Grounding Paradigmatic Shifts In Newspaper Reporting In Big Data. Analysing Journalism History By Using Transparent Automatic Genre Classification.

XML

1. Introduction

This paper shows how the systematic and quantitative study of genre in large digitized newspaper collections sheds light on the development of journalism discourse. We argue that genre conventions that can be discerned in a newspaper text signal the underlying discursive norms and practices of journalism as a profession (Broersma, 2010). However, digital newspaper archives do not contain fine-grained genre information on the article level. As such, this paper adopts a machine learning approach to add genre labels to newspaper articles. Classifying genre in a standardized and reliable manner is challenging, though, because genre is a typical example of a latent content category, which needs considerable interpretation. To ensure the reliability of the results and to evaluate the machine learning approach, it is crucial to make the methodological impact of various machine learning pipelines transparent. This paper therefore discusses how the distribution of news genres developed over time and how transparent classification empowers journalism history researchers to benefit from the computational methods for doing large-scale content analysis.

2. Genre and journalistic discourse

The grand narrative of journalism history, prevalent in journalism historiography, claims that journalism from the end of the 19 th century has moved away from partisan journalism and opinion-oriented reporting practices, towards critical and autonomous journalism that is event-centred and factbased. For the Netherlands specifically, this journalistic development has taken place between the 1950s and the 1990s as part of the depillarization of society (Harbers, 2014). Since genre connects textual features to journalism's underlying discursive norms and routines (Broersma, 2010), a historical analysis of newspaper genres can elucidate how journalism's professional practice has developed over time. We focus particularly on modal genres, which are defined based according to their formal features, e.g. an interview. This definition has been dominant in how journalists and academics have defined journalistic genres in the Dutch context. We have formulated 16 classes of journalistic genres (see figure 1).

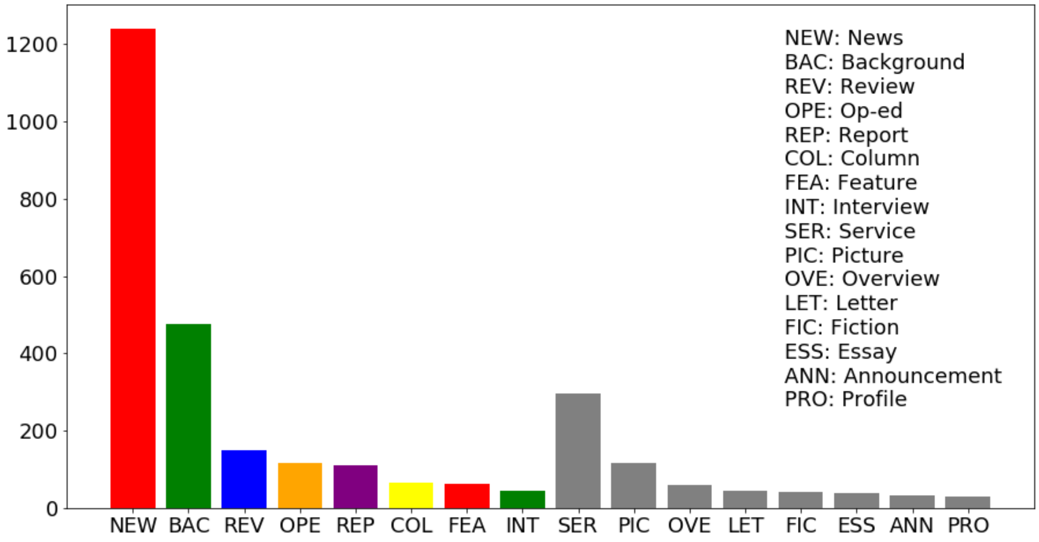

Figure 1 : The distribution of 16 news genres used in this study in a selection of newspaper articles of the Dutch newspaper NRC. Genres of interest to our study are represented by bright colors while other genres are colored grey. The large difference in genre frequencies caused the machine learner to focus on predicting frequent genres correctly while ignoring the others. In order to avoid this problem, we use a balanced training set in which all genres occur equally often

3. Transparent automatic genre classification

Classifying the genres of newspaper articles retrieved from digital newspaper archives raises multiple conceptual and methodological challenges. Firstly, genres are prototypical constructs: some articles do not necessarily match the characteristics of the genre perfectly, nor can they always be neatly delineated from other genres. Furthermore, the distribution of genres over specific topics is skewed, e.g. reports are often about sport matches and interviews about human interest. However, topic is not the defining feature of modal genres and it may introduce unwanted bias in the automatic classification process. To make an informed decision on the right pipeline and understand its inherent biases, we have developed a dashboard that allows the scholar to explore the underlying decision-making process of the machine learning pipeline. On this platform, transparency is achieved through data visualisations that show the performance of the classifier per genre, by offering article-level and classifier-level explanations for the performance, as well as by providing comparison between various machine learning pipelines.

4. The workflow

The platform supports data preprocessing and feature extraction, pipeline configuration, training and testing of the pipeline, pipeline evaluation and comparison, and hypothesis testing (Bilgin et al., 2018). Regarding the data pre-processing, we remove quotes from the documents using regular expressions, to avoid confusion between practices of sources and journalists (the latter being relevant to the genre distinctions). For the feature extraction in order to represent the data, we consider three methods: Bag-of-words (TF-IDF) features, manually curated linguistic features (extracted using a natural language processing toolkit named FROG), such as the number of sentences and the number of first person pronouns, and pooled features combining word features with linguistic features. As for pipeline configuration and training, the platform currently supports 7 machine learning algorithms: GradientBoost, LightGBM, Multi-Layer Perceptron, Naïve Bayes, Random Forest, Support Vector Machines and XGBoost.

5. Pipeline evaluation

The machine learning pipelines are trained on a balanced data set containing 960 annotated articles from 9 different newspapers evenly spread over time, and across topic domains. Through an iterative process, which involves comparing pipelines that use different pre-processing settings in addition to various machine learning algorithms and assessing their underlying decision-making criteria given by interpretability tools, the best pipeline is chosen to perform the genre classification. We measure the performance by commonly used classification metrics such as precision, recall, and accuracy. The accuracy scores vary between 0.41 and 0.70 (compared to 77% inter-annotator agreement). This paper shows, however, that choosing a pipeline based on accuracy score alone does not necessarily result in choosing the best pipeline for our research purpose. Not all genres are equally difficult for a machine learning algorithm to distinguish. Service messages, such as television guides or stock market information, are particularly easy, for example. Yet, these are not the genres we are most interested in for our research questions, since these do not reflect the developing journalistic practices we are interested in.

Pipelines that score higher accuracy scores because they are outperforming others on classifying these less relevant genres, are thus not the best choice for our purpose. More relevant are pipelines that are performing well on the genres we are most interested in. We thus need to evaluate the pipelines beyond their accuracy scores.

Furthermore, looking into the explanations, we noticed that some pipelines tend to do topic classification instead of genre classification, as can be observed by the recurrence of topic-related words in the feature importance rankings for some genres. The predictive features for the decisions of other pipelines, however, do line up more closely with how genres are defined in journalism studies: first personal pronouns (“ik”, “me” and “mij” in Dutch) being predictive for an article to be a column, for example (see Figure 2). The pipelines that combine a bag-of-words approach with manually curated linguistic features give the most promising results. They are able to identify both distinctive words for certain genres - such as “to be continued” (“wordt vervolgd” in Dutch) for literary fiction - and the relevance of linguistic features - such as the number of adjectives which can be predictive for the reportage. Furthermore, these pipelines can recognize words that point at the direction of relevant (linguistic) features that had not been considered yet, contributing to the understanding of genre in journalism history by uncovering patterns that were not known before. Based on the pipeline evaluation and comparison, the preferred pipeline for our studies is chosen.

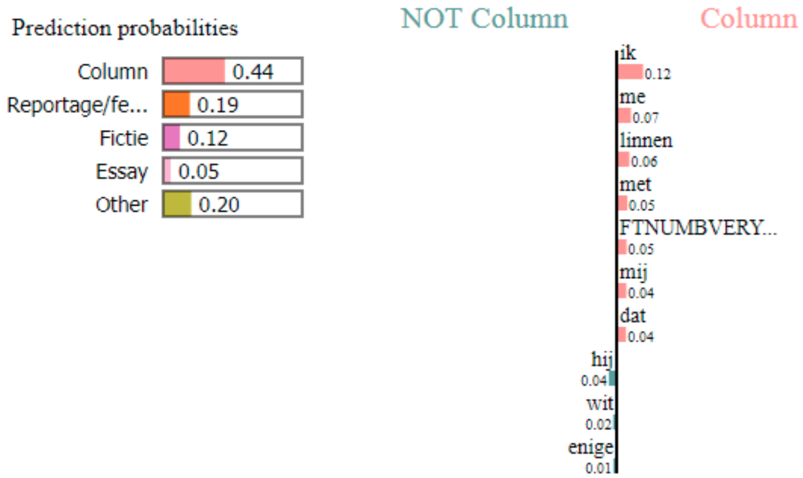

Figure 2 : Local feature explanation by LIME. On the left it shows the probability that the article is a particular genre. On the right it shows the bag of word and curated linguistic features of this article that have been most predictive in classifying it as column, and which have been suggesting that it is not a column. The features, especially the first personal pronouns (“ik” (I), “me” (me) and “mij” (me), line up with the journalism studies understanding of column as an article that is focused on the personal impressions of the author

6. Application of transparent automatic genre classification to journalism history

In order to gain more insight into the development of genre in journalism history, we aim to gain insight in how the machine learning pipeline's output compares to the distribution of the manually coded golden standard data. In other words, the goal is to observe if the machine learner can reproduce the genre distribution for the years for which we have gold standard data. From previous research, we have inherited a data set with approximately 10.000 manually annotated articles for two constructed weeks in 1965 and in 1985 for four newspapers: *Algemeen Handelsblad/ NRC Handelsblad*, *De Telegraaf* and *De Volkskrant*. This data set is used as a golden standard for the distribution of genres in 1965 and 1985. The study compares the distribution of genres in this data set to the distribution of machine-labelled data. In this way, we test the trustworthiness of the results: does the distribution of the machine-labelled data correspond to the distribution noted in the gold standard data?

The future work for this study is to apply the most trusted machine learning pipeline to large-scale unlabelled data to comprehend the development of genre distribution between 1950 and 1995. The results of such work will shed light on the shift from opinion-oriented to fact-centred news in Dutch journalism, marking advances in Dutch media history.

Appendix A

Bibliography

1. **Bilgin, A., Hollink, L., Ossenbruggen, J. van, Tjong Kim Sang, E., Smeenk, K., Harbers, F. and Broersma, M.** (2018). Utilizing a Transparency-driven Environment toward Trusted Automatic Genre Classification: A Case Study in Journalism History. *Proceedings of the 14 th IEEE eScience Conference*, Amsterdam, November 2018.
2. **Broersma, M. J.** (2010). Journalism as Performative Discourse. The Importance of Form and Style in Journalism. In Rupar, V. (Ed.), *Journalism and Meaning-making: Reading the Newspaper*. Hampton Press, pp. 15 – 35.
3. **Harbers, F.** (2014). *Between personal experience and detached information: The development of reporting and the reportage in Great Britain, the Netherlands and France, 1880-2005*. [S.l.]: [S.n.].

Kim Smeenk (k.s.p.smeenk@rug.nl), University of Groningen and Aysenur Bilgin (aysenur.bilgin@cwi.nl), CWI, Amsterdam and Tom Klaver (t.klaver@esciencecenter.nl), Netherlands eScience Center and Erik Tjong Kim

Sang (e.tjongkimsang@esciencecenter.nl), Netherlands eScience Center and Laura Hollink (l.hollink@cwi.nl), CWI, Amsterdam and Jacco van Ossenbruggen (jacco.van.ossenbruggen@cwi.nl), CWI, Amsterdam and Frank Harbers (f.harbers@rug.nl), University of Groningen and Marcel Broersma (m.j.broersma@rug.nl), University of Groningen