Examining Temporality in Document Classification

Xiaolei Huang and **Michael J. Paul**
Information Science
University of Colorado
Boulder, CO 80309, USA
{xiaolei.huang,mpaul}@colorado.edu

Abstract

Many corpora span broad periods of time. Language processing models trained during one time period may not work well in future time periods, and the best model may depend on specific times of year (e.g., people might describe hotels differently in reviews during the winter versus the summer). This study investigates how document classifiers trained on documents from certain time intervals perform on documents from other time intervals, considering both seasonal intervals (intervals that repeat across years, e.g., winter) and non-seasonal intervals (e.g., specific years). We show experimentally that classification performance varies over time, and that performance can be improved by using a standard domain adaptation approach to adjust for changes in time.

1 Introduction

Language, and therefore data derived from language, changes over time (Ullmann, 1962). Word senses can shift over long periods of time (Wilkins, 1993; Wijaya and Yeniterzi, 2011; Hamilton et al., 2016), and written language can change rapidly in online platforms (Eisenstein et al., 2014; Goel et al., 2016). However, little is known about how shifts in text over time affect the performance of language processing systems.

This paper focuses on a standard text processing task, document classification, to provide insight into how classification performance varies with time. We consider both long-term variations in text over time and seasonal variations which change throughout a year but repeat across years. Our empirical study considers corpora containing formal text spanning decades as well as user-generated content spanning only a few years.

After describing the datasets and experiment design, this paper has two main sections, respectively addressing the following research questions:

1. In what ways does document classification depend on the timestamps of the documents?
2. Can document classifiers be adapted to perform better in time-varying corpora?

To address question 1, we train and test on data from different time periods, to understand how performance varies with time. To address question 2, we apply a domain adaptation approach, treating time intervals as domains. We show that in most cases this approach can lead to improvements in classification performance, even on future time intervals.

1.1 Related Work

Time is implicitly embedded in the classification process: classifiers are often built to be applied to future data that doesn't yet exist, and performance on held-out data is measured to estimate performance on future data whose distribution may have changed. Methods exist to adjust for changes in the data distribution (*covariate shift*) (Shimodaira, 2000; Bickel et al., 2009), but time is not typically incorporated into such methods explicitly.

One line of work that explicitly studies the relationship between time and the distribution of data is work on classifying the time period in which a document was written (*document dating*) (Kanhabua and Nørvåg, 2008; Chambers, 2012; Kotsakos et al., 2014). However, this task is directed differently from our work: predicting timestamps given documents, rather than predicting information about documents given timestamps.

Dataset	Time intervals (non-seasonal)	Time intervals (seasonal)	Size
Reviews (music)	1997-99, 2000-02, 2003-05, 2006-08, 2009-11, 2012-14	Jan-Mar, Apr-Jun, Jul-Sep, Oct-Dec	653K
Reviews (hotels)	2005-08, 2009-11, 2012-14, 2015-17	Jan-Mar, Apr-Jun, Jul-Sep, Oct-Dec	78.6K
Reviews (restaurants)	2005-08, 2009-11, 2012-14, 2015-17	Jan-Mar, Apr-Jun, Jul-Sep, Oct-Dec	1.16M
News (economy)	1950-70, 1971-85, 1986-2000, 2001-14	Jan-Mar, Apr-Jun, Jul-Sep, Oct-Dec	6.29K
Politics (platforms)	1948-56, 1960-68, 1972-80, 1984-92, 1996-2004, 2008-16	n/a	35.8K
Twitter (vaccines)	2013, 2014, 2015, 2016	Jan-Mar, Apr-Jun, Jul-Sep, Oct-Dec	9.83K

Table 1: Descriptions of corpora spanning multiple time intervals. Size is the number of documents.

2 Datasets and Experimental Setup

Our study experiments with six corpora:

- **Reviews:** Three corpora containing reviews labeled with sentiment: music reviews from Amazon (He and McAuley, 2016), and hotel reviews and restaurant reviews from Yelp.[1] We discarded reviews that had fewer than 10 tokens or a helpfulness/usefulness score of zero. The reviews with neutral scores were removed.
- **Politics:** Sentences from the American party platforms of Republicans and Democrats from 1948 to 2016, available every four years.[2]
- **News:** Newspaper articles from 1950-2014, labeled with whether the article is relevant to the US economy.[3]
- **Twitter:** Tweets labeled with whether they indicate that the user received an influenza vaccination (i.e., a flu shot) (Huang et al., 2017).

Our experiments require documents to be grouped into time intervals. Table 1 shows the intervals for each corpus. Documents that fall outside of these time intervals were removed. We grouped documents into two types of intervals:

- **Seasonal:** Time intervals within a year (e.g., January through March) that may be repeated across years.
- **Non-seasonal:** Time intervals that do not repeat (e.g., 1997-1999).

For each dataset, we performed binary classification, implemented in `sklearn` (Pedregosa et al., 2011). We built logistic regression classifiers with TF-IDF weighted n-gram features ($n \in \{1, 2, 3\}$), removing features that appeared in less than 2 documents. Except when otherwise specified, we held out a random 10% of documents as validation data for each dataset. We used Elastic Net (combined ℓ_1 and ℓ_2) regularization (Zou and Hastie, 2005), and tuned the regularization parameters to maximize performance on the validation data. We evaluated the performance using weighted F1 scores.

[1] https://www.yelp.com/dataset

[2] https://www.comparativeagendas.net/datasets_codebooks

[3] https://www.crowdflower.com/data-for-everyone/

3 How Does Classification Performance Vary with Time?

We first conduct an analysis of how classifier performance depends on the time intervals in which it is trained and applied. For each corpus, we train the classifier on each time interval and test on each time interval. We downsampled the training data within each time interval to match the number of documents in the smallest interval, so that differences in performance are not due to the size of the training data.

In all experiments, we train a classifier on a partition of 80% of the documents in the time interval, and repeat this five times on different partitions, averaging the five F1 scores to produce the final estimate. When training and testing on the same interval, we test on the held-out 20% of documents in that interval (standard cross-validation). When testing on different time intervals, we test on all documents, since they are all held-out from the training interval; however, we still train on five subsets of 80% of documents, so that the training data is identical across all experiments.

Finally, to understand why performance varies, we also qualitatively examined how the distribution of content changes across time intervals. To measure the distribution of content, we trained a topic model with 20 topics using `gensim` (Řehůřek and Sojka, 2010) with default parameters. We associated each document with one topic (the most probable topic in the document), and then calculated the proportion of each topic within a time period as the proportion of documents in that time period assigned to that topic. We can then visualize the extent to which the distribution of 20 topics varies by time.

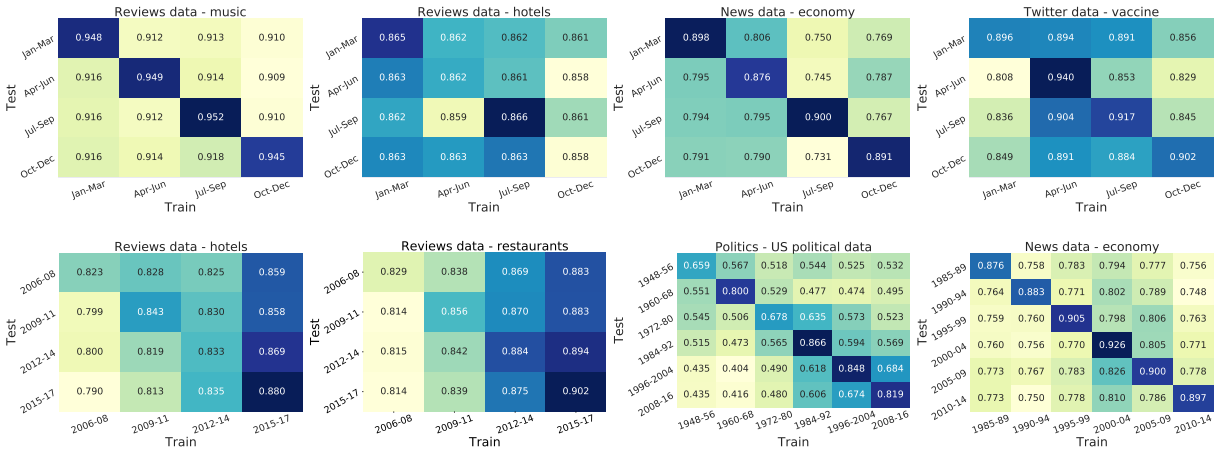

Figure 1: Document classification performance when training and testing on different times of year (top) and different years (bottom). Some corpora are omitted for space.

3.1 Seasonal Variability

The top row of Figure 1 shows the test scores from training and testing on each pair of seasonal time intervals for four of the datasets. We observe very strong seasonal variations in the economic news corpus, with a drop in F1 score on the order of 10 when there is a mismatch in the season between training and testing. There is a similar, but weaker, effect on performance in the music reviews from Amazon and the vaccine tweets. There was virtually no difference in performance in any of the pairs in both review corpora from Yelp (restaurants, not pictured, and hotels).

To help understand why the performance varies, Figure 2 (left) shows the distribution of topics in each seasonal interval for two corpora: Amazon music reviews and Twitter. We observe very little variation in the topic distribution across seasons in the Amazon corpus, but some variation in the Twitter corpus, which may explain the large performance differences when testing on held-out seasons in the Twitter data as compared to the Amazon corpus.

For space, we do not show the descriptions of the topics, but instead only the shape of the distributions to show the degree of variability. We did qualitatively examine the differences in word features across the time periods, but had difficulty interpreting the observations and were unable to draw clear conclusions. Thus, characterizing the ways in which content distributions vary over time, and why this affects performance, is still an open question.

3.2 Non-seasonal Variability

The bottom row of Figure 1 shows the test scores from training and testing on each pair of non-seasonal time intervals. A strong pattern emerges in the political parties corpus: F1 scores can drop by as much as 40 points when testing on different time intervals. This is perhaps unsurprising, as this collection spans decades, and US party positions have substantially changed over time. The performance declines more when testing on time intervals that are further away in time from the training interval, suggesting that changes in party platforms shift gradually over time. In contrast, while there was a performance drop when testing outside the training interval in the economic news corpus, the drop was not gradual. In the Twitter dataset (not pictured), F1 dropped by an average of 4.9 points outside the training interval.

We observe an intriguing non-seasonal pattern that is consistent in both of the review corpora from Yelp, but not in the music review corpus from Amazon (not pictured), which is that the classification performance fairly consistently increases over time. Since we sampled the dataset so that the time intervals have the same number of reviews, this suggests something else changed over time about the way reviews are written that makes the sentiment easier to detect.

The right side of Figure 2 shows the topic distribution in the Amazon and Twitter datasets across non-seasonal intervals. We observe higher levels of variability across time in the non-seasonal intervals as compared to the seasonal intervals.

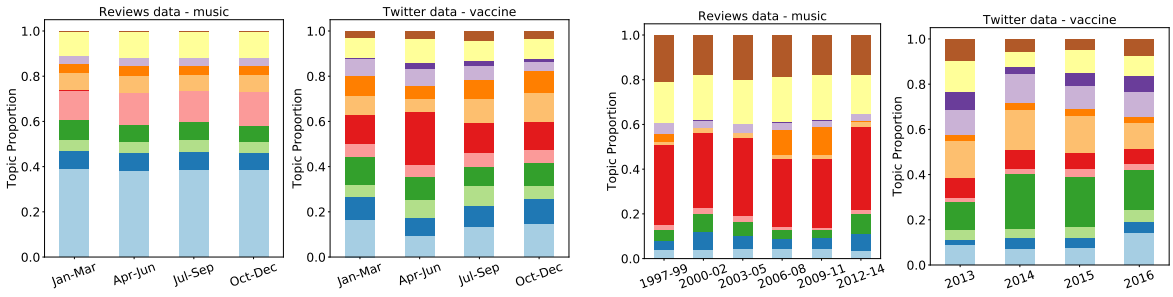

Figure 2: Topic distributions in each time of year (left) and each span of years (right). Topic models are trained independently in the seasonal vs. non-seasonal settings and are not aligned.

3.3 Discussion

Overall, it is clear that classifiers generally perform best when applied to the same time interval they were trained. Performance diminishes when applied to different time intervals, although different corpora exhibit differ patterns in the way in which the performance diminishes. This kind of analysis can be applied to any corpus and could provide insights into characteristics of the corpus that may be helpful when designing a classifier.

4 Making Classification Robust to Temporality

We now consider how to improve classifiers when working with datasets that span different time intervals. We propose to treat this as a *domain adaptation* problem. In domain adaptation, any partition of data that is expected to have a different distribution of features can be treated as a domain (Joshi et al., 2013). Traditionally, domain adaptation is used to adapt models to a common task across rather different sets of data, e.g., a sentiment classifier for different types of products (Blitzer et al., 2007). Recent work has also applied domain adaptation to adjust for potentially more subtle differences in data, such as adapting for differences in the demographics of authors (Volkova et al., 2013; Lynn et al., 2017). We follow the same approach, treating time intervals as domains.

In our experiments, we use the feature augmentation approach of Daumé III (2007) to perform domain adaptation. Each feature is duplicated to have a specific version of the feature for every domain, as well as a domain-independent version of the feature. In each instance, the domain-independent feature and the domain-specific feature for that instance's domain have the same feature value, while the value is zeroed out for the domain-specific features for the other domains.

Data (Seasonal)	Baseline	Adaptation
Reviews (music)	.901	**.919**
Reviews (hotels)	.867	**.881**
Reviews (restaurants)	.874	**.898**
News (economy)	.782	.782
Twitter (vaccines)	**.881**	.880

Table 2: F1 scores when treating each seasonal time interval as a domain and applying domain adaptation compared to using no adaptation.

This is equivalent to a model where the feature weights are domain specific but share a Gaussian prior across domains (Finkel and Manning, 2009). This approach is widely used due to its simplicity, and derivatives of this approach have been used in similar work (e.g., (Lynn et al., 2017)). Following Finkel and Manning (2009), we separately adjust the regularization strength for the domain-independent feature weights and the domain-specific feature weights.

4.1 Seasonal Adaptation

We first examine classification performance on the datasets when grouping the seasonal time intervals (January-March, April-June, July-August, September-December) as domains and applying the feature augmentation approach for domain adaptation. As a baseline comparison, we apply the same classifier, but without domain adaptation.

Results are shown in Table 2. We see that applying domain adaptation provides a small boost in three of the datasets, and has no effect on two of the datasets. If this pattern holds in other corpora, then this suggests that it does not hurt performance to apply domain adaptation across different times of year, and in some cases can lead to a small performance boost.

Data (Non-seasonal)	Baseline	Adaptation	Adapt.+seasons
Reviews (music)	.895	**.924**	.910
Reviews (hotels)	.886	.892	**.920**
Reviews (restaurants)	.831	.879	**.889**
News (economy)	.763	.780	**.859**
Politics (platforms)	.661	**.665**	n/a
Twitter (vaccines)	.910	.903	**.920**

Table 3: F1 scores when testing on the final time interval after training on all previous intervals.

4.2 Non-seasonal Adaptation

We now consider the non-seasonal time intervals (spans of years). In particular, we consider the scenario when one wants to apply a classifier trained on older data to *future* data. This requires a modification to the domain adaptation approach, because future data includes domains that did not exist in the training data, and thus we cannot learn domain-specific feature weights. To solve this, we train in the usual way, but when testing on future data, we only include the domain-independent features. The intuition is that the domain-independent parameters should be applicable to all domains, and so using only these features should lead to better generalizability to new domains. We test this hypothesis by training the classifiers on all but the last time interval, and testing on the final interval. For hyperparameter tuning, we used the final time interval of the training data (i.e., the penultimate interval) as the validation set. The intuition is that the penultimate interval is the closest to the test data and thus is expected to be most similar to it.

Results are shown in the first three columns of Table 3. We see that this approach leads to a small performance boost in all cases except the Twitter dataset. This means that this simple feature augmentation approach has the potential to make classifiers more robust to future changes in data.

How to apply the feature augmentation technique to unseen domains is not well understood. By removing the domain-specific features, as we did here, the prediction model has changed, and so its behavior may be hard to predict. Nonetheless, this appears to be a successful approach.

4.2.1 Adding Seasonal Features

We also experimented with including the seasonal features when performing non-seasonal adaptation. In this setting, we train the models with two domain-specific features in addition to the domain-independent features: one for the season, and one for the non-seasonal interval. As above, we remove the non-seasonal features at test time; however, we retain the season-specific features in addition to the domain-independent features, as they can be reused in future years.

The results of this approach are shown in the last column of Table 3. We find that combining seasonal and non-seasonal features together leads to an additional performance gain in most cases.

5 Conclusion

Our experiments suggest that time can substantially affect the performance of document classification, and practitioners should be cognizant of this variable when developing classifiers. A simple analysis comparing pairs of time intervals can provide insights into how performance varies with time, which could be a good practice to do when initially working with a corpus. Our experiments also suggest that simple domain adaptation techniques can help account for this variation.[4]

We make two practical recommendations following the insights from this work. First, evaluation will be most accurate if the test data is as similar as possible to whatever future data the classifier will be applied to, and one way to achieve this is to select test data from the chronological end of the corpus, rather than randomly sampling data without regard to time. Second, we observed that performance on future data tends to increase when hyperparameter tuning is conducted on later data; thus, we also recommend sampling validation data from the chronological end of the corpus.

Acknowledgements

The authors thank the anonymous reviews for their insightful comments and suggestions. This work was supported in part by the National Science Foundation under award number IIS-1657338.

[4]Our code is available at: https://github.com/xiaoleihuang/Domain_Adaptation_ACL2018

References

Steffen Bickel, Michael Brckner, and Tobias Scheffer. 2009. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10:2137–2155.

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

Nathanael Chambers. 2012. Labeling documents with timestamps: Learning from their time expressions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 98–106.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Association for Computational Linguistics (ACL)*.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE*, 9.

Jenny R. Finkel and Christopher D. Manning. 2009. Hierarchical Bayesian domain adaptation. In *North American Chapter of the Association for Computational Linguistics (ACL)*.

Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. 2016. The social dynamics of language change in online networks. In *The International Conference on Social Informatics (SocInfo)*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Association for Computational Linguistics (ACL)*.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pages 507–517. International World Wide Web Conferences Steering Committee.

Xiaolei Huang, Michael C Smith, Michael J Paul, Dmytro Ryzhkov, Sandra C Quinn, David A Broniatowski, and Mark Dredze. 2017. Examining patterns of influenza vaccination in social media. In *Proceedings of the AAAI Joint Workshop on Health Intelligence (W3PHIAI), San Francisco, CA, USA*, pages 4–5.

Mahesh Joshi, Mark Dredze, William W. Cohen, and Carolyn P. Rose. 2013. What's in a domain? multi-domain learning for multi-attribute data. In *North American Chapter of the Association for Computational Linguistics (NAACL) (short paper)*, pages 685–690.

N. Kanhabua and K. Nørvåg. 2008. Improving temporal language models for determining time of non-timestamped documents. In *European Conference on Digital Libraries (ECDL)*.

Dimitrios Kotsakos, Theodoros Lappas, Dimitrios Kotzias, Dimitrios Gunopulos, Nattiya Kanhabua, and Kjetil Nørvåg. 2014. A burstiness-aware approach for document dating. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR)*, pages 1003–1006.

Veronica E. Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human centered NLP with user-factor adaptation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1155.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227 – 244.

Stephen Ullmann. 1962. *Semantics: an introduction to the science of meaning*. Basil Blackwell, Oxford.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1815–1827.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web*.

D.P. Wilkins. 1993. *From Part to Person: Natural Tendencies of Semantic Change and the Search for Cognates*. Cognitive Anthropology Research Group at the Max Planck Institute for Psycholinguistics.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.