Creating A Disability Corpus for Literary Analysis: Pilot Classification Experiments

Ryan Dubnicek[1] , Ted Underwood[1], and J. Stephen Downie[1]

[1] School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL USA
rdubnic2@illinois.edu

Abstract. As literary text opens to researchers for distant reading, the computational analysis of large corpora of text for literary scholarship, problems beyond typical data science roadblocks, such as data scale and statistical significance of findings have emerged. For scholars studying character and social representation in literature, the identification of characters within the given classes of study is crucial, painstaking, and often a manual process. However, for characters with disabilities, manual identification is prohibitively difficult to undertake at scale, and especially challenging given the coded textual markers that can be used to refer to disability. There currently exists no corpus of characters in fiction with disabilities, which is the first step to at-scale computational study of this topic. This project seeks to pilot a classification process using manually assigned ground truth on a subset of volumes from the HathiTrust. Having successfully built and evaluated a Naïve Bayes classifier, we suggest full-scale deployment of a statistical classifier on a large corpus of literature in order to assemble a disability corpus. This project also covers preliminary exploratory textual analysis of characters with disabilities to yield potential research questions for further exploration.

Keywords: Distant Reading, Digital Humanities, Disability in Fiction, HathiTrust

1 Context

The emergence of distant reading—a term coined by Franco Moretti, and generally defined as computational analysis of large corpora of text for literary scholarship [1]—has allowed the study of character depictions in literature to extend to general investigation across thousands of books published over hundreds of years, as opposed to tens of books over a specific literary era. This project is focused on using distant reading techniques to identify and study characters with disabilities in fiction, a relatively unexplored area of study, and one with unique challenges. Beyond the general snares of data availability and format, the study of characters with disabilities has an additional significant impediment: less-reliable manual methods of identification and the absence of any substantial list of such characters in fiction. In fact, much of the focus on disability in literature is confined to the area of children's and young adult fiction. Representations of individuals with disabilities in media can affect societal views of disability [2], and the absence of a corpus of fiction with

characters with disabilities prevents further study into this potentially impactful realm. This project seeks to pilot a potential process to fill the gap, using a statistical classification approach to programmatically identify and analyze characters with disabilities in fiction, and thus better unlock these characters and volumes to literary study, both traditional and computational.

2 Methods

Using manual identification of characters with disabilities, we compiled a list of 28 novels that had at least one character with a disability. For this project, we're using the World Health Organization's broad definition of disability [3], which includes physical and mental/intellectual disabilities, chosen to help ensure enough data could be gathered for training the classifier and to allow for more generalized exploratory analysis and future study across sub-classes of disability. The volumes in question span publication dates from 1831 to 2004, with a representative volume from each decade 1900 to 2010. Example characters include Lennie Small (*Of Mice and Men*), Quasimodo (*Notre-Dame de Paris*) and Christopher John Francis Boone (*The Curious Incident of the Dog in the Night Time*).

Once a set of volumes and characters was assembled, BookNLP [4]—a Natural Language Processing (NLP) pipeline specifically scaled to long pieces of text, such as books—was used to extract all characters, and the words most associated with them, from each volume. BookNLP extracts words associated with each character in five categories—modifiers, dialogue, verbs performed by the character, objects of the character and verbs used in reference to the character. Output from BookNLP was manually spot-checked to ensure accuracy.

Sets of characters with and without disabilities were then balanced based on both character count and word count between each classification. This yielded a final dataset with a total of 119 characters, including 7,079 words for characters with disabilities and 5,819 words for characters without. A Naïve Bayes classifier was then built, trained and employed, using a five-fold cross-validation process to ensure accuracy. Naïve Bayes was chosen both for its relative simplicity in construction and for its generally high accuracy [5], the latter a potential strength in identifying characters with uncertain textual markers [6].

3 Findings

The classifier was able to identify the characters with disabilities to a high level of accuracy. On average, a character was correctly identified 92% of the time as having or not having a disability. As a control, character classification based on gender (also manually assigned for ground truth) was conducted with a 100% success rate, unsurprising for a smaller test corpus. These results suggest deploying a statistical classifier over a large-scale literary corpus, such as that made available by the HathiTrust, may be a fruitful next step in building a corpus of literary characters with disabilities, and volumes in which they appear.

TOP VALUES:	BOTTOM VALUES:
mustache 18.472522071976197	mother -195.4466664158344
bargain 18.472522071976197	thought -68.26984216214623
dad 16.8769550880802	father -65.98077233986645
throat 15.750153629790725	wondered -43.82090999202654
gaze 11.40223703226242	room -37.37774220703629
husband 10.791808350060874	asked -34.78775554608738
tie 10.514841379831875	cried -27.14499733824231
responded 10.514841379831875	bed -23.79884462181867
girlfriend 10.514841379831875	sure -20.002466446793548
beard 10.514841379831875	rang -17.988994366829676
boots 10.396339323018275	looks -17.671480206029642
chest 9.674982526579017	aunt -15.649644173063631
house 9.080016359952728	doing -14.209596105961015
spectacles 8.932228983961625	wheeled -13.314330036640637
rasped 8.932228983961625	cap -13.263943662451524

Fig. 1. Top and bottom words associated with characters without disabilities.

Having extracted the character data using BookNLP, we then performed exploratory term frequency analysis between characters with and without disabilities, using Dunning's log likelihood, as well as sentiment analysis using TextBlob [7]. Within the top 50 words associated with each class, the findings unexpectedly showed an overrepresentation of verbs associated with characters with disabilities and a comparable number of references to the body for both classes, as shown in Fig. 1.

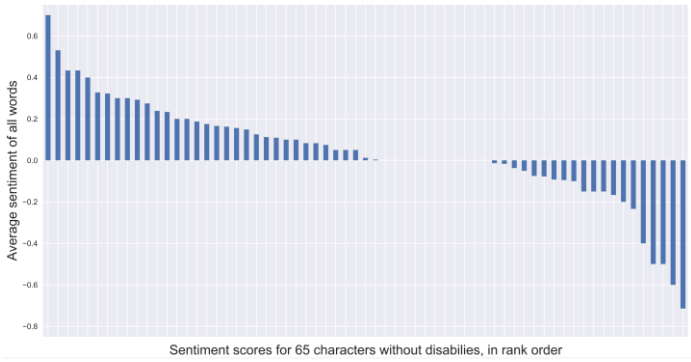

Fig. 2. Average sentiment of words for 65 characters without disabilities.

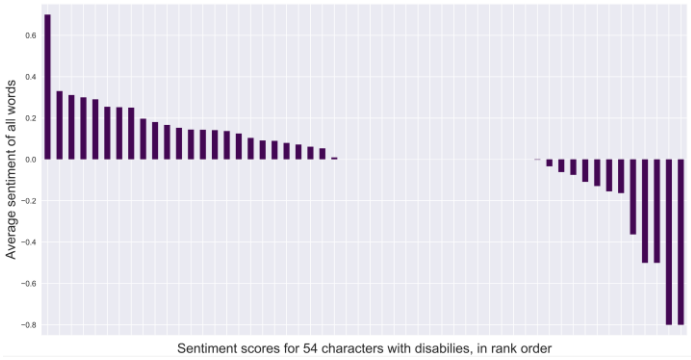

Fig. 3. Average sentiment of words for 54 characters with disabilities.

Sentiment analysis demonstrated a starkly different average sentiment for words associated with each class. Characters with disabilities have an average sentiment 2.5 times more negative than characters without. Full sentiment distribution, by character and class are shown in Fig. 2 and Fig. 3. Distribution of sentiment for both classes was relatively irregular, and TextBlob's objectivity score for all sentiment values was nearly even between classes, with overall average objectivity for both below 50%.

4 Future Work

The obvious next step is to run full-scale classification of every character in the fiction subset of HathiTrust and release of results, both character data and volume identifiers, after verification. Such a release could enable wide and general use of the data for diverse research. Further, having a larger dataset to explore would allow for more substantial testing of the pilot analysis of sentiment and term frequency between classes, which could lead to broader statements and observations about trends in fiction that primarily have been impossible to explore at scale. This pilot project and any resulting corpus could also serve as a first step to finer-grained identification of characters with disabilities in literature.

While literary scholars may be most interested in the direct findings of a large-scale application of this pilot study, the resulting textual analysis of this data would have broad appeal across the social science and humanities study. Some scholars even point to depictions of characters with disability in media as a factor that fosters real-world discrimination [8], an interesting claim difficult to test without first enabling and understanding characterization at scale.

Specific to data science in the humanities, this process would provide a chance to evaluate Naïve Bayes, or other statistical classifiers, when deployed against a large-scale dataset where ground truth may not be possible to assign. Employing this type of semi-supervised machine learning could prove a useful test case for other types of classification where textual markers are unclear, such as sexuality or race.

References

1. Moretti, F.: Distant Reading. 1st edn. Verso Books, London (2013).
2. Oyebode, F.: Fictional Narrative and Psychiatry. Advances in Psychiatric Treatment. 10(2), 140-145 (2004).
3. WHO | Disabilities, https://www.who.int/topics/disabilities/en/, last accessed 2017/09/18.
4. Bamman, D., Underwood, T., Smith, N.: A Bayesian Mixed Effects Model of Literary Character. In: ACL Proceedings, pp. 370-389. Publisher, Baltimore, MD USA (2014).
5. Zhang, H., F.: Exploring Conditions for the Optimality of Naïve Bayes. International Journal of Pattern Recognition & Artificial Intelligence. 19(2), 183–198 (March 2005).
6. Iyer, A.: Depiction of Intellectual Disability in Fiction. Advances in Psychiatric Treatment. 13(2), 127-133 (2007).
7. TextBlob 0.13.0 Documentation, https://textblob.readthedocs.io/en/dev/, last accessed 2017/9/15.
8. Mitchell, D.T. and Snyder, S.L.: Narrative Prosthesis: Disability and the Dependencies of Discourse, 15-46. University of Michigan Press, Ann Arbor (2000).