Cross-Domain Learning for Classifying Propaganda in Online Contents

Liqiang Wang\textsuperscript{1,2}, Xiaoyu Shen\textsuperscript{1}, Gerard de Melo\textsuperscript{3}, Gerhard Weikum\textsuperscript{1}
Max Planck Institute for Informatic, Saarbrücken, Germany\textsuperscript{1}
Shandong University, Jinan, China\textsuperscript{2}
Hasso Plattner Institute, University of Potsdam, Potsdam, Germany\textsuperscript{3}
\{lwang, xshen, weikum\}@mpi-inf.mpg.de, gdm@demelo.org

Abstract

As news and social media exhibit an increasing amount of manipulative polarized content, detecting such propaganda has received attention as a new task for content analysis. Prior work has focused on supervised learning with training data from the same domain. However, as propaganda can be subtle and keeps evolving, manual identification and proper labeling are very demanding. As a consequence, training data is a major bottleneck.

In this paper, we tackle this bottleneck and present an approach to leverage cross-domain learning, based on labeled documents and sentences from news and tweets, as well as political speeches with a clear difference in their degrees of being propagandistic. We devise informative features and build various classifiers for propaganda labeling, using cross-domain learning. Our experiments demonstrate the usefulness of this approach, and identify difficulties and limitations in various configurations of sources and targets for the transfer step. We further analyze the influence of various features, and characterize salient indicators of propaganda.

1 Introduction

1.1 Motivation and Problem

Propaganda can be loosely defined as “misleading information that is spread deliberately to deceive and manipulate its recipients” (see, e.g., Jowett and O’Donnell, 2018 and www.britannica.com/topic/propaganda). Various factors of propaganda have been studied in the humanities, including emotionality of language, biased selection of information and deviation from facts, manipulation of cognition, and more (Ellul and Kellen, 1973; Silverstein, 1987; Jowett and O’Donnell, 2018). However, there is no consensus on the decisive factors that tell whether a given article or speech is propagandistic or not.

In the modern digital world, the influence of propaganda on society has drastically increased. Hence, there is also a major increase in computer science, computational linguistics and computational sociology research on analyzing, characterizing and, ultimately, automatically detecting propaganda (Da San Martino et al., 2020).

To a first degree, one may think of propaganda as a variation of fake news, and some works investigate propaganda as a refined type of disinformation (see, e.g., Rashkin et al., 2017; Wang et al., 2019; Shu et al., 2017). While false claims can be an element of propaganda, we think that fake news is merely the tip of the iceberg, and that the persuasive and manipulative nature of propagandistic contents requires deeper approaches. Classifiers for propaganda detection need to better capture how propaganda is expressed in subtle ways by language style and rhetoric or even demagogic wording. This holds for news as well as social media posts and speeches. In all these cases, correct information may be presented in incomplete form or placed in distorted contexts, along with manipulative phrases, in order to mislead the audience.

Prior work has mostly looked into news articles (e.g., Saleh et al., 2019; Barrón-Cedeño et al., 2019; Da San Martino et al., 2019) and tweets, and has typically focused on strongly polarized topics like the 2016 US election and the related Russian Internet Research Agency (IRA) affair, the UK Brexit discussion, or political extremism. All these approaches consider propaganda detection as a classification task assuming sufficient amounts of labeled in-domain training data. For example, in the “Hack the News” datathon chal-
a large number of news articles (Barrón-Cedeño et al., 2019) and sentences (Da San Martino et al., 2019) from such articles were annotated by distant supervision and human judgment, respectively, to train a variety of machine learning methods. The resulting F1 scores on the leaderboard of this benchmark are amazingly high, around 90%. This may give the impression that propaganda detection is a solved problem. However, most of the positively labeled samples are simple cases of “loaded language” with strong linguistic cues independent of the topic. Moreover, the learned classifiers benefit from ample training data, which is all but self-guaranteed in general.

In this paper, we question these prior assumptions, hypothesizing that propagandistic sources and speakers are sophisticated and creative and will find new forms of deception evading the trained classifiers. The overall approach is still text classification; the novelty of our approach lies in cross-domain learning, where domains denote different kinds of sources, such as news articles vs. social media posts vs. public speeches. We acknowledge that there is often a shortage of perfectly fitting labeled data, and instead tap into alternative sources that require a transfer step. Specifically, we consider speeches and tweets, in addition to news articles, at both article and sentence levels.

1.2 Approach and Contribution

Our goal is to build more general propaganda detectors, which can leverage different kinds of data sources. In particular, we tap on political speeches of notorious propagandists, such as Joseph Goebbels (the Nazi’s Reich Minister of Propaganda). As it is very difficult (and often subjective) to label speeches and their sentences in a binary manner, we pursue a pairwise ordinal approach where training data merely ranks samples of a strongly propagandistic speaker against those of a relatively temperate speaker. We investigate to what extent models learned from such data can be transferred to classifying news and tweets, and we also study the inverse direction of learning from news and tweets to cope with speeches.

Figure 1 illustrates our framework towards generalizable propaganda detection that overcomes the bottleneck of directly applicable training labels and instead leverages cross-domain learning.

Figure 1: Framework for propaganda detection and analysis

The salient contributions of this paper are as follows.

1. We introduce a framework for cross-domain learning of propaganda classification. This comprises data collection, feature selection, training procedures, different learning methods, and analysis.

2. We devise a pairwise ranking LSTM model, called LSTMR, that uses a judiciously designed loss function to enhance cross-domain performance.

3. We present experiments and analyses that provide insights into the advantages as well as limitations of cross-domain learning for propaganda classification.

4. All our datasets and code have been made publicly available to support further research on understanding and countering advanced forms of propaganda.2

2 Related Work

Propaganda Detection. We summarize the state-of-the-art according to the type of underlying data sources: the news, tweets or data to be fact-checked. Saleh et al. (2019) detect propagandistic content using a variety of well-engineered features. The study is based on a dataset of news articles annotated for the SemEval 2019 task on “Hyperpartisan News Detection”. Barrón-Cedeño et al. (2019) generated training and test data by collecting news from blacklisted sites as propaganda, while considering the Gigaword News corpus as trusted news. Further research related to

https://www.datasciencesociety.net/hack-news-datathon/
propaganda detection on news articles includes work on multi-class labeling (Da San Martino et al., 2019) and on the related topic of hyperpartisanship (Potthast et al., 2018).

For Twitter as a data source, several datasets are widely used, most notably the tweets on the Russia-based IRA (Miller, 2019; Farkas and Bastos, 2018; Badawy et al., 2019), contents spread by trolls or bots (Caldarelli et al., 2019; Williamson III and Scrofani, 2019), and extremist tweets (Nizzoli et al., 2019; Johnston and Weiss, 2017). Most of these studies rely on classifiers developed for bot detection to filter the propagandistic contents. Deep learning methods are also used to detect propaganda (Johnston and Weiss, 2017; Nizzoli et al., 2019). As more emphasis is on the analytics, the above detection techniques are still in the formative stage, by applying basic approaches like logistic regression, support vector machines, LSTMs, etc. These methods motivate us with regard to our feature engineering.

As mentioned above, some works take propaganda as a fine-grained type of fake news (Rashkin et al., 2017). Some approaches (Rashkin et al., 2017; Wang et al., 2019; Shu et al., 2017, 2019) considered fact-checked statements from sources such as PolitiFact, Snopes, BuzzFeeD, etc. This data encompasses trusted content, satire, hoaxes, as well as propaganda, but the approaches do not develop propaganda-specific techniques. In our view, sophisticated forms of propaganda are quite different from hoaxes or plainly wrong statements, and call for custom-tailored approaches.

**Propaganda Analysis.** Analyses of propagandistic content have mostly looked into quantifying its influence and spreading across networks. Timothy (2017) studied how propaganda influences public opinion by means of simulations. Some studies (Caldarelli et al., 2019; Williamson III and Scrofani, 2019) analyzed the role of bots in spreading propagandistic posts. Gorrell et al. (2019) examined strategies of troll accounts. Farkas and Bastos (2018) reported findings on propaganda in IRA tweets. The observations suggest that techniques are customized to the targeted political agenda. As for feature analysis, the work of Barrón-Cedeño et al. (2019) and Potthast et al. (2018) investigated the effectiveness of linguistic and stylistic features. Bisgin et al. (2019) analyzed extremism propaganda contents in terms of entities, topics and targets.

Fake news analysis techniques (Resende et al., 2019; Volkova and Jang, 2018; Shen et al., 2017; Wang et al., 2019; Zhang et al., 2018) can as well serve as inspiration to better understand propaganda. In particular, Budak (2019) studied the prevalence and focus of fake news based on related tweets, news and interviews. Yang et al. (2019) used visualization techniques for analyzing fake news to make the detector more explainable.

3 Datasets

For our study, we compile five datasets from three domains with the aim of exploring cross-domain characteristics and performance of propaganda detection.

**Speeches.** We collected transcripts of speeches from four politicians, organized as ordered pairs. Trump\(^3\) and Obama\(^4\) are considered as contemporary speakers, largely talking about the same or related topics. We consider the former as more propagandistic than the latter. We use Joseph Goebbels (the Nazis’ Minister of Propaganda, human-translated to English by the data provider)\(^5\) and Winston Churchill (the Prime Minister of the UK\(^6\) as prominent figures from the World War II era, with the former being more propagandistic than the latter. We realize that all four of these politicians have given some propagandistic speeches. Our assumption is that, collectively and relatively, two of the speakers exhibit substantially less propaganda than the other two. The data is organized at two different levels of granularity: articles (SPE\(_A\)) and sentences (SPE\(_S\)).

**News.** The news dataset comes from the Hack the News Datathon\(^1\). We combined and reorganized different datasets (Barrón-Cedeño et al., 2019; Da San Martino et al., 2019) from this source, to construct an article-level news corpus (NEW\(_A\)) and a sentence-level corpus (NEW\(_S\)), both comprehensively annotated with binary labels: propagandistic or normal. Note that the articles for the sentence-level corpus are completely disjoint from the ones in the article-level corpus. So learning on one and testing on the other entails a challenging cross-domain transfer as well.

---
\(^1\)https://factba.se
\(^2\)http://obamaspeeches.com
\(^3\)https://research.calvin.edu/german-propaganda-archive/goebmain.htm
\(^4\)https://winstonchurchill.org/resources/speeches/
Tweets. We combine two pre-existing collections of tweets to construct this TWE dataset. We consider the Twitter IRA corpus (Edgett, 2017) with time period of 2016 as propagandistic, and the “twitter7” data from SNAP (Yang and Leskovec, 2011) in 2009 as regular. As “twitter7” encompasses around 476 million tweets, we undersampled 8,963 tweets to maintain a balance with the IRA data. The tweets are randomly sampled from the June 2009 collection within the dataset. As with speeches, the data is cast into ordered pairs rather than using absolute labels as ground truth.

As some of the data initially comes with a strong label skew (with way more negative than positive samples), we apply under-sampling to construct corpora with balanced positive and negative samples. Table 1 summarizes our five datasets.

Table 1: Dataset sizes. Subscripts A and S indicate article and sentence granularity, respectively.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Speeches</th>
<th>News</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPE_A</td>
<td>288</td>
<td>7,798</td>
<td>17,926</td>
</tr>
<tr>
<td>SPE_S</td>
<td>24,934</td>
<td>7,876</td>
<td></td>
</tr>
</tbody>
</table>

4 Methods

In this section, we first introduce commonly used methods for propaganda detection, and subsequently introduce our proposed pairwise ranking model that aims at enhancing the effectiveness in the cross-domain setting.

We investigate three widely used methods as propaganda classifiers: logistic regression (LR) (Hosmer Jr et al., 2013), support vector machines (SVM) (Cortes and Vapnik, 1995) and bidirectional long short-term memory (LSTM) (Graves and Schmidhuber, 2005) neural networks. In addition, we devise an enhanced form of LSTM-based network that incorporates pairwise ranking information and is equipped with a specifically designed loss function. We denote this method as LSTMR, and present it in the Subsection 4.3.

4.1 Feature-based Models

LR and SVM are feature-driven learners. We consider various informative features on the language characteristics of articles and sentences:

- word counts (WC),
- word level TF-IDF scores (TI-W),
- N-gram level TF-IDF scores for $N=2$ or 3 (TI-G),
- occurrence statistics of word categories, such as first-person pronouns or negative-emotion words, from the widely used Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2015),
- the combination of all the features (ALL).

4.2 LSTM Baseline Model

Unlike LR and SVM, the LSTM-based methods do not rely on feature modeling. Instead they can use pre-trained word embeddings to capture the text structure and semantic features. We adopt a basic LSTM classifier as a baseline. The structure of the LSTM baseline model is illustrated in Figure 2. On top of the bi-LSTM layers, the networks have a dense layer and a final sigmoid function to learn scores and yield classifier labels. The cross-entropy loss function is applied for the LSTM classifier.

4.3 Pairwise Ranking Model

Model Architecture. Figure 2 depicts our pairwise propaganda ranking model LSTMR. The motivation for this design is that training labels may be subjective and noisy, especially but not only in a cross-domain setting where training and test data do not come from the same distribution. Traditional supervised learning models strictly distinguish the samples according to their labels, which leads to models over-fitting the training set, while their cross-domain generalizability diminishes. Supervised learning typically interprets the labels of data samples as a strict ground truth. In certain settings, this can lead to over-fitting to the training data distribution while hampering the ability for cross-domain learning. The LSTMR model aims to mitigate this over-fitting and enhance cross-domain applicability by relaxing constraints from strict labeling to rankings. Unavoidably, the in-domain performance may decrease. As Figure 2 shows, the model operates in two phases:

1. The first phase is to train the model based on the ranked pairs of data samples, where one is considered more propagandistic than the other. The bi-LSTM and dense layers serve to learn numeric scores for each of the two data points in each pair. Then the two scores are
constrained by a rule that the more propagandistic sample is supposed to obtain a higher score than the other sample that it is paired up with. We can think of the former as a positive sample and the latter as a negative sample, in terms of a binary classifier. This stage results in an improved scoring function, with awareness of the constraints.

2. The second phase addresses the scoring of previously unseen data points with the trained LSTMR model receiving a single input text instead of a pair. The label is then determined according to the constraints-based scoring function.

Pairwise Constraints. The constraints are instantiated using the following loss functions, where \( y \) is the ground-truth label (0 and 1 corresponding to the less propagandistic and the more propagandistic sample, respectively), \( \hat{y} \) is the predicted score, and \( \mathbb{I}(\cdot) \) is the indicator function with output 1(constraints are satisfied) or 0:

- Logistic loss (LOG) (Pasumarthi et al., 2019):
  \[
  \ell(y, \hat{y}) = \sum_{j=1}^{n} \sum_{k=1}^{n} \mathbb{I}(y_j > y_k) \cdot \log(1 + \exp(\hat{y}_k - \hat{y}_j))
  \]

- Linear loss (LIN) (Hanselowski et al., 2018):
  \[
  \ell(y, \hat{y}) = \sum_{j=1}^{n} \sum_{k=1}^{n} \mathbb{I}(y_j > y_k) \cdot (1 + \hat{y}_k - \hat{y}_j)
  \]

- Threshold loss (THR), where \( \theta \) is the threshold to control the intensity of the constraints:
  \[
  \ell(y, \hat{y}) = \sum_{j=1}^{n} \sum_{k=1}^{n} \mathbb{I}(y_j > y_k) \cdot \mathbb{I}\left(\frac{1 + \hat{y}_k - \hat{y}_j}{2} > \theta\right) \cdot \frac{1 + \hat{y}_k - \hat{y}_j}{2}
  \]

- Counting loss (COU) to count the discordant pairs:
  \[
  \ell(y, \hat{y}) = \sum_{j=1}^{n} \sum_{k=1}^{n} \mathbb{I}(y_j > y_k) \cdot \mathbb{I}(\hat{y}_k > \hat{y}_j)
  \]

Among the above loss functions, the threshold loss is designed to further relax the constraints by explicitly disregarding those pairs with confidence below the threshold. We compare the model performance using different loss functions in Section 6. In the subsequent evaluation experiments we adopt THR loss function for LSTMR by reason of its better performance.

Sampling Method. We apply different methods for constructing and sampling the ranked pairs. We consider data from all three source types: news, tweets and speeches. Whenever we have 0/1-labeled data points, we randomly combine a positive sample (1) and a negative sample (0) into an ordered pair for training the LSTMR model. This is the case for news and tweets, as these datasets have been manually annotated. When we have merely relatively ordered data points where some are more propagandistic than others, we sample such pairs. This is the case for the speeches, where we assume that Goebbels is more propagandistic than Churchill and the same ranking holds for Trump versus Obama.

For news and tweets, we need to cope with skewed label distributions: way more negative than positive samples. As we consider cross-domain learning where the eventual test set has an a-priori unknown distribution that could be fairly
different from the prior label distribution at training time, we generally re-balance all datasets. We consider two strategies to this end. The first is to under-sample the entire dataset, and the second is to over-sample the smaller class. Note that the over-sampling is per data point, so we still construct new pairs of ranked samples (i.e., there are no duplicate pairs).

5 Experimental Evaluation

Setup. The feature dimensions of WC, TI-W, TI-G are truncated to their respective top 5,000, and we use LIWC 2015 (Pennebaker et al., 2015) with 76 categories including first person singular, negation, sexual, swear, etc. Optionally, we eliminate proper nouns (PN), as tagged using the NLTK toolkit (Bird et al., 2009). This is to reduce the influence of particular names and their respective topics (e.g., “Iran”). The rationale is that texts could be about the very same topic but differ in their degree of propaganda.

For LSTM and LSTMR, we use the pre-trained 100-dimensional embeddings from GloVe (Pennington et al., 2014), 128 bi-LSTM cells and 200×50×1 dense layers. The embeddings are frozen during training. For LSTMR, the THR loss with a threshold of 0.4 and complete sampling (neither under- nor over-sampling) is adopted. Note that LSTMR is a ranking model rather than a classifier. As the ranking scores are comparable and range from 0 to 1, we empirically consider ≥ 0.5 as a threshold to classify test samples as positive or negative.

For in-domain learning, where training and test data are from the same source, we adopt 5-fold cross validation. Hyper-parameters are tuned by this setting, and then kept fixed for cross-domain experiments. Note that in the cross-domain case, cross-validation is impossible as the target-domain data is treated completely unseen. We report on precision (P), recall (R) and F1 (F) scores, with propaganda being the positive class.

In-Domain Performance. The results are shown in Table 2. Overall, we achieve acceptable in-domain performance on most datasets (over 80%) except NEWs (around 65%). Results for article-level classification (SPEA and NEWA) are better than for the sparser, and perhaps noisier, sentence-level data (SPEs, TWE and NEWs). We observe a larger drop for NEWs than SPEs regarding the LR and SVM methods but not for the LSTM and LSTMR methods. This suggests a greater robustness of semantic features compared to linguistic features.

For the LR and SVM methods, the WC, TI-W and TI-G features work better than the LIWC features. However, the combination of all four feature groups yields the best results. When omitting proper nouns, performance drops, which indicates that the models also learn some topic features instead of propaganda itself. However, the LIWC representations are not affected much when excluding proper nouns due to their focus on word categories.

The LSTM and LSTMR methods perform slightly worse than LR and SVM on all datasets except NEWs. Due to the influence of the pairwise ranking and the loss function, LSTMR performs slightly worse than LSTM on some datasets. As a trade-off, it gets better cross-domain performance (see Table 3). Overall, feature-based learning works best for the standard case of in-domain classification.

Cross-Domain Performance. The cross-domain results are given in Table 3. As expected, cross-domain classification is much more challenging than in-domain classification. Still, many cross-domain settings show reasonably good performance. The best cross-domain results are observed for the classifier trained on NEWs and tested on NEWs. The most balanced precision–recall performance is obtained by the classifier trained on TWE and tested on NEWs. Overall, LSTMR compares favorably against other methods, especially on TWE, NEWs and NEWs.

Combining Datasets for Training. As a final configuration, we combine several datasets, excluding NEWs, into a single training set and test on NEWs. The results are given in Table 5. As we see, the larger amount of training data does not help the classifiers to improve their cross-domain performance. On the contrary, the F1 scores slightly drop when more data is combined. The likely reason is the high variability in the underlying features and data characteristics that comes from such highly heterogeneous sources. These settings require further research.

6 Analysis

Datasets and Granularity. The best cross-domain results are obtained for training on news...
Table 2: F1 score (%) for in-domain classification with different methods, features and datasets. PN stands for proper noun.

<table>
<thead>
<tr>
<th>Method</th>
<th>SPEA</th>
<th>SPEe</th>
<th>TWE</th>
<th>NEWA</th>
<th>NEWe</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>YES</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>964</td>
<td>971</td>
<td>963</td>
<td>964</td>
<td>865</td>
</tr>
<tr>
<td>SVM</td>
<td>960</td>
<td>982</td>
<td>960</td>
<td>900</td>
<td>982</td>
</tr>
<tr>
<td>LSTM</td>
<td>908</td>
<td>751</td>
<td>807</td>
<td>866</td>
<td>751</td>
</tr>
<tr>
<td><strong>NO</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>960</td>
<td>971</td>
<td>963</td>
<td>964</td>
<td>865</td>
</tr>
<tr>
<td>SVM</td>
<td>953</td>
<td>982</td>
<td>960</td>
<td>900</td>
<td>975</td>
</tr>
</tbody>
</table>

Table 3: Precision (P), recall (R) and F1 (F) score (%) on cross-domain classification with the features. Results are highlighted when LSTM is better than LSTMR.

Table 4: F1 score (%) for cross-domain classification with proper nouns removed.

Table 5: Cross-domain performance (%) of classifiers trained on combined datasets and tested on NEWs.

and applying the learned models to speeches or tweets. This suggests that high-quality training labels are still crucial. Performance for articles is better than for sentences, underlining the difficulty of dealing with very short texts without placing them into their full context. From a model perspective, this makes features such as TF-IDF and applying the learned models to speeches or tweets. This suggests that high-quality training labels are still crucial. Performance for articles is better than for sentences, underlining the difficulty of dealing with very short texts without placing them into their full context. From a model perspective, this makes features such as TF-IDF...
again the big challenge in cross-domain classification.

Table 6: Cross-domain performance (F1 score, %) with training on NEW$_S$, using different features. ‘1’, ‘2’, ‘3’ and ‘4’ stand for WC, TI-W, TI-G and LIWC.

<table>
<thead>
<tr>
<th>Method</th>
<th>SPE$_A$</th>
<th>SPE$_S$</th>
<th>TWE</th>
<th>NEW$_S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>63 673 671 387</td>
<td>657 458 452 491</td>
<td>672 487 418 491</td>
<td>672 678 623 603</td>
</tr>
<tr>
<td>SVM</td>
<td>674 673 678 390</td>
<td>428 467 458 403</td>
<td>459 498 416 443</td>
<td>687 688 662 596</td>
</tr>
</tbody>
</table>

**Informative Language Features.** To analyze to what extent word-level features can yield interpretable cues that characterize propaganda in different kinds of sources, we trained the LR classifier on the NEW$_S$ data with TI-W and LIWC features, excluding proper nouns. We observed the following most distinctive words, characteristic for propaganda:


The most salient LIWC features (word categories) were:

affect, neg-emo, anger, swear, certain

This suggests that exaggerations (e.g., “absolutely”) and negative emotions (e.g., “lies” or “devastating”) play a key role in manipulating the audience. As for LIWC features, words that express negative emotions are typical for propaganda, as well as being strongly self-confident.

**LSTM R Loss Functions.** For our LSTM R model, the selection of the loss function has a notable influence on performance. The previous experiments all used THR as loss function. Training on NEW$_S$ and with cross-domain test data, Figure 3 compares the effectiveness of all four loss functions that we studied. We notice that the specifically designed THR loss function outperforms all alternatives, especially on the SPE$_S$ and TWE datasets. This observation is in line with our design goal that THR helps the model training to discount ranked pairs of low confidence. We sacrifice in-domain performance to some extent, but improve the cross-domain classifier.

![Figure 3: F1 of LSTM R trained on NEW$_S$ and tested on other datasets, with different loss functions.](image)

**Sampling Strategies.** The training of LSTM R requires ranked pairs of samples. Over-sampling constructs these pairs by repeatedly drawing from the positive and negative classes. The effect of the amount of over-sampling is plotted in Fig-

![Figure 4: F1 of LSTM R trained on NEW$_S$ and tested on other data, with varying rate of over-sampling.](image)

![Figure 5: F1 of LSTM R trained on NEW$_S$ and tested on other data, with varying percentage of under-sampling.](image)
Figure 4, for the setting with training on NEW$S$ and cross-domain testing on the other datasets. The results show the limitations of this sampling strategy: there is hardly any increase in performance and even a drop for some cases. This calls for further research towards overcoming training bottlenecks.

For the under-sampling strategy, Figure 5 shows results (for the same train-test setting) with varying fractions of training samples. On the SPE$A$ data, the performance increases with more training samples, as expected. However, on other test datasets, performance remains unchanged or drops or fluctuates. This points out the limitations of cross-domain classification. Our approach counters the training bottleneck to a good degree, but there is quite some room for improvement and the need for further research.

7 Conclusion and Future Work

Although propaganda has become a pervasive challenge in online media, previous work has mostly treated it as a variation of fake news, or considered unrealistic settings where the test distribution precisely matches the training data distribution.

In this paper, we present a first and preliminary analysis of the problem of propaganda detection in cross-domain learning settings. This encompasses several novel aspects, ranging from data collection methods, feature computation, designing different classifiers, and the corresponding analysis. We tap into a previously unexplored content source: speeches by politicians who are known for different levels of propaganda, using them as collective and relative signals. On the methodology side, we devise a pairwise ranking method with customized loss functions to improve the classification. The experimental results demonstrate the effectiveness of this method. Furthermore, we conduct a series of experiments to explore the most salient factors for cross-domain generalizability of propaganda detection learning. The observations and analysis reveal insightful patterns and lessons for building more general propaganda detectors.

As our datasets are still fairly small, our findings are of preliminary nature and our methodology is subject to ongoing research. We believe that cross-domain learning is a crucial asset for the important topic of propaganda detection, and hope that our initial results are useful for further research along these lines.

Acknowledgments

The work is partly supported by “The Fundamental Research Funds of Shandong University”. Gerhard Weikum’s work is partly supported by the ERC Synergy Grant 610150 (imPACT). Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References


