

BERT Goes Off-Topic: Investigating the Domain Transfer Challenge in Genre Classification

Anonymous ACL submission

Abstract

While performance of many text classification tasks has been recently improved due to pre-trained language models (e.g. BERT), in this paper we show that they still suffer from a performance gap when the underlying distribution of topics changes. For example, a genre classifier trained on *political* topics often fails when tested on documents in the same genre, but about *sport* or *medicine*. In this work, 1) We develop methods to quantify this phenomenon empirically, 2) We verify that domain transfer in genre classification remains challenging even for the pre-trained models, and 3) we develop a data augmentation approach by generating texts in any desired genre and on any desired topic, even when there are no documents in the training corpus that are both in that particular genre and on that particular topic. We empirically verify that augmenting the training dataset with the topically-controlled synthetic documents facilitates domain transfer. Our F1 classification metric has shown improvements of up to 50% for some topics, approaching on-topic training, while showing next to no improvement for others.

1 Introduction

Automatic genre classification is a critical task in natural language processing, enabling proper understanding, summarization, archiving, and retrieval of documents across various domains, such as legal and medical. It is crucial to differentiate between different genres of texts, such as academic articles, experimental protocols, regulatory documents, and patient leaflets. People can easily recognize document genres from just a few examples even if those examples are from a different domain (Crowston et al., 2010). Thus, in order to create true artificial general intelligence (AGI), we need to make sure that the trained computer models can do that as well.

Text classification research often contrasts the properties of topic vs. those of style (Dewdney

et al., 2001). However, this contrast is difficult to maintain, as the training sets in most corpora for style or genre prediction are biased with respect to topics specific to individual styles or genres, so that classifiers do not transfer across corpora in case of variation between their topics. For example, a model identifying FAQs can learn to pay attention to such keywords as *hurricane* and *tax advice* in case these topics are common for FAQs in a specific training corpus (Sharoff et al., 2010).

However, this cross-influence of topics and styles has not been studied in the context of pre-trained language models such as BERT (Devlin et al., 2018), T5 (Raffel et al., 2020) or GPTs (Brown et al., 2020). There has also been no quantification of the gap in transferring genre/style classifiers to new domains, for example, to study the performance degradation of when trained on *political* topics but to texts about *sport* or *medicine*.

In this work, we claim the following original contributions¹:

- We have created a large **corpus** with “natural genre annotation” covering a range of topics;
- We empirically **quantify** the domain transfer gap on our corpus, demonstrating drops in F1 classification performance by 20-30 absolute percentage points;
- We propose a data **augmentation** approach which involves training text generators that can produce synthetic documents in any of the **genres** present in the genre training corpus and on any **topic**, out of those identified by neural topic-modeling algorithm (Dieng et al., 2020) trained on an unrelated topically diverse large corpus.
- We verify that augmenting the training dataset with synthetics tests generated by our approach **facilitates** domain transfer by im-

¹The tools and the experimental setups will be available.

082 proving F1 classification metric by 2-6 ab-
083 solute percentage points in average and on
084 some topics as much as from 57.6 to 73.0.
085 This improves over a recent data augmentation
086 baseline, which generate synthetic doc-
087 uments but do not apply any domain transfer
088 mechanisms that we propose here.

- 089 • Through **ablation studies**, we verify that
090 all the components of our augmentation ap-
091 proach are crucial. Also, through variation of
092 hyper-parameters we can detect the optimal
093 augmentation setup to avoid gradual degra-
094 dation of its performance.
- 095 • Through a qualitative exploratory study with
096 **ChatGPT** we were able to confirm that even
097 a much larger language model still suffers a
098 domain transfer gap.

099 2 Related studies and baselines

100 Automatic genre classification has been recog-
101 nised as an important task since the 1990s, for
102 more information see [Santini et al. \(2010\)](#). The
103 effect of topical biases has been estimated em-
104 pirically by considering the reduction in per-
105 formance of genre classifiers across topics in the New
106 York Times corpus ([Petrenz and Webber, 2010](#)). Several studies have also demonstrated the
107 success of PLMs with respect to the genre classifi-
108 cation tasks ([Rönnqvist et al., 2021; Kuzman et al., 2022](#)). However, there have been no studies of topo-
109 ical biases for these models. The split between
110 topics and styles has been studied for a related
111 task, including disentangled representation ([John et al., 2019](#)) and other methods of topic-style de-
112 composition ([Romanov et al., 2019; Subramanian et al., 2019](#)). However, our study focuses on the
113 numerical estimates of the topic transfer gap on
114 large samples diverse in topics and in genres.

115 A related research area concerns the use of
116 causal models for interpreting the biases of neu-
117 ral predictions, for example, with respect to gen-
118 der ([Vig et al., 2020](#)). There have been studies
119 to investigate biases in neural models by
120 adding counter-factuals ([Hall Maudslay et al., 2019; Kaushik et al., 2020](#)). Also, there have been
121 studies that looked at impact of out-of-domain
122 training data on other kinds of PLM-based clas-
123 sifiers. In particular, [Hendrycks et al. \(2020\)](#) no-
124 ticed that while in general PLMs are more robust
125 than previous models, they still suffer from spuri-
126 ous clues. However, they tested the transfer gap

127 only on a few hand-picked datasets with similar
128 tasks but different data distributions (e.g. sen-
129 timent analysis trained on *book* reviews applied
130 to *movie* reviews), while here we are presenting
131 an original methodology based on a neural topic
132 model to investigate domain transfer between a
133 wide variety of topics. Also, none of those prior
134 works looked at domain transfer for genre/style
135 classification tasks which we do here.

136 It has been noted that well-established data aug-
137 gmentation (DA) methods in domains such as com-
138 puter vision and speech recognition ([Anaby-Tavor et al., 2020; Giridhara et al., 2019; Krizhevsky et al., 2017](#)), relying on simple transformations of
139 existing samples, cannot be easily applied to nat-
140 ural text since they can lead to syntactic and se-
141 mantic distortions. For a survey of DA approaches
142 for various natural language processing tasks we
143 refer a reader to [Feng et al. \(2021\)](#). They survey
144 mentions several studies showing that DA is gen-
145 erally much less beneficial when applied to out-
146 of-domain data (as studied here), likely because
147 “the distribution of augmented data can substan-
148 tially differ from the original data.” While only
149 a few of the surveyed works involved pre-trained
150 language models (PLMs), the survey points out
151 that PLMs have made many previously useful DA
152 techniques *obsolete* since fine-tuned PLM-based
153 classifiers already achieve high performance, as
154 they have been pre-trained on large and diverse
155 corpora. For those reasons, we decided not to con-
156 trast our approach with any of the classical pre-
157 PLM domain transfer techniques, such as [Blitzer et al. \(2007\)](#) or [Daumé III et al. \(2010\)](#).

158 While up to our knowledge, none of the prior
159 works has specifically looked into domain trans-
160 fer gap for genre (or style) classification, it is
161 still worth to note several closely related works,
162 some of them included in [Feng et al. \(2021\)](#)
163 survey that involved PLMs not only as classi-
164 fiers but also as generators for augmented data.
165 This includes [Kumar et al. \(2020\)](#) who looked at
166 sentiment/intent/question classification, [Lee et al. \(2021\)](#) who targeted under-represented categories,
167 [Edwards et al. \(2021\)](#) who focused on selecting
168 the seeds examples to train augmentation genera-
169 tion in the context of few-shot classification, and
170 [Yang et al. \(2020\)](#) focused on low-resource in
171 commonsense reasoning. Since the augmentation
172 approach tried in those works is based on straight-
173 forward training (fine-tuning) a PLM-based text

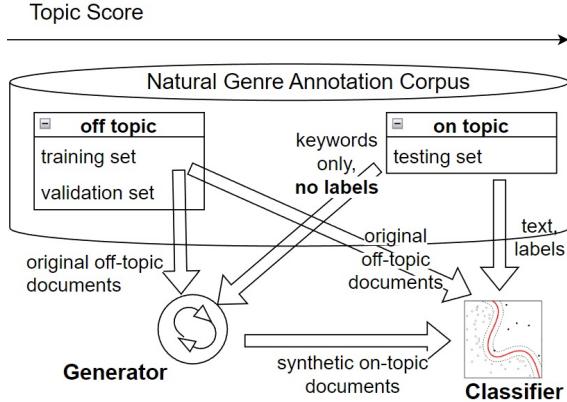


Figure 1: Experimental workflow

generator using the existing data (without exercising any topical control), we include the results from this general approach in “aug baseline” column in addition to the baseline that does not attempt any augmentation (“off-topic” column in Table 5). Since the above mentioned works also have demonstrated that classical “back-translation” augmentation approach is substantially inferior to the PLM-based text generation, we decided not to include the former in our experiments.

Jin et al. (2022) provides an overview of recent research in a closely related task of text style transfer (TST). Unlike TST, we are interested in keeping the topic, but not specifically concerned with preserving the content as long as the generated documents aid in domain transfer. The challenges maintaining coherent style and topic within longer texts (that exceed the current transformers’ input limits of 500-4000 tokens) have been proposed to address by progressive generation (Tan et al., 2020). Here, we are not as much concerned with the output quality, but rather with their help in domain adaptation.

3 Methodology

3.1 Assessing domain biases

First, we have elaborated ad-hoc studies of domain biases in genre classification (Petrenz and Weber, 2010; Sharoff et al., 2010), which used a small number of hand-picked datasets with similar tasks but somewhat different distributions, and developed a general methodology for assessing domain transfer. The main idea is to simulate the situation

when a classifier is trained on the documents that lack a certain topic, e.g. *medicine*, and then test it on the documents where such topic is well represented. This performance is compared with the situation when the classifier is trained on the documents where this topic is represented as well. We involve two classes of models: 1) a topic model produced from a diverse corpus (even though it might be biased with respect to its genres), and 2) genre-classification models based on a pre-trained language model (PLM, such as Bert) fine-tuned on a genre-diverse corpus (even though each individual genre might be biased with respect to its topics). Figure 1 illustrates the overall workflow for our experiments as described below.

3.1.1 Estimation of Topic Models

For our experiments, we needed as diverse topic model as possible so that we can assess the performance gaps when transferring between the topics. The topic model in this study was produced by a neural model (Dieng et al., 2020) which can achieve better interpretability in comparison to traditional Latent Dirichlet Allocation (LDA) models (Blei et al., 2003). More specifically, the Embedding Topic Model (ETM) differs from LDA by estimating the distribution of words over topics as:

$$w_{dn} \sim \text{softmax}(\rho^\top \alpha_{z_{dn}})$$

where ρ are word embeddings and $\alpha_{z_{dn}}$ are topic embeddings, dn refers to iteration over documents and topics, see Dieng et al. (2020) for the full description of ETM. For estimating the topic model, we used a topically-diverse corpus of ukWac (Baroni et al., 2009) created by wide crawling of web pages from the .uk top level domain name (the total size of ukWac is 2 billion words, 2.3 million Web pages). As suggested by Dieng et al. (2020), the number of topics of a topic model can be selected by maximising the product of topic coherence (the average pointwise mutual information of the top words for a topic) by its diversity (the rate of unique words in the top k words of all topics). In this way we arrived at choosing 25 topics for the ukWac corpus, see Table 3, Topic Coherence of this model is 0.195, Topic Diversity is 0.781. In the absence of a gold test set for an unsupervised method, all of the topics are interpretable (the topic labels in Table 3 in Appendix have been assigned by inspecting the keywords and a sample of documents). Topic 8 applies to short documents with residual fragments from HTML boilerplate

cleaning in ukWac, so that the date and time indicators remain the only identifiable keywords for such documents.

3.1.2 Genre Corpus

We also needed a corpus with good coverage of several genres. Up to our knowledge, there is no large corpus for that purpose, so we combined several data sources into a corpus of “natural genre annotation” so that each source is relatively homogeneous with respect to its genres. The list of our genres follows other studies which detect text types which are common on the Web (Sharoff, 2018). They have been matched to commonly used datasets, such as a portion of the Giga News corpus to represent News reporting and the Hyperpartisan corpus to represent news articles expressing opinions. The composition of the natural genre corpus is listed in Table 1 in the Appendix. The corpus of natural genres is large, but it is biased with respect to its topics. For example, the Amazon reviews dataset contains a large number of book and music reviews, and a small number of reviews of office products and musical instruments. However, these are not the topics inferred by the topic model, as this division into topics exists only with the reviews dataset, while other sources of natural annotation do not have any topics listed in the Amazon reviews or have a very different structure of annotated topics. For example, the categories assigned to the pages in Wikipedia are different from both the Amazon review labels and for the inferred ukWac topics as listed in Table 3 in Appendix. Having the topics for all documents inferred by the topic model and the documents annotated with their genres gives two views on the same document, for example, a document which starts with

(1) There's little need to review this CD after Daniel Hamlow's thoughtful and informative critique above, but I loved the CD so much I had to weigh in. In case you aren't familiar with his citations, he mentions the big three Brazilian music classics: Astrud Gilberto's "Jazz Masters 9" from Verve, "Jazz Samba" ...

can be described as a Review from its provenance from the Amazon reviews dataset and as primarily belonging to Topic 1 (Entertainment, [Table 3](#)) from its ETM inference.

3.1.3 Transfer Assessment

This subsection describes the methodology that we have developed to test the effect of a topic change. While this methodology is applicable to

any non-topical classification, here, we describe how we use it with document genres. Our main goal here is being able to create training, validation and test sets on particular topics to experiment with a genre classification task, specifically knowledge transfer between the topics. The methodology on a topic model estimation. For each topic (e.g., “Entertainment”), we create a dataset, that we label as *off-topic*. For this, we take N documents of each class (document *genre* in our case) from our genre corpus. For example, for $N = 100$ we take 100 argumentative texts, 100 instructions, 100 news reports, etc. such that the selected documents have the lowest scores with respect to that topic, e.g. documents not about entertainment. Through our experiments, we compare the classification results trained on the *off-topic* datasets with those trained on *on-topic* datasets. The latter are constructed in exactly the same way except by selecting the documents with the highest scores on the topic, e.g. those most relevant to entertainment. For each topic, we also created an *on-topic* test set making sure it does not overlap with the training sets. validation sets were off-topic since within a domain transfer setting there isn’t any on-topic training data available. Specifically, in the experiments below, we used 300 documents of each genre in a test set, 300 documents of each genre in a validation set, and varied the sizes of the training sets as stated in our section 4. This way we assess the “domain transfer”: a scenario when a model trained on *off-topic* data needs to be applied to an *on-topic* dataset. Structuring our datasets that way has several advantages: 1) both *on-topic* and *off-topic* sets have same number of documents in each class (genre) and the same total size, which allows us to determine the transfer gap under the same conditions, and 2) the datasets are automatically balanced with respect to each class (genre), even while our original corpus is not, thus the comparison metrics are more reliable and interpretable.

To build the genre classifiers, we fine-tune the ROBERTA-large (Liu et al., 2019) and BERT-large (Devlin et al., 2018) models from the Hugging-Face library² with the common in the prior research learning rate of 10^{-5} for 6 epochs, using its Adam optimizer. Following the standard validation procedure, we report the F1 score computed on the respective test set for the number of

²<https://huggingface.co/>

Table 1: Corpus of natural genre annotation

Genre	General prototypes	Texts	Natural sources
ARGument	Expressing opinions, editorials	126755	Hyperpartisan (Kiesel et al., 2019)
INSTRUCTION	Tutorials, FAQs, manuals	127472	A sample of StackExchange
NEWS	Reporting newswires	16389	Giga News (Cieri and Liberman, 2002)
PERSONal	Diary entries, travel blogs	16432	ICWSM collection (Gordon and Swanson, 2009)
INFORMATION	Encyclopedic articles	97575	A sample of Wikipedia
Review	Product reviews	1302495	Amazon reviews (Blitzer et al., 2007)
Total		1687118	

356 epochs that showed the best score on the validation
 357 (development) set.

358 As a comprise between the reliability of our
 359 results and the processing time, after preliminary
 360 investigation we settled on working with the window
 361 of 1000 characters randomly positioned within a
 362 document. Random positioning mitigates the im-
 363 pact of document structure (e.g. an introductory
 364 question positioned at the start of the StackEx-
 365 change dataset). The windows obtained this way
 366 still provide sufficient information to determine
 367 the topic and genre when read by a human.

368 In order to mitigate the superficial differences
 369 between the sources, when training and applying
 370 our classifiers, we remove all the numbers and
 371 punctuation. We do not apply this filtering when
 372 training our text generators to preserve readability.
 373 We apply it to the generated texts instead.

374 3.2 Data augmentation

375 3.2.1 Our Keyword Extraction Algorithm

376 Our domain adaptation approach involves gener-
 377 ating synthetic documents on a given topic. Thus,
 378 the generator is trained to receive a sequence of
 379 keywords and to generate a document in the genre
 380 needed. We experimented with several variations
 381 of a heuristic algorithm to select the keywords and
 382 settled on the following approach after manually
 383 inspecting the quality of the generations and their
 384 topical relatedness. We are not much concerned
 385 how truthfully the keywords represent the content
 386 of the document, but rather how well they repre-
 387 sent the topic to enable topic-focused generation.
 388 Thus, when deciding which words to extract as
 389 keywords, we promote those that are strong rep-
 390 resentatives of the document topic, which is quan-
 391 titatively assessed by our topic model. It assigns
 392 each word (in the corpus) a score with respect
 393 to each topic between 0 and 1. The higher the
 394 score the stronger the word is related to the topic.
 395 Since some documents mix several topics, at times
 396 with numerically similar proportions, we accord-

397 ingly weight the individual word scores with the
 398 overall topic scores in the document. Finally, we
 399 also want to adjust for repeated occurrences of
 400 the same word. Thus, our word scoring formula
 401 (within a document) simply iterates through all the
 402 topics and through all the word occurrences in the
 403 document while adding up the word scores with
 404 respect to the corresponding topic:

$$score(w, D) = \sum_{i \in D_w} \sum_t L(D, t) \cdot L(w, t)$$

405 where i goes over all the occurrences of the word
 406 w in the document D , t goes over all topics (25
 407 in the study here), $L(D, t)$ is the score of the doc-
 408 ument with respect to topic t and $L(w, t)$ is the
 409 score of the word w with respect to topic t .

410 We preserve only 10 top-scoring words in each
 411 document, so all the other words are discarded and
 412 the original sequence of the remaining words be-
 413 comes the keyword sequence for the generator. Ta-
 414 ble 4 in Appendix shows an example of extracted
 415 keywords along with how they are used to gen-
 416 erate new synthetic documents, as detailed in the
 417 following subsection.

418 3.2.2 Our Topical Augmentation Control

419 Our suggested method of improving domain trans-
 420 fer proceeds by augmenting the *off-topic* training
 421 set with automatically generated *on-topic* doc-
 422 uments. Thus, in a practical scenario, the test top-
 423 ics (keywords) don't have to be known in advance
 424 but can be extracted from previously unseen test
 425 documents from the target domain. The only tool
 426 required for this is an existing topic model, which
 427 can be built similarly to as we did here on any gen-
 428 eral corpus of a modest size (e.g. ukWac, Barony
 429 et.al., 2009), which is not resource-consuming.

430 To achieve this we fine-tune a pre-trained lan-
 431 guage model into a separate generator for each of
 432 our genres (listed in Table 1 in Appendix). Our
 433 earlier experimenting with using a single model
 434 for all genres and a special token to specify the
 435 desired genre resulted in weaker results. For this

407 fine-tuning, we use exactly the same $N \cdot 6$ documents
408 as are in our *off-topic* training set, thus operating
409 in a practical scenario when on-topic documents
410 are not available. Each generator is fine-
411 tuned to take a sequence of keywords extracted
412 according to the algorithm detailed above as input
413 and to generate a document in the genre corre-
414 sponding to this generator and of the topic defined
415 by the keywords. During fine-tuning, the genera-
416 tors learn to associate the input keywords with the
417 content of the output document, which becomes
418 an important mechanism of topic control and fa-
419 cilitating the domain transfer.

420 We specifically used T5 as our generating
421 model (Raffel et al., 2020). It is a unified text-
422 to-text transformer, trained on the Colossal Com-
423 mon Crawl Corpus to predict the next word based
424 on the preceding words in an auto-regressive way.
425 We used the small version since we did not ob-
426 serve any advantage in using the Base or Large
427 T5 model in our early experiments, so we kept
428 the less computationally intensive model. Its input
429 format requires a prefix to indicate which down-
430 stream task is being fine-tuned, so we used the
431 word “generate.” We trained each model for 16
432 epochs using Simple Transformers library³ with
433 a default learning rate of .001 and its Adam op-
434 timizer. For generating, we also use the follow-
435 ing T5 hyper-parameters, specifically number of
436 beams = 1, top k = 50, top p = .95. The
437 selected hyper-parameters were chosen after pre-
438 liminary experimentation by inspecting the pro-
439 duced quality of generations in terms of both top-
440 ical and genre fit. Table 4 in Appendix illustrates
441 our domain adaptation approach by examples of
442 extracted keywords and synthetic documents gen-
443 erated from those keywords in different genres.

444 One of our overall hyper-parameters is how
445 many documents to generate. Our preliminary ex-
446 perimentation suggested that 1:1 was a near opti-
447 mal ratio: the same number of original and syn-
448 thetic documents. We include several other com-
449 binations in our empirical results below.

450 4 Experiments

451 The most time-consuming part of our experi-
452 ments were fine-tuning the generators (T5) and
453 the classifiers at the cost of roughly 6000 hours
454 of NVIDIA GeForce RTX 2080.

³<https://simpletransformers.ai/>

455 4.1 Comparison Results

456 We assess the effect of domain mismatch and our
457 approach to improving domain transfer by aug-
458 menting the training sets with synthetic on-topic
459 documents. The difference between the accuracy
460 obtained before and after generation demonstrates
461 the efficiency of the augmentation model. Ta-
462 ble 2 shows the comparison results for 3 differ-
463 ent sizes of training data: 1000, 100 or 30 docu-
464 ments per genre. As we can see, the topic mis-
465 match effect is *extremely significant*: the aver-
466 age absolute F1 drop from *on-topic* to *off-topic*
467 training set is around 20% for $N = 1000$ and
468 30% for smaller N s. The average *on-topic* F1
469 score for the largest size is 86.4%, which is only
470 somewhat below 93% average human-level per-
471 formance⁴. The average *off-topic* performance for
472 that size drops to 66.8%. All three configura-
473 tions (“aug adapt” columns) demonstrate 2-6 per-
474 centage point increases in F1 from non-augmented
475 off-topic training sets (“off-topic” columns). At
476 the same time, the straightforward “augmenta-
477 tion by generating” approach from prior works (“aug
478 baseline” columns) does not show any noticeable
479 improvement, even though it was found by prior
480 work somewhat effective in several tasks not in-
481 volving domain transfer. We hypothesise that this
482 is because the general approach does not provide
483 a mechanism to facilitate *domain transfer*, while
484 our approach does. All the differences between
485 our approach and the baselines are statistically sig-
486 nificant at the level of alpha 0.01 according to
487 a pairwise t-test. This confirms empirically with
488 high confidence that our augmentation procedure
489 is beneficial for genre classification. While in this
490 current study we prioritized reporting metrics av-
491 eraged across all 25 topics rather than on individ-
492 ual topic level, we still can observe that the mag-
493 nitude of the transfer gap and the augmentation ef-
494 ffects are normally consistent across all the con-
495 figurations and models used. Still, there are some
496 exceptions due to a large number random factors
497 involved including the choice of off-topic docu-
498 ments, the quality of synthetic documents in terms
499 of both genre and topic, the optimality of hyperpa-
500 rameters, and others.

501 4.2 Ablation Studies

502 This subsection reports several *ablation* exper-
503 iments that we conducted to additionally verify the

⁴as independently estimated by the authors

Table 2: Averaged across topics F1 scores for testing genre classification domain transfer gaps and our augmentation approach. The “on-topic” columns show the performance when training and testing on in-domain documents. The “off-topic” columns present training on the off-topic documents and testing “on-topic”. “aug baseline” is the result of augmentation by generation without domain adaptation. Our domain adaptation augmentation results are in the last column for each N (“aug adapt”). The results for separate topics are included in Appendix. All our results are statistically significantly different from the baselines at the level $p < 0.01$. There is no statistically significant difference at that level between “aug baseline” and “off-topic”.

Topics:	N=30				N=100				N=1000			
	on-topic	off-topic	aug baseline	aug adapt	on-topic	off-topic	aug baseline	aug adapt	on-topic	off topic	aug baseline	aug adapt
Model:												
Roberta Large	77.0	52.1	52.4	58.1	61.2	61.4	65.1	94.9	78.8	79.0	81.7	
Bert large	79.5	49.9	51.1	55.2	84.9	55.6	58.8	85.1	68.8	68.7	71.9	

504 effects reported above and to gain the insight into
 505 the phenomena studied. In order to verify that
 506 the genre labels in our synthetic texts were im-
 507 portant we randomly shuffled them. This way,
 508 the augmented data became to act only as noise.
 509 Not surprisingly, the average scores dropped to
 510 the baseline levels which verified that using the
 511 proper model for each genre to generate the syn-
 512 thetic augmenting texts is important, and that the
 513 improvements reported above were not due to sim-
 514 plly the change in the statistical properties of the
 515 training and validation sets or due to addition of
 516 noise.

517 We also looked at several ways of mixing the
 518 original and augmented data. Table 6 presents
 519 the average across topics scores for various sizes
 520 used. It can be observed that while some small
 521 improvements can be achieved by generating more
 522 documents, those gains are not statistically signif-
 523 icant. On the other side, very small numbers of
 524 added documents indeed result in statistically de-
 525 tectable drops. Using only synthetic documents
 526 results in drops to the levels only slightly above
 527 or even below baselines. We also observed that
 528 using keywords from randomly selected off-topic
 529 documents is significantly worse than using those
 530 from the on-topic documents, which confirms that
 531 using domain adaptation mechanism such as sug-
 532 gested here by us is crucial. The details are in the
 533 last rows for each N in Table 6 in Appendix.

534 We have also looked at the optimal choice of
 535 number of keywords. While the details are pre-
 536 sented on Figure 2 in Appendix, it is worth noting
 537 here that the optimal number is indeed around 10-
 538 20 keywords. Also, the augmentation affect drops
 539 to 0 on both ends: Too few keywords means no
 540 topical control is performed. 100+ keywords re-
 541 sult in practically all the non-stop words treated as
 542 keywords. This means the model does not really

543 learn how to generate a document on a topic spec-
 544 ified by a set of keywords but it rather learns how
 545 to restore deleted stop-words from the given text.

4.3 Qualitative Exploratory Study with ChatGPT

546 As a further qualitative investigation into the prob-
 547 lem, we have also confirmed that a much larger
 548 language model still suffers a domain transfer gap
 549 when tasked with genre classification. We have
 550 randomly sampled 72 triples consisting of a pair of
 551 non-identical genres and a topic. Then, we com-
 552 pared binary classification accuracy by entering
 553 specially crafted prompts into ChatGPT⁵, which
 554 is built on top of GPT-3.5 model with approxi-
 555 mately 355 billion parameters. An example of a
 556 prompt is presented in the Appendix. Each in-
 557 cludes 5 randomly selected document examples of
 558 each genre (5-shot). The choice of those num-
 559 bers was dictated by the combination of input size
 560 limitation, our early experience and advice from
 561 prior works. For assessing a domain transfer gap,
 562 we followed the same methodology as described
 563 in section 3: we compared the binary classifi-
 564 cation performance when *off-topic* documents were
 565 used as prompt examples with when *on-topic doc-
 566 uments* were used. We have indeed verified that
 567 the domain gap exists even in a language model of
 568 that size: the average accuracy with on-topic ex-
 569 amples was 83% while the average accuracy when
 570 using off-topic examples was 42%. We also es-
 571 timated human accuracy in this setup as 88%.

572 When experimenting with our prompts, we dis-
 573 covered that it was crucial to use chain-of-thought
 574 (CoT) approach (e.g. Wei et al., 2022): After pre-
 575 senting examples of both classes, we asked the
 576 model to “list at least three criteria by which Class

577 ⁵<https://chat.openai.com/> accessed throughout March-
 578 April 2023

579 1 and Class 2 texts are different from each other."
580 Examples of the criteria generated by the model
581 can be found in Table 8 in Appendix. We have
582 qualitatively (informally) observed that: 1) Chat-
583 GPT was able to use both on-topic and off-topic
584 examples to produce criteria that looked poten-
585 tially useful for genre classification, e.g. "Class
586 1 texts appear to be informational or factual,
587 whereas Class 2 texts appear to be more conver-
588 sational or personal in nature." or "Class 1 texts
589 are typically more objective and neutral in tone,
590 while Class 2 texts tend to be more subjective
591 and expressive." 2) Both on-topic and off-topic
592 examples occasionally resulted in the criteria that
593 are topic-reliant, e.g. "Class 1 texts provided are
594 about musicians and their careers" or "Class 2 uses
595 words like *position*, *certified gold*, and *innovation*."
596 3) The presence of topically-reliant criteria
597 was stronger with off-topic examples.

598 Next, within our prompt, we separately asked
599 to apply each of the three criteria to the given
600 test document, followed by a request to com-
601 bine the criteria to make a classification de-
602 cision. Examples can be found in Table 9 in
603 Appendix. By inspecting the model's responses, we
604 have observed that using off-topic examples re-
605 sulted in the following types of chain-of-thought
606 "confusion" to happen more often than using on-
607 topic examples: 1) applying not the same criteria
608 that originally stated 2) applying a criterion incor-
609 rectly. 3) erroneously "swapping" classes when
610 combining. This suggests that while ChatGPT
611 has strong "emerging" capabilities for recogniz-
612 ing genres (see another confirmation at [Kuzman et al. \(2023\)](#)), they are weaker when the examples
613 are off-topic and so are more likely to "break" the
614 chains of thoughts.

616 5 Conclusions

617 We have demonstrated the impact of a topic trans-
618 fer gap for a PLM-based *document genre* classi-
619 fier: a severe degradation when trained on one
620 topic (e.g. *politics*) and tested on another (e.g.
621 *healthcare*). Rather than following the prior em-
622 pirical studies on the impact of domain transfer
623 that involved only a few hand-picked datasets with
624 similar tasks but somewhat different data distri-
625 butions, we have developed an original method-
626 ology based on a neural topic model to assess do-
627 main transfer between a wide variety of topics and
628 with a diverse set of genre labels. The topical

629 gap can be mitigated by means of proper topic
630 control while generating additional training doc-
631 uments (augmentation). As a result of our ap-
632 proach, a system can be trained on the documents
633 in one topic (e.g. *politics*) and applied to an-
634 other (e.g. *healthcare*) even when there are no
635 healthcare-related documents in the training cor-
636 pus that represent all possible class labels (genres
637 in the case here). We have created a large
638 corpus with natural genre annotation and a very
639 general/diverse topic model. Both can be used
640 in follow-up studies. Our methodology allows
641 to investigate other non-topical classification tasks
642 such as gender, authorship, or sentiment classifi-
643 cation. Still, our study has certain limitations. The
644 degree of improvements from augmentation is not
645 uniform. For some topics we obtain much better
646 results than for others, while occasionally the per-
647 formance on the augmented set is even lower than
648 on the original off-topic training set. We were not
649 able to observe any clear patterns of relationships
650 between the properties of separate topics, the do-
651 main gaps and augmentation effects. For those
652 reasons, this work has prioritized reporting met-
653 rrics averaged across all 25 topics. More research
654 is needed to investigate topic-level conditions for
655 more successful transfer, so we are leaving this for
656 future. However, through a qualitative exploratory
657 study with ChatGPT we were able to confirm that
658 even a much larger language model still suffers
659 a domain transfer gap. Even while our approach
660 does not solve this very challenging domain trans-
661 fer problem completely, it still suggests a direc-
662 tion in which a small but productive step can be
663 made. Larger pre-trained language models can be
664 tried in future such as GPT-3, for both generation
665 and classification. Also, larger training sets can be
666 explored, as well the "few-shot" settings. A num-
667 ber of approaches improving the quality of gener-
668 ated text, e.g. those based on Generative Adver-
669 sarial Networks ([Goodfellow et al., 2020](#)) or meta
670 learning ([Lee et al., 2022](#)) can be explored, as well
671 as various methods for controlling the quality and
672 topical fit of the generated texts.

673 References

674 Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich,
675 Amir Kantor, George Kour, Segev Shlomov, Naama
676 Tepper, and Naama Zwerdling. 2020. Do not have
677 enough data? deep learning to the rescue! In *Pro-*

678	ceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 7383–7390.	734
679		735
680	Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. <i>Language Resources and Evaluation</i> , 43(3):209–226.	736
681		737
682		
683		
684		
685	David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. <i>Journal of Machine Learning Research</i> , 3:993–1022.	738
686		739
687		
688	John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification . In <i>Proc ACL</i> , pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.	740
689		741
690		
691		
692		
693		
694	Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. <i>Advances in neural information processing systems</i> , 33:1877–1901.	742
695		743
696		
697		
698		
699		
700	Christopher Cieri and Mark Liberman. 2002. Language resources creation and distribution at the Linguistic Data Consortium. In <i>Proc LREC</i> , pages 1327–1333, Las Palmas, Spain.	744
701		745
702		
703		
704	Kevin Crowston, Barbara Kwasnik, and Joseph Rubleske. 2010. Problems in the use-centered development of a taxonomy of web genres. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, <i>Genres on the Web: Computational Models and Empirical Studies</i> . Springer.	746
705		747
706		
707		
708		
709		
710	Hal Daumé III, Abhishek Kumar, and Avishek Saha. 2010. Frustratingly easy semi-supervised domain adaptation . In <i>Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing</i> , pages 53–59, Uppsala, Sweden. Association for Computational Linguistics.	748
711		749
712		
713		
714		
715		
716	Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. <i>arXiv preprint arXiv:1810.04805</i> .	750
717		751
718		
719		
720	Nigel Dewdney, Carol VanEss-Dykema, and Richard MacMillan. 2001. The form is the substance: classification of genres in text . In <i>Proc. Human Language Technology and Knowledge Management</i> , pages 1–8.	752
721		753
722		
723		
724		
725	Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces . <i>Transactions of the Association for Computational Linguistics</i> , 8:439–453.	754
726		755
727		
728		
729	Aleksandra Edwards, Asahi Ushio, Jose Camacho-Collados, Hélène de Ribaupierre, and Alun Preece. 2021. Guiding generative language models for data augmentation in few-shot text classification. <i>arXiv preprint arXiv:2111.09064</i> .	756
730		757
731		
732		
733		
678	Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. <i>arXiv preprint arXiv:2105.03075</i> .	758
679		759
680		
681		
682		
683		
684		
685	Praveen Kumar Badimala Giridhara, Chinmaya Mishra, Reddy Kumar Modam Venkataramana, Syed Saqib Bukhari, and Andreas Dengel. 2019. A study of various text augmentation techniques for relation classification in free text. <i>ICPRAM</i> , 3:5.	760
686		761
687		
688	Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. <i>Communications of the ACM</i> , 63(11):139–144.	762
689		763
690		
691		
692		
693		
694	Andrew Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries . In <i>Proceedings of International Conference on Weblogs and Social Media</i> , San Jose, CA.	764
695		765
696		
697		
698		
699		
700	Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It’s all in the name: Mitigating gender bias with name-based counterfactual data substitution . In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 5267–5275, Hong Kong, China. Association for Computational Linguistics.	766
701		767
702		
703		
704	Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness . In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 2744–2751, Online. Association for Computational Linguistics.	768
705		769
706		
707		
708		
709		
710	Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. <i>Computational Linguistics</i> , 48(1):155–205.	770
711		771
712		
713		
714		
715		
716	Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer . In <i>Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics</i> , pages 424–434, Florence, Italy. Association for Computational Linguistics.	772
717		773
718		
719		
720	Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data . In <i>International Conference on Learning Representations</i> .	774
721		775
722		
723		
724		
725	Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection . In <i>Proceedings of the 13th International Workshop on</i>	776
726		777
727		
728		
729		
730		
731		
732		
733		

789	Semantic Evaluation, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.	844
790		845
791		846
792	Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. <i>Communications of the ACM</i> , 60(6):84–90.	847
793		848
794		849
795		850
796	Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. <i>arXiv preprint arXiv:2003.02245</i> .	851
797		852
798		
799	Taja Kuzman, Nikola Ljubešić, and Igor Mozetič. 2023. Chatgpt: Beginning of an end of manual annotation? use case of automatic genre identification. <i>arXiv preprint arXiv:2303.03953</i> .	853
800		854
801		855
802		
803	Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2022. The GINCO training dataset for web genre identification of documents out in the wild. In <i>Proceedings of the Thirteenth Language Resources and Evaluation Conference</i> , pages 1584–1594, Marseille, France. European Language Resources Association.	860
804		861
805		862
806		863
807		864
808		
809		
810	Hung-yi Lee, Shang-Wen Li, and Thang Vu. 2022. Meta learning for natural language processing: A survey. In <i>Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 666–684, Seattle, United States. Association for Computational Linguistics.	865
811		866
812		867
813		
814		
815		
816		
817	Kenton Lee, Kelvin Guu, Luheng He, Tim Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation. <i>arXiv preprint arXiv:2102.01335</i> .	868
818		869
819		870
820		871
821	Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. <i>arXiv preprint arXiv:1907.11692</i> .	872
822		873
823		
824		
825		
826	Philipp Petrenz and Bonnie Webber. 2010. Stable classification of text genres. <i>Computational Linguistics</i> , 34(4):285–293.	874
827		875
828		876
829		877
830	Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. <i>Journal of Machine Learning Research</i> , 21(140):1–67.	878
831		879
832		880
833		881
834		882
835	Alexey Romanov, Anna Rumshisky, Anna Rogers, and David Donahue. 2019. Adversarial decomposition of text representation. In <i>Proc NAACL</i> , pages 815–825, Minneapolis, Minnesota. Association for Computational Linguistics.	883
836		
837		
838		
839		
840	Samuel Rönnqvist, Valtteri Skantsi, Miika Oinonen, and Veronika Laippala. 2021. Multilingual and zero-shot is closing in on monolingual web register classification. In <i>Proceedings of the 23rd Nordic</i>	884
841		
842		
843		
789	Conference on Computational Linguistics (NoDaLiDa), pages 157–165, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.	844
790		845
791		846
792	Marina Santini, Alexander Mehler, and Serge Sharoff. 2010. Riding the rough waves of genre on the web. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, <i>Genres on the Web: Computational Models and Empirical Studies</i> . Springer, Berlin/New York.	847
793		848
794		849
795		850
796	Serge Sharoff. 2018. Functional text dimensions for the annotation of Web corpora. <i>Corpora</i> , 13(1):65–95.	851
797		852
798		
799	Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The Web library of Babel: evaluating genre collections. In <i>Proc Seventh Language Resources and Evaluation Conference, LREC</i> , Malta.	853
800		854
801		855
802		
803	Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc’Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text style transfer. <i>arXiv preprint arXiv:1811.00552</i> .	860
804		861
805		862
806		863
807		864
808		
809		
810	Bowen Tan, Zichao Yang, Maruan AI-Shedivat, Eric P Xing, and Zhiting Hu. 2020. Progressive generation of long text. <i>arXiv preprint arXiv:2006.15720</i> .	865
811		866
812		867
813		
814		
815		
816		
817	Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural NLP: The case of gender bias. <i>arXiv preprint arXiv:2004.12265</i> .	868
818		869
819		870
820		871
821		872
822		873
823		
824		
825		
826	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. <i>arXiv preprint arXiv:2201.11903</i> .	874
827		875
828		876
829		877
830		
831		
832		
833		
834		
835	Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. <i>arXiv preprint arXiv:2004.11546</i> .	878
836		879
837		880
838		881
839		882
840		883

6 Appendix

Table 3: Keywords from ukWac for the topic model with 25 topics

Label: Nr	Top keywords
Finances: 0	insurance, property, pay, credit, home, money, card, order, payment, make, tax, cost, time, service, loan
Entertain: 1	music, film, band, show, album, theatre, festival, play, live, sound, radio, song, dance, songs, tv, series
Geography: 2	road, london, centre, transport, park, area, street, station, car, north, east, city, west, south, council, local
Business: 3	business, management, company, service, customers, development, companies, team, experience, industry
University: 4	students, university, research, learning, skills, education, training, teaching, study, work, programme
Markets: 5	year, market, million, energy, waste, years, cent, industry, investment, government, financial, increase
Web: 6	information, site, web, website, page, online, search, email, click, internet, details, links, free, find, sites
Science: 7	data, research, system, analysis, model, results, number, time, science, methods, surface, cell, energy, test
*Cleaning: 8	2006, 2005, posted, 2004, june, july, october, march, april, september, 2003, august, january, november, post
Politics1: 9	government, world, people, international, war, party, countries, political, european, country, labour, british
Travel: 10	hotel, room, day, area, house, accommodation, holiday, visit, city, centre, facilities, town, great, tour
Health: 11	health, patients, treatment, care, medical, hospital, clinical, disease, cancer, patient, nhs, risk, drug
Councils: 12	development, local, community, council, project, services, public, national, planning, work, government
Life1: 13	people, time, questions, work, make, important, question, problem, change, good, problems, understand
Software: 14	software, system, file, computer, data, user, windows, digital, set, files, server, users, pc, video, mobile
Sports: 15	game, club, team, games, play, race, players, time, season, back, football, win, world, poker, sports, sport
Religion: 16	god, life, church, people, lord, world, man, jesus, christian, time, love, day, great, death, faith, men, christ
Arts: 17	book, art, history, published, work, collection, world, library, author, london, museum, review, gallery
Law: 18	law, act, legal, court, information, case, made, public, order, safety, section, rights, regulations, authority
Nature: 19	food, water, species, fish, plants, garden, plant, animals, animal, birds, small, dogs, dog, tree, red, wildlife
History: 20	years, century, house, st, john, royal, family, early, war, time, built, church, building, william, great, history
Engineering: 21	range, design, light, front, high, car, made, water, power, colour, quality, designed, price, equipment, top
Politics2: 22	members, meeting, mr, committee, conference, year, group, event, scottish, council, member, association
Life2: 23	time, back, good, people, day, things, make, bit, thing, big, lot, can, long, night, feel, thought, great, find
School: 24	people, children, school, support, young, work, schools, child, community, education, parents, local, care

Table 4: Domain Adaptation: examples of documents generated in different genres from the same keywords on the topic of Entertainment (topic 0 in Table 3). Keywords (shown in *italic*) are taken from an “on-topic” document in a test set, but its genre label (product review in this case) is not used since the task is to predict it. Instead, synthetic documents are generated from those keywords in each genre. Those documents augment the training set. The generated documents respect the constraints of topics, keywords and genres.

On-topic document: As a serious *form of music* outside of Jamaica, on par with American and British rock and roll and R&B, Before this breakthrough Reggae was seen as something of a joke even though Jimmy Cliff’s “The Harder They Come” *set* the stage “Catch a *Fire*” clinched it. The *album* is a solid classic and a masterpiece with lead vocals not just from Marley but band-mate Peter Tosh as well on the tracks *years* and *stop* that train the deluxe edition features both the *original* “Catch a *Fire*” *album* released only in Jamaica in December of with its two bonus tracks as well as the *original album* released internationally to the *public*.

keywords: *form music set fire album years original fire album original album public date album album fire*

Generated texts for the topic of Entertainment with these keywords:

A Wikipedia article: He wrote poetry in a Broadway, and was a semi - finalist. As a performer, Eve was renowned for his strong wit and strong vocals, as well as his striking delivery of heavy metal, a set of wits and his physical presence on stage, earned him the title “Master Of Song”. Eve was also the master of wits’ composition as well as some of his later works. He opened *fire* to wits in. Eve appeared in two of his first two albums as a performer and then mixed wits and wits, gaining his first *album*, a release from his old band. Eve was the sound of a song for eve, which remained unfinished.

A StackExchange discussion: Do you have a video on repeat since the original hits them on the radio every evening for the first two pieces sold separately by the label? Does anyone have an official release *date* or should i be? Thank you in advance. On this *date* on your public air *date* the entire *album* is sold separately, which is released through all tv series. the original release *date* is released on the public register for all tv series and can be purchased on every *album*.

A product review: He was the king of rock and roll, legendary guitarist Peter Tosh, and drummer Neal Schon. Of course, if you really like your *music*, then you should buy this *album*. from this CD, you’ll see why. from the infectious, surrealistic guitar riffs that set the stage, at the beach or in the hall of the ancient mariner, to the funky “Catch a *Fire*”, to the insanely catchy catchy summer pop staple.

A Hyperpartisan article: I’m an atheist myself, and I’m just a British freak! maybe if it was only about pop culture then it would be rude to you to leave the UK, and do something about it. I don’t feel like am American player doing the same with a *Fire-fox*. You can add to that one if you’re serious about it. I don’t hold an open exchange about anything. I was an apologist to the *album* you were talking about.

Table 5: F1 score results for fine-tuned Roberta Large classifier reported by separate topics. The “on-topic” columns show the performance when training and testing on in-domain documents. The “off-topic” columns present training on the off-topic documents and testing “on-topic”. Our augmentation results are in the last column for each N.

Topics:	N=30			N=100			N=1000		
	on-topic	off-topic	aug adapt	on-topic	off-topic	aug adapt	on-topic	off-topic	aug adapt
Finances: 0	73.9	52.0	62.0	94.2	60.2	65.2	95.6	84.3	83.1
Entertain: 1	76.8	51.1	61.2	96.0	65.0	63.0	97.1	79.0	79.5
Geography: 2	84.2	55.5	73.9	95.5	65.1	64.6	97.3	95.3	93.1
Business: 3	81.9	49.1	52.3	93.8	56.7	67.6	97.2	93.8	88.5
University: 4	80.5	58.5	61.3	89.2	67.7	71.2	97.5	81.0	83.2
Markets: 5	78.9	46.6	47.3	87.9	50.0	53.7	89.0	66.8	70.8
Web: 6	74.6	42.7	56.5	92.2	57.4	64.5	97.2	94.5	93.1
Science: 7	78.1	56.6	54.7	86.6	61.5	66.0	92.3	82.3	84.0
*Cleaning: 8	76.8	63.3	73.0	84.5	66.2	72.2	93.0	73.7	73.2
Politics1: 9	66.7	39.7	58.3	85.1	51.0	56.9	92.4	59.4	63.5
Travel: 10	89.5	65.5	71.0	93.1	63.4	71.7	96.6	69.1	81.8
Health: 11	72.7	50.1	57.3	88.0	57.2	61.1	85.0	69.2	67.3
Councils: 12	82.0	49.6	53.4	94.2	58.9	67.0	95.5	87.2	87.5
Life1: 13	80.6	45.6	54.1	94.5	45.4	57.2	97.0	89.9	84.8
Software: 14	85.8	52.4	52.8	94.7	64.1	65.0	96.0	62.3	79.1
Sports: 15	88.4	49.0	75.5	93.4	70.2	64.6	94.3	65.9	79.3
Religion: 16	73.2	48.9	66.6	87.0	56.0	64.8	95.7	79.4	80.4
Arts: 17	72.1	39.8	65.1	89.6	72.0	76.3	94.5	66.6	69.1
Law: 18	65.5	51.3	47.0	81.2	51.0	56.0	93.0	75.2	83.2
Nature: 19	77.1	55.9	61.1	94.1	65.0	75.1	95.3	81.4	93.4
History: 20	69.2	49.6	51.3	88.7	64.7	62.2	93.9	86.6	84.3
Engineering: 21	86.6	55.0	47.5	96.8	66.3	64.7	97.7	86.2	93.0
Politics2: 22	71.2	49.6	55.7	85.8	62.2	63.1	94.9	66.4	67.0
Life2: 23	75.4	47.1	47.2	93.5	66.0	66.6	96.5	73.4	83.3
School: 24	73.2	59.1	60.2	94.8	64.6	78.3	97.2	83.8	85.5
Average	77.0	52.1	58.1	91.0	61.2	65.1	94.9	78.8	81.7

Table 6: Ablations: average performance for mixing original and synthetic documents. The statistical differences at the level of .05 from the best configuration within each N are marked with ⁺⁺.

Original	Augmented	Roberta Large	Bert Large
1000	0 (baseline)	78.8 ⁺⁺	68.8 ⁺⁺
1000	10	79.0 ⁺⁺	69.1 ⁺⁺
1000	100	80.1 ⁺⁺	70.8 ⁺⁺
1000	1000	81.7	71.9
1000	3000	81.9	72.0
1000	5000	81.5	71.8
0	1000	78.6 ⁺⁺	68.9 ⁺⁺
1000	1000 random	78.6 ⁺⁺	68.5 ⁺⁺
100	0 (baseline)	61.2 ⁺⁺	55.0 ⁺⁺
100	10	61.3 ⁺⁺	56.3 ⁺⁺
100	100	65.1	58.8
100	300	65.4	58.7
100	500	65.2	58.7
0	100	60.8 ⁺⁺	54.7 ⁺⁺
100	100 random	61.4 ⁺⁺	55.3 ⁺⁺
30	0 (baseline)	52.1 ⁺⁺	49.9 ⁺⁺
30	10	56.9 ⁺⁺	53.3 ⁺⁺
30	30	58.1	55.2
30	100	58.3	55.1
30	150	58.2	55.1
0	30	52.9 ⁺⁺	51.2 ⁺⁺
30	30 random	52.4 ⁺⁺	50.8 ⁺⁺

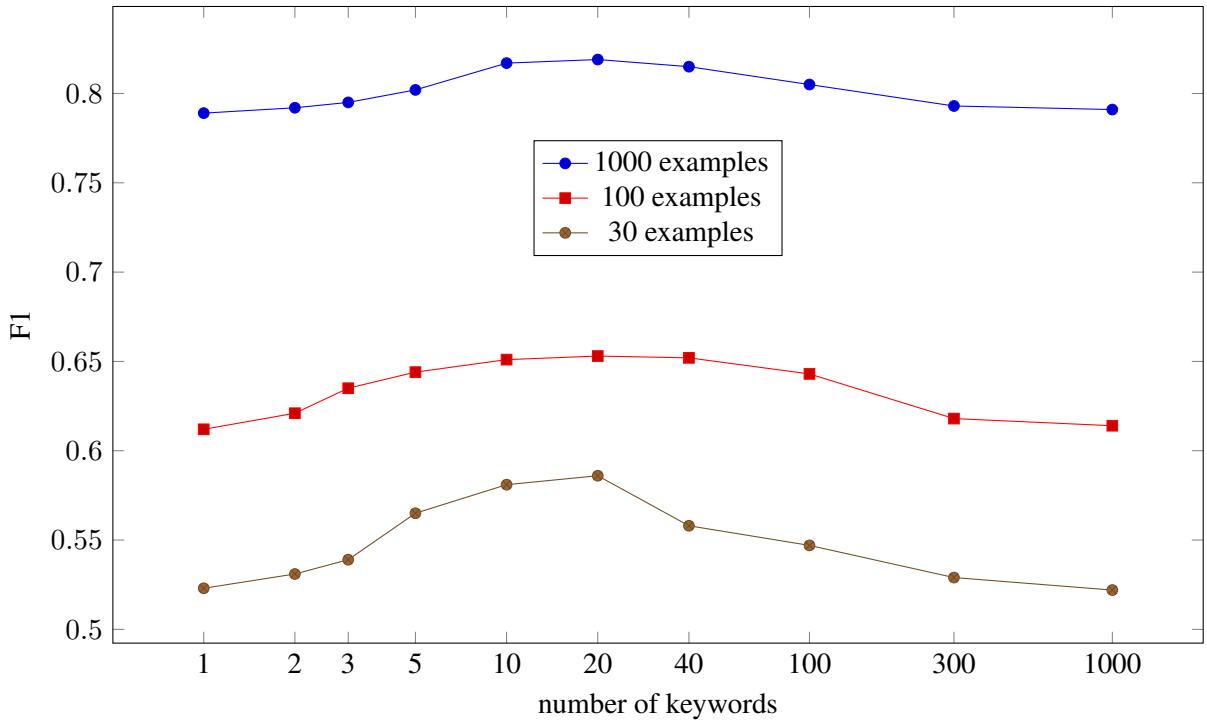


Figure 2: F1 metric of performance for various numbers of keywords and data sizes with Roberta Large classifier.

Table 7: Example of ChatGPT prompt used in our study. Class 1 is Wikipedia. Class 2 is Giga News. The topic is "Entertainment". Off-topic class examples.

Based on the examples of texts of Class 1 and texts of Class 2 below, list at least three criteria by which Class 1 and Class 2 texts are different from each other.

Here are some example texts of Class 1:

Example 1: World Darts Championship: He defeated number five seed Tony Eccles in the first round but lost to Shaun Greatbatch in round two. PDC career: Laursen became the first Dane to play in the PDC World Darts Championship. In the competition, he beat Colin Monk in the first round but lost to Dennis Priestley in the second round. Despite the fact that Laursen was up and missed eight darts to win the match before losing. He came through the Danish qualifying system for the second time for the PDC World Darts Championship but lost to Alan Tabern in the first round. Laursen has had some success in tournaments in his own country, reaching the final of the Danish Open (losing to Vincent van der Voort) and winning the Danish National Championships in 20. Laursen once again represented his country in the PDC World Darts Championship, having

...

Here are some example texts of Class 2: Liverpool, Manchester United, Arsenal, and West Ham in recent weeks, at least finished the half on a high. Blackburn captain Tim Sherwood just shot past the left-hand post in the 33rd minute after breaking through from a deep position and receiving an accurate pass from Jason Wilcox. After Asprilla shot over the bar and saw another effort pushed away by Flowers, Blackburn had another superb opportunity from Sherwood in the 38th minute. Wilcox again fed Sherwood, but his powerful shot could only find the crossbar via a deflection. Then Batty received a square pass from the right from substitute Keith Gillespie before firing home with a rare left-foot shot into the right-hand corner of Flowers' goal. Then, four minutes from time, Shearer fed Graham Fenton who charged into the area and volleyed first time past Hislop, who could only knock the ball high into the net. With a draw seemingly on the cards in the dying seconds, Fenton broke through once more to chip Flowers

...

Now, apply the first criteria to the following text:

Kuhn originated the role of Helen Bechdel in the off-Broadway Public Theater production of the musical *Fun Home*, which began its run September and opened officially on October. The run was extended multiple times and closed on January. She played the same role in the Broadway production, which ran from April to September at the Circle in the Square. Kuhn played the role of Golde in the Broadway revival of *Fiddler on the Roof*, starting on November. She plays Golde in the Menier Chocolate Factory London production of *Fiddler on the Roof*, which began on December and runs to March. Her television credits include *Law & Order* and *Law & Order: SVU*, *All My Children*, and two PBS shows: *My Favorite Broadway: The Leading Ladies*, recorded, released and in performance at the White House, and *A Tribute to Broadway*. The shows in March, Kuhn sang

Now, apply your second criteria to the same text.

Now, apply your third criteria to the same text.

Now, combine the criteria to decide which of those two classes the same text is more likely to belong.

Table 8: Examples of criteria generated by ChatGPT. Class 1 is Wikipedia. Class 2 is ICWSM collection.

Class 1 texts appear to be informational or factual, whereas Class 2 texts appear to be more conversational or personal in nature. Based on the examples provided, here are three criteria that differentiate Class 1 and Class 2 texts:

Language style: Class 1 texts tend to use more formal language and technical jargon, while Class 2 texts tend to use more informal language, colloquialisms, and slang. For example, Class 1 uses words like "position," "certified gold," and "innovation," whereas Class 2 uses phrases like "thrown down a little twist," "gone crazy," and "sooth your soul."

Content: Class 1 texts are often more focused on specific topics or subjects, while Class 2 texts tend to be more varied in content and topic. For example, the Class 1 texts provided are about musicians and their careers, while the Class 2 texts cover a range of topics, such as the Olympics, Chinese music, and personal relationships.

Tone: Class 1 texts are typically more objective and neutral in tone, while Class 2 texts tend to be more subjective and expressive. For example, the Class 1 texts provide information about musicians and their achievements without expressing personal opinions, while the Class 2 texts express personal feelings and opinions, such as excitement about the Olympics or admiration for Chinese music.

Table 9: Examples of ChatGPT applying a criteria generated previously to a test document from Wikipedia. The second class was ICWSM collection.

Criteria Generated:
 Language style: Class 1 texts tend to use more formal language and technical jargon, while Class 2 texts tend to use more informal language, colloquialisms, and slang. For example, Class 1 uses words like "position," "certified gold," and "innovation," whereas Class 2 uses phrases like "thrown down a little twist," "gone crazy," and "sooth your soul."

Test text:
 Orchestras of Ray Ventura and Jacques Hélian, Adison's band, often billed as Fred Adison and his Collegians, was one of the principal French backing groups for singers and films in the 1930s. Adison also recorded copiously and released many RPM commercial recordings during this time. After the onset of World War II, he toured with Django Reinhardt in September-October 1939 and continued writing music for film. He was imprisoned in a Nazi war camp in 1940. Following the war, he led a new big band. In 1956, Charles Spiessert, owner of Cirque Pinder, hired Fred Adison to conduct the circus orchestra, a position he held until 1964. He accompanied Luis Mariano and Gloria Lasso during this time and recorded several albums of circus music. Following this, Adison went into semi-retirement, occasionally playing with small ensembles and dance orchestras on television. Filmography: Chantal Brunschwig, Louis-Jean Calvet, Jean-Claude Klein, "Cent Ans De Chanson."

Generated output when the criteria applied:
 Based on the application of Criteria 1, which is language style, the text appears to belong to Class 1 - informational or factual, as it is written in a formal style and uses technical terms and proper nouns related to music and history.