Topic Modelling vs Distant Supervision: A Comparative Evaluation based on the Classification of Parliamentary Enquiries

DOI:
[978-3-030-30760-8](https://doi.org/978-3-030-30760-8)

Document Version

Accepted author manuscript

[Link to publication record in Manchester Research Explorer](#)

Citation for published version (APA):

Batista-Navarro, R. T., & Hawkins, O. (2019). Topic Modelling vs Distant Supervision: A Comparative Evaluation based on the Classification of Parliamentary Enquiries. In *Digital Libraries for Open Knowledge* (Vol. 11799). (Lecture Notes in Computer Science). Springer Nature. <https://doi.org/978-3-030-30760-8>

Published in:

Digital Libraries for Open Knowledge

Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

Topic Modelling vs Distant Supervision: A Comparative Evaluation based on the Classification of Parliamentary Enquiries^{*}

Riza Batista-Navarro¹ and Oliver Hawkins²

¹ School of Computer Science, University of Manchester, UK

² House of Commons Library, Parliament, UK

riza.batista@manchester.ac.uk, hawkinso@parliament.uk

Abstract. We investigate two different approaches to text classification, by categorising enquiries submitted to the House of Commons Library from elected Members of the UK Parliament. One is an unsupervised approach, i.e. topic modelling, and the other is a supervised approach based on weakly labelled data, i.e. distant supervision. Models were trained on two types of feature sets: one based only on bag of words, and the other combining bag of words with structured metadata attached to enquiries. Our results show that topic modelling obtains superior performance on this task, and that the incorporation of structured metadata as learning features contributes insignificantly to improved model performance.

Keywords: Text Classification · Topic Modelling · Distant Supervision
· Parliamentary Enquiries · House of Commons Library.

1 Introduction

The House of Commons Library is the research service for elected Members of the UK Parliament (MPs). The Library produces independent and impartial research for MPs and their staff on a wide range of public policy issues. Some of this research exists as publicly available briefing material, including in-depth research briefings, short online articles, and interactive data tools. MPs are also able to commission research directly from the Library through a confidential enquiry service. Responding to these enquiries forms the bulk of the Library's day-to-day research work.

MPs' offices can submit enquiries to the Library in person, by phone, by email, or by post. Each enquiry is received by a member of the House of Commons (HoC) staff and is recorded in the Enquiries Database, an internally accessible platform for searching and tracking enquiries. The Enquiries Database has existed in some form since the 1990s, but data has been recorded using the current structure since 2005. For every enquiry, the following fields are recorded:

^{*} Supported by an EPSRC Impact Acceleration Account awarded to the University of Manchester.

date, title, description, the name of the MP whose office has submitted the enquiry, the name of the individual who submitted the enquiry, the purpose or reason for the enquiry, the names of the Library researchers who will respond to the enquiry and their research sections within the Library³.

The title and description fields allow for free-text input and are thus written in natural language. The database does not currently support the manual categorisation or tagging of each enquiry according to any taxonomy of subjects or topics. While the user interface allows for searching by exact keyword matching against words in the free-text fields of the enquiry (i.e. title and description), there is no way to search the database for enquiries by specifying topics of interest. For example, if a database user is interested in retrieving enquiries pertaining to the topic of the European Union, only enquiries that mention “European Union” verbatim will be retrieved (and miss those that use alternative terms, e.g. “Brexit”, “EU Exit”, “EU”).

In this work, we seek to enrich the information stored in the Enquiries Database by automatically tagging each enquiry using the 20 highest-level topics in the UK Parliamentary topic taxonomy⁴. To this end, we explore and compare two approaches to text classification, one unsupervised and the other supervised, namely, topic modelling and distant supervision.

2 Methodology

In this section, we first provide details of the Enquiries Corpus that we constructed for the purposes of developing our text classification methods. This is then followed by an overview of how our machine learning-based models were developed.

2.1 The Enquiries Corpus

From the Enquiries Database, we retrieved enquiries recorded over 10 years, i.e. from the 1st January 2008 up to the 31st December 2017. Out of these enquiries, only those that were allocated to research sections of the Library were retained. This was to eliminate enquiries that could not be classified under topics that typically require research (e.g. simple reference enquiries for things like telephone numbers and addresses, or library book loans). The final corpus contains a total of 125,071 enquiries.

Each enquiry comes with free-text fields (i.e. enquiry title and description) and the following metadata: date, a unique identifier corresponding to the MP whose office submitted the enquiry, a unique identifier corresponding to the

³ There are eight research sections in the Library: Business and Transport Section (BTS), Economic Policy and Statistics (EPAS), Home Affairs Section (HAS), International Affairs and Defence Section (IADS), Parliament and Constitution Centre (PCC), Science and Environment Section (SES), Social and General Statistics (SGS), Social Policy Section (SPS).

⁴ <https://www.parliament.uk/topics/topical-issues.htm>

person who submitted the enquiry, the reason for the enquiry (any of Frontbench Duties, Backbench Duties, Constituent Issue and Unknown), and the names of the Library staff to which the enquiry was allocated (subject specialists who are tasked with responding to the enquiry).

2.2 Topic Modelling

We firstly developed topic models based on Latent Dirichlet Allocation [1]. Here, we explored various values for the number of topics (from $k = 20$ to $k = 45$ in increments of 5), taking into consideration that we are aiming to classify enquiries according to the 20 highest-level topics in the Parliamentary taxonomy. We made use of the Structural Topic Modelling (STM) package [4] in R⁵, which allowed us to investigate the effect of incorporating enquiry metadata as additional features for learning into the topic models.

The resulting topics, each of which is represented as a list of the 20 most frequently occurring keywords within that topic, were then manually assigned labels from the Parliamentary taxonomy by an expert (the second author of this paper), who is a long-time member of the Library’s research staff. This step was done with the topics generated using $k = 45$, as these provided the best coverage of the 20 topics in the taxonomy (i.e. some topics were missing when lower values of k were specified for modelling).

2.3 Distant Supervision

In investigating supervised approaches to text classification, we employed distant supervision, which involves the assignment of ‘weak labels’ to originally unlabelled data [3]. In this case, we obtained weak labels by making use of the Library’s Subject Specialists Directory. This contains a mapping between the names of Library researchers (to whom enquiries are allocated) and their subject areas of specialisation, which in turn can be mapped to the Parliamentary taxonomy. Casting the problem as a multi-class, multi-label classification task, we trained support vector machine (SVM) models [2] with a linear kernel⁶, using a one-vs-all setting. Here, we trained models based on bag of words (BoW), with and without enquiry metadata as additional features for learning.

3 Results

In this section, we present the results of the evaluation of our two proposed approaches. The evaluation was performed against a validation corpus containing 500 enquiries submitted in 2018, which were manually labelled with topics in the taxonomy by our expert⁷. To allow us to assess the reliability of these labels, a

⁵ <http://www.structuraltopicmodel.com/>

⁶ <http://topepo.github.io/caret/index.html>

⁷ Each enquiry was assigned at most two labels or topics from the taxonomy.

subset of 200 enquiries were also manually labelled independently by the first author of this paper. The inter-annotator agreement on this subset is satisfactory, i.e. 85% in terms of F-score.

The results obtained by our methods on the validation corpus are summarised in Table 1. In training our models, two different feature sets were used: one using only bag of words (BoW) and the other combining bag of words with metadata attached to enquiries; namely, reason for enquiry and political party⁸, which are both represented as categorical features (BoW+Metadata).

Table 1. Comparison of Topic Modelling and Distant Supervision Results

	Features	Precision	Recall	F-score
Topic Modelling	BoW	50.25	59.80	54.61
	BoW+Metadata	49.92	60.80	54.82
Distant Supervision	BoW	46.69	56.80	51.07
	BoW+Metadata	47.66	55.00	51.25

Our results show that in general, topic modelling obtained superior performance over distant supervision, in terms of all of the metrics: precision, recall and F-score. Specifically, the topic models trained on features that incorporated not only textual features (BoW) but also enquiry metadata, yielded optimal performance, although the improvement over using only BoW is insignificant.

A similar observation can be made when it comes to the models trained using distant supervision: there is only a slight improvement in the performance of SVM models trained on BoW+Metadata, compared with models trained only on BoW.

4 Conclusion

In this paper, we describe our investigation into the application of one unsupervised and one supervised approach to a text classification task; specifically, the categorisation of MPs’ enquiries according to a 20-class topic taxonomy. In particular, we trained topic models (unsupervised) and support vector machine models (supervised), while examining the effect of incorporating structured metadata as additional features for learning. Our results demonstrate that for the task of classifying enquiries: (1) topic modelling obtains better performance over distant supervision, and (2) the incorporation of structured metadata as features contributes very little to the improvement of model performance.

⁸ The political affiliation of the MP’s office at the time they submitted an enquiry, which was obtained from the UK Parliament’s data platform using the pdpr R package: <https://github.com/olihawkins/pdpr>.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (Mar 2003)
2. Joachims, T.: Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) *Machine Learning: ECML-98*. pp. 137–142. Springer Berlin Heidelberg, Berlin, Heidelberg (1998)
3. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant Supervision for Relation Extraction Without Labeled Data. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*. pp. 1003–1011. ACL '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009)
4. Roberts, M.E., Stewart, B.M., Tingley, D., et al.: STM: R Package for Structural Topic Models. *Journal of Statistical Software* **10**, 1–40 (2014)