Surviving the Legal Jungle: Text Classification of Italian Laws in extremely Noisy conditions

Riccardo Coltrinari
Computer Science dept.
University of Camerino
Camerino (MC)

Alessandro Antinori
Computer Science dept.
University of Camerino
Camerino (MC)

{riccardo.coltrinari, alessandro.antinori}@studenti.unicam.it

Fabio Celli
Research and Development
Maggioli S.p.A.
Santarcangelo (RN)
fabio.celli@maggioli.it

Abstract

In this paper, we present a method based on Linear Discriminant Analysis for legal text classification of extremely noisy data, such as duplicated documents classified in different classes. The results show that Linear Discriminant Analysis obtains very good performances both in clean and noisy conditions, if used as classifier in ensemble learning and in multi-label text classification.

1 Motivation and Background

We address text categorization of business-oriented legal documents in Italian, but with a custom and overlapping hierarchy of product categories. A typical approach to tackle similar tasks is to exploit resources such as EUROVOC (Daudaravicius, 2012), a multilingual thesaurus consisting of over 6700 hierarchically-organised class descriptors used by many organizations of the European Union (EU) for the classification and retrieval of official documents. Our editorial system has a hierarchy of 23 product categories and more than 20600 labels, manually annotated and customized for different clients in more than 15 years, hence it is not possible to exploit resources like EUROVOC to categorize documents.

In this paper, we propose a fast and efficient method for document classification for noisy data based on Linear Discriminant Analysis, a dimensionality reduction technique that has been employed successfully in many domains, including neuroimaging and medicine. We believe that our contribution will be useful to the NLP community in the context of document categorization as well as automatic ontology population, in particular when dealing with very noisy data.

The paper is structured as follows: in Section 1.1 we present the related works in the field of text classification and the potential of Linear Discriminant Analysis, in Section 2 we describe the datasets we used, in Section 3 we report and discuss the result of our classification experiments and in Section 4 we draw our conclusions.

1.1 Related Work

There are many applications of NLP in the legal text domain, such as the creation of ontologies for knowledge extraction (Lenci et al., 2009) or legal reasoning (Palmirani et al., 2018), other tasks include dependency parsing (Dell'Orletta et al., 2012), deception detection (Fornaciari et al., 2013) and semantic annotation exploiting external resources like FrameNet (Venturi, 2011). In this domain, the most popular way to perform text categorization is to use ontologies: for example many used EUROVOC to label documents in several languages (Steinberger et al., 2013) with one label for each document, in order to train SVMs (Boella et al., 2013) or deep learning models (Caled et al., 2019), for the prediction of labels at different levels of granularity in the label hierarchy. Another approach is to use the judgments of the Supreme Court as gold standard labels, thus reducing the complexity of the task, and then train machine learning models, such as SVMs, to perform classification (Sulea et al., 2017). It is known that active learning does not reach a good performance in the legal domain (Cardellino et al., 2015), but it is possible to align different resources to perform ontology population or expansion (Cardellino et al., 2017). The state-of-the-art in text classification ranges from 40% to 85% or more, depending on the complexity and size of the dataset, and from the number of document classes (Adhikari et al., 2019). The results of a noise introduction simulation study revealed that substituting up to 40% of words with random text strings yields to a small decrease in text classification performance, while the substitution of more than 40% of the text yields a dramatic decrease in classification performance (Agarwal et al., 2007).

A similar task, Extreme Multi-Label Text Classification (XMTC), consists in the classification of documents annotated with multiple tags. Recent experiments of XMTC with Convolutional Neural Networks on a dataset of 57k legal documents annotated with multiple concepts from EUROVOC, revealed that word embeddings extracted with label-wise attention Networks (Mullenbach et al., 2018) leads to the best overall performance, compared pre-trained word embeddings, Hierarchical word embedding and Max-Pooling Scorers that produce section-based word embeddings (Chalkidis et al., 2019). It has been demonstrated in more than one context that cNNs perform well for text categorization, but also that there is no single algorithm that performed the best across the combination of data sets and training sample sizes (Keeling et al., 2019). The rationale behind the good performance of label-wise attention networks is their ability to maximise the difference of the words/features associated to different labels. A very similar -but faster- approach is Linear Discriminant Analysis (Balakrishnama and Ganapathiraju, 1998), a feature selection and classification technique that has been successfully used for the incremental classification of large streams of data (Pang et al., 2005), to find identity patterns in images before the advent of deep learning (Prince and Elder, 2007) and as feature selection technique for discriminating fMRI response patterns to visual stimuli (Mandelkow et al., 2016).

Linear Discriminant Analysis (henceforth LDA) is a widely accepted dimensionality reduction and classification method, which aims to find a transformation matrix to convert a feature space to a smaller space by maximising the between-class scatter matrix while minimising the within-class scatter matrix (Boroujeni et al., 2018). The criticism towards this technique emphasize the fact that it suffers from the domination of the largest objectives, in particular when close class pairs tend to overlap in a feature subspace, but this can be solved with various optimizations, including eigenvalue decomposition, among others (Li et al., 2017).

2 Data

Our dataset consists of 2030 legal italian documents with an average of 800 words each. We have 23 classes representing products manually annotated over 15 years, every document is categorized in one or more classes. Classes are not balanced, but their distribution is proportional to the whole editorial system, that consists of 443.7k documents. We extracted such a small dataset from the editorial system because we plan to update our models very frequently, using a small portion of documents each time in order to save computational power and time. Figure 1 reports the distribution of the classes in our dataset.

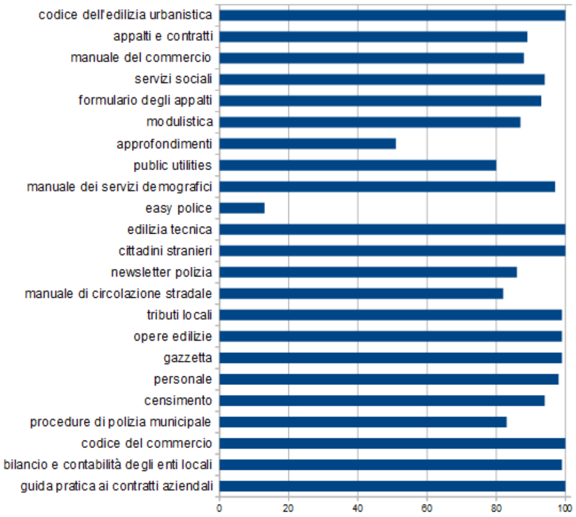

Figure 1: Distribution of the classes in our dataset.

Since documents can fall under more than one class, we have 43% of documents repeated under different classes. We tested the performance of different classifiers under two different conditions: noisy (with repeated documents) and clean (without the repeated documents).

3 Experiments and Discussion

In both cases (noisy and clean) we performed preprocessing on text, deleting punctuation and Italian stopwords. We did not use stemming or lemmatization since their usage has led to a degradation of results. We formalize the task in two ways: a simple multinomial classification, where we train a classifier to predict one class per document, and a multi-label classification, where we produce a score ranking of labels for each document and evaluate if the gold standard label occurs in the first N positions.

		noisy	noisy	clean	clean
features	algorithm	acc(10f-cv)	acc(split)	acc(10f-cv)	acc(split)
baseline	majority (zeroR)	4.6%	2.9%	8.3%	6.9%
200 glove embeddings	cNN	5.1%	4.6%	10.6%	10.8%
200 glove embeddings	rNN	19.6%	21.7%	33.6%	31.0%
200 glove embeddings	bayesian network	9.6%	9.6%	25.1%	23.8%
200 glove embeddings	naïve bayes	7.6%	9.0%	13.6%	14.9%
200 glove embeddings	SVM	6.4%	3.9%	11.0%	10.2%
200 glove embeddings	random forest	**26.4%**	**27.7%**	49.5%	**50.4%**
200 glove embeddings	LDA	24.8%	25.9%	**52.5%**	49.0%
4700 words tf-idf	bayesian network	34.4%	32.5%	58.7%	55.6%
4700 words tf-idf	naïve bayes	28.6%	26.9%	51.9%	46.8%
4700 words tf-idf	SVM	**38.7%**	**38.9%**	**76.9%**	**74.2%**
4700 words tf-idf	random forest	36.8%	36.9%	69.9%	67.3%
200 words selected LDA	bayesian network	34.2%	32.5%	56.2%	56.5%
200 words selected LDA	naïve bayes	29.0%	26.4%	45.6%	48.1%
200 words selected LDA	SVM	**37.9%**	38.0%	71.0%	70.3%
200 words selected LDA	random forest	37.5%	**39.0%**	**73.5%**	**71.7%**
200 words selected LDA	LDA	35.2%	34.3%	66.9%	64.5%
200 words selected corr	bayesian network	31.5%	30.2%	55.4%	57.3%
200 words selected corr	naïve bayes	22.3%	19.2%	41.8%	46.8%
200 words selected corr	SVM	34.4%	33.6%	69.4%	68.6%
200 words selected corr	random forest	**36.1%**	**36.2%**	**70.7%**	**70.3%**
200 words selected corr	LDA	33.5%	32.6%	64.1%	60.3%
23 feat LDAclass from 200 w selected LDA	naïve bayes	46.4%	47.6%	67.2%	57.0%
23 feat LDAclass from 200 w selected LDA	SVM	**60.0%**	**58.9%**	87.5%	**88.3%**
23 feat LDAclass from 200 w selected LDA	random forest	52.5%	54.5%	**89.3%**	**88.3%**

Table 1: Results of the multinomial text classification with different settings (200 GloVe embeddings, 4700 words tf-idf, 200 words LDA feature selection, 200 words correlation feature selection, 23 LDA predictions as features), algorithms (cNN, rNN, bayesian network, naïve bayes, SVM, random forest and LDA classifier), datasets (noisy, clean) and evaluation methods (10-fold Cross Validation, 70%-30% training test split). The best results for each feature setting are marked in bold.

3.1 Multinomial Classification

We tested different feature settings and algorithms with 10-fold cross validation (10f-cv) and 70%-30% training-testing split in the clean and noisy dataset conditions. Table 1 reports the results in terms of accuracy, that is to say the percentage of documents correctly classified. In both conditions the majority baseline is very low, ranging from 4.6% to 8.3%. First we experimented with pre-trained GloVe word vectors as features (vector size 200). As a matter of fact the GloVe Project provides word vectors of different dimensions for words representation trained on massive web datasets (Pennington et al., 2014). For instance the word vectors we used here have been pre-trained by the GloVe Project from two massive corpora, Wikipedia 2014 and Gigaword 5. As we can see in Table 1 in the GloVe embeddings setting we used the following classification algorithms: cNN (with 2 convolutional layers with ReLU activation, 1 pooling layer and 1 output layer), rNN (with 1 rNN sequence layer, 1 LSTM layer with tanH activation and 1 rNN outpur layer), bayesian networks, naïve bayes, SVMs, random forest and LDA. In general, Deep Learning algorithms suffer from the small data used for the experiment, but surprisingly, cNNs performed badly and rNNs worked better, indicating that the sequentiality of text plays an important role. Among the other classification algorithms it turned out that random forest and LDA obtained the best performances, proving that the ability of the algorithm to generalize is crucial. The general low accuracies obtained with these features might indicate that the contexts of our documents represented by word embeddings are not very discriminative. The results increased significantly in the classification with the TF-IDF scores of 4700 words, especially with SVMs as algorithms. This suggests that using more features brings better results without overfittng the data, as shown by the similar accuracies obtained with a 10-fold cross validation and with training-test split. Next we experimented with feature selection, using LDA and Pearsons' correlations to select the best 200 words for the prediction. Results show that, in this feature setting, random forests are the best classification algorithm and that LDA outperforms correlations as feature selection algorithm. Furthermore, as can be seen in the last part of Table 1, we were able to reach state-of-the-art results with an ensemble learning scheme: using LDA as a classifier we transformed

		noisy	noisy	clean	clean
features	algorithm	acc(10f-cv)	acc(split)	acc(10f-cv)	acc(split)
baseline	majority (zeroR)	4.6%	2.9%	8.3%	6.9%
500 words per label tf-idf selected	scoreranking LDA (1 label)	56.7%	59.9%	52.9%	52.9%
500 words per label tf-idf selected	scoreranking LDA (2 labels)	62.5%	64.2%	63.9%	62.5%
500 words per label tf-idf selected	scoreranking LDA (3 labels)	66.8%	67.5%	68.2%	67.5%
500 words per label tf-idf selected	scoreranking LDA (4 labels)	70.7%	70.3%	73.2%	73.0%
500 words per label tf-idf selected	scoreranking LDA (5 labels)	74.2%	74.4%	76.3%	76.7%
500 words per label tf-idf selected	scoreranking LDA (6 labels)	**79.4%**	77.0%	**79.4%**	78.1%
1000 words per label tf-idf selected	scoreranking LDA (1 label)	56.7%	54.7%	53.9%	62.2%
1000 words per label tf-idf selected	scoreranking LDA (2 labels)	63.6%	61.0%	63.8%	67.2%
1000 words per label tf-idf selected	scoreranking LDA (3 labels)	67.9%	65.4%	68.4%	72.9%
1000 words per label tf-idf selected	scoreranking LDA (4 labels)	71.7%	69.2%	72.8%	78.9%
1000 words per label tf-idf selected	scoreranking LDA (5 labels)	75.0%	73.1%	76.6%	82.6%
1000 words per label tf-idf selected	scoreranking LDA (6 labels)	77.9%	77.4%	**83.7%**	**85.2%**

Table 2: Results of the text classification with different feature settings (500 or 1000 words per label), number of labels in the ranking evaluated (1 to 6 labels), datasets (noisy, clean) and evaluation methods (10-fold Cross Validation, 70%-30% training test split). The best results for each feature setting are marked in bold.

the initial space of 200 word features, previously selected with LDA, in a space of 23 binary features corresponding to the final classes. On top of that we applied different classification algorithms, finding that SVM is the best performing one in the noisy dataset while random forest obtained the best performance in the clean dataset.

3.2 Multi-Label Classification

The Multi-Label classification task is structured as follows: for each document label in the training set, we create a Bag-of-Words (BoW) from the words of its associated documents, then we use TF-IDF scores to weight every word within the BoW obtaining a word ranking that we use for feature selection, since words with higher values better characterize a particular label. Then we apply LDA classification, but unlike the previous experiment, here the prediction returns a list of all the labels, ordered by the total score achieved, we call score ranking this algorithm. Since the classifier returns a list as an outcome, but the editors (our customers) want to choose one or more label from this list, we have to evaluate if the gold standard label occurs in the returned list, thus we can assign multiple labels to a document and test whether the original one is present or not. In this sense, the Score Ranking classifier is evaluated as a Multi-Class classifier (so the metrics in Table 2 are actually Hit@N metrics where N is the size of the returned list), but the returned list is used by the end users to simulate a Multi-Label functionality, leaving to the editors the choice of the best labels to assign among the ones returned. The result of this experiment, reported in Table 2, shows that the performance with 1 label is in line with the ensemble learning setting of the Multinomial classification, but the score ranking system only works well in the noisy dataset, as the results are very similar in both noisy and clean conditions. The performance increases at an average of +3.9% when keeping more than one label. In general, we observe that using 500 or 1000 words per label yield similar results in our small dataset, but using more words can help to capture more nuances in text, that might be useful in larger sets of documents. We also observe that 1000 words per label increase the results in the clean condition, while 500 words per label are enough in the noisy condition.

4 Conclusion and Future

We experimented with various settings, feature selection methods and classification algorithms, and we found a method to extract good models in extremely noisy conditions, even with documents repeated under different labels. LDA proved to be a valuable classification and feature selection technique, but we obtained the best performances when LDA is combined with other algorithms. The results we obtained with the score ranking classification are in line with the state-of-the-art, but our method is more suitable for small and noisy datasets. In the future we plan to apply the score ranking algorithm on a larger dataset and to use it in a real multi-label environment comparing the results with the state-of-the-art of Extreme Multi-Label Document Classification (Chalkidis et al., 2019). We also plan to make comparisons with the more recent state of the art deep learning techniques and to apply semantic indexing to the documents to check for improvements.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*.

Sumeet Agarwal, Shantanu Godbole, Diwakar Punjani, and Shourya Roy. 2007. How much noise is too much: A study in automatic text classification. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 3–12. IEEE.

Suresh Balakrishnama and Aravind Ganapathiraju. 1998. Linear discriminant analysis-a brief tutorial. In *Institute for Signal and information Processing*, volume 18, pages 1–8.

Guido Boella, Luigi Di Caro, Daniele Rispoli, and Livio Robaldo. 2013. A system for classifying multi-label text into eurovoc. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pages 239–240.

Forough Rezaei Boroujeni, Sen Wang, Zhihui Li, Nicholas West, Bela Stantic, Lina Yao, and Guodong Long. 2018. Trace ratio optimization with feature correlation mining for multiclass discriminant analysis. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Danielle Caled, Miguel Won, Bruno Martins, and Mário J Silva. 2019. A hierarchical label network for multi-label eurovoc classification of legislative contents. In *International Conference on Theory and Practice of Digital Libraries*, pages 238–252. Springer.

Cristian Cardellino, Serena Villata, Laura Alonso Alemany, and Elena Cabrio. 2015. Information extraction with active learning: A case study in legal text. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 483–494. Springer.

Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. Ontology population and alignment for the legal domain: Yago, wikipedia and lkif. In *International Semantic Web Conference: Posters Demos and Industry Tracks*, pages 1–4.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019. Extreme multi-label legal text classification: A case study in eu legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–87.

Vidas Daudaravicius. 2012. Automatic multilingual annotation of eu legislation with eurovoc descriptors. In *EEOP2012: Exploring and Exploiting Official Publications Workshop Programme*, page 14.

Felice Dell'Orletta, Simone Marchi, Simonetta Montemagni, Barbara Plank, and Giulia Venturi. 2012. The splet–2012 shared task on dependency parsing of legal texts. In *Semantic Processing of Legal Texts (SPLeT-2012) Workshop Programme*, page 42.

Tommaso Fornaciari, Fabio Celli, and Massimo Poesio. 2013. The effect of personality type on deceptive communication style. In *Intelligence and Security Informatics Conference (EISIC), 2013 European*, pages 1–6. IEEE.

Robert Keeling, Rishi Chhatwal, Nathaniel Huber-Fliflet, Jianping Zhang, Fusheng Wei, Haozhen Zhao, Shi Ye, and Han Qin. 2019. Empirical comparisons of cnn with other learning algorithms for text classification in legal document review. *arXiv preprint arXiv:1912.09499*.

Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, and Giulia Venturi. 2009. Ontology learning from italian legal texts. *Law, Ontologies and the Semantic Web*, 188:75–94.

Zhihui Li, Feiping Nie, Xiaojun Chang, and Yi Yang. 2017. Beyond trace ratio: weighted harmonic mean of trace ratios for multiclass discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2100–2110.

Hendrik Mandelkow, Jacco A de Zwart, and Jeff H Duyn. 2016. Linear discriminant analysis achieves high classification accuracy for the bold fmri response to naturalistic movie stimuli. *Frontiers in human neuroscience*, 10:128.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Monica Palmirani, Michele Martoni, Arianna Rossi, Cesare Bartolini, and Livio Robaldo. 2018. Pronto: Privacy ontology for legal reasoning. In *International Conference on Electronic Government and the Information Systems Perspective*, pages 139–152. Springer.

Shaoning Pang, Seiichi Ozawa, and Nikola Kasabov. 2005. Incremental linear discriminant analysis for classification of data streams. *IEEE transactions on Systems, Man, and Cybernetics, part B (Cybernetics)*, 35(5):905–914.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Simon JD Prince and James H Elder. 2007. Probabilistic linear discriminant analysis for inferences about identity. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE.

Ralf Steinberger, Mohamed Ebrahim, and Marco Turchi. 2013. Jrc eurovoc indexer jex-a freely available multi-label categorisation tool. *arXiv preprint arXiv:1309.5223*.

Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef Van Genabith. 2017. Exploring the use of text classification in the legal domain. *arXiv preprint arXiv:1710.09306*.

Giulia Venturi. 2011. Semantic annotation of italian legal texts: a framenet-based approach. *Constructions and Frames*, 3(1):46–79.