An Adaptive Methodology: Machine Learning and Literary Adaptation

Grant Glass and Saksham Jain

Abstract:

Using one of the most adapted texts in history, *Robinson Crusoe,* we ask whether or not a computer can find adaptations that scholars have yet to identify. Through testing the effectiveness of different machine learning techniques for text embedding on small groups of full-length texts, we determine the best model for our task, the universal sentence encoder, and then use it to build a deep neural network based binary classifier trained on a large dataset of adaptation and random texts. We attempt to *implicitly* teach the computer the plot of Crusoe, instead of making decisions based on stylistic details, as is a pitfall of traditional techniques. We hope this novel pipeline will help other scholars work with large units of text at the plot - level.

Problem Description:

In light of recent shifts in undergraduate enrolment and funding, the humanities are in crisis. While this sort of forecast might prompt the humanities to reevaluate its methodology, there hasn't been a widely accepted shift in the methodology of literary analysis since the 1980s. Works like, Daniel Shore's *Cyberformalism: Histories of Linguistic Forms in the Digital Archive,* Andrew Piper's. *Enumerations: data and literary study,* and Ted Underwood's *Distant horizons: digital evidence and literary change* all have attempted to change the literary methodology by using algorithms to find patterns and features in texts. While these methodologies utilize many machine learning techniques, these methods have met with massive pushback from the larger humanities community.[1] At the same time, these new methodologies force scholars to think about modeling and conceiving of literary texts differently (McCarty). In this shift of modeling literary text, the question that often comes up is how we can frame these literary questions in a format that a machine learning algorithm can understand.

Early "distant reading" (reading books through algorithms) like Franco Moretti's work use a broad scope of 15,000 novels produced in English to talk about shifting word use and popular works. However, the results did not add anything new to literary knowledge, they just validated existing theories of literature. Underwood's study of genre is another "distant reading" methodology, but the data used needs to be hard coded, since the modern understanding of genre has only emerged in the last half-century and libraries do not capture genres in their metadata. Also, Underwood's results have been more enlightening for the fields of information science rather than literature. In more traditional methods (non-computational) looking at large literary history is using something like Northrupt Fyre's archetypes or Thomas Foster's allusions. The

[1] See Da, Nan Z. "The computational case against computational literary studies." *Critical inquiry* 45.3 (2019): 601-639.

main issue with applying these theories to texts is that each scholar only ends up utilizing only a handful of literary examples to make their arguments and use exceptional, often canonical texts as models for the world. The main problem that this project started with is how you can get a vast enough corpus of material, but still be able to ask specific questions that can actually enlighten literary scholarship.

This problem lead us to look at Daniel Defoe's *The Life and Surprising Adventures of Robinson Crusoe*[2], which has never been out of print in its over three-hundred year print history and has amassed thousands of editions – not to mention the plethora of movies and T.V. shows. Using this text allows us to gain enough material to make these machine learning algorithms viable. The next problem was creating a research question easy enough to compute, but complex enough for humanities scholars to see the merit in this type of analysis. When one begins to understand how pervasive the story of Crusoe is today, showing up in movies like *Robinson Crusoe* (1997), *The Wild Life* (2016), *Robinson Crusoe on Mars* (1964), *Cast Away* (2000), *Swiss Family Robinson* (1960), *Lost in Space* (1998), *Robinson Crusoe* (1954), *His Girl Friday* (1940), *Man Friday* (1970), and *Crusoe* (1988), one begins to question, how many stories are like Crusoe? As we continue to create new stories each day, how do we know if we are borrowing or appropriating[3] the text of Crusoe? It might be easy to identify an adaptation of Crusoe that uses any of the key characters (like Friday) or setting (shipwrecked on an island), but what if the character and settings change (like *The Martian)*? Teaching a machine learning algorithm the story of Crusoe (by feeding it different adaptations) could it start to distinguish between a random story and a Crusoe like story? Could it identify new adaptations of Crusoe that have yet to be discovered? Our project hopes to begin to answer these questions.

Data Description

In our early experiments to determine the suitable text embedding technique, we used four different texts: the "Original Text" which is the 1719 first edition of *Robinson Crusoe* by Daniel Defoe, a "close" adaptation[4] of a 1918 text called, *The Dog Crusoe,* a "far" science fiction adaptation[5] called *The Happy Castaway* (1965)*,* and a random text, *Pride and Prejudice* by Jane Austen (1813). We chose the text based on Grant's scholarship of Robinson Crusoe: the "close" adaptation is one that follows the exact storyline, but changes the human Crusoe into a talking dog, the far adaptation takes the same plot, but everything about the text is changed, and the random text is something similar stylistically, but has no character or plot similarity to Crusoe.

[2] While this project shares the same source texts as the 2018 Data+ Project, "Pirating Texts." We will not be dealing with different editions that purport to be *Robinson Crusoe* by Daniel Defoe, but rather we will be looking at adaptations, which do not mention the source text at all. The methods developed in that project were developed on texts that were extremely similar to one another, so we will not be borrowing the word2vec, distributed memory, and bag of words approaches outlined in the original project.
[3] Julie Sander's refers to appropriation when a text does not indicate (through character or setting) that it is an adaptation of another work.
[4] "Close" refers to the similarity in setting, characters, and plot to the original.
[5] "Far" refers to only the plot being loosely similar to the original text.

The final project utilized two different large datasets. The first corpus was a random pooling of 2,188 texts from the Eighteenth Century Collections Online (ECCO) Text Creation Partnership (TCP)[6], which the metadata for the files is contained in the ECCOTCP.csv file. This data is freely available through the ECCO-TCP website and was verified through the corresponding CSV file, a preview of which is included in Table 1. These works draw from a variety of 18th century and even 19th century works, most 20th century works are still in copyright restrictions and therefore are much harder to obtain. All of these works are in English originally and have their modern spellings intact. An advantage of using the ECCO-TCP corpra is that it is cleaned up by librarians and scholars, so there is very little Optical Character Recognition (OCR) issues with wrong transcriptions of the texts.

TCP	EEBO	VID	STC	Status	Author	Date	Title
K000039.000				Free	Centlivre, Susanna, 1667?-1723.	1705	The basset-table: A comedy. As it is acted at the Theater Royal in Drury Lane, by Her Majesty's servants. By the author of The gamester.
K000122.000				Free	Jones, Henry, 1721-1770.	1745	The bricklayer's poem to the Countess of Chesterfield, on Her Ladyship's saving the soldiers from being shot
K000152.000				Free	Congreve, William, 1670-1729.	1705	Prologue to the court: on the Queen's birthday, 1704.
K000180.000				Free	Cumberland, Richard, 1732-1811.	1772	The fashionable lover; a comedy: as it is acted at the Theater Royal in Drury Lane.
K000266.000				Free	Defoe, Daniel, 1661?-1731.; Pittis, William, 1674-1724.	1705	The diet of Poland: a satyr. Considered paragraph by paragraph. To which is added a key to the whole, ...
K000268.000				Free	Duck, Stephen, 1705-1756.	1741	Every man in his own way: An epistle to a friend. By Stephen Duck.
K000335.000				Free	Cadogan, William, 1711-1797.	1748	An essay upon nursing: and the management of children, from their birth to three years of age. By a physician. In a letter to one of the governors of the Foundling Hospital. Published by order of the General Committee ...
K000343.000				Free	Alderson, John, 1757-1829.	1794	An essay on the rhus toxicodendron: pubescent poison oak, or sumach, with cases showing it is efficacy in the cure of paralysis, ... By John Alderson, ...
K000379.000				Free	Duck, Stephen, 1705-1756.	1741	Hints to a schoolmaster: Addressed to the Revd. Dr. Turnbull. By Stephen Duck.
K000406.000				Free	Hill, John, 1714?-1775.	1775	Hypochondriasis. A practical treatise on the nature and cure of that disorder; commonly called the hyp and hypo. By Sir John Hill
K000415.000				Free	Kelly, Hugh.	1765	An elegy to the memory of the Right Honorable William, late Earl of Bath
K000454.000				Free	Carleton, George, fl. 1728.; Defoe, Daniel, attributed name. 1661?-1731.	1728	The memoirs of an English officer: who served in the Dutch war in 1672. to the peace of Utrecht, in 1713. ... By Capt. George Carleton.
K000532.000				Free	Parnell, Thomas, 1679-1718.	1713	The horse and the olive: or, war and peace.
K000637.000				Free	Arbuthnot, John, 1667-1735.	1712	John Bull still in his senses: being the third part of Law is a bottomless-pit. Printed from a manuscript found in the cabinet of the famous Sir Humphry Polesworth: and published, (as well as the two former parts) by the author
K000983.000				Free	Centlivre, Susanna, 1667?-1723.	1715	An epistle to Mrs. Wallup: now in the train of Her Royal Highness, the Princess of Wales. As it was sent to her to the Hague. Written by Mrs. Susanna Centlivre.
K000685.000				Free	Hill, Aaron, 1685-1750.	1707	Camillus: a poem: humbly inscribed to the Right Honorable Charles Earl of Peterborough and Monmouth. By Aaron Hill, gent.
K000691.000				Free	Fielding, Henry, 1707-1754.	1730	The coffee-house politician: or, the justice caught in his own trap. A comedy. As it is acted at the Theater Royal in Lincoln's End Fields. Written by Mr. Fielding.; Rape upon rape
K000780.000				Free	Cumberland, Richard, 1732-1811.	1754	An elegy written on Saint Mark's eve
K000791.000				Free	Gentleman, Francis, 1728-1784.	1764	The general. A poem: Respectfully inscribed to the Right Honorable the Marquis of Granby. By the author of A trip to the moon.
K000841.000				Free	Fordyce, George, 1736-1802.	1770	Elements of the practice of physic: Part the first. Containing the natural history of the human body. By George Fordyce, ...; Elements of the practice of physic.
K000913.000				Free	Savage, Richard, d. 1743.	1732	An epistle to the Right Honorable Sir Robert Walpole: ... By Richard Savage esquire;.
K000923.000				Free	Morley, John, d. 1776.	1763	An essay on the nature and cure of the king's evil: deduced from observations and practice. The second edition: with an addition of remarkable cases of poor sufferers, cured by the author. ...
K000924.000				Free	Pope, Alexander, 1688-1744.	1728	The history of the Norfolk steward continued: In two parts. Part I. Containing an account of Mr. Lyn's private character, ... Part II. Containing some farther account of Mr. Lyn's management, ...
K000934.000				Free	Pain, Thomas, 1737-1809.	1796	The American crisis, and a letter to Sir Guy Carleton, on the murder of Captain Huddy, and the intended retaliation on Captain Asgill, of the Guards. By Thomas Pain,
K000944.000				Free	Churchill, Charles, 1731-1764.	1762	An epistle to the author of The four farthing candles: By the author of The Rosciad of C-v-nt-G-rd-n.
K000958.000				Free	Kemble, John Philip, 1757-1823.; Shadwell, Thomas, 1642?-1692.	1763	The female officer: or the humors of the army, a comedy. Altered from Shadwell.
K001031.000				Free	Pott, Percivall, 1714-1788.	1765	An account of a particular kind of rupture, frequently attendant upon newborn children: and sometimes met with in adults; viz. that in which the intestine, ... is found in the same cavity, and in contact with the testicle. By Perci
K001036.000				Free	Nolan, William, 18th cent.	1786	An essay on humanity: or a view of abuses in hospitals. With a plan for correcting them. By William Nolan.
K001048.000				Free	Defoe, Daniel, 1661?-1731.	1710	An answer to The tale of a nettle: Written by D. D'Foe.
K001056.000				Free		1772	An essay on the force of imagination in pregnant women: Addressed to the ladies.
K001133.000				Free	More, Hannah, 1745-1833.	1795	The carpenter: or, the danger of evil company.
K001176.000				Free	Murphy, Arthur, 1727-1805.	1769	Genuine memoirs of the life and adventures of the celebrated Miss Ann Elliot: Written by a gentleman intimately acquainted with her; ...
K001204.000				Free	Francis, Philip, Sir, 1740-1818.	1787	House of Commons, Tuesday, 11th December, 1787. Mr. Francis. Mr. Speaker, before I offer any thing to the consideration of the House
K001297.000				Free	Cheyne, George, 1673-1743.	1721	An essay on the gout: with an account of the nature and qualities of the Bath waters. Intended for the benefit of Richard Tennison, Esquire; By Geo. Cheyne, ...; Observations concerning the nature and due method of treating
K001310.000				Free	Defoe, Daniel, 1661?-1731.	1705	Advice to all parties: By the author of The true-born Englishman.
K001312.000				Free	O'Keeffe, John, 1747-1833.	1783	The agreeable surprise: A comic opera. In two acts. By Mr. O'Keefe. The music composed by Dr. Arnold.
K001334.001				Free	Goldsmith, Oliver, 1730?-1774.	1762	The citizen of the world: or, letters from a Chinese philosopher, residing in London, to his friends in the east. ...
K001334.002				Free	Goldsmith, Oliver, 1730?-1774.	1762	The citizen of the world: or, letters from a Chinese philosopher, residing in London, to his friends in the east. ...
K001365.000				Free	Berkenhout, John, 1730?-1791.	1783	An essay on the bite of a mad dog: in which the claim to infallibility of the principal preservative remedies against the hydrophobia is examined. By John Berkenhout, M.D.
K001382.000				Free	Jephson, Robert, 1736-1803.	1783	The hotel: or, the servant with two masters. As it was performed at the Theater Royal, Smock-Alley, with distinguished applause
K001415.000				Free	Hill, Aaron, 1685-1750.	1708	The invasion: a poem to the Queen: By Mr. Hill.

Table 1: ECCO-TCP CSV File describing all the data in the corpus.

The next corpus included 1,484 texts drawn from a variety of variations of *Robinson Crusoe,* pulled from HathiTrust[7] using the HATHI_SCRAPE.py script and Hathitrust's Rsync. The metadata for these files is located in the RC.csv file and provides original and translated English texts from 1719-1903. Many of these texts include the word Robinson Crusoe, but vary their telling, sometimes Robinson Crusoe is from New York City instead of England, sometimes he never finds Friday, sometimes he just gets stuck in the middle of nowhere (not on an island).

[6] https://textcreationpartnership.org/tcp-texts/ecco-tcp-eighteenth-century-collections-online/ see Welzenbach, Rebecca. "Making the Most of Free, Unrestricted Texts: a first look at the promise of the Text Creation Partnership."

[7] https://www.hathitrust.org see Christenson, Heather, "HathiTrust."

Preprocessing

After downloading the Robinson Crusoe texts from the Hathitrust Research Center, we used unzipFiles.py and unzipping.py to unzip the files and ensure all the file names were standardized to the way they showed up on the CSV. Additionally,we used Ted Underwood's Python library for cleaning up 18th century texts[8] since the OCR of the scanned pages of the original books oftentimes does not pick up the long S.

Examples of this sort of error look like this:

```
felecterrors = ['fee', 'fea', 'fay', 'fays', 'fame', 'fell', 'funk', 'fold', 'haft', 'fat', 'fix', 'chafe', 'loft']
selecttruths = ['see', 'sea', 'say', 'says', 'same', 'sell', 'sunk', 'sold', 'hast', 'sat', 'six', 'chase', 'lost']
```

After pulling the texts through the cleanup, we further processed by removing the title page and other paratextual material, then removing any strange not UTF-8 characters and inappropriate spacing. By using regular expressions, we removed any weird digits in the string (process.py and in the dataset.ipynb) caused by OCR or other encoding issues. Example of this code includes: `re.sub(r'[^a-zA-Z0-9. /"]', r'', txt_file_as_string)`

Core Methodology

All of our initial methods are outlined in the similarity.ipynb notebook. Our first experiment to determine which method would generate the best text embeddings for our task was with the sklearn TfidfVectorizer to build the embeddings of our training data[9]. We calculated the cosine similarity scores for each of the texts to the reference text (i.e. the Original Text) using sklearn's metrics-pairwise package. Then we experimented with Google Research's Universal Sentence Encoder (USE)[10]. While originally meant for generation of sentence-level embeddings, the model does not actually require a set maximum sequence length, which is a useful functionality that allows us to represent full-length texts of varying lengths as a fixed-dimensional embedding layer. The USE directly uses the encoding sub-graph of the original transformer architecture. Lastly, we used BERT, described in Jacob Devin et al.'s "Bert: Pre-training of deep bidirectional transformers for language understanding." to obtain sentence-level embeddings. Specifically, we used UKPLab's implementation[11]. In this method, we chose to tokenize each text into sentences first (because important information might be lost

[8] https://github.com/tedunderwood/DataMunging see Underwood, Ted. *Distant horizons: digital evidence and literary change*.

[9] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_extraction.text We used the S

[10] https://tfhub.dev/google/universal-sentence-encoder/4 -the latest pretrained model available, updated 2020. See Yang, Yinfei, et al. "Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax."

[11] https://github.com/UKPLab/sentence-transformers see Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding."

if we allow BERT to automatically truncate the input sequence after a max length of 512 tokens), and then averaged the resulting similarity scores over the entire text.

In the end, we chose to work with the USE embeddings because it gave more context-aware, and as a result, discriminatory embeddings than the other candidates. Notice (in Figure 1) that the text determined as 'close' to the reference text by Grant (human expert), while indeed the closest, still showed a cosine similarity of only 0.528. Further, the texts determined as 'random' and 'far' were also significantly further from 'close' as well as the reference text, but very close to each other - which is what we might expect from a model which has learnt semantic relationships particularly well (after all, why should *Pride and Prejudice* be closer to *Robinson Crusoe* than *The Happy Castaway* - both are unrelated by plot). Note that the BERT embeddings, *Pride and Prejudice* turned out to be closer to *Robinson Crusoe* which we posit is due to the nature of the sentence-level embeddings - the representation learnt is more about the similarity in the stylistic/linguistic/grammatical/lexical sense than about the plot.

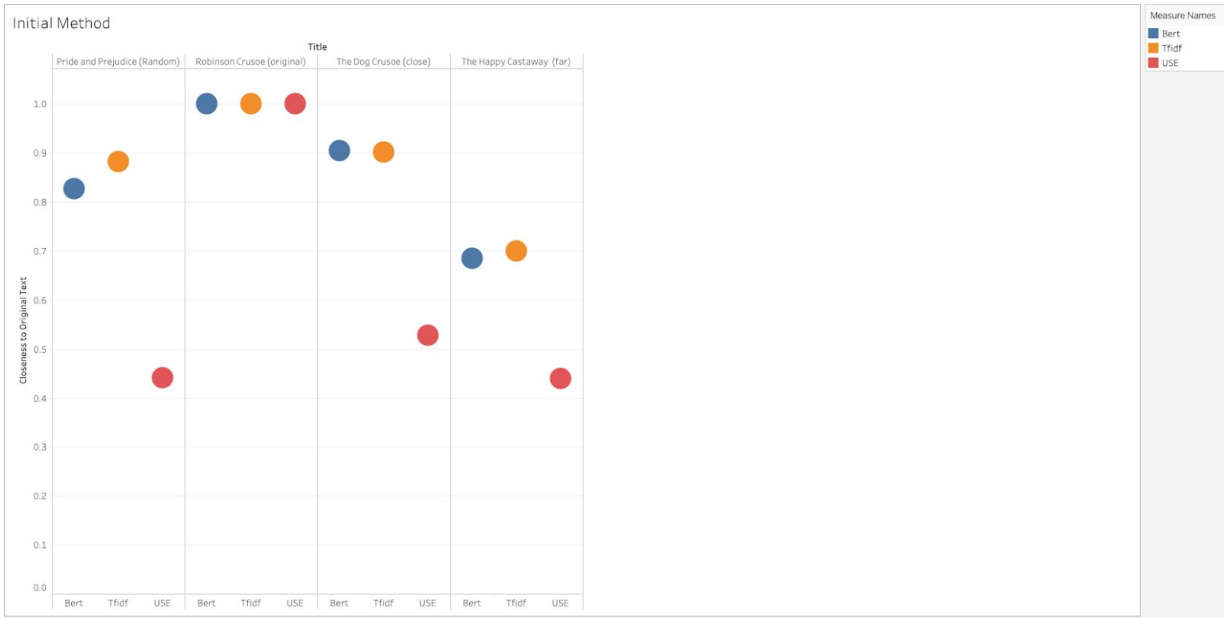

Figure 1: Results of Initial Method Across Different Texts (1.0 being closest to the original text) USE-Universal Sentence Encoder.

Ultimately, we built a neural network classifier with 2 hidden layers on top of the Universal Sentence Encoder embedding layer in order to classify between the Adaptations and Random (i.e. Not-Adaptation texts). First, we build a dataset with an equal number of Adaptations and Non-Adaptations are present) through the code in dataset.ipynb with both classes containing 1484 samples each. Then we randomly split it into 70% training data, and 15% each of validation and testing data. We implement the classifier model in Keras in the train.ipynb notebook, which takes each of the texts in the training dataset as an input string, outputs a 512-dimensional embedding layer through USE, which is then passed through 2

densely connected layers with Relu activations and 256 and 32 neurons each before outputting the resulting class through the final softmax layer. We use the default learning rate and the Adam optimizer and notice that the model converges very quickly (within 10 epochs).

Final Results

The model performs exceptionally well on the validation and test sets, identifying the adaptations (denoted by class 1 in Figure 2) with near perfect precision and recall.

	precision	recall	f1-score	support
0	0.98	1.00	0.99	137
1	1.00	0.99	1.00	309
accuracy			0.99	446
macro avg	0.99	1.00	0.99	446
weighted avg	0.99	0.99	0.99	446

Figure 2: The classification results from the mode. 0 denotes non-adaptations and 1 denotes adaptations

Current Conclusions and Future Work

The potential pitfall with this technique is that we will not be able to measure how similar a text is to *Robinson Crusoe*, but that is what the Data+ team did already, by using this technique to look at a larger window of text than a sentence, we can find works that share a similar plot, which would begin to make a new model of adaptation centered around plot rather than setting or characters. The challenge becomes where exactly the plot gets figured out, what unit of text can tell us that? If we can begin to think about where the plot gets encoded in the text and we can make the window of analysis the same, then we can begin to move forward.

A future endeavor of this project would be to work through the book covers and movie posters of these adaptations to see if a computer vision model works better at identifying similarities than a textual one, and see if a multimodal model works even better. Another concern here is to see if a computer can reliably understand the similarities of loneliness generated by an island to one of a planet (like many of the science fiction versions of Crusoe) - neural networks are notoriously seen as black boxes so we must also look towards explainable models. Another potential pitfall is the change of styles throughout the years, however, book covers are more consistently replaced with old stories and movies being updated not through the content, but through its cover.

References

Chaudhary, Vishrav, et al. "Low-Resource Corpus Filtering using Multilingual Sentence Embeddings." *arXiv preprint arXiv:1906.08885* (2019).

Christenson, Heather. "HathiTrust." *Library Resources & Technical Services* 55.2 (2011): 93-102

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Foster, Thomas C., David De Vries, and © 2012 by Harper Collins Publishers. *How to read literature like a professor*. Harper Collins Publishers, 2012.

Frye, Northrop. *Anatomy of criticism: Four essays*. Princeton University Press, 2020.

McCarty, Willard. *Humanities computing*. 2014.

Moretti, Franco. *Distant reading*. Verso Books, 2013.

Piper, Andrew. *Enumerations: data and literary study*. University of Chicago Press, 2018.

Rae, Jack W., et al. "Compressive transformers for long-range sequence modelling." *arXiv preprint arXiv:1911.05507* (2019).

Sanders, Julie. "Adaptation/Appropriation." *The Encyclopedia of the Novel* (2010).

Shore, Daniel. *Cyberformalism: Histories of Linguistic Forms in the Digital Archive*. JHU Press, 2018.

Underwood, Ted. *Distant horizons: digital evidence and literary change*. University of Chicago Press, 2019.

Watt, Ian. *The rise of the novel*. Univ of California Press, 2001.

Welzenbach, Rebecca. "Making the Most of Free, Unrestricted Texts: a first look at the promise of the Text Creation Partnership." (2011).

Yang, Yinfei, et al. "Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax." *arXiv preprint arXiv:1902.08564* (2019).

We both affirm that we adhered to the honor code in the report.