Synthetic text augmentation for non-topical classification: a case of document genre

Anonymous ACL submission

Abstract

While the task of non topical text classification (e.g. document genre, author profile, sentiment, etc.) has been recently improved due to pre-trained language models (e.g. Bert), it has been observed that the resulting classifiers suffer from a performance gap when applied to new domains. E.g. a genre classifier trained on *political* topics often fails when tested on documents about *sport* or *medicine*. In this work, 1) We develop a robust method to quantify this phenomenon empirically. 2) We verify that domain transfer in non-topical classification remains challenging even for the modern pre-trained models, and 3) we test a data augmentation approach which involves training texts generators in any desired genre and on any 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 synthetics documents generated by our approach facilitates domain transfer, so that the model can correctly predict genres that don't have "on-topic" examples in the training set. The "off-topic" F1 score can be improved for some topics as much as from 57.6 to 73.0.

1 Introduction

Linguistic research often contrasts the properties of topic vs. those of style, which is also reflected in text classification research (Dewdney et al., 2001). However, this contrast is difficult to maintain, as the training sets in most corpora for style or genre prediction are dominated by topics specific to individual styles, so that transfer learning across corpora is limited in case of variation between their topics. For example, a model identifying FAQs can learn to pay attention to such words as *hurricane* and *tax advice* since these are common topics of FAQs in the training corpus (Sharoff et al., 2010). The effect of such contamination has been also shown empirically by considering topic influence for genre prediction over the New York Times corpus (Petrenz and Webber, 2010).

Up to our knowledge, this cross-influence of topics and styles has not being 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 non-topical style classifiers to new domains, for example, to study the performance degradation of a genre classifier trained on *political* topics when it is applied to texts on *sport* or *medicine*.

In this work, we claim the following original contributions:

- 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 metric 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 training corpus and on any **topic**, which is controlled by the keywords extracted by our original algorithm;
- We verify that augmenting the training dataset with synthetics tests generated by our approach **facilitates** domain transfer by improving F1 classification metric by 2-4 absolute percentage points in average and on some topics as much as from 57.6 to 73.0.

Non-topical text classification (including document genre) is an important practical language processing task. It aids in proper understanding, summarizing, archiving and retrieving documents in many different domains, including such important ones as legal and medical. Research has shown that people can easily recognize document genres from just a few examples even if those

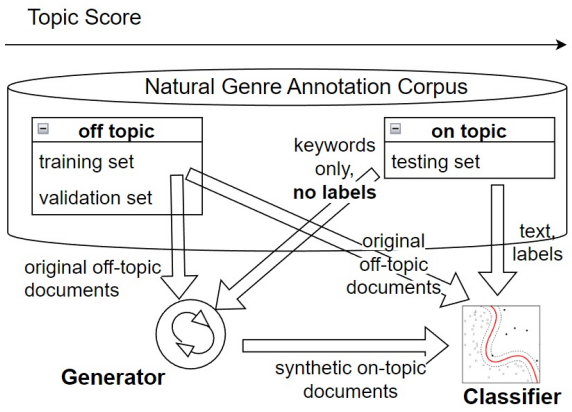

Figure 1: Experimental workflow

examples are from a different domain (Crowston et al., 2010). Thus, in order to create artificial general intelligence (AGI) at some point in future, we need to find ways to train computers to be able to perform that as well.

The tools and the experimental setup are available.[1]

2 Methodology

Since we are not aware of any standard solutions for genre and topical control during text generation and for assessing domain transfer in non-topical classification, we have developed several original solutions presented in this section. We separate the contributions of topics and genres by having two models, a topic model produced from a topically diverse corpus (even though it might be biased with respect to its genres), and a genre model which is a classifier based on a pre-trained language model (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 in the sections below.

2.1 Topic Model

For our experiments, we needed a topic model so that we can assess the performance gaps when transferring between the topics in our corpus. The topic model in this study was produced by a neural topic model (Dieng et al., 2020) which can achieve better interpretability in comparison to traditional 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 in (Dieng et al., 2020), the number of topics can be selected by maximising the product of topic coherence of a model (the average pointwise mutual information of the top words for a topic) by its topic diversity (the rate of unique words in the top words of all topics). In this way we arrived at choosing 25 topics for the ukWac corpus, see Table 1, Topic Coherence of this model is 0.195, Topic Diversity is 0.781. All of the topics are interpretable (the topic labels in Table 1 have been assigned by inspecting the keywords and a sample of documents). In the absence of a gold test set for an unsupervised method the most likely topics assigned to documents in the test set are reasonable. 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.

2.2 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 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 2. 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

[1] Anonymous at the submission stage

Table 1: 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

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 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 1. 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 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 1) from its ETM inference.

2.3 Assessing Domain Transfer

This subsection describes the methodology that we have developed to test the effect of a topic change on non-topical classification. 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- (development) and testing- sets on particular topics to experiment with a genre classification task, specifically knowledge transfer between the topics. The methodology that we involved relies on our topic model described above and proceeds as following. For each of our topics (e.g. "Science"), 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. the least "scientific" documents. 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. the most "scientific" documents. For each topic, we also created an *on-topic* test-set making sure it does not overlap with the training sets. Validation- (development) sets were off-topic since within a domain transfer setting no training on-topic data is available. Specifically, in experiments below, we used 300 documents of each

Table 2: 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)
PROMOtion	Adverts, promotional postings	10993	Promotional websites
INFOrmation	Encyclopedic articles	97575	A sample of Wikipedia
Review	Product reviews	1302495	Amazon reviews (Blitzer et al., 2007)
	Total	1687118	

genre in a test-set, 300 documents of each genre in a validation- (development) set, and varied the sizes of the training sets as stated in our empirical results section. 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 BERT-base (Devlin et al., 2018) from the Hugging-Face library with the default 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 epochs that showed the best score on the validation (development) set.

Often pre-trained transformer models make the right decisions for wrong reasons, for example, by detecting differences in formatting of the StackExchange questions in comparison to the format of hyperlinks in the Wikipedia entries. Given that in reality either FAQs or texts providing reference information can be formatted in other ways, performance on the natural genres corpus without preprocessing can be unrealistic.

As a comprise between the reliability of our results and the processing time, after preliminary investigation we settled on working with the window of 1000 characters randomly positioned within a document. Random positioning mitigates the impact of document structure (e.g. an introductory question positioned at the start of the StackExchange dataset). The windows obtained this way still provide sufficient text to determine the topic and genre when read by a human.

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

2.4 Keyword Extraction Algorithm

We had experimented with several variations of a heuristic algorithm to select the keywords and settled on the following approach after manually inspecting the quality of the generations and their topical relatedness. We are not much concerned how truthfully the keywords represent the content of the document, but rather how well they represent the topic to enable topic-focused generation. Thus, when deciding which words to extract as keywords, we promote those that are strong representatives of the document topic, which is quantitatively assessed by our topic model. It assigns each word (in the corpus) a score with respect to each topic between 0 and 1. The higher the score the stronger the word is related to the topic. Since some documents mix several topics, at times with numerically similar proportions, we accordingly weight the individual word scores with the overall topic scores in the document. Finally, we also want to adjust for repeated occurrences of the same word. Thus, our word scoring formula (within a document) simply iterates through all the topics and through all the word occurrences in the document while adding up the word scores with respect to the corresponding topic:

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

where i goes over all the occurrences of the word w in the document D, t goes over all topics (25 in the study here), $L(D, t)$ is the score of the document with respect to topic t and $L(w, t)$ is the score of the word w with respect to topic t.

We preserve only 10 top-scoring words in each document, so all the other words are discarded and the original sequence of the remaining words becomes the keyword sequence for the generator. Table 3 shows an example of extracted keywords along with how they are used to generate new synthetic documents, as detailed in the following subsection.

2.5 Augmenting by Generating

Our suggested method of improving domain transfer proceeds by augmenting the *off-topic* training set with automatically generated *on-topic* documents.

To achieve this we fine-tune a pre-trained language model into a separate generator for each genre listed in Table 2. Our earlier experimenting with using a single model for all genres and a special token to specify the desired genre resulted in weaker results. For this fine-tuning, we use exactly the same $N \cdot 6$ documents as are in our *off-topic* training set, thus operating in a practical scenario when on-topic documents are not available. Each generator is fine-tuned to take a sequence of keywords extracted according to the algorithm detailed above as input and to generate a document in the genre corresponding to this generator and of the topic defined by the keywords. During fine-tuning, the generators learn to associate the input keywords with the content of the output document, which becomes an important mechanism of topic control and facilitating the domain transfer.

We specifically used T5 as our generating model (Raffel et al., 2020). It is a unified text-to-text transformer, trained on the Colossal Common Crawl Corpus to predict the next word based on the preceding words in an auto-regressive way. We used the small version since we did not observe any advantage in using the Base or Large T5 model in our early experiments, so we kept the less computationally intensive model. Its input format requires a prefix to indicate which downstream task is being fine-tuned, so we used the word "generate." We trained each model for 16 epochs using Simple Transformers library[2] with a default learning rate of .001 and its Adam optimizer. For generating, we also use the following T5 hyper-parameters, specifically number of beams $= 1$, top $k = 50$, top $p = .95$. The selected hyper-parameters were chosen after preliminary experimentation by inspecting the produced quality of generations in terms of both topical and genre fit. Table 3 illustrates our domain transfer approach by examples of extracted keywords and synthetic documents generated from those keywords in different genres.

When using the generators for augmentation, we do provide them the *on-topic* keywords, but not the class (genre) labels, so that they generate the same number of synthetic documents for each label. Thus the use of on-topic test-set keywords for augmentation does not give any unfair advantage to the augmented model and is methodologically acceptable as a common practice of **inference-time** optimization.

One of our overall hyper-parameters is how many documents to generate. Our preliminary experimentation suggested that 1:1 was a near optimal ratio: the same number of original and synthetic documents. We include several other combinations in our empirical results below.

3 Experiments

3.1 Comparison Results

We assess the effect of domain mismatch and our approach to improving domain transfer by augmenting the training sets with synthetic on-topic documents. The difference between the accuracy obtained before and after generation demonstrates the efficiency of the augmentation model. Table 4 shows the comparison results for 3 different sizes of training data: 1000, 100 and 30 documents per genre accordingly. As we can see, the topic mismatch effect is *extremely significant*: the average drop from *on-topic* to *off-topic* training set is around 20% for N = 1000 and 30% for smaller Ns. The average *on-topic* F1 score for the largest size is 86.4%, which is only somewhat below 93% average human-level performance as independently estimated by the authors.The average *off-topic* performance for that size drops to 66.8%.

All three size configurations demonstrate statistically significant increases from non-augmented off-topic sets (baselines here) ranging from 2.5% on the smaller sizes and 4% on N = 1000. All the differences are statistically significant at the level of alpha 0.01 according to a pair-wise t-test. This empirically confirms with high confidence that our augmentation procedure is beneficial for non-topical text classification.

The most time-consuming part of our experi-

[2] https://simpletransformers.ai/

Table 3: Domain Transfer: examples of documents generated in different genres from the same keywords on the topic of Entertainment (topic 0 in Table 1). 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. You can see that the generated documents are topically similar, but are in fact in different 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 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.
Generated a **stack exchange 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*.
Generated 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.
Generated 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.

ments were fine-tuning the generators (T5) and the classifiers (Bert) with the cost of roughly 5000 hours of NVIDIA GeForce RTX 2080.

3.2 Ablation Studies

This subsection reports several *ablation* experiments that we conducted to additionally verify the effects reported above and to gain the insight into the phenomena studied. In order to verify that the genre labels in our synthetic texts were important we randomly shuffled them. This way, the augmented data became to act only as noise. Not surprisingly, the average scores dropped to the baseline levels which verified that using the proper model for each genre to generate the synthetic augmenting texts is important, and that the improvements reported above were not due to simply the change in the statistical properties of the training and validation sets or due to addition of noise.

We also looked at several ways of mixing the original and augmented data. Table 5 presents the average across topics scores for various sizes used. It can be observed that while some small improvements can be achieved by generating more documents, those gains are not statistically significant. On the other side, very small numbers of added documents indeed result in statistically detectable drops. Using only synthetic documents results in drops to the levels only slightly above or even below baselines. The last rows for each N in Table 5 show the results when augmenting documents were obtained using keywords from randomly selected off-topic documents, and thus not attempting any domain transfer. You can see that they are significantly worse than those with the transfer. The last column shows the results when T5-small was used as a classifier instead of Bert. While the overall classification accuracy is lower, T5 results follow the same pattern as with Bert and thus additionally support that our augmentation facilitates domain transfer.

4 Related studies

The split between topics and styles has been studied for transformer models, including disentangled representation (John et al., 2019) and other methods of topic-style decomposition (Romanov et al., 2019; Subramanian et al., 2019). However, our study focuses on the numerical estimates of the topic transfer gap on large samples diverse in topics and in genres. This makes our study similar to the suggested controls in object recognition for

Table 4: F1 score comparison for testing genre classification topic gaps and our transfer 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". Our augmentation results are in the last column for each N. All configurations show improvements statistically significant at the level 0.01.

	N=30			N=100			N=1000		
Topics:	on-topic	off-topic	augmented	on-topic	off-topic	augmented	on-topic	off-topic	augmented
Finances: 0	81.2	56.2	62.1	83.6	48.1	58.1	86.2	64.5	67.0
Entertain: 1	76.1	35.2	48.8	78.2	40.0	57.0	75.8	57.6	73.0
Geography: 2	86.3	57.9	61.0	88.5	62.2	59.6	92.3	82.5	83.8
Business: 3	79.1	41.1	44.9	85.8	46.0	54.3	87.0	76.3	72.1
University: 4	84.1	51.8	51.9	87.0	55.4	55.4	89.5	70.2	78.3
Markets: 5	71.1	44.0	42.7	77.8	42.5	44.5	81.0	55.2	59.0
Web: 6	79.8	44.1	50.3	83.3	43.9	51.6	91.0	81.0	80.5
Science: 7	77.8	47.1	46.6	82.9	50.3	50.5	83.5	68.3	72.4
*Cleaning: 8	73.3	62.4	62.6	80.9	65.3	62.2	86.2	69.1	70.0
Politics1: 9	73.9	33.2	39.4	78.9	37.6	39.2	83.4	36.3	42.7
Travel: 10	84.5	50.8	65.8	89.2	56.0	51.8	92.1	65.7	77.9
Health: 11	74.2	40.5	41.8	78.6	46.4	53.3	76.9	54.7	59.8
Councils: 12	81.0	47.6	44.2	86.6	51.2	47.8	89.7	71.4	75.4
Life1: 13	81.7	43.7	39.9	86.1	41.9	43.5	92.3	68.1	72.7
Software: 14	77.9	42.3	47.6	84.2	44.6	52.4	87.5	62.7	67.5
Sports: 15	85.9	57.7	63.6	85.4	51.2	59.0	87.0	70.5	64.7
Religion: 16	73.7	37.4	40.0	78.4	40.1	38.8	84.5	61.9	65.0
Arts: 17	79.2	55.2	45.4	80.1	51.2	61.1	82.8	63.5	73.3
Law: 18	75.0	45.1	45.4	77.5	46.2	48.5	73.7	83.4	70.1
Nature: 19	77.0	50.2	53.3	81.2	50.2	59.4	85.1	76.9	80.0
History: 20	75.5	46.8	49.7	79.6	50.9	53.4	84.4	68.3	74.0
Engineering: 21	87.4	49.7	52.5	90.0	54.0	53.9	89.9	74.7	80.4
Politics2: 22	75.4	47.7	50.5	83.3	52.3	57.8	86.9	52.6	58.9
Life2: 23	81.5	45.8	38.2	85.2	47.0	46.7	92.1	55.9	66.2
School: 24	80.0	51.4	52.8	85.7	53.6	53.7	90.1	70.2	79.2
Average	78.9	47.4	49.9	83.2	49.2	51.9	86.4	66.8	70.9

Table 5: 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	Bert F1	T5 F1
1000	0 (baseline)	66.8^{++}	54.4^{++}
1000	10	67.2^{++}	55.3^{++}
1000	100	68.1^{++}	56.1^{++}
1000	1000	70.9	60.5
1000	3000	**71.1**	**60.6**
1000	5000	71.0	60.2
0	1000	67.1^{++}	54.5^{++}
1000	1000 random	66.6 $^{++}$	54.1^{++}
100	0 (baseline)	49.2^{++}	40.0^{++}
100	10	49.5^{++}	40.7^{++}
100	100	51.9	**44.3**
100	300	**52.1**	44.1
100	500	52.0	44.1
0	100	48.3^{++}	40.9^{++}
100	100 random	49.5 $^{++}$	40.5^{++}
30	0 (baseline)	47.4^{++}	38.6^{++}
30	10	49.1^{++}	38.8^{++}
30	30	49.9	**42.6**
30	100	**50.4**	42.1
30	150	50.1	42.3
0	30	48.4 $^{++}$	37.5^{++}
30	30 random	47.9 $^{++}$	38.6^{++}

generalizable, robust, and more human-like computer vision (Barbu et al., 2019).

A related research area concerns the use of causal models for interpreting the biases of neural predictions, for example, with respect to gender (Vig et al., 2020). There have been studies to investigate biases in neural models via adding counter-factuals (Hall Maudslay et al., 2019; Kaushik et al., 2020). Our focus in this study is different: we want to investigate the possibility of correcting biases by generation of appropriate texts.

Both traditional feature-based and neural approaches in domain transfer assume a semi-supervised procedure by inferring a shared representation space which takes into account both labeled “out-of-domain” data and unlabeled “in-domain” data (Daumé III et al., 2010; Bengio, 2012). This has been largely superseded through the use of pre-training for transformer models.

For an overview of the works on a closely related task of text style transfer (TST) we refer the reader to Jin et al. (2022). Unlike TST, we are not specifically concerned with preserving the content as long as the generated documents aid

in domain transfer. Also, the reviewed works did not involve pre-trained language models. Additional in-domain pre-training was suggested for the approaches based on cloze-style patterns for a number of few-shot downstream tasks (Schick and Schütze, 2021), but genre classification task considered here does not suggest any obvious prompt patterns to use. A review of recent works on generating prompts for the pre-trained language models can be found in Liu et al. (2021). Several ways to control text generation including its style have been suggested (Keskar et al., 2019) but they required pre-training a custom language model from scratch rather than fine-tuning an existing model as we do here. Some earlier works looked at topical control during text generation, e.g. Hu et al. (2017), but they did not use pre-trained language models. The challenges maintaining coherent style and topic within longer texts (exceeding 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 that much concerned with the output quality but rather their help in domain adaptation. Also, we perform our experiments on the text windows of the sizes easily fitting transformers' limitations. Recently demonstrated ability of GPT line of models (Brown et al., 2020) to generate text often indistinguishable from human has been tried for various applications (Floridi and Chiriatti, 2020). GPT-based and other text generators have been successfully used for anonymization of data to address privacy concerns (Guan et al., 2018). Augmenting training sets with synthetic documents has been also proposed for the tasks of classifying flight reservation requests, open-domain question answering and customer support in a few-shot scenario (Anaby-Tavor et al., 2020) but they did not involve any topical control.

5 Conclusions, Limitations and Future Work

We have demonstrated the impact of document topics on non-topical text classification performed with the help of pre-trained language models. We have also shown how we can mitigate this impact by means of proper selection of keywords and fine-tuning a pre-trained language model for generation. This allowed us to augment training data for non-topical document classification (specifically *document genre*) to reduce the loss in performance during topical domain transfer. As a result, a system can be trained on the documents in one topic (e.g. *politics*) and applied to another (e.g. *healthcare*) even when there are no healthcare-related documents in the training corpus that represent all possible class labels (genres). In order to assess the impact of domain switch on classification accuracy and our suggested way of alleviating it, we have developed an original methodology based on a topic model. We have also created a large corpus with "natural genre annotation" that can be used in follow-up studies.

Still, our study has certain limitations. Specifically, additional pre-trained language models can be tried in future (including the largest ones like GPT-3), for both generation and classification, and more formal comparison between various keyword-selection algorithms and their hyperparameters can be performed. Larger training sets can be explored, as well smaller ones in a "few-shot" setting.

At the same time, the degree of improvements from augmentation is not uniform. For some topics we obtain much better results than for others, while occasionally the performance on the augmented set is lower than on the original off-topic training set. More research is needed to investigate the conditions under which this happens in comparison to more successful examples of transfer. A number of approaches improving the quality of generated text, e.g. those based on Generative Adversarial Networks (Goodfellow et al., 2020) or meta learning (Lee et al., 2022) can be explored, as well as various methods for controlled generation.

References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Yoshua Bengio. 2012. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27 of *Proceedings of Machine Learning Research*, pages 17–36, Bellevue, Washington, USA. PMLR.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proc ACL*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

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. *Advances in neural information processing systems*, 33:1877–1901.

Christopher Cieri and Mark Liberman. 2002. Language resources creation and distribution at the Linguistic Data Consortium. In *Proc LREC*, pages 1327–1333. Las Palmas, Spain.

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, *Genres on the Web: Computational Models and Empirical Studies*. Springer.

Hal Daumé III, Abhishek Kumar, and Avishek Saha. 2010. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 53–59, Uppsala, Sweden. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nigel Dewdney, Carol VanEss-Dykema, and Richard MacMillan. 2001. The form is the substance: classification of genres in text. In *Proc. Human Language Technology and Knowledge Management*, pages 1–8.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Andrew Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Proceedings of International Conference on Weblogs and Social Media*, San Jose, CA.

Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. 2018. Generation of synthetic electronic medical record text. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 374–380. IEEE.

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 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China. Association for Computational Linguistics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

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 *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Hung-yi Lee, Shang-Wen Li, and Thang Vu. 2022. Meta learning for natural language processing: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 666–684, Seattle, United States. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Philipp Petrenz and Bonnie Webber. 2010. Stable classification of text genres. *Computational Linguistics*, 34(4):285–293.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Alexey Romanov, Anna Rumshisky, Anna Rogers, and David Donahue. 2019. Adversarial decomposition of text representation. In *Proc NAACL*, pages 815–825, Minneapolis, Minnesota. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Serge Sharoff. 2018. Functional text dimensions for the annotation of Web corpora. *Corpora*, 13(1):65–95.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The Web library of Babel: evaluating genre collections. In *Proc Seventh Language Resources and Evaluation Conference, LREC*, Malta.

Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text style transfer. *arXiv preprint arXiv:1811.00552*.

Bowen Tan, Zichao Yang, Maruan AI-Shedivat, Eric P Xing, and Zhiting Hu. 2020. Progressive generation of long text. *arXiv preprint arXiv:2006.15720*.

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. *arXiv preprint arXiv:2004.12265*.