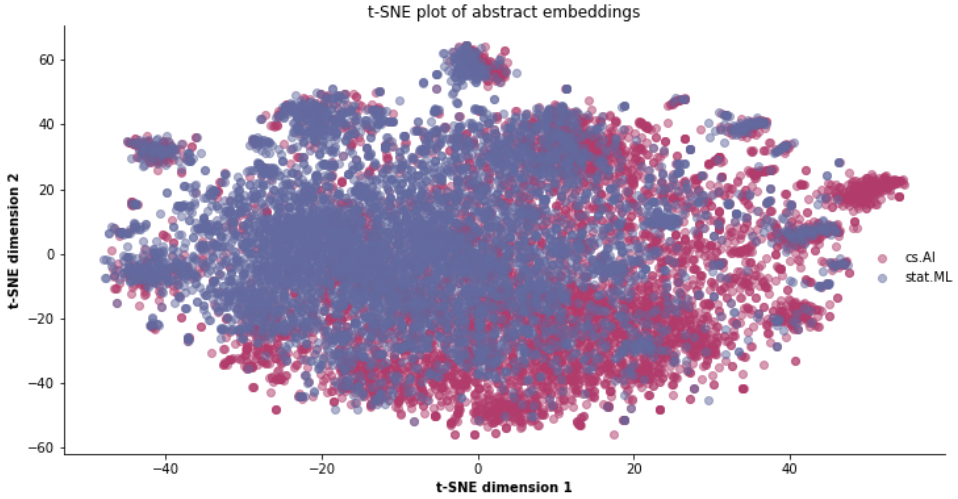

Figure 9: Dimensionality-reduced abstract vector embeddings faceted by subject

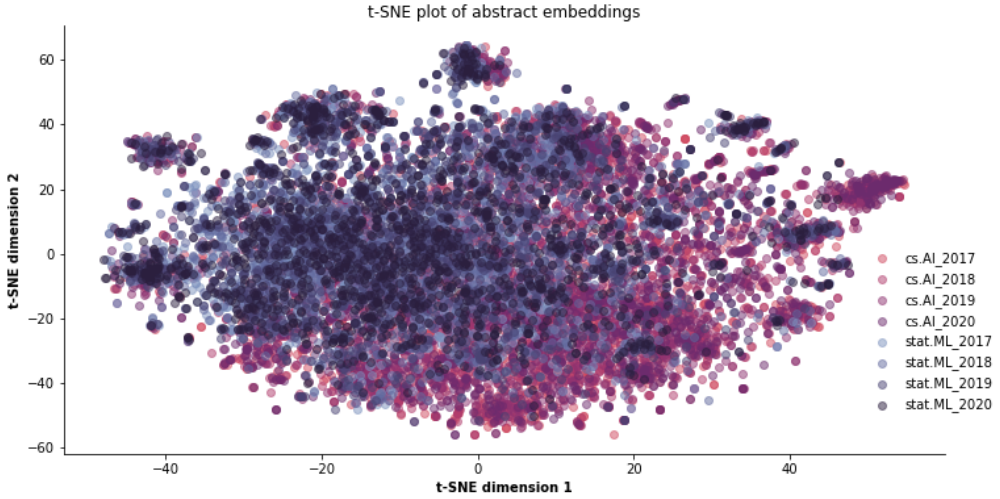

Figure 10: Dimensionality-reduced abstract vector embeddings faceted by subject and year

Abstracts are much more likely to have neighbours pertaining to the same subject than neighbours coinciding in both, subject and year. Subject purity ranges from 40% ("cs.AI_2018") to 74% ("stat.ML_2019") whereas subject-year purity assumes low-range values between 6% ("cs.AI_2019") and 17% ("cs.AI_2017"). With an average subject purity of 47.25%, AI-related abstracts are consistently less likely to surround themselves with kindred neighbours than ML-related abstracts averaging at 69.25% which hints at hierarchical association rather than a levelled relationship between the two subject matters.

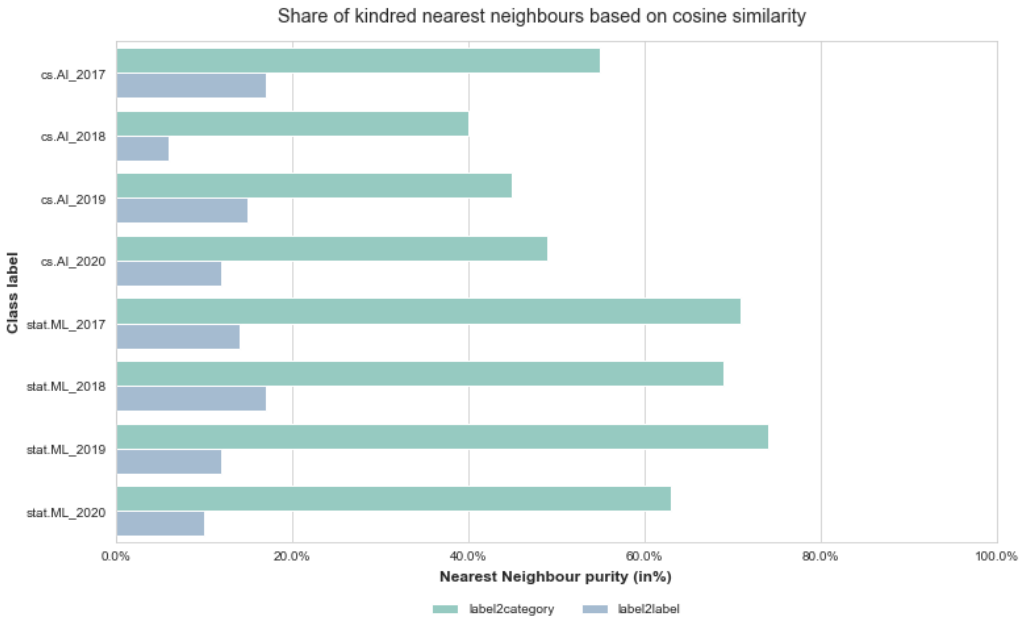

Figure 11: Share of kindred nearest neighbours based on *cosine similarity*

3.2 Classification Experiment Outcomes

Results notation remark: Model names represent concatenations of parameters in place. 2-class classification is marked with "cat". 8-class classification is marked with "cat_year". For instance, code *mutual_info-unigram-cat-5000* qualifies a model classifying between subjects (2 classes) using 5000 unigram features pruned with *MI*.

3.2.1 Experiment 1 - Full Abstract Classification

Grid searching for optimal feature size and pruning technique for either unigrams or bigrams yields two batches of 28 models trained for the classification of either 2 classes (subjects) or 8 classes (subjects by year). Figure 12 summarizes confusion matrix results for the best performing uni- and bigram models per classification task.

Classification results validate suspicions raised during exploratory analyses. The best model notably exceeds the random baseline of 50% by distinguishing between AI and ML-related abstracts with 83% accuracy (Figure 12,Subplot 1). Comparatively low type I and type II errors translate into elevated precision and recall rates in the lower 80%'s. By contrast, the best model on the 8-class classification task distinguishes between subject and year with 29% accuracy only (Figure 12,Subplot 2).

Outcompeting the random baseline of 12.5%, the model yet commits numerous misclassifications which disproportionally allot to the time axis within a given subject irrespective of gram (Figure 12,Subplot 3 % 4). Apart from a few exceptions ("stat.ML_2020" for unigram-based classification; "stat.ML_2019" & "cs.AI_2017" for bigram-based classification), the model is inclined to classify the year immediately before or after the true class label which invites the hypothesis that increasing temporal proximity coincides with topical similarity.

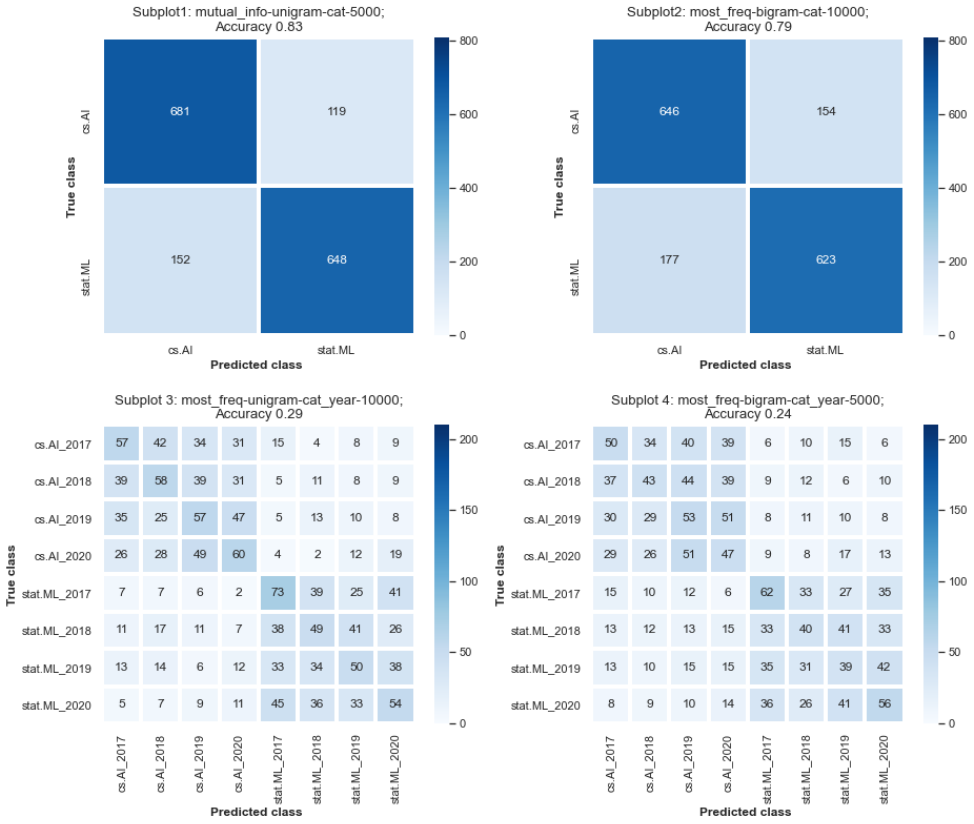

Figure 12: Confusion Matrix results for best performing models

Figure 13 further shows that the inability to disambiguate abstracts within subjects by year is largely configuration agnostic. While increasing the number of features tends to ease the challenge, accuracy remains below 30% irrespective of pruning methods and gram.

To scrutinize model-inflicted shortcomings, one-shot attempts to train a two-layer *feed forward neural network* (*FFNN*) and a *LSTM* are made. Surprisingly, neither allowing for non-linearity (*FFNN*) nor the consideration of potential long-term dependencies (*LSTM*) outperformed *MLM*. *FFNN* test accuracy results remain constantly below 27.88% across different hidden dimensions considered whereas *LSTM*

validation accuracies remain at the level of chance across epochs. Investigating model output value distributions of reveals that the *LSTM* recurrently predicts the same class irrespective of input sequence, indicating that the model fails to learn from the underlying data. Since this exercise was reactive in nature and not formally a part of the originally conceived methodology, *FFNN* and *LSTM* model outputs are presented in the enclosed *Juypter Notebook* only.

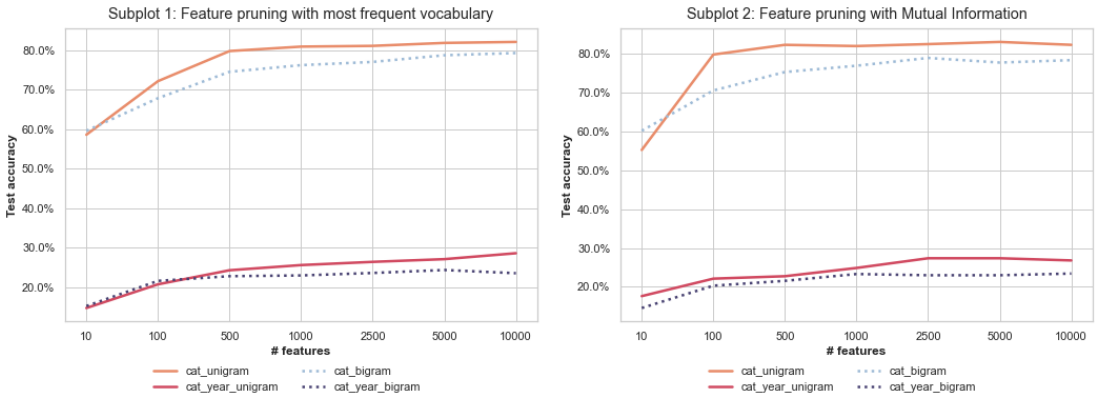

Figure 13: Test accuracy as function of pruning method and feature set size

Two additional insights emerge.

First, classification based on unigrams consistently outcompetes bigram-based modelling attempts. Second, only 1 out of the 4 winning models grounds on MI-driven feature selection (Figure 12,Subplot 1). A closer look at the candidate model longlist yet puts results in perspective (see attached *Jupyter Notebook* for reference). MI-based selection of only 500 features achieves higher accuracy on the two-class classification task (82.31%) than word-frequency based pruning which exhausts the entire hyper-parameter bandwidth of 10,000 features to achieve comparable performance (82.13%). At the same time, the remaining 3 best models showcased in the confusion matrix (Figure 12, Subplots 2-4) are closely followed by MI-based model alternatives achieving marginally lower accuracy with considerably less features.

The efficacy of *MI*-based feature pruning resurfaces in Figure 13 where models grounding on *MI* do not only achieve higher accuracy from the outset but also witness steeper increases as more features are selected. This observation is particularly pronounced for 2-class classification.

Considering these findings, model predictor appraisals are limited to MI-based candidate models trained to differentiate between both subject matters in the aggregate. Reading out the 5 most predictive uni- and bigrams across feature sizes, we are able

to confirm preliminary insights from exploratory analyses. Top predictors affiliated with ML are predominantly technical. Vocabulary can be broadly segmented into statistical jargon (e.g. *stochastic gradient,posterior distribution,random variable,total variation,sampling algorithm*), terminology descriptive of machine learning model classes (e.g. *topic model,lstm model,pca*) and evaluative vocabulary (*screening,attain,outperforms previous,loss,predictive performance*). AI-related features are comparatively difficult to categorize. At the same time, they are less specific (e.g. *programming,analysis technique*) and less technical in nature. We recognize a tendency to substitute jargon with simplifying terminology which "abstracts away" technical complexities (e.g. *action space,analysis technique,ai systems,intelligent systems*).In addition, a "two-sides of the same coin" dynamic looms in the predictor set. Employing words such as *lately, recent* or *possibility*, research in AI accentuates novelty and opportunity whereas ML, using words such as *challenge*, *problem* or *problems*, emphasizes problem solving.

3.2.2 Experiment 2 - Expanding Window Classification

We conduct a gating experiment subjecting the classification tasks to progressively expanding windows of text. Each run considers abstracts' text body only up to a certain % thresholds. In acknowledgement of experiment 1 results, each model configuration was executed using MI feature pruning and 5000 features. Two batches of 10 candidate models are trained for the classification of either 2 classes (i.e. subjects) or 8 classes (i.e. subjects by year). Within each batch, 5 models each are trained using either unigram or bigram features across all % thresholds.

Figure 14 illustrates that training classifiers on larger portions of abstract texts generally coincides with increasing accuracy.

Monotonic performance increases are recorded for uni- and bigram classifiers for the 2-class prediction task whereas non-monotonic increases manifest among classifiers trained to disambiguate between subjects and years. Interestingly, accuracy gains are modest in magnitude. When differentiating between AI and ML in the aggregate, expanding from 10% to 100% abstract text translates into a 9.75% uplift for unigram-based classification (from 73.31% to 83.06%) and a 10.07% uplift for bigram-based classification (from 67.81% to 77.88%). Accuracy gains barely exceed 5% for 8-class classification. Findings may be attributable to both, the efficacy of *MI*-based feature selection or subject salience found on the outset of research paper abstracts. The latter suspicion is further investigated in the third and last experiment.

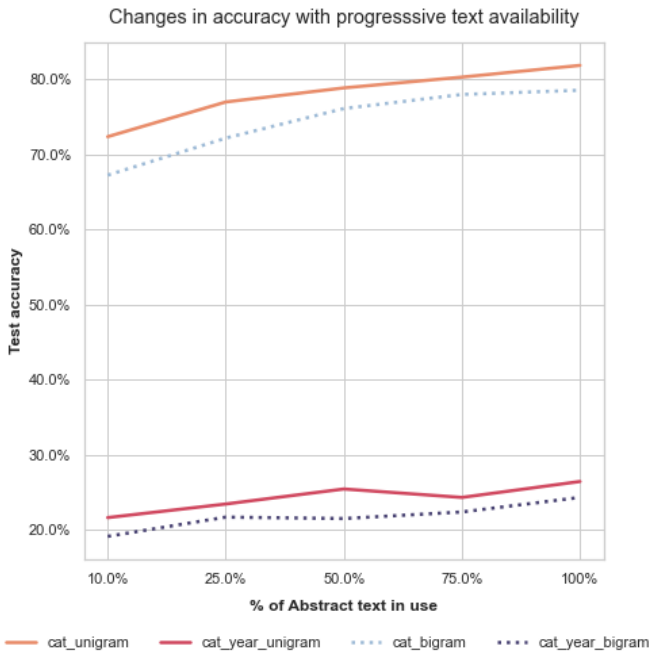

Figure 14: Test Accuracy as function of expanding window size

3.2.3 Experiment 3 - Sliding Window Classification

In a second gating experiment, classification is subjected to sliding windows of text. Abstracts are partitioned in five segments containing roughly 20% of text. During each run, a different segment is used. Modelling once again grounds on *MI* feature pruning selecting the top 5000 most informative features. As in the previous gating experiment, exhausting all combinatory possibilities yields 20 models.

With the exception of distinguishing between both subjects based on unigrams irrespective of year (referred to as "cat_unigram" in Figure 15), highest prediction accuracies are achieved when classifying based on either the top or bottom 20% percent of an abstract texts for all remaining series. For the former, predicting on the first 2 20% folds yields the best performance. Results suggest that subject matter cues tend to concentrate at the beginning and end of an abstract as opposed to the middle sections.

It is worth noting that changes in accuracy are modest in magnitude across partitions. Examining the 2-class classification model series based on bigrams ("cat_bigram" in Figure 15) which shows the strongest curvature, accuracy maximum (Window 1: 71.81%) and minimum (Window 3: 65.50%) differ by 6.3%. The 3 remaining model series witness even less performance discrepancies between the most and least predictive window. While between-window differences are evident, no partition hence deteriorates classification capacities to an extreme.

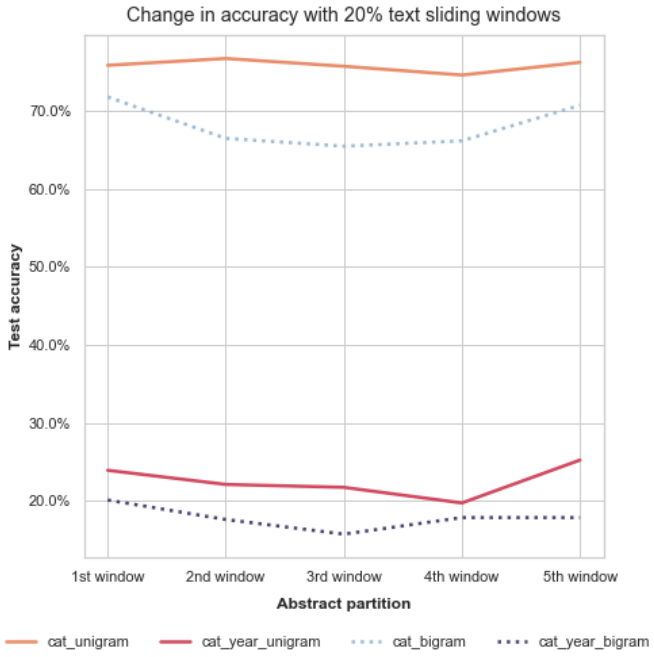

Figure 15: Test Accuracy as function of expanding window size

4 Discussion and Conclusion

By performing faceted exploratory analyses and several classification experiments on a sample of 16,000 research abstracts from *arXiv*, between- and within-subject matter differences concerning the fields of AI & ML and the predictive power of abstracts has been explored.

While findings of comparatively lower nearest neighbour homogeneities for abstracts relating to AI than ML across years allude to a hierarchical relationship that conforms with Martínez-Plumed et al.'s (2018) (2018) taxonomy, regression results disambiguating subject with 83% (unigrams) and 79% (bigrams) accuracy affirm salient differences between ML and AI in response to RQ_1. Results are not only supportive of previous scholarship asserting that abstracts vary by discipline (Melander et al., 2011) but also affiliate with the sorts of lexical differences surfaced in literature. Top mutually informative n-grams and most predictive regressors assert the soft-defined, abstract nature of AI while invoking strong associations between ML and highly technical jargon ranging from model terminology to statistical concepts (Boutaba et al., 2018; Martínez-Plumed et al., 2018). Yet, it is worthwhile noting that n-gram-based analyses, while amenable to large sample sizes, are a complement rather than a substitute of qualitative, in-depth investigations advocated in traditional abstract research (e.g. Cross Oppenheim, 2006). While sequences render spurious results with increasing window sizes (e.g. *question answering vqa*), stop-

word removals meant to ease classification hamper the reliable interpretations of resulting features at times (e.g. bigram feature "challenge 2018" may originate from "challenge **in** 2018. . . " or "challenge **before** 2018..").

Regarding RQ_2, looming difficulties to differentiate either subject by year during exploratory investigations have materialized across all classification experiments. Subsequent interpretations are tentative and constitute promising avenues for future research.

The inability to capture long-range dependencies may *MLR* from valuable cues. Spot checks on abstract texts reveal that opening sentences name a problem prevalent in a field of interest at the outset and mention how the proposed method does things differently a few sentences later. In result, two abstracts from different years may be classified alike for researching the same field (e.g. *Bayesian Statistics*) even though one extends the latter with a revamped methodology.

Then again, a singular, yet unsuccessful attempt to improve 8-class classification performance with a *LSTM* gives rise to a competing hypothesis that the celebrated brevity of abstracts ultimately limits their capacity to be more than a succinct summary of the actual publications where the devil often lies in the details. Such details might be found in the body of the paper which has not been available for modelling in this study. Future research is encouraged to bring a bandwidth of different *LSTM* hyper-parameter configurations to trial but also to follow (Cross Oppenheim's (2006) suggestion to study congruencies between abstracts and their parent documents in order to inquire into the aforementioned suspicion. Taking other, readily-available input features such as title into account offer additional means for enrichment. Lastly, *Jaccard* similarity and confusion matrix outcomes welcome a third hypothesis. While the former foreshadows that lexical overlaps decrease over time, the latter insinuates that misclassifications tend to allot to years in immediate vicinity to the true label. In result, a continuous time-frame of 4 years might have been too narrow to detect systemic changes for these disciplines. Since the pace with which AI and ML evolves intuitively opposes the just-mentioned chain-of-thought, future research should validate this hypothesis by experimenting with disconnected time-frames and different lags.

Regarding RQ_3, we extend Pang Lee's (2004) findings to the field of abstract research. Both gating experiments conducted suggest that certain abstract portions are more apt to point out the underlying research discipline than others. While training the classifier for the two-class classification problem on expanding windows of text yields high baseline accuracies from the start, imposing sliding windows calls out

the predictive power of abstracts' opening and closing sections. Resorting to Cross Oppenheim's (2006) finding of a five-move pattern governing the macro-structure of scientific abstracts, the moves of situating a piece of research with existing literature and the discussion of results may be particularly distinguishing pointers. The overarching question whether the five-move framework applies to ML-and AI-related research abstracts on *arXiv* calls for future research. Moderate uplifts recorded for larger windows as well as between best and worst performing sliding windows yet allude to the virtue of abstracts of being extremely dense surrogates of the parent paper where healthy levels of topical salience arises everywhere from start to end.

This paper concludes with a final reflection: While our attempt to distinguish between AI and ML was successful, model predictor analyses leave us with the impression that both subjects address similar contents from different angles. Irrespective of differences uncovered, we hope that this study provides compelling grounds for more joint research endeavors in the future than one witnesses on *arXiv* today.

5 Appendices

5.1 Top 10 n-grams based on Mutual Information

Top 10 Unigrams		**Top 10 Bigrams**	
stat.ML	cs.AI	stat.ML	cs.AI
(data,)	(ai,)	(neural, networks)	(artificial, intelligence)
(gaussian,)	(intelligence,)	(machine, learning)	(reinforcement, learning)
(regression,)	(agents,)	(gradient, descent)	(intelligence, ai)
(matrix,)	(reasoning,)	(stochastic, gradient)	(deep, reinforcement)
(gradient,)	(human,)	(numerical, experiments)	(natural, language)
(convergence,)	(agent,)	(real, data)	(question, answering)
(linear,)	(language,)	(gaussian, process)	(knowledge, base)
(distribution,)	(logic,)	(variational, inference)	(ai, systems)
(descent,)	(reinforcement,)	(component, analysis)	(knowledge, representation)
(stochastic,)	(planning,)	(convergence, rate)	(tree, search)

Table: Top 10 unigrams and bigrams per subject ranked by Mutual Information

Top 10 Trigrams	
stat.ML	cs.AI
(stochastic, gradient, descent)	(artificial, intelligence, ai)
(deep, neural, networks)	(deep, reinforcement, learning)
(principal, component, analysis)	(reinforcement, learning, rl)
(machine, learning, models)	(monte, carlo, tree)
(empirical, risk, minimization)	(carlo, tree, search)
(synthetic, real, data)	(visual, question, answering)
(markov, chain, monte)	(neural, machine, translation)
(chain, monte, carlo)	(question, answering, vqa)
(generative, adversarial, networks)	(reinforcement, learning, drl)
(convolutional, neural, networks)	(natural, language, understanding)

Table: Top 10 trigrams per subject ranked by Mutual Information

5.2 Top 5 unigram & bigram predictors across feature sizes

Feat.	label	Top unigrams	Top bigrams
10	cs.AI	human, regression, matrix, ai, data	(intelligence, ai), (natural, language), (reinforcement, learning), (deep, reinforcement), (machine, learning)
10	stat.ML	matrix, regression, gaussian, gradient, agents,	(gradient, descent), (neural, networks), (machine, learning), (stochastic, gradient), (numerical, experiments)
100	cs.AI	programming, artificial, logic, intelligence	(intelligent, systems),(knowledge, bases), (action, space),(ai, systems)
100	stat.ML	variational, loss, synthetic, stochastic, high-dimensional	(predictive, performance), (model, selection),(high-dimensional, data), (optimal, transport), (posterior, distribution)
500	cs.AI	propositional, checking, rules, fuzzy, describes	(probabilistic, reasoning), (social, choice),(visual, features), (tasks, requiring)
500	stat.ML	bandits, statistics, variational, pca, smooth	(problem, estimating), (random, variable),(total, variation), (high, dimensional), (inference, problems)
1000	cs.AI	creates, compilation, conflict, fuzzy, express	(bayesian, networks), (learning, artificial),(dempster-shafer, theory), (makes, difficult), (recent, deep)
1000	stat.ML	screening, package, finance, initialized, attain	(sampling, algorithm), (update, rule), (high, confidence), (algorithm, provide), (networks, provide)
2500	cs.AI	metaheuristic, detected, counter, unintended	(real, human), (paper, aim),(program, analysis), (analysis, technique)
2500	stat.ML	wifi, framed, investors, epidemiology, posits	(image, domain), (unsupervised, fashion), (topic, model), (lstm, model)
5000	cs.AI	reliant, neuromorphic, pv, 256, 5x	(separate, data), (neighborhood, information),(space, data), (possibility, using)
5000	stat.ML	transposed, functions, payments, hides	(directly, raw), (value, information),(accuracy, using), (asr, systems), (models, usually)
10000	cs.AI	lately, low-confidence, shape-restricted,criticality	(training, achieve), (confidence, values),(data, restricted), (networks, currently), (network, optimization)
10000	stat.ML	interfaced, scrambled, finite-state, winter	(outperforms, previous), (ability, generalize),(tasks, showing), (challenge, 2018)

References

arXiv. (2021a). *arxiv.org.* Retrieved from `https://arxiv.org`

arXiv. (2021b). *Title and abstract fields (metadata).* Retrieved from `https://arxiv.org/help/prep`

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Boutaba, R., Salahuddin, M. A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F., & Caicedo, O. M. (2018). A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. *Journal of Internet Services and Applications*, *9*(1), 1–99.

Cleveland, D., & Cleveland, A. (1983). Introduction to indexing and abstracting. libraries unlimited. *Inc., Littleton, Colorado*.

Cross, C., & Oppenheim, C. (2006). A genre analysis of scientific abstracts. *Journal of documentation*.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Text classification and naive bayes. *Introduction to information retrieval*, *1*(6).

Martınez-Plumed, F., Loe, B. S., Flach, P., O hEigeartaigh, S., Vold, K., & Hernández-Orallo, J. (2018). The facets of artificial intelligence: a framework to track the evolution of ai. In *International joint conferences on artificial intelligence* (pp. 5180–5187).

Martın, P. M. (2003). A genre analysis of english and spanish research paper abstracts in experimental social sciences. *English for specific purposes*, *22*(1), 25–43.

Melander, B., Swales, J. M., & Fredrickson, K. M. (2011). Journal abstracts from three academic fields in the united states and sweden: National or disciplinary proclivities? In *Culture and styles of academic discourse* (pp. 251–272). De Gruyter Mouton.

Orasan, C. (2001). Patterns in scientific abstracts. In *Proceedings of corpus linguistics 2001 conference* (pp. 433–443).

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*.

Paraschiv, I. C., Dascalu, M., Trausan-Matu, S., & Dessus, P. (2015). Analyzing the semantic relatedness of paper abstracts: An application to the educational research field. In *2015 20th international conference on control systems and computer science* (pp. 759–764).

Ruiying, Y., & Allison, D. (2003). Research articles in applied linguistics: Moving from results to conclusions. *English for specific purposes*, *22*(4), 365–385.

Sasaki, S., Suzuki, J., & Inui, K. (2019). Subword-based compact reconstruction of word embeddings. 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)* (pp. 3498–3508).

Tapanainen, P., & Jarvinen, T. (1997). A non-projective dependency parser. In *Fifth conference on applied natural language processing* (pp. 64–71).