ISSN 2007-9737

Evaluation of Feature Extraction Techniques in Automatic Authorship Attribution

Germán Ríos-Toledo[1], Erick Velázquez-Lozada[2], Juan Pablo Francisco Posadas-Durán[2], Saúl Prado Becerra[1], Fernando Pech May[3], María Guadalupe Monjarás Velasco[1]

[1] Tecnólogico Nacional de México, Campus Tuxtla Gutiérrez, México

[2] Instituto Politécnico Nacional, Escuela Superior de Ingeniería Mecánica y Eléctrica, México

[3] Instituto Tecnológico Superior de los Ríos, Balancán, México

{german.rt, maria.mv}@tuxtla.tecnm.mx {evelazquezl, jposadasd}@ipn.mx, saulpradobecerra@gmail.com, fernando.pech@cinvestav.mx,

Abstract. There are two main approaches to automatic text classification: content-based classification and style-based classification. With content-based text classification, the topic of a document (politics, sports, health) or fake news is detected. On the other hand, Style-based text classification is used to detect the gender or age of an author, author identification, and authorship attribution. In style-based classification, the set of words defines the author's vocabulary, which contains several hundred words. In this work, the words are known as dimensions. Texts generate high-dimensional vectors. Multiple works have shown that a large number of dimensions decreases the performance of classifiers. To reduce dimensions there are selection and extraction techniques. This article discusses the use of extraction techniques, which create low-dimensional vectors from combinations of the high-dimensional vector. Due to the development of Deep Learning networks, the use of dimension reduction techniques has decreased because these networks perform dimension reduction automatically. However, in Machine Learning such techniques are still used intensively. Motivated by the above, in this paper, the Principal Component Analysis (PCA) and Latent Semantic Analysis (LSA) dimension reduction algorithms are proposed for the identification of texts written by 14 authors of the Corpus PAN 2012. The texts were divided into sequences of 10, 20, and 30 words called sentences. Likewise, blocks of texts made up of 100 sentences were created. The supervised classification was performed with the Nearest Neighbors (KNN), Support Vector Machines (SVM) and Logistic Regression (LR) algorithms using the accuracy metric. The results showed that the reduction of dimensions with PCA and the LR and SVM classifiers achieved better results than other similar works of the state of the art using the same corpus.

Keywords. Dimension reduction, feature extraction, authorship attribution, machine learning.

1 Introduction

From the machine learning approach, the authorship attribution task is a multiclass classification problem with a single label. For automatic style-based classification, texts are

Table 1. Dimensions according to sentence length

Set	Texts	10w	20w	30w
Train	280	19,421	26,431	31,169
Test	140	13,628	19,031	22,735

represented as words (Bag of Words), char n-grams, word n-grams, POS tags, dependency relationships, among others.

Any text representation generates high dimensional vectors (features). The vectors store the frequency of use of the features in the text. This information is stored in a two-dimensional matrix where rows represent texts and columns features or dimensions. Some features have very high frequency but most appear very infrequently.

According to [10], dimensionality reduction is a process that removes irrelevant features and retains the most important ones related to the predictive modeling problem.

At first glance, adding more and more features to the model improves the classifier metrics but the effect is not as expected. This phenomenon is known as the curse of dimensionality [7]. Increasing the dimensionality without increasing the number of samples causes the density of the vectors to become sparse.

Because of this, the classifier will find a perfect solution to the machine learning model, which leads to overfitting: the model overmatches a particular data set and does not generalize well. Dimensionality reduction is performed by using feature selection and extraction techniques.

2 Related Work

Zhou et al. [16] used Term Frequency and Inverse Document Frequency (TF-IDF) and Latent Dirichlet Assignment (LDA) extraction methods in fault diagnosis texts. The authors proposed a combination of both methods and called it TI-LDA.

They concluded that their method improves intraclass and interclass compactness compared to methods using TF-IDF and LDA independently. Avinash and Sivasankar [1] used the same Term Frequency and Inverse Document Frequency (TF-IDF) and Document-to-Vector (Doc2vec) [5] extraction techniques.

They also used the Logistic Regression (LR), Support Vector Machines (SVM), Nearest Neighbors (KNN) and Decision Trees (DT) classifiers. They reported that both extraction techniques achieved satisfactory performance on different data sets but that Doc2vec's accuracy scores are better than TF-IDF.

Similarly, Singh et al. [11] proposed the TF-IDF method and GloVe word embedding. They compared their method with Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Latent Semantic Indexing (LSI) and a hybrid PCA+LDA approach and the Naive Bayes classifier. They claimed that their method gives better classification results than existing dimension reduction techniques.

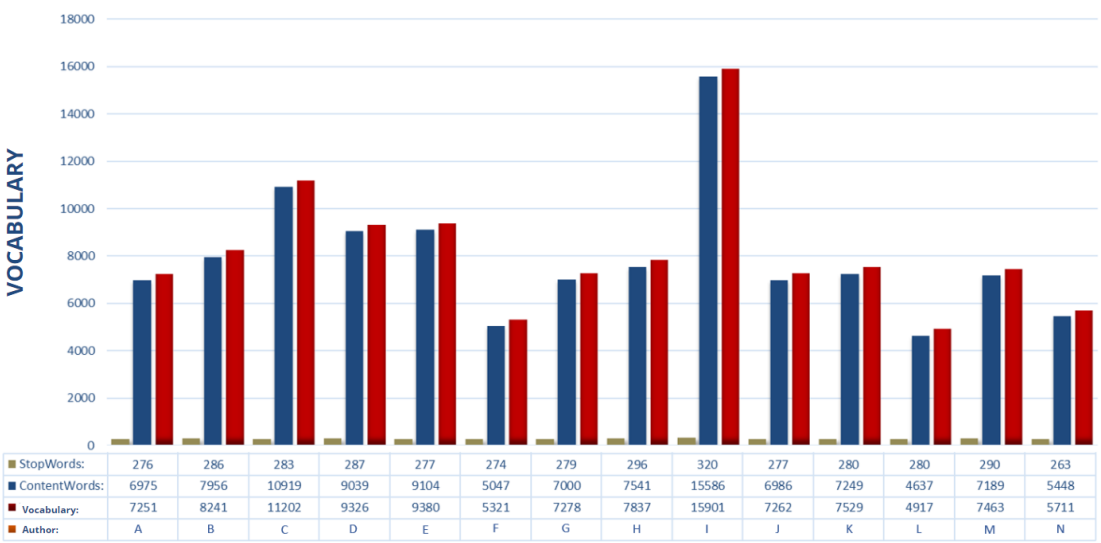
Wadud et al. [13] classified offensive texts with a model called LSTM-BOOST, which uses the modified AdaBoost algorithm with Principal Component Analysis (PCA) and LSTM networks. They compared their method against approaches such as Bag of Words (BoW), TF-IDF, Word2Vec, and fastText [2]. Wadud et al. reported that their method outperformed most reference architectures with an F1 of 92.61% on the offending text corpus.

Su et al. [12] proposed the Tree Structure Multilinear Principal Component Analysis (TMPCA) method. The authors stated that this technique reduces the dimension of input sequences and sentences to simplify subsequent text classification tasks. Based on their results, the authors concluded that the SVM method applied to the data processed by TMPCA achieves better performance than the state-of-the-art Recurrent Neural Network (RNN) approach.

3 Corpus Description

To evaluate the proposed method, the 2012 PAN Corpus was used in task I of closed class authorship attribution. It is called a closed class because there is a closed set of candidate authors and the system must identify to which author an anonymous text belongs. PAN is a series of scientific events and shared tasks on forensic analysis and stylometry of digital texts[1].

[1] https://pan.webis.de/index.html



StopWords:	276	286	283	287	277	274	279	296	320	277	280	280	290	263
ContentWords:	6975	7956	10919	9039	9104	5047	7000	7541	15586	6986	7249	4637	7189	5448
Vocabulary:	7251	8241	11202	9326	9380	5321	7278	7837	15901	7262	7529	4917	7463	5711
Author:	A	B	C	D	E	F	G	H	I	J	K	L	M	N

Fig. 1. Stop-words and content-words per author in PAN 2012

The Corpus contains texts by 14 authors, the names of the authors are identified with letters from A to N. The texts are classified into training and test sets. In the training set there are 2 texts per author and in the test set there is one text per author.

Table 2 shows the number of words identified with the Spacy[2] tokenizer. The Training column shows the word average of the two novels. The vocabulary words of each author are organized in a dictionary, which contains stop-words and content-words.

Content-words are words that provide information on the topic that a text is addressing: nouns, verbs, adjectives and adverbs.

On the other hand, stop-words are used to interconnect content-words, they are meaningless but they are crucial to build sentences: articles, pronouns, prepositions and auxiliary verbs.

The number of stop-words is much smaller than the content-words. Figure 1 shows the distribution of content-words and stop-words by each author in the test set.

[2]https://spacy.io/

4 Proposed Method

4.1 Text Preprocessing

Continuous sequences of words of different length called sentences were obtained. Sentences contained 10, 20 and 30 words. With these sentences, texts of 100 sentences were created to increase the number of texts [4].

The first 10 texts of each novel were used. According to sentence length, each text contained 1,000, 2,000, and 3,000 words. Sentences are identified by the notation 10w, 20w, and 30w, where w indicates words.

All the authors had 20 texts in the training set and 10 in the test set, ensuring that the 14 classes are balanced in terms of the number of instances. Subsequently, term-document matrices were created to store the frequency of use of the words (dimension). Table 1 shows the dimensions in the training and test sets.

4.2 Feature Extraction

Representing data in low dimensions tends to overcome the problem of the curse of dimensionality , and allows easy processing and visualization of that data [15].

Table 2. Word averages in PAN corpus 2012

Author	Number of words	
	Train Set	Test Set
A	75,048	70,130
B	148,874	82,211
C	139,929	150,769
D	81,808	93,075
E	125,544	96,382
F	54,040	42,751
G	70,795	84,940
H	109,738	94,730
I	53,924	194,441
J	61,199	60,999
K	51,036	80,212
L	57,029	50,555
M	93,468	77,804
N	80,570	53,295

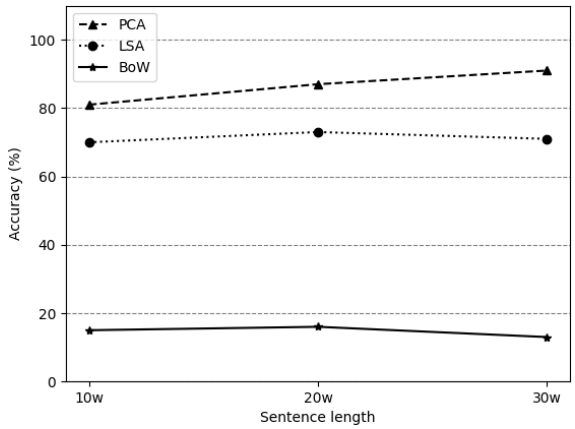


Fig. 2. Classifier accuracy vs sentences and dimensions

In this study, two methods of dimension reduction by extraction were used: Principal Component Analysis (PCA) [14] and Latent Semantic Analysis (LSA) [6]. According to [9], the goal of PCA is to find an optimal position for the best variance reduction of the data.

PCA is an unsupervised learning method that reduces the dimensionality of a data set with a large number of variables while preserving as much variation as possible.

LSA is a method that uses the statistical approach to identify the association between words in the text. The technique produces a set of concepts smaller than the original set.

It is an unsupervised learning technique, unlike PCA, LSA does not center the data before calculating the singular value decomposition. Both algorithms need the new number of dimensions of the term-document matrix.

The number of components tested on the two algorithms were 20, 50, 100, 200, and 280. PCA and LSA algorithms are implementations of scikit-learn[3].

4.3 Supervised Learning Algorithms

The K-Nearest Neighbors (KNN) algorithm is a nonparametric supervised learning classifier that uses the clustering proximity of an individual data point for predictions. It is used for regression or classification problems [8].

The Support Vector Machines (SVM) algorithm is a supervised learning model used for classification problems and regression analysis. In the training stage, SVM assigns examples to points in space by maximizing the width of the gap between the two categories.

In the testing stage, new examples are assigned and predict the category they belong to according to the side of the gap they were assigned to.

The Logistic Regression algorithm is a classifier based on the Maximum Entropy Modeling Framework, which considers all probability distributions that are empirically consistent with the training data; and choose the distribution with the highest entropy.

All three classifiers are implementations of scikit-learn. The training data set was used to perform an exhaustive search for the best parameters for each classifier.

[3] https://scikit-learn.org/stable/index.html

Table 3. Classifier performance with PCA

Components	10w			20w			30w		
	LR	SVM	KNN	LR	SVM	KNN	LR	SVM	KNN
20	67	**69**	64	**79**	75	75	**85**	82	82
50	75	**77**	65	**82**	81	79	**85**	82	77
100	**78**	76	62	**83**	82	72	87	**88**	70
200	**80**	75	58	**87**	85	70	**88**	**88**	74
280	80	**81**	47	**84**	83	58	**91**	90	51

Table 4. Classifier performance with LSA

Components	10w			20w			30w		
	LR	SVM	KNN	LR	SVM	KNN	LR	SVM	KNN
20	**57**	54	28	62	**65**	44	**71**	67	52
50	**68**	59	22	**72**	67	26	**67**	60	30
100	**66**	50	15	**67**	55	26	**66**	60	23
200	**65**	45	10	**64**	50	26	**64**	45	14
280	**70**	56	12	**73**	72	10	**69**	57	14

5 Results

Table 3 shows the accuracy of classifiers with PCA reduction algorithm. In 10w sentences, the LR and SVM classifiers achieved better results with 200 and 280 components, 80% and 81% in each case.

This last data represents all the texts in the training set. The number of components determines the percentage of variation retained from the original data. In the 20w sentences the number of words in the texts is greater.

Regardless of the number of components, the LR classifier achieved the highest accuracy. Highlighting 87% with 200 components. The SVM and KNN classifiers also present favorable results of at least 75%. In the 30w sentences, the LR and SVM classifiers obtain at least 82% accuracy.

Furthermore, with 280 components, LR achieves 91%. and SVM 88% with 100 and 200 components. On the other hand, the KNN classifier obtained 82% accuracy with 20 components. However, as the components increased, the accuracy decreased.

Table 4 shows the accuracy of classifiers with LSA reduction algorithm. All classifiers showed lower accuracy percentages with respect to PCA.

The LR classifier outperformed SVM and KNN in the different experiments. The number of words in the texts was not an important factor for the performance of the classifiers.

The highest percentages were obtained with 280 components. It is worth noting that in 30w sentences and 280 components, the accuracy of the KNN classifier decreased to 14%.

In addition, an experiment was carried out without applying reduction techniques with the Bag of Words (Bag of Word, BoW) model. Table 5 shows the average accuracy obtained by each classifier.

Figure 2 shows the highest precision obtained in the different sentences and dimension reduction techniques. The highest accuracy is obtained using the PCA technique and 30w sentences.

6 Discussions

In this paper, a method was proposed to solve the Authorship Attribution problem using dimension reduction techniques by extraction. The task was approached as a supervised machine learning-based classification problem with the Corpus of the PAN 2012 competition and subtask I.

Table 5. Classifier performance without PCA and LSA

Sentence	Dimensions	LR(%)	SVM(%)	KNN(%)
10w	19,421	8	7	15
20w	26,431	7	7	16
30w	31,169	7	7	13

Table 6. Proposed method vs related works

	Hitschler et al.	Jafariakinabad et al.	Ríos et al.
Text Size	1,500 words	100 sentences with 30 words	100 sentences with 30 words
Tools	frequency based selection, CNN	POS CNN-LSTM, SoftMax	PCA, Logistic regression
Accuracy(%)	52.73	78.76	91.00

The texts were divided into sentences of 10, 20 and 30 words. With them, text blocks made up of 100 sentences were created. Unlike the original PAN 2012 task, we focused on the classification of the proposed blocks and not on the complete novels to reproduce the results of the related works. This paper reports the accuracy metric used in the PAN 2012 competition.

Tables 3, 4 and 5 show the accuracy achieved by the LR, SVM and KNN classifiers in the test set. The best results were achieved with sentences of 20 or 30 words.

This is because these texts contain more information, allowing classifiers to improve the accuracy of identifying an author's writing style-based on word frequency. Figure 2 shows that dimension reduction techniques generate an optimized model compared to the bag of words (BoW) model.

The best PCA result beats the best BoW result by approximately 75% and the best LSA result by approximately 57%. PCA performed better than LSA in all experiments.

The best result with PCA is approximately 18% higher than the best result with LSA. The use of a selection technique based on information variance proved to be more efficient than that based on Information Retrieval strategies.

The following articles also propose strategies to solve the same problem of Authorship Attribution with the PAN 2012 corpus. In [3] they used a Convolutional Neural Network (CNN) and grammar tags (POS).

They used segments of 1,500 words and feature (dimensions) selection based on the frequency of occurrence testing different cut-offs. On the other hand, in [4] uses a Recurrent Neural Network and a Convolutional Neural Network with a Long Term Short Term Memory (LSTM) to learn the syntactic information of the occurrence of POS labels.

This work carried out tests with segments of 20, 50, 100 and 200 sentences. Likewise, the sentences were of different sizes (10, 20, 30 and 40 words).

The best results were obtained with segments of 100 sentences and sentences of 30 words. Table 6 shows the best configurations of these works and the results they obtained with the corpus of the PAN 2012 competition.

The results correspond to the accuracy obtained when classifying each text block of the corpus independently test. The proposed method overcome both previous works.

The use of traditional techniques such as PCA and the Logistic Regression classifier achieves competitive results in texts where information is scarce. That is, segments much smaller than the length of the original text.

7 Conclusions

In this work, the use of PCA and LSA dimension reduction techniques in the Authorship Attribution problem was evaluated. Both algorithms are frequently used in previous works related to this task.

The PCA technique achieved the best results. In general, the use of feature extraction techniques allows to obtain better than the BoW model.

The use of lexical information proved to be more relevant for the development of models that allow identifying the writing styles of an author compared to the use of syntactic information (POS tags).

In addition, it was verified that a text segment between 2,000 and 3,000 words is enough for classifiers to learn the style of a particular author. It is not ruled out that the use of syntactic information is useful to identify an author's writing style.

References

1. **Avinash, M., Sivasankar, E. (2019).** A study of feature extraction techniques for sentiment analysis. Emerging Technologies in Data Mining and Information Security, pp. 475–486. DOI: 10.1007/978-981-13-1501-5_41.
2. **Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017).** Enriching word vectors with subword information. Transactions of the association for computational linguistics, Vol. 5, pp. 135–146. DOI: 10.1162/tacl_a_00051.
3. **Hitschler, J., Van Den Berg, E., Rehbein, I. (2018).** Authorship attribution with convolutional neural networks and pos-eliding. Proceedings of the Workshop on Stylistic Variation, pp. 53–28. DOI: 10.18653/v1/W17-4907.
4. **Jafariakinabad, F., Tarnpradab, S., Hua, K. A. (2020).** Syntactic neural model for authorship attribution. The Thirty-Third International Flairs Conference.
5. **Le, Q. V., Mikolov, T. (2014).** Distributed representations of sentences and documents. International conference on machine learning, pp. 1188–1196.
6. **Mohammed, S. H., Al-augby, S. (2020).** LSA and LDA topic modeling classification: Comparison study on e-books. Indonesian Journal of Electrical Engineering and Computer Science, Vol. 19, No. 1, pp. 353. DOI: 10.11591/ijeecs.v19.i1.pp353-362.
7. **Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., Liao, Q. (2017).** Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. International Journal of Automation and Computing, Vol. 14, No. 5, pp. 503–519. DOI: 10.1007/s11633-017-1054-2.
8. **Raschka, S. (2018).** Stat 479: Machine learning lecture notes. Vol. 38.
9. **Salih-Hasan, B. M., Adnan Mohsin, A. (2021).** A review of principal component analysis algorithm for dimensionality reduction. Journal of Soft Computing and Data Mining, Vol. 2, No. 1, pp. 20–30.
10. **Salo, F., Nassif, A. B., Essex, A. (2019).** Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. Computer Networks, Vol. 148, pp. 164–175. DOI: 10.1016/j.comnet.2018.11.010.
11. **Singh, K. N., Devi, S. D., Devi, H. M., Mahanta, A. K. (2022).** A novel approach for dimension reduction using word embedding: An enhanced text classification approach. International Journal of Information Management Data Insights, Vol. 2, No. 1. DOI: 10.1016/j.jjimei.2022.100061.
12. **Su, Y., Huang, Y., Kuo, C. C. J. (2018).** Efficient text classification using tree-structured multi-linear principal component analysis. 24th international conference on pattern recognition (ICPR), pp. 585–590. DOI: 10.1109/ICPR.2018.8545832.
13. **Wadud, M. A. H., Kabir, M. M., Mridha, M., Ali, M. A., Hamid, M. A., Monowar, M. M. (2022).** How can we manage offensive text in social media-a text classification approach using lstm-boost. International Journal of Information Management Data Insights, Vol. 2, No. 2, pp. 100095. DOI: 10.1016/j.jjimei.2022.100095.
14. **Wang, Z., Mekala, D., Shang, J. (2020).** X-class: Text classification with extremely weak supervision. arXiv preprint arXiv:2010.12794. DOI: 10.48550/arXiv.2010.12794.
15. **Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., Saeed, J. (2020).** A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. Journal of Applied Science and Technology Trends, Vol. 1, No. 2, pp. 56–70. DOI: 10.38094/jastt1224.
16. **Zhou, S., Chen, B., Zhang, Y., Liu, H., Xiao, Y., Pan, X. (2020).** A feature extraction method based on feature fusion and its application in the text-driven failure diagnosis field. Vol. 4, No. 6. DOI: 10.9781/ijimai.2020.11.006.

Article received on 04/10/2022; accepted on 15/12/2022.
Corresponding author is Juan Pablo Francisco Posadas-Durán.