Détermination du parti politique auquel appartient l'orateur

Eve Sauvage
Université Paris Nanterre
41017970@parisnanterre.fr

Abstract

Ce document explore l'analyse d'opinion à partir d'un corpus regroupant un ensemble de débats au Parlement européen en trois langues, le français, l'anglais et l'italien. L'objectif de la tâche est d'assigner correctement les partis politiques aux interventions des parlementaires. Nous utiliserons pour cela plusieurs modèles d'apprentissage artificiel afin de déterminer un modèle idéal.

1 Introduction

La tâche visée par ce document est proposée par l'édition 2009 du défi fouille de texte organisé par l'université Paris Saclay. Cette édition se concentre sur l'analyse d'opinion multilingue et notamment en explorant l'affectation des interventions au parlement européen à des partis politiques. Les corpus utilisés pour l'entraînement et la vérification des modèles sont fournis en français, anglais et italien.

Dans nos expériences, on suppose que la différences entre les langues nécessite que les modèles soient entraînés séparemments. Nous chercherons d'abord à déterminer le meilleur modèle sur le français avant d'essayer d'observer son efficacité sur l'anglais et l'italien. Le modèle choisi sera réentraîné sur les langues en question pour une plus grande adaptabilité.

2 Traitement des données

Les fichiers de corpus sont disponibles au format xml ce qui permet de récupérer aisement leur contenu à l'aide de la bibliothèque python `xml.etree.ElementTree` qui permet d'accéder directement aux noeuds du fichier grâce à leur nom. Les données d'entraînement sont séparées en deux listes ordonnées contenant le texte d'un côté et les étiquettes correspondantes de l'autre.

Nous decidons d'entraîner le modèle sur la totalité des données d'entraînement et de récupérer les annotations de référence pour vérifier nos résultats. Les deux nouvelles listes obtenues sont nettoyées des résultats vides afin d'éviter les erreurs lors du décompte des résultats.

Dans un premier temps, nous realiserons la vectorisation du texte à l'aide du TF-IDF proposé par la bibliothèque `scikit-learn` sans lemmatisation du texte d'origine. Nous réviserons cette approche au vu des résultats.

3 Résultats

3.1 Arbres de décisions

Nous entraînons d'abord les données à l'aide d'un modèle d'arbre de décision : un modèle simple et permettant une visualisation. Ce premier modèle simple constitura notre modèle baseline avec lequel nous comparerons nos prochains essais.

	precision	recall	f-score	support
ELDR	0.72	0.67	0.70	1339
GUE-NGL	0.75	0.75	0.75	1793
PPE-DE	0.76	0.78	0.77	4571
PSE	0.73	0.73	0.73	3627
Verts-ALE	0.69	0.67	0.68	1585
accuracy			0.74	12917
macro avg	0.73	0.72	0.73	12917
weighted avg	0.74	0.74	0.74	12917

Table 1: rapport de classification pour le modèle arbre de décision

Les résultats obtenus sont satisfaisants étant donné la difficulté de la tâche et les résultats d'annotation manuelle présentées par (Groin, 2009). Toutefois, il semble, en visualisant l'arbre de décision obtenu, que le modèle surapprend les données fournies. Ce surapprentissage est lié à

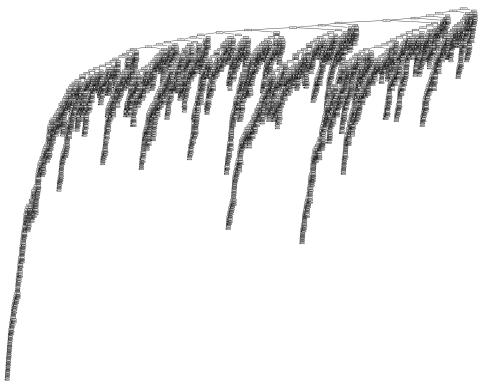

Figure 1: représentation du modèle

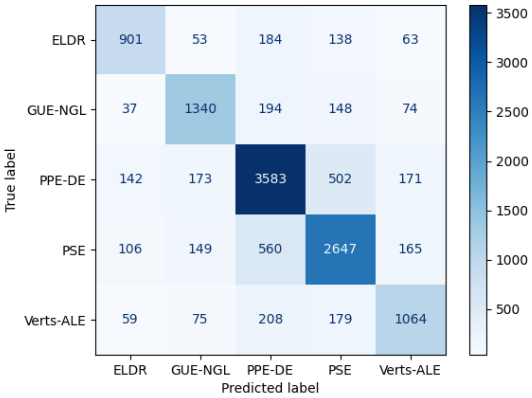

Figure 2: Matrice de confusion pour l'arbre de décision

l'absence de restriction de profondeur ainsi qu'à la multiplicité des paramètres entaîné par l'utilisation d'un TF-IDF pour la vectorisation des données. Il est donc pertinent de réduire le nombre de paramètres pour utiliser pleinement le potentiel de visualisation du modèle.

3.1.1 lemmatisation et suppression des mots stops

Nous supposons que la suppression de la ponctuation ainsi que sa lemmatisation et la suppression des mots stops amélioreront la profondeur du modèle d'arbre de décision. Toutefois, les expériences montrent que les résultats se détériorent et que la profondeur augmente.

3.1.2 Réequilibrage des données

On essaie également de réequilibrer les données d'entrainement afin de supprimer le biais du modèle mais si la précision et le rappel s'améliorent dans les classes minoritaires, les très mauvais résultats dans les classes réequilibrées nous invite à conserver le biais de départ afin de conserver une f-mesure acceptable. Les mauvais résultats dans les classes majoritaires persistent lorsque l'on rééquilibre également les données de tests. En revanche, le rééquilibrage des données permet d'effectivement diminuer la profondeur de l'arbre en utilisant un rééquilibrage par downsampling en passant d'un profondeur de 146 à une prodondeur de 96 grâce au downsampling.

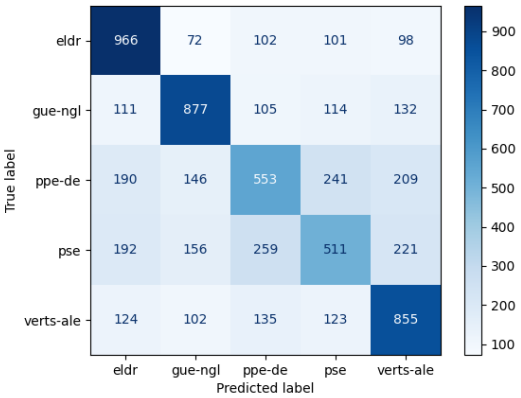

Figure 3: Matrice de confusion avec apprentissage sur des données rééquilibrées en downsampling

Le upsampling permet de moins déteriorer les résultats pour les classes précédemment majoritaire et diminue de ce fait, moins le résultat général. Le modèle entraîné avec le réequilibrage en upsampling présente une f mesure moyenne de 0.65.

3.2 Random Forest

Face à ces résultats peu fructueux, nous essayons d'autres modèles d'apprentissage en commençant par Random Forest qui devrait présenter de meilleurs résultats. En effet, le modèle Random Forest permet de combiner plusieurs arbres de décisions augmentant ainsi leur efficacité.

La f-mesure moyenne de random Forest sans modification des données est effectivement meilleure que celle des arbres de décisions testés précedemment avec un score de 77%. Toutefois, la progression est médiocre, l'augmentation n'étant que d'2%.

Dans le cas de random Forest, la simplification des données semble avoir moins d'impact en particulier dans le cas de la lemmatisation qui ne fait perdre aucun point à la f-mesure moyenne.

	precision	recall	f-score
ELDR	1.00	0.61	0.76
GUE-NGL	0.97	0.71	0.82
PPE-DE	0.64	0.95	0.76
PSE	0.83	0.69	0.75
Verts-ALE	1.00	0.61	0.76
accuracy			0.77
macro avg	0.89	0.71	0.77
weighted avg	0.82	0.77	0.77

Table 2: rapport de classification pour le modèle random forest sans pretraitement

3.3 LinearSVC

On arrive à augmenter le score grâce à un modèle de Classification par Support de Vecteur linéaire.

```
\documentclass[11pt]{article}
```

To load the style file in the review version:

```
\usepackage[review]{acl}
```

For the final version, omit the `review` option:

```
\usepackage{acl}
```

To use Times Roman, put the following in the preamble:

```
\usepackage{times}
```

(Alternatives like txfonts or newtx are also acceptable.)

Please see the LaTeX source of this document for comments on other packages that may be useful.

Set the title and author using `\title` and `\author`. Within the author list, format multiple authors using `\and` and `\And` and `\AND`; please see the LaTeX source for examples.

By default, the box containing the title and author names is set to the minimum of 5 cm. If you need more space, include the following in the preamble:

```
\setlength\titlebox{<dim>}
```

where `<dim>` is replaced with a length. Do not set this length smaller than 5 cm.

4 Document Body

4.1 Footnotes

Footnotes are inserted with the `\footnote` command.[1]

[1]This is a footnote.

Command	Output	Command	Output
`{\"a}`	ä	`{\c c}`	ç
`{\^e}`	ê	`{\u g}`	ğ
`{\`i}`	ì	`{\l}`	ł
`{\.I}`	İ	`{\~n}`	ñ
`{\o}`	ø	`{\H o}`	ő
`{\'u}`	ú	`{\v r}`	ř
`{\aa}`	å	`{\ss}`	ß

Table 3: Example commands for accented characters, to be used in, *e.g.*, BibTeX entries.

4.2 Tables and figures

See Table 3 for an example of a table and its caption. **Do not override the default caption sizes.**

4.3 Hyperlinks

Users of older versions of LaTeX may encounter the following error during compilation:

```
\pdfendlink ended up in
different nesting level
than \pdfstartlink.
```

This happens when pdfLaTeX is used and a citation splits across a page boundary. The best way to fix this is to upgrade LaTeX to 2018-12-01 or later.

4.4 Citations

Table 4 shows the syntax supported by the style files. We encourage you to use the natbib styles. You can use the command `\citet` (cite in text) to get "author (year)" citations, like this citation to a paper by **?**. You can use the command `\citep` (cite in parentheses) to get "(author, year)" citations (**?**). You can use the command `\citealp` (alternative cite without parentheses) to get "author, year" citations, which is useful for using citations within parentheses (e.g. **?**).

4.5 References

The LaTeX and BibTeX style files provided roughly follow the American Psychological Association format. If your own bib file is named `custom.bib`, then placing the following before any appendices in your LaTeX file will generate the references section for you:

```
\bibliography{custom}
```

You can obtain the complete ACL Anthology as a BibTeX file from https://aclweb.org/anthology/anthology.bib.gz. To

Output	natbib command	Old ACL-style command
(?)	\citep	\cite
?	\citealp	no equivalent
?	\citet	\newcite
(?)	\citeyearpar	\shortcite

Table 4: Citation commands supported by the style file. The style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

include both the Anthology and your own .bib file, use the following instead of the above.

```
\bibliography{anthology,custom}
```

Please see Section 5 for information on preparing BibTeX files.

4.6 Appendices

Use `\appendix` before any appendix section to switch the section numbering over to letters. See Appendix A for an example.

5 BibTeX Files

Unicode cannot be used in BibTeX entries, and some ways of typing special characters can disrupt BibTeX's alphabetization. The recommended way of typing special characters is shown in Table 3.

Please ensure that BibTeX records contain DOIs or URLs when possible, and for all the ACL materials that you reference. Use the `doi` field for DOIs and the `url` field for URLs. If a BibTeX entry has a URL or DOI field, the paper title in the references section will appear as a hyperlink to the paper, using the hyperref LaTeX package.

Acknowledgements

This document has been adapted by Steven Bethard, Ryan Cotterell and Rui Yan from the instructions for earlier ACL and NAACL proceedings, including those for ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTeX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

A Example Appendix

This is an appendix.