Diversity Checker: Toward Recommendations for Improving Journalism with Respect to Diversity

Jeroen Peperkamp
KU Leuven, Department of Computer Science
https://people.cs.kuleuven.be/~jeroen.peperkamp

Bettina Berendt
KU Leuven, Department of Computer Science
https://people.cs.kuleuven.be/~bettina.berendt

ABSTRACT

The Diversity Checker is a tool that aims to make it easier for journalists to author their texts with diversity in mind. To provide helpful hints for them in this respect, it is necessary to define how to quantify diversity so that this can be programmed into the tool. At this early stage in the development of the tool, we present a twofold contribution. First, we offer an analysis on what we mean by "improving diversity". Second, we present the first version of the Diversity Checker, along with some analysis of its current performance.

ACM Reference Format:
Jeroen Peperkamp and Bettina Berendt. 2018. Diversity Checker: Toward Recommendations for Improving Journalism with Respect to Diversity. In *UMAP'18 Adjunct: 26th Conference on User Modeling, Adaptation and Personalization Adjunct, July 8–11, 2018, Singapore, Singapore.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3213586.3226208

1 INTRODUCTION

Neutrality is a widely cherished quality in journalism, and a well-balanced media offering is in many ways like a search engine with diversity[1] built into its result selection and ranking. A unit of media content (for example, an article) can be viewed as a recommendation, to its readers, to learn about, consider, compare and contrast a diverse set of facts and opinions on some issue. It is therefore no coincidence that machine learning and data mining researchers have developed many tools that analyze corpora with respect to some notion of diversity — the idea being that a presentation of the most striking differences, for example in visual form [9, 36], can enhance a reader's perception of differences and critical stance towards "the news".

However, these tools are summative with respect to articles, in the sense that they have already been written, and the task of the search and recommender engine is to encourage readers to choose a diverse media diet. We believe that this approach should be complemented by a formative approach: tools that give recommendations to journalists and other media producers while they produce content, with a view to helping them produce diverse content.

[1]Note, however, that diversity and neutrality are not the same: one can cite a diversity of sources, but still end up skewing the facts through bias.

UMAP'18 Adjunct, July 8–11, 2018, Singapore, Singapore

ACM ISBN 978-1-4503-5784-5/18/07…$15.00
https://doi.org/10.1145/3213586.3226208

The Diversity Checker, whose first prototype is presented in the present paper, is designed to be such a recommender system. It is not realistic to expect to present a complete recommender system that can support journalists in the full generality of their workload, so we limit ourselves here to Dutch language newspaper articles surrounding the theme of the European migrant crisis.

To make things concrete, we offer the following excerpt from an article that appeared in the Flemish quality newspaper *De Standaard* on 25 August 2015[2].

> *Daily record of asylum seekers increases pressure on government*
>
> After the weekend, there are always more asylum seekers who want to register at the Foreigners' Office (DVZ), but yesterday the line was as long as never before. 540 candidate asylum seekers were there. The DVZ's waiting hall can only hold 250–260 people. Result: about as many were sent back onto the street. They received a letter — in Dutch — with the request to register again today. Families with children as well. Normally, these are prioritized, but due to the influx it was not even possible to help them.
>
> The Foreigners' Office even had to call in the police yesterday morning to calm the people in line. "They started fistfights with each other to just get into our waiting hall," says Geert De Vulder, spokesperson of the DVZ. […]
>
> The need for an additional or larger waiting hall is therefore growing by the day. […]

Based on this example, we identify several issues and indicate in which sections we go into more detail about them.

Two questions. When discussing diversity with the example in mind, two questions arise immediately: what is the right unit of analysis, and with respect to what do we examine the diversity?

If we can make the traditional assumption that a set of articles is read together, as in a newspaper, each individual article need not be equally balanced, as long as the ensemble is. Indeed, such a pattern was found by Masini et al.[23], and was also present in the edition from which the above example was taken. On the other hand, a balance is sought also within individual articles in many journalistic genres. In addition, with modern news consumption patterns (e.g. use of social or personalized news aggregators), readers have less of a chance to see a full edition's context. Thus, more diversity needs to be guaranteed at the article level.

[2]Obtained from the Flemish media archive Gopress; own translation.

As for the second question, we note that while a purely lexical analysis can produce interesting results (e.g. Fortuna et al. [13]), it makes sense to defer to the experts: how do media and communications studies experts measure diversity? Humprecht and Esser [15] note that the understanding of diversity in the news is based on the normative claim that a diversity of perspectives should be offered, which can be gauged by analyzing the diversity of political actors, who are usually deemed to have a leading role in determining the direction of the debate (although of course we will not restrict ourselves to political actors). More broadly, the question of which types of actors' views are represented is an important one [4, 23]. For this reason we focus on actor diversity with the current version of our tool, which will be described in more detail in Section 3.

Analyzing the example. With the above in mind, and using a codebook developed by communiction scientists (see below for more details), we can explore the actor diversity of the excerpt and of the surrounding article (not shown here). The article contains a limited number of actors. Only the spokesperson of the Foreigners' Office appears as an individual that is quoted directly. The responsible state secretary is mentioned, but only in that he "was not available (for comment)". Further actors that are active (in the grammatical sense) are institutions (the Foreigners' Office and the police). Asylum seekers appear only as collectives, and when they act, it is by starting fights (otherwise, they "are there", or they "want to register"). Mostly, they are passive (they are being sent onto the street, they are being prioritized, they receive a letter). The popular water metaphor (in the quoted passage: "influx") is used. Arguably, the phrasing of the language mismatch emphasizes the asylum seekers' deficit (they don't even understand Dutch) rather than the Office's deficit (they don't even write their letters in English, French, Arabic, ...). Summarizing, the framing is that of a bureaucratic unit in need, with a bias towards the plight of the public employees in the unit[3].

Conclusions from the example. This short analysis shows three things. First, one could, in a bottom-up or unsupervised fashion, identify that there are diverse actors and could count how often they appear. Second, at least as important as this bottom-up processing is a top-down background ontology that shows what is *not* there (cf. Humprecht and Esser [15]), such as asylum seekers as individuals and as active, or the government (which is the receiving actor in the title, but then does not appear again). Especially if the goal is to give recommendations to an author while they are designing/writing a piece, this knowledge of what is not (yet) there is important. Obviously, for this task to be semi-automated, measures of diversity that operate on such an ontology are required as well, and we will offer an example of how such a measure may be designed in Section 3.

Third, we note that a recommender system for journalists has a normative aspect that may be stronger and certainly more ethically debatable than the normative face of other recommender systems: diversity must not be understood mechanically; it is not a goal in itself. There is generally something, some actor, some viewpoint, that is *not* represented. This choice is a value judgment that may depend on the local context, but also on personal or editorial choices, or society-wide consensus. For example, in current reporting on refugees, xenophobic positions of the far right are not always, but regularly included (and the extreme right is also an actor in the coding ontology we use), while in reporting on the second world war, holocaust deniers and their positions are generally omitted. For the purposes of the present paper, we take a given ontology, see Figure 2 and the next section.

In Section 3, we will describe our approach that rests on three pillars: (a) an ontology, (b) natural-language processing, and (c) diversity measures on ontology and instances, as well as the first version of the Diversity Checker tool. Before that, Section 2 will give a brief overview of related work. Section 4 describes a case study and first evaluation results of the tool. We conclude with an outlook on future work in Section 5.

2 RELATED WORK

Given the above, the task we have set for ourselves is to identify actors in the text and determine whether there is an acceptable balance in the types of actors present. This builds on the task of Named Entity Recognition (NER). This extracts more than just actors, however, it glosses over cases where entities are referred to using generic terms like common nouns or pronouns (e.g. "he", "the police", "the women"). Fortunately, NER is one of the NLP tasks that has seen results published for many languages, not just English [25]. Besides tools that work on multiple languages [1, 19, 29], there are things specifically focused on Dutch [7, 8].

Supposing, then, that we can identify entities in Dutch text, another issue arises, namely that of granularity. Frequently, NER tools have classified entities in very coarse classes of entities, usually including *person*, *organization* and *location*, e.g. [7]. Attempts are being made to classify more fine-grained categories, especially for the person category, e.g. [8, 10, 12, 22]. The way these works arrive at their fine-grained classes is not generally agreed upon [22]; Ling and Weld, for instance, base themselves on tags from Freebase, while Sekine and Nobata [31] constructed a hierarchy using a combination of labels manually assigned to entities extracted from a corpus, old results, and thesauri.

There are also tools like DBpedia Spotlight [24], which annotate entities with links to DBpedia URIs and thus provide an extensive ontology. Another tool called Enrycher [32] also links entities recognized to ontology concepts. Another tool in a similar vein is OpenCalais, see [14]. None of these tools currently works for Dutch, however. We notice two trends: the methods by which fine-grained categories are constructed are often somewhat ad hoc. Particularly the assumptions going into the specific categorization are often left unspecified. Second, naturally the authors of these structures attempt to strike a balance when constructing them between generality and learnability (e.g. Desmet and Hoste [8]).

The fine-grained ontologies referred to above offer some specificity, but they are not necessarily appropriate for our domain. To make meaningful statements about our domain, we need an ontology of social roles and functions. This makes the ontology a key component of our recommender system. We therefore believe it is

[3]Framing and bias are technical terms that open the door to future work; we will discuss them briefly in Section 5.

necessary for someone with expert knowledge on the subject to inform it. In this case, we obtained the ontology from our colleagues in communication science and used it in a two-step process to refine the relevant entities identified by a generic NER system.

3 DIVERSITY CHECKER

Before we can operationalize actor diversity as discussed so far, we need to answer the two questions posed in the introduction: what is the right unit of analysis, and what do we compare to arrive at a verdict about diversity? We choose to analyze individual articles, given that we cannot assume that an ensemble of articles is served as a unit like a traditional newspaper (see Section 1).

As for the second question, we can refer to the literature to get a better grip on how diversity may be measured. For instance, Stirling [33] proposed a general framework in which three individually insufficient qualities jointly give rise to diversity: variety, disparity, balance. He combines these three into a general score that also supports weights. In our case we can fill in the three as follows, with respect to actors: variety is the number of different categories of actors, disparity is the degree to which different categories are different, and balance is the degree to which the set of actors is evenly distributed over the categories. The score is then computed as follows:

$$\Delta = \sum_{i \neq j} (d_{ij})^{\alpha} (p_i p_j)^{\beta},$$

with d_{ij} the disparity between actor types i and j, p_i and p_j the respective fraction of the total number of actors of type i and j, and α and β weights between 0 and 1. For the present paper, we set $\alpha = \beta = 1$. The disparity d_{ij} can be defined as follows, using a technique inspired by Navigli and Velardi [26]. Given an ontology of actor types, d_{ij} is defined as $1 - \ell$, where ℓ is the length of the shortest path between the categories of actors i and j scaled by the length of the longest possible path between any two categories[4]. The latter is computable in linear time because the ontology is a DAG.

A "diversity score" on a single article makes no sense without reference to some standard [15]. A standard one might use could be a set of reference texts which are manually analyzed and deemed to be of sufficient diversity. One could then compute the scores of these texts and use them as a threshold (e.g. via mean or median) on the diversity score of an arbitrary text. This ties in with our earlier comments about normative choices: even if the score is below the threshold, normative arguments may well be able to justify the choices made in constructing the text. We do not speak out about such issues, of course; our tool is merely a recommender, not a decider. Also note that quality is not solely dependent on a diversity score but depends on other aspects as well, like factual accuracy.

Concretely, we set up the following pipeline, see also Figure 1. First, a given text needs to be preprocessed, which is done using Frog [5], a tool that provides tokenization, lemmatization, POS tags, morphological information, NER and both shallow and full parsing for Dutch. The result is a well-organized XML structure in the FoLiA (Format for Linguistic Annotation [35]). Once the input has been structured, it becomes straightforward to extract actor candidates from the text, as these are the entities with the coarse type person. Next, we need to classify the fine-grained actor type according to the ontology. To perform this classification, we apply kNN classification (using Scikit-learn [28]) using a set of articles that were annotated manually. The features used for this classifier are listed in Table 1. Categorical features are mapped to $[0, n-1]$, with n the number of categories. With Boolean features mapped to $[0, 1]$, we obtain a numerical feature vector that can be used to compute similarities according to the Euclidean distance.

Essentially, we use two groups of features: one contains features derived from the grammatical annotations generated by Frog, the other is a set of Boolean variables indicating membership in some word list. Several remarks need to be made here. Firstly, an entity as recognized in the coarse NER step can consist of multiple words, but the grammatical features are defined on single words. In case any entity consists of multiple words, the head h (the word closest to the root of the dependency tree) is used. We also use the head to check for membership of one of the word lists.

The second remark concerns the word lists. We constructed these lists in consultation with our communication science colleagues. They are meant to contain words that correlate with the most common and homogeneous categories. For instance, "police and security services" is a category that occurs frequently and most of the entities with this category are either literally the word 'police' or otherwise contain this word. Hence, this is one of the words on the list for this category. Conversely, the way a single person is individuated is often by their name and/or occupation, making the generic category of "other" under citizen unsuited for the construction of a word list.

Finally, the diversity score is computed as outlined above. As mentioned, p_i is the number of occurrences of actor category i and d_{ij} is the minimum length of a path in our ontology between categories i and j; both p_i and d_{ij} are normalized. For each recognized actor, the classifier determines the category. Once all the categories occurring in a given article are known, the sum for Δ shown above can be computed. The diversity score for an article is this Δ, which we normalize by the number of words in the article.

4 CASE STUDY

We now report on a case study to test the performance of the tool and demonstrate its abilities. To evaluate the classification performance, we need an annotated corpus and an ontology. We obtained a small corpus from our colleagues in communication science. This corpus consists of 99 articles from the Belgian newspaper *De Standaard*, all in some way related to the refugee and migration crisis. The articles were published between 6 November 2015 and 30 January 2016. Actors and their types are marked in the text. One of the articles was excluded due to poor formatting. The rest of the corpus contains 57,207 tokens in 3,264 sentences. Our colleagues also provided us with an ontology, partially shown in Figure 2.

The full ontology consists of a tree containing 235 nodes and 176 leaves. Given the large number of leaves, we performed several tests in which we cut off the tree at progressively lower levels, giving runs on 2-, 3- and 4-level versions, as well as the full ontology. We ran the analysis on each of the articles in the corpus,

[4] Stirling makes no assumptions about any structuring of the available categories. The disparity factor enables us to capture this structure. Note, however, that just a naive path length as used here will not give the best results when the hierarchy of categories is lopsided.

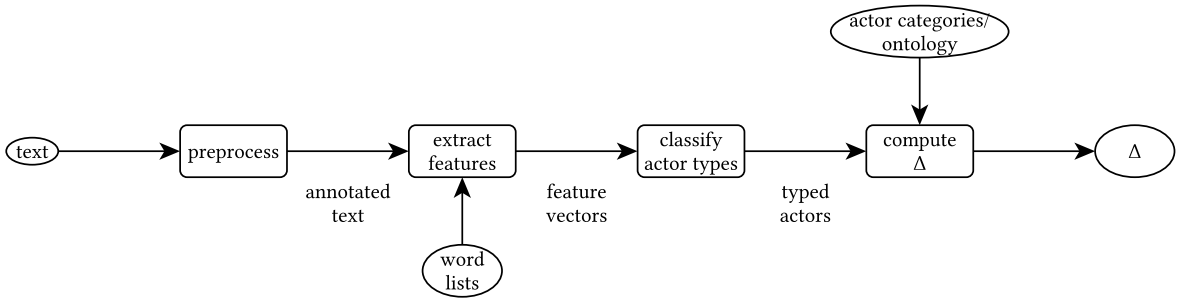

Figure 1: The architecture of the Diversity Checker. Text goes in, is preprocessed by Frog and has features extracted based on the annotated text and the word lists. The feature vectors are then used to determine actor types, which are then analyzed according to the ontology to compute Δ. For the ontology of actor categories, see Figure 2.

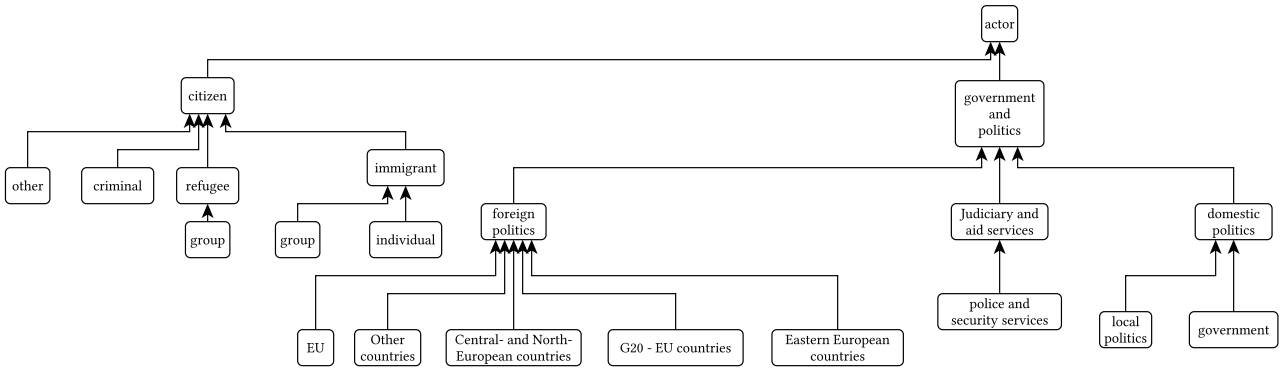

Figure 2: A cropped version of our ontology showing the categories with non-zero F_1 scores in the case study in Section 4 down to the third level.

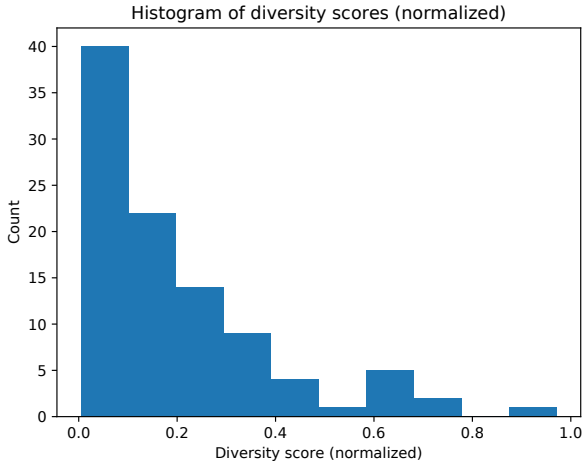

Figure 3: Diversity score Δ divided by number of words.

then computed the precision, recall and F_1 scores. In Table 2, we report these scores. Some categories had insufficient support (in some cases none at all), and $F_1 = 0$. These have been omitted.

Several things are worth noting. Since we use a simple kNN classifier, the hierarchical structure of the ontology is not taken into account in classification or scoring: the right category is the most specific one available, and any other category counts as an error. Furthermore, the hierarchy is not balanced, as there are many more categories under politics than under the business and citizens. This results in some of the shallow leaves having scores for multiple runs of the experiment with different ontology depths. These scores differ by only small amounts, which shows the algorithm is robust as its classification does not change when more options are added. (In the table, we only show the score for the first run during which a category occurs.)

We also give an overview of the averages of the scores reported in Table 2, see Table 3. These are macro-averaged scores, weighed by support because the categories are unbalanced. Comparing the two tables, we can see that the average values quickly diminish when the ontology becomes deeper, as can be expected when its structure is not taken into account. We note, however, that several

Table 1: Features used to classify an entity *e*'s actor type. The head of a multi-word entity is *h*.

Feature	Description
entity POS	(use *h*'s POS for multi-word entities)
POS of parent	the POS of *e*'s parent in the dependency tree
dependency relation	the relationship *e* has with its parent, can be unknown if the parser failed
word count	number of words *e* consists of
aid services	*e* is on the word list for aid services
army	*e* is on the word list for the armed forces
criminal	*e* is on the word list for criminals
judiciary	*e* is on the word list for the judiciary
law enforcement	*e* is on the word list for law enforcement
migrant group	*e* is on the word list for migrants as group
other citizen	*e* is on the word list for citizens not belonging to any other category
protester	*e* is on the word list for protesters
refugee	*e* is on the word list for refugees
EU bodies	*e* is on the word list for EU bodies
vox pop	*e* is on the word list for vox pops

classes maintain a relatively high score, for which two explanations are possible. Firstly, there is quite a strong correlation between the number of examples available for a class and its score, which is quite logical. Second, there are some classes that have a surprisingly high score, given the number of samples for them, which occurs particularly with the class of police and security services. We inspected the annotations and found that most actors with that category were simply the word 'politie' (police).

The classification of actors is but the first step in measuring diversity, and given the current recognition quality, the values computed for Δ would contain a substantial margin of error. To determine an upper bound for performance and to understand the diversity characteristics of our corpus, we therefore systematically computed Δ for each article from the ground-truth annotations. Figure 3 shows Δ normalized by article length. This shows that many articles score quite poorly on diversity. We manually inspected some of the articles with the lowest and highest values of Δ. One with a very low score turned out to be an interview, for which the focus is naturally narrower. Another was a wordy piece citing various people at the station in Cologne after the infamous mass sexual assault during the New Year's Eve celebrations, where most of the people cited were miscellaneous citizens. Again the setting may explain some of the low diversity, but the fact remains that the people cited were not very diverse. This could also be an issue with the ontology used, which offers relatively few subcategories of citizens so that a generic cross-section (which is what the article seemed to aim at) is mostly categorized under "other citizen". Note that this choice, made when constructing the ontology, represents a normative decision, namely that it matters less to discern different types of citizens than it does to discern certain other types of actors in a more fine-grained fashion (notably, political actors).

Table 2: Precision (p), recall (r), F_1 and support (s) after running the tool on a progressively deeper ontology. The level column gives the depth of the ontology at which a given category appears.

category	p	r	F_1	s	level
government + politics	0.73	0.85	0.78	326	2
domestic politics	0.23	0.33	0.27	67	3
local politics	0.03	0.07	0.04	14	4
government	0.16	0.16	0.16	49	4
Secretary for Asylum and Migration	0.12	0.19	0.15	27	6
Belgian chamber of people's representatives	0.50	0.33	0.40	3	5
foreign politics	0.57	0.73	0.64	225	3
EU	0.25	0.14	0.18	38	4
EU commission	0.20	0.25	0.22	12	5
other EU bodies	0.06	0.14	0.08	7	5
G20 - EU countries	0.20	0.39	0.26	66	4
politician Germany	0.09	0.26	0.13	23	6
government Denmark	0.33	0.17	0.22	6	6
Eastern European countries	0.09	0.23	0.13	22	4
Central and Northern European countries	0.08	0.05	0.06	22	4
judiciary and aid services	0.75	0.44	0.56	34	3
police and security services	0.71	0.63	0.67	27	4
citizens	0.77	0.71	0.74	257	2
refugee	0.51	0.61	0.55	94	3
group of refugees	0.50	0.70	0.58	82	4
immigrant	0.49	0.37	0.42	105	3
group of immigrants	0.41	0.42	0.42	80	4
individual immigrant	0.41	0.20	0.27	25	4
criminal	0.75	0.12	0.20	26	3
other citizen	0.18	0.08	0.11	25	3

On the positive side, several articles with a high score did indeed cite sources from multiple different corners of the ontology. We do note, however, that with a complex issue like the migrant crisis, it can be easy to simply mention a wide variety of actors while still ending up representing a relatively small number of viewpoints[5]. Of course we saw examples of this as well.

5 SUMMARY, CONCLUSIONS AND FUTURE WORK

In this paper, we have outlined one way of studying diversity in the media, namely by looking at the distribution of actors in the content on offer. To that end we have introduced the first version of the Diversity Checker, a tool that is meant to analyze this distribution and that will ultimately be able to recommend improvements on a given text to help its 'diversity score'. We performed a case study on a small corpus of Flemish newspaper articles about the

[5]After all, while viewpoint diversity can be measured through actor diversity, the latter is not a perfect proxy for the former [4].

Table 3: Average score and missed values for each depth d tested. We also list the number of available actors that were missed by the system due to low support in the m column. d is the depth of the ontology for the given values.

	p	r	F_1	s	m
$d = 2$	0.70	0.74	0.72	620	37
$d = 3$	0.47	0.49	0.46	620	44
$d = 4$	0.27	0.27	0.25	620	112
$d =$ full	0.23	0.23	0.21	620	233

European migrant crisis, and found that indeed there is room for improvement of the diversity, but also of the Diversity Checker.

What do we learn from the case study? Superficially it turns out there is indeed a need for a Diversity Checker, or some tool at least to improve the poor scores we see in Figure 3. We do note, however, given what we saw in the manual inspections, that the tool in its current form is not sufficiently intelligent. It became clear enough that sometimes, there are clear explanations for a low diversity score, including deficiencies in the tool itself.

We see several points for improvement. For instance, we will need to employ coreference resolution to detect actors more accurately. This will improve counting actors by enabling us to more accurately include oblique references such as pronouns. It is generally sensible to look at the context in which actors are mentioned. For instance, when someone is cited, it could be positively (e.g. to support an argument) or negatively (e.g. as something being argued against). Furthermore, different machine learning techniques can be explored, such as support vector machines.

However, there are also a few more general features we would want to add. When it comes to journalistic text, there are two phenomena we are interested in that play a role in affecting the quality of the content with respect to diversity, namely framing and bias. Framing is the phenomenon where frames are used to highlight the salient aspects of an issue [11], thereby potentially de-emphasizing other aspects. In the example above, the difficulty to process asylum seekers is highlighted while the effect on the people themselves is largely ignored: asylum seekers are mostly passive and attention is shifted to bureaucrats (e.g. "*our* waiting hall", emphasis ours). Automatic frame analysis has been studied [2, 3, 6, 16, 18, 34], and we would continue that line of research.

Bias is a subtly distinct phenomenon that is sometimes conflated with framing (e.g. "framing bias" is one type of bias identified in [30]). We take bias to manifest in low-level things like word choice, which can contribute to the overall framing. In the example, for instance, we would call 'influx' a biased term, as it encodes refers to the idea that the country is being drowned by a tidal wave of foreigners. Automatic bias detection has also been studied [13, 20, 21, 27, 30], and we would like to contribute to that area by disentangling the concepts of framing and bias.

A key practical question will be how to proceed from an NLP analysis that outputs an annotated text and an evaluative diversity score to a tool that is actually used by journalists. Our target users are highly skilled professionals whose very self-concept derives from what and how they write. A healthy self-confidence of "I know how to write, and my reporting is fair and accurate" (see https://fair.org/) is a prerequisite of the job, even if it should be — and usually is — balanced by a healthy self-consciousness about the inherent limits of such statements. So a recommender approach must be chosen, and a user interface designed, with care. An autocompleting recommender that may work well for spell checking or searches would probably not be appropriate in this setting. But what would be? We note that everyone has blind spots, and it would be useful to be able to alert journalists to things they may have missed. In preliminary interviews with some professionals, we found they would indeed be interested to see if they missed things, and also whether there are trends over time (e.g. relative underrepresentation of a minority group's interests).

Approaches that come to mind and that do not need to exclude each other are (a) dashboard-like interfaces that can show, for the current text version, the annotations as well as a diversity score derived from them, (b) reference texts and thresholds for comparison, as suggested in Section 3 above, or (c) more open-ended interfaces such as on-the-fly searches for texts that are similar in content but different in diversity (inspired by [13]). The displays could focus attention on what is there and/or what is *not* there, namely (d) ontological categories that are missing — where the choice of what to display as missing-but-complementary would need to be well calibrated and tested. Given any of these, (e) should analysis components allow for configuration and what-if analyses? Besides the preliminary interviews mentioned above, we will have more in-depth consultations to fine-tune the tool as it matures.

While we have stated that we want to build "a recommender, not a decider", we are aware that however non-directive we try to be (e.g. by aiming for a transparency tool or a nudge, see [17]), we cannot avoid baking value judgments and perceptions of the world into the ontologies, formulae, and functions we use, and we will transport certain values by the very fact of offering such a helper tool in the first place. Open-ended discussions in the requirements elicitation and evaluation sessions will be one way to work on this, but we would also like to build into the tool itself invitations to challenge the tool's assumptions and us who — whether through explicit or implicit bias — built these assumptions into the software.

Finally, it is conceivable that authors could game the algorithm in its current form, writing a piece with many types of actors that humans would not judge as diverse. This shows the multi-faceted nature of diversity: in such a situation, diversity of social classes could for instance obscure a lack of diversity of viewpoints, or a biased way of representing the viewpoints. The current version of the algorithm fails to detect such cases, but the planned improvements should increase robustness against them.

Through this range of investigations at different levels from text-analysis algorithms to usage settings, we hope to contribute to, and help both broaden and focus, the currently ongoing discussions about just what diversity is and should be.

6 ACKNOWLEDGEMENTS

This work is part of the DIAMOND project, funded by FWO. We thank our colleagues Kathleen Beckers and Peter Van Aelst for providing the annotated corpus, and the ontology.

REFERENCES

[1] Rodrigo Agerri and German Rigau. 2016. Robust multilingual Named Entity Recognition with shallow semi-supervised features. *Artificial Intelligence* 238 (2016), 63–82. https://doi.org/10.1016/j.artint.2016.05.003

[2] F. C. Barboza, H. Jeong, K. Kobayashi, and S. Shiramatsu. 2013. An Ontology-Based Computational Framework for Analyzing Public Opinion Framing in News Media. In *2013 IEEE Int. Conf, Systems, Man, and Cybernetics* (2013-10). 2420–2426. https://doi.org/10.1109/SMC.2013.413

[3] Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proc. 2015 Conf. NAACL: HLT*. 1472–1482.

[4] Rodney Benson and Tim Wood. 2015. Who Says What or Nothing at All? Speakers, Frames, and Frameless Quotes in Unauthorized Immigration News in the United States, Norway, and France. *American Behavioral Scientist* 59, 7 (June 2015), 802–821. https://doi.org/10.1177/0002764215573257

[5] Antal van den Bosch, Bertjan Busser, Sander Canisius, and Walter Daelemans. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. *LOT Occasional Series* 7 (2007), 191–206.

[6] Björn Burscher, Daan Odijk, Rens Vliegenthart, Maarten de Rijke, and Claes H. de Vreese. 2014. Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis. *Communication Methods and Measures* 8, 3 (2014), 190–206. https://doi.org/10.1080/19312458.2014.937527 arXiv:https://doi.org/10.1080/19312458.2014.937527

[7] Fien De Meulder, Walter Daelemans, and Veronique Hoste. 2002. A named entity recognition system for Dutch. In *Language and Computers : Studies in Practical Linguistics* (2002), Vol. 45. Rodopi, 77–88. http://hdl.handle.net/1854/LU-598016

[8] Bart Desmet and Véronique Hoste. 2014. Fine-grained Dutch named entity recognition. *Language resources and evaluation* 48, 2 (2014), 307–343.

[9] Nicholas Diakopoulos, Amy X. Zhang, and Andrew Salway. 2013. Visual analytics of media frames in online news and blogs. In *In Proc. IEEE InfoVis Workshop on Text Visualization* (2013).

[10] Asif Ekbal, Eva Sourjikova, Anette Frank, and Simone Paolo Ponzetto. 2010. Assessing the Challenge of Fine-grained Named Entity Recognition and Classification. In *Proc. 2010 Named Entities Workshop* (2010) *(NEWS '10)*. ACL, 93–101. http://dl.acm.org/citation.cfm?id=1870457.1870472

[11] Robert M. Entman. 1993. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication* 43, 4 (December 1993), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

[12] Michael Fleischman and Eduard Hovy. 2002. Fine Grained Classification of Named Entities. In *Proc. 19th Int. Conf. on Comp. Ling. - Vol. 1* (2002) *(COLING '02)*. ACL, 1–7. https://doi.org/10.3115/1072228.1072358

[13] Blaž Fortuna, Carolina Galleguillos, and Nello Cristianini. 2009. Detection of Bias in Media Outlets with Statistical Learning Methods. In *Text mining: classification, clustering, and applications*. 27–50.

[14] Aldo Gangemi. 2013. A Comparison of Knowledge Extraction Tools for the Semantic Web. In *The Semantic Web: Semantics and Big Data*, Philipp Cimiano, Oscar Corcho, Valentina Presutti, Laura Hollink, and Sebastian Rudolph (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 351–366.

[15] Edda Humprecht and Frank Esser. 2017. Diversity in Online News. *Journalism Studies* (April 2017), 1–23. https://doi.org/10.1080/1461670X.2017.1308229

[16] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political Ideology Detection Using Recursive Neural Networks. In *Proc. 52nd Annual Meeting ACL (Volume 1: Long Papers)*. ACL. docs/2014_acl_rnn_ideology.pdf

[17] Anthony Jameson, Bettina Berendt, Silvia Gabrielli, Federica Cena, Cristina Gena, Fabiana Vernero, Katharina Reinecke, et al. 2014. Choice architecture for human-computer interaction. *Foundations and Trends® in Human–Computer Interaction* 7, 1–2 (2014), 1–235.

[18] Soo-Min Kim and Eduard Hovy. 2006. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text (SST '06)*. ACL, Stroudsburg, PA, USA, 1–8. http://dl.acm.org/citation.cfm?id=1654641.1654642

[19] Michal Konkol, Tomáš Brychcín, and Miloslav Konopík. 2015. Latent semantics in named entity recognition. *Expert Systems with Applications* 42, 7 (2015), 3470–3479. https://doi.org/10.1016/j.eswa.2014.12.015

[20] Sicong Kuang and Brian D Davison. 2016. Semantic and context-aware linguistic model for bias detection. In *Proc. of the Natural Language Processing meets Journalism IJCAI-16 Workshop*. 57–62.

[21] K. Lazaridou, R. Krestel, and F. Naumann. 2017. Identifying Media Bias by Analyzing Reported Speech. In *2017 IEEE Int. Conf. on Data Mining (ICDM)*. 943–948. https://doi.org/10.1109/ICDM.2017.119

[22] Xiao Ling and Daniel S. Weld. 2012. Fine-grained entity recognition. In *In Proc. 26th AAAI Conf. Artificial Intelligence* (2012).

[23] Andrea Masini, Peter Van Aelst, Thomas Zerback, Carsten Reinemann, Paolo Mancini, Marco Mazzoni, Marco Damiani, and Sharon Coen. 2017. Measuring and Explaining the Diversity of Voices and Viewpoints in the News: A comparative study on the determinants of content diversity of immigration news. *Journalism Studies* (July 2017), 1–20. https://doi.org/10.1080/1461670X.2017.1343650

[24] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*. ACM, 1–8.

[25] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30, 1 (2007), 3–26. https://doi.org/10.1075/li.30.1.03nad

[26] R. Navigli and P. Velardi. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Trans. Pattern Analysis and Machine Intelligence* 27, 7 (July 2005), 1075–1086. https://doi.org/10.1109/TPAMI.2005.149

[27] Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. QUOTUS: The Structure of Political Media Coverage As Revealed by Quoting Patterns. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 798–808. https://doi.org/10.1145/2736277.2741688

[28] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. 12 (October 2011), 2825–2830. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[29] Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, Irina Temnikova, Anna Widiger, Wajdi Zaghouani, and Jan Zizka. 2006. Multilingual person name recognition and transliteration. *arXiv preprint cs/0609051* (2006).

[30] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language.. In *ACL (1)*. 1650–1659.

[31] Satoshi Sekine and Chikashi Nobata. 2004. Definition, dictionaries and tagger for Extended Named Entity Hierarchy. In *Proc. Conf. on language resources and evaluation* (2004). 1977–1980.

[32] Tadej Štajner, Delia Rusu, Lorand Dali, Blaž Fortuna, Dunja Mladenić, and Marko Grobelnik. 2009. Enrycher: service oriented text enrichment. *Proc. of SiKDD* (2009).

[33] Andy Stirling. 2007. A general framework for analysing diversity in science, technology and society. 4, 15 (August 2007), 707–719. https://doi.org/10.1098/rsif.2007.0213

[34] Oren Tsur, Dan Calacci, and David Lazer. 2015. A Frame of Mind: Using Statistical Models for Detection of Framing and Agenda Setting Campaigns.. In *Proc. 53rd Annual Meeting ACL and 7th Int. Joint Conf. NLP*. ACL, 1629–1638.

[35] Maarten van Gompel and Martin Reynaert. 2013. FoLiA: A practical XML format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal* 3 (12/2013 2013), 63–81.

[36] Jasna Škrbec, Marko Grobelnik, and Blaž Fortuna. 2012. Exploring History Through Newspaper Archives. In *The Semantic Web: ESWC 2012 Satellite Events* (2012-05-27) *(Lecture Notes in Computer Science)*. Springer, Berlin, Heidelberg, 366–370. https://doi.org/10.1007/978-3-662-46641-4_28