A Lightweight yet Robust Approach to Textual Anomaly Detection

Anonymous ACL submission

Abstract

Highly imbalanced textual datasets continue to pose a challenge for supervised learning models, especially when the minority class is multi-topical. Viewing such imbalanced text data as an anomaly detection (AD) problem however has advantages for certain tasks such as detecting hate speech, or inappropriate and/or offensive language in large social media feeds. There the unwanted content tends to be both rare and non-uniform with respect to its thematic character, and better fits the definition of an anomaly than a class. Several recent approaches to textual AD use transformer models, achieving good results but with trade-offs in pre-training and inflexibility to new domains. In this paper we compare two linear models within the NMF family, which also have a recent history in textual AD. We introduce a new approach based on an alternative regularization of the NMF objective. Our results surpass other linear AD models and are on par with deep models, performing comparably well even in very small outlier concentrations.

1 Introduction

Anomaly detection (AD), also known as Outlier Detection is a well-researched area of machine learning. Traditional machine learning approaches to AD include proximity-based models where points that are separated from the rest of the data by a certain distance are considered outliers. These fall into several subclasses. There are cluster-based methods, such as k-means (MacQueen, 1967), where the point is an outlier if they have a large distance between the point and the nearest cluster, density-based methods, such as LOF (Breunig et al., 2000) and DBSCAN (Ester et al., 1996), where an object is an outlier if its density is lower than that of its neighbors and distance-based methods, such as K-NN (Cover and Hart, 1967), where the outlier neighborhood has few other points.

Most recently, Transformer models (Manolache et al., 2021) and word embedding with multi-head self-attention (Ruff et al., 2019) have been applied in textual AD models, surpassing previously top-performing reconstruction-based approaches using Non-negative Matrix Factorization (NMF) as in (Kannan et al., 2017).

We propose a new NMF-based approach as an alternative to recent transformer models. This approach, we argue, is not only well-suited to the task of AD due to its lightweight architecture and flexibility but is also the better choice versus recent supervised models for hate-speech detection. Detecting hate speech and offensive language in general is a challenging task because these tend to take various forms, change dynamically and be found in only a small minority of relatively short texts. Recent studies (Yin and Zubiaga, 2021) have pointed to concerns about generalizing results where even the best performing models show large variances in quality from one dataset to another in this domain.

This paper is organized as follows: Previous approaches are discussed in Section 2, Data and Methods are discussed in Section 3, our results are in Section 4 and the Conclusion and plans for future work in section 5. Code to reproduce our results can be found here: (github repo provided upon acceptance)

2 Previous Work

While Anomaly detection in text does not have a long history in the literature, there are some notable exceptions. For example, Guthrie (2008) and Guthrie et al. (2007) consider texts that are unusual because of author, genre, style or emotional tone.

Peng et al. (2014), analyzed idiom recognition as a type of outlier detection. Idioms have certain key properties that make detection more likely using methods for finding outliers. Examples in English include “kick the bucket” or “have a cow”.

Other studies (Manevitz and Yousef (2002), Kannan et al. (2017), Barrett et al. (2019), Ruff et al.

(2019), Manolache et al. (2021)), treat textual anomalies as topical intrusions, where the texts from one topic constitute the "inliers" and a smaller set of intrusion texts constitute the "outliers". We use this data definition for our anomaly detection task.

Among topic-intrusion type models, the currently best-performing is the transformer approach in Manolache et al. (2021), a discriminator-generator model that outperformed the previously top performing OCSVM approach in Ruff et al. (2019). A non-negative matrix factorization model was used in Kannan et al. (2017). All three approaches have outperformed traditional AD models like Isolation Forests (Désir et al., 2013) on text.

3 Proposed Methods

We propose a lightweight alternative Non-negative Matrix Factorization (NMF) model that improves upon the results of Kannan and also provides comparable results to deep models without pre-training, or attention layers. We use simple frequency-based document representations and do not rely on trained embeddings. We show results on benchmark datasets and also on a dataset of hate speech in order to show the power and adaptability of our approach to an important NLP problem. Overall, our model is tested on four datasets in multiple combinations with different outlier-inlier concentrations.

Matrix factorization models like TONMF find outliers through a reconstruction process that isolates outlier documents as residual noise. The quality of the result depends on manipulating norms on both the residual matrix and the low-rank approximation of the input matrix. Kannan et al. (2017) for example use the following optimization:

$$\arg\min_{W\geq 0, H\geq 0; Z} \frac{1}{2}||A - WH - Z||_F^2 + \alpha||Z||_{1,2} \quad (1)$$

where the Frobenius norm is applied to the main divergence function, alpha is a weight parameter applied to the residual matrix Z and Z has a special norm. The norm in this case is designed to minimize the column values representing all the outliers in a document. An additional term $\beta||H||_1$ is added to produce a more interpretable low rank matrix WH with sparse coefficients.

3.1 Matrix Factorization with Additional Constraints

We used the basic model architecture in Kannan et al. (2017) to gauge the effect of changing the main objective function. This design includes a residual matrix representing the outlying points not reproducible by the main factorization process.

We set up two competing NMF-based models. Our baseline model is a hierarchical least-squares (HALS) approach (Cichocki et al., 2008), which is the base model architecture of Kannan et al. (2017) . HALS solves the non-negative least squares sub-problem by updating each column of W separately, and generally can converge to a stationary point. Each column of W is successively updated, using gradient descent to solve each column-wise sub-problem. This has been shown to converge faster than a matrix-wise iterative updating procedure (Cichocki et al., 2008). We refer to this approach as H-NMF, henceforth in this paper.

3.2 Alternative Updating

Our experimental model uses a different updating approach entirely, replacing the squared error function with an alternative. We use an NMF approach leveraging the Correntropy-induced metric (Liu et al., 2006) in which the similarity between two variables (or sub-matrices in the NMF case) is determined through applying the Gaussian kernel to the error term:

$$V_\sigma(x, y) = \frac{1}{n}\sum_{i=1}^{n} k_\sigma(x_i - y_i) \quad (2)$$

where k_σ is the kernel function. CIM-based NMF substitutes the squared error on each entry with the kernel function. We take this a step farther following Du et al. (2012), wherein the CIM-based NMF optimizes on the row level, substituting the squared residuals on each row rather than each entry. We combine this optimization with the constrained residual matrix in the objective function as follows:

$$\frac{1}{2}\sum_{i=1}^{n}[w_i||(A-Z)_i * -W_i * H^T||^2 + \phi(w_i)] + \alpha||Z||_{1,2} \quad (3)$$

where the weight factor is defined as:

$$w_i = exp(-\frac{||(A - Z)_i * -W_i * H^T||^2}{2\sigma^2}) \quad (4)$$

The half-quadratic optimization method used here and in Du et al. (2012) has been used in the past to detect and correct errors in facial recognition problems (He et al., 2014). This method sets up a robust strategy for identifying text segments that are topically anomalous not just because of bursty word distributions but because of the topicality of the entire segment. We refer to this approach as R-NMF, henceforth in this paper.

Both our baseline and experimental models leave the residual matrix constraints fixed and focus on the main objective function, in an effort to improve the quality of outliers that are passed as residuals.

4 Experimental Results

Below we describe the datasets and preparation. All models were run on four public datasets representing distinct genres (listserv, news, wiki and hate speech). We used three outlier-inlier concentrations for each.

4.1 Data and Experimental Design

The 20Newsgroups is a publicly available collection of approximately 20,000 newsgroup documents organized into 20 topical subgroups[1]. Some newsgroups are similar (e.g., IBM/Mac hardware), while others are highly unrelated (e.g., for sale/Christian religion).

Reuters-21578 is a publicly available dataset of stories appearing on Reuters' newswire in 1987[2]. It contains 21,578 documents indexed and assigned categories by members of the Reuters Ltd. staff.

WikiPeople is the subset of the English language Wikipedia dump[3] consisting of the 945,662 articles in the category "living people".

Our dataset of Hate Speech is from de Gibert et al. (2018) and contains 9,916 samples in total of forum posts from Stormfront, a white-supremacy based forum where the "hate" class represents 11 percent of the corpus.

For each dataset, we blend the inlier classes listed in Table 1 with a sample from the outlier class to achieve three concentrations: .01, .025 and .05. When such a sample is too small, we omit the .01 concentration.

For both NMF models, we parse the input text into word count vectors using sklearn's CountVectorizer with all default parameters.We call the factorization routine on the sparse word-document matrix to obtain low-rank matrices W and H and outlier matrix Z. Following the methodology in Kannan et al. (2017), we then use the L2 norm of each column in the Z matrix as the outlier score for every document. For both models, we use 3 CPU cores with 8Gb RAM.

We also train the DATE model (Manolache et al., 2021) on our data as a benchmark. We use the the code provided by the authors[4] to run experiments. We use a learning rate of $1e^{-5}$ and sequences of maximum length 128. Training is stopped at convergence, which occurs after 5000 steps on average. We use the same evaluation framework as proposed in the paper to report results. For the DATE experiments, we use 2 Tesla V100 GPU nodes each with 32 GB RAM and 6 CPU cores.

4.2 Model Results

We show results from H-NMF, R-NMF and the DATE model of Manolache et al. (2021). Model results are shown in Table 1. We list the results for each dataset for each sample and concentration, along with the inlier and outlier classes we used to create each sample. The size of the inlier class is listed in parenthesis below the inlier class name. The outliers are sampled at random from the outlier class so as to achieve the specified outlier/inlier concentration. Winners are shown in bold.

The results are the best from a sweep of eight values of the hyper-parameter k within the range [1,128] and 5 values of alpha within the range [1,16], for both the H-NMF baseline and R-NMF. The beta parameter, commonly used for the degree of sparseness is only used for H-NMF, and there we use a sweep in the range [1,16][5].

4.3 Results Analysis

The results show that the rCIM model (R-NMF) outperforms baseline (H-NMF) overall and in particular on Reuters and WikiPeople but is outperformed by DATE on 20Newsgroups and Reuters in the larger concentrations using the 'trade' class as outliers. For the Hate Speech corpus, rCIM

[1] https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups

[2] https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

[3] https://dumps.wikimedia.org/

[4] https://github.com/bit-ml/date

[5] Du et al. (2012) find that using an L1 norm would cause the rCIM objective function to be dominated by the datapoints with near-zero fitting error and actually reduce the quality of row-based outliers.

Dataset	Inliers	Outliers	Concentration	H-NMF	R-NMF	DATE
20Newsgroups	pc/mac.hardware	ms-windows.misc	0.025	0.600	0.592	**0.650**
	(2000)		0.05	0.543	0.559	**0.767**
20Newsgroups	pc/mac.hardware	comp.windows.x	0.025	0.567	0.595	**0.691**
	(2000)		0.05	0.557	0.555	**0.712**
Reuters-21578	earn+acq	interest	0.01	0.741	**0.769**	0.691
	(5795)		0.025	0.725	**0.766**	0.712
			0.05	0.716	**0.777**	0.725
Reuters-21578	earn+acq	trade	0.01	0.871	**0.889**	0.886
	(5795)		0.025	0.826	0.859	**0.905**
			0.05	0.848	0.877	**0.894**
WikiPeople	life	career	0.025	0.675	**0.694**	0.548
	(5000)		0.05	0.690	**0.707**	0.617
Hate Speech	noHate	hate	0.01	0.688	**0.702**	0.508
	(9507)		0.025	**0.697**	0.693	0.499
			0.05	**0.679**	0.675	0.505

Table 1: AUROC Results. Bolded values indicate the best performance for each dataset blend.

does better in the lowest concentration, whereas HALS has a slight edge in larger concentrations. Both NMF-based models outperform DATE on this dataset in all concentrations.

All models achieved the best AUC on the Reuters data, with the more challenging datasets being WikiPeople and 20Newsgroups.The greatest difference between the two NMF-based approaches is found on the Reuters data where rCIM has the stronger results. Note that the results are better for all three models when the outlier class is "trade" than when it is "interest", possibly because the "interest" topic is more closely related to and thus harder to distinguish from the inlier topics "earn" and acq".

In the Hate Speech data, both NMF-based models outperform the transformer-based model. In addition our model required considerably fewer compute resources, running on 3 CPU cores, compared to 2 GPUs and 6 cores for the transformer. Other recent supervised models trained on Hate Speech alone (not developed for AD) (Wullach et al. (2021), show poor F1 scores for hate speech data sets including the de Gibert et al. (2018) data set in particular, even using all the hate samples in full concentration. Our rCIM model on the other hand shows the best performance on very small concentrations. Since posts containing hate speech or offensive language tend to be in a small minority in the real world, our model is ideally suited and does not have to compensate for data imbalance issues.

5 Conclusion and Future Work

Although recent approaches to textual Anomaly Detection using deep models are very robust, our model performs comparably and even outperforms the state of the art on the majority of AD datasets including a hate speech dataset. We also improve upon recent NMF-based AD by combining a row-centric approach with a separate residual matrix. Our approach requires no pretraining or fine tuning, making it highly adaptable to different data sets with different concentrations of anomalous texts in a low compute resource setting. The model is well-suited in particular to the task of identifying hate speech or offensive language, where supervised approaches have a poor performance history due to extreme imbalance and the unpredictable nature of such language.

We plan to continue further experiments on new AD data sets, including those containing hate speech and offensive language.

Acknowledgements

References

Leslie Barrett, Sidney Fletcher, Robert Kingan, Mrinal Kumar, Alexandra Ortan, Siddarth Parikh, Anu Pradhan, and Ryon Smey. 2019. Textual outlier detection and anomalies in financial reporting. 2nd KDD Workshop on Anomaly Detection in Finance, KDD '19. Association for Computing Machinery.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. Lof: Identifying density-based local outliers. 29(2).

A. Cichocki, R. Zdunek, and S. Amari. 2008. Nonnegative matrix and tensor factorization [lecture notes]. *IEEE Signal Processing Magazine*, 25(1):142–145.

T. Cover and P. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

Ona de Gibert, Naiara Pérez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *ALW*.

Chesner Désir, Simon Bernard, Caroline Petitjean, and Laurent Heutte. 2013. One class random forests. *Pattern Recogn.*, 46(12):3490–3506.

L. Du, X. Li, and Y. Shen. 2012. Robust nonnegative matrix factorization via half-quadratic minimization. In *2012 IEEE 12th International Conference on Data Mining*, pages 201–210.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press.

David Guthrie. 2008. Unsupervised detection of anomalous text. In *Doctoral Dissertation*. University of Sheffield.

David Guthrie, Louise Guthrie, Ben Allison, and Yorick Wilks. 2007. Unsupervised anomaly detection. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pages 1624–1628.

Ran He, Wei-Shi Zheng, Tieniu Tan, and Zhenan Sun. 2014. Half-quadratic-based iterative minimization for robust sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(2):261–275.

Ramakrishnan Kannan, Hyenkyun Woo, Charu C. Aggarwal, and Haesun Park. 2017. Outlier detection for text data : An extended version.

Weifeng Liu, P. P. Pokharel, and J. Príncipe. 2006. Correntropy: A localized similarity measure. *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 4919–4924.

J. B. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Larry M. Manevitz and Malik Yousef. 2002. One-class SVMs for document classification. *J. Mach. Learn. Res.*, 2:139–154.

Andrei Manolache, Florin Brad, and Elena Burceanu. 2021. DATE: Detecting anomalies in text via self-supervision of transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2014. Classifying idiomatic and literal expressions using topic models and intensity of emotions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2019–2027, Doha, Qatar. Association for Computational Linguistics.

Lukas Ruff, Yury Zemlyanskiy, Robert A. Vandermeulen, Thomas Schnake, and M. Kloft. 2019. Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In *ACL*.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Towards hate speech detection at large via deep generative modeling. *IEEE Internet Computing*, 25:48–57.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7.

A Appendix

Examples of Hate Speech from the de Gibert et al. (2018) corpus:

TEXT	CLASS
As of March 13th , 2014 the booklet had been downloaded over 18,300 times and counting .	no_hate
In order to help increase it would be great if all Stormfronters who had YouTube accounts , could display the following text in the description boxes of their uploaded YouTube videos .	no_hate
Simply copy and paste the following text into your YouTube videos description boxes	no_hate
Click below for a FREE download of a colorfully illustrated 132 page e-book on the Zionist-engineered INTENTIONAL destruction of Western civilization .	hate
She may or may not be a Jew she seems to think the Blacks wo n't kill her alongside every other White they can get their dirty hands on , what a muppet !	hate
Thank you for posting your story .	no_hate
I think you should write a book as well	no_hate
And the sad thing is the white students at those schools will act like that too .	hate

B Hyper-parameters

The hyper-parameter values that yielded the best results for each dataset blend. These were obtained from a sweep of eight values of k within the range [1,128], 5 values of alpha within the range [1,16], and 5 values of beta within the range [1,16]. The beta parameter is only used for H-NMF.

C Risk Statement

Anomaly detection is a type of classification model which may have imperfect Precision and Recall. As such it may classify hateful or toxic language incorrectly and should be subject to human review in contexts of high risk.

Dataset	Inliers/Outliers	Concentration	Best Model	k	alpha	beta
20Newsgroups	pc/mac.hardware	0.025	H-NMF	100	1	16
	ms-windows.misc	0.05	R-NMF	64	1	
20Newsgroups	pc/mac.hardware	0.025	R-NMF	64	2	
	comp.windows.x	0.05	H-NMF	32	1	1
Reuters-21578	earn+acq	0.01	R-NMF	16	16	
	interest	0.025	R-NMF	16	16	
		0.05	R-NMF	16	16	
Reuters-21578	earn+acq	0.01	R-NMF	16	16	
	trade	0.025	R-NMF	8	16	
		0.05	R-NMF	8	8	
WikiPeople	life	0.025	R-NMF	8	16	
	career	0.05	R-NMF	8	8	
Hate Speech	noHate	0.01	R-NMF	8	8	
	hate	0.025	H-NMF	8	8	1
		0.05	H-NMF	16	8	1

Table 2: Best hyper-parameters for each dataset blend.